Title
Spherical Encoding for Osteoarthritis Biomarker Discovery

Author
Morales Martinez, Alejandro Guillermo

Publication Date
2021

Peer reviewed|Thesis/dissertation

Spherical Encoding for Osteoarthritis Biomarker Discovery

by

Alejandro Guillermo Morales Martinez

DISSERTATION

Submitted in partial satisfaction of the requirements for degree of

DOCTOR OF PHILOSOPHY

in

Bioengineering

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO
AND
UNIVERSITY OF CALIFORNIA, BERKELEY

Approved:

DocuSigned by:
Valentina Pedoia
CB0AB38D57D842C...

Valentina Pedoia

Chair

DocuSigned by:
Srikantan Nagarajan

Srikantan Nagarajan

DocuSigned by:
Thomas Link
DD70B4488CA24F4...

Thomas Link

Committee Members

Acknowledgements

I have a lot of people to thank for helping me through this incredible journey. My advisor Valentina Pedoia has been a patient and encouraging mentor every step of the way. Even though I had no background in MRI or deep learning research when I started, she saw the potential in me just from our initial meeting. She gave me the space to learn at my own pace all the while pushing me to challenge my expectations. Her constant trust in my skills also buoyed me throughout my PhD and left me no choice but to believe in myself. I especially appreciate her endless scientific creativity and exceptional ability to spread her enthusiasm about our projects to others, including me. I am proud to have been her first PhD student and I am looking forward to the incredible things she will accomplish throughout her career.

I would like to thank the other members of my dissertation committee, who happen to be the members of my qualifying exam committee. The qualifying exam was the hardest challenge of my PhD and I had the privilege of sharing it with some amazing scientists. Srikantan Nagarajan is a venerable fountain of knowledge about statistics and machine learning, always friendly and inquisitive, making me a better scholar. Thomas Link has a musculoskeletal expertise second to none, as well as one of the most pleasant people I have ever met. Peder Larson knows more about MRI physics than I can learn in my lifetime, yet he was a patient mentor throughout my qualifying exam, answering even the most basic of questions. The quality I admired the most from these accomplished scientists was their unassuming nature. They made me feel like a peer during my qualifying exam preparation and presentation, which boosts my confidence to this day. I will follow their example and emulate their qualities for the rest of my career.

My work was possible thanks to the support from my group and collaborators at the Center for Intelligent Imaging in the Department of Radiology and Biomedical Imaging at UCSF. Francesco Caliva, thank you for your friendship and for your help in our projects. I consider persuading you to join us one of my proudest achievements in these last four years. Claudia Iriondo, thank you for recruiting me into the group when I needed that direction at the start of this experience, I am proud to graduate alongside you. Aniket Topaldi and Kenneth Gao, you are both now the face of the group and we entrust all of the social and recruitment activities to you. Felix Liu, thank you for your boundless knowledge of the OAI and always being so responsive. Jenny Lee, thank you for all of your assistance designing statistical tests, I appreciate it. Madeline Hess, thank you for your curiosity and the excitement you bring to the group. Pablo Damasceno, thank you for being a soundboard for all my programming ideas. Sharmila Majumdar, thank you for being a fearless leader and inspiring the next generation of scientists.

Another group of people I want to thank is the administrative and IT staff at UC Berkeley and UCSF for their support. The paperwork for the UC Berkeley and UCSF Bioengineering Graduate Program can be hard to navigate given the two campuses but SarahJane Taylor, Barbara Green, Cathy Devine, Kristin Olson, Victoria Ross, and Rocio Sanchez are all administrative superstars. Jed Chan, Peter Storey, and Luis Torres were always responsive to my IT problems and kept my work running, even when I pushed the radiology servers to their limit.

I am also thankful for my fellow students in the Bioengineering program. You all made this experience fun and enjoyable, especially during the first two years when I did not know

anybody. Special thanks to Andrew Leynes for being an early friend and guide on this journey. Thanks to Natalie Korn for introducing me to teaching and being a mentor. To anyone else I missed, I am sorry and please know that I am thankful for your help.

I cannot sufficiently express how grateful I am for the support of my family. They fostered my scientific curiosity from a young age and encouraged me to relentlessly pursue my dreams. I truly would not be where I am today without their sacrifices. To my mother, Marta Nieves Martinez Solis, I thank you for being a role model and one of my biggest supporters. To my father, Jose Antonio Morales Monzon, thank you for your impeccable work ethic and your encouragement when I was deciding to do a PhD. To my sister, Graciela Morales Martinez, thank you for keeping me humble all these years with your continuous ribbing. To my aunt, Martha Zita Morales Monzon, and the rest of my Cuban family, I want to thank you for your love and I cannot wait to celebrate this and many other accomplishments in person.

Last but certainly not least, I want to thank my loving wife and partner in life, Emily Gustafson. Meeting you during the first year of my PhD is the biggest blessing of my life. I am incredibly fortunate getting to share this experience with you, your unwavering support has been a source of inspiration that has carried me through the hardest challenges I have faced. Thank you for making me smile every day and for making sure I prioritized my physical, mental, and spiritual wellbeing throughout these last four years. **I dedicate this work to you**.

Contributions

Chapter 3 of this work was adapted from the following publication:

Martinez AM, Caliva F, Flament I, et al. Learning osteoarthritis imaging biomarkers from bone surface spherical encoding. Magnetic Resonance in Medicine. n/a(n/a). doi:10.1002/mrm.28251

Yo vengo de todas partes, y hacia todas partes voy…

- Jose Martí

Spherical Encoding for Osteoarthritis Biomarker Discovery

Alejandro Guillermo Morales Martinez

Abstract

Knee osteoarthritis is a degenerative musculoskeletal disorder marked by gradual cartilage breakdown, and involving all tissues of the joint. It is one of the leading worldwide causes of chronic disability in older populations, with the prevalence expected to increase. There is currently no available treatment to reverse the degenerative damage characteristic in osteoarthritis, with the only option available for end stages of the disease being a partial or total knee replacement. Furthermore, the clinical standard for osteoarthritis diagnosis is a radiographic score which reflects advanced pathological stages, often with irreversible damage. The lack of therapies has generated a need for osteoarthritis imaging biomarkers capable of detecting and monitoring the progression of the disease. This dissertation aims to bridge this gap by defining a novel spherical encoding representation for known quantitative imaging biomarkers for osteoarthritis. In this work, we leverage the superior cartilage sensitivity of MRI, a large retrospective labeled imaging dataset, and the superior feature-learning ability of convolutional neural networks to define novel OA imaging biomarkers based on spherical maps. Large-scale quantitative analysis using convolutional neural networks uncovered new associations between bone shape, cartilage thickness, and cartilage T_2 relaxation time values and OA symptoms.

Table of Contents

List of Figures

List of Tables

Chapter 1: Osteoarthritis

1.1 Impact and pathophysiology of OA

Knee osteoarthritis (OA) is a complex degenerative joint disease that is characterized by the progressive degeneration of articular cartilage, as well as changes in tissues within the joint such as the synovium, subchondral bone, and ligaments. OA is the most common joint disease, with US estimates of the disease prevalence approaching 14% of the population and expected to keep increasing as the US population becomes more obese and sedentary[1], both risk factors for OA. This prevalence places a large burden in the US economy, with the total arthritis-attributable US medical care expenditures and earnings losses amounting to 1% of the 2013 US gross domestic product[2]. At an individual level, OA can manifest through symptoms like pain and loss of joint function, both of which can lead to chronic disability and reduced quality of life depending on their severity[3]. Additionally, there is no current treatment available to revert the progression of OA, with partial or total arthroplasty being the only treatment available for end-stage OA. Current treatment of OA mainly consists of palliative measures aimed to reduce the clinical symptoms of OA, such as pain and loss of joint function, as well as disease progression management through modifiable risk factors like body weight and exercise.

Given this substantial impact in society, the OA disease etiology has been the focus of significant research efforts in order to discover ways to curb and monitor the disease progression. The development of OA has been associated with risk factors such as age, obesity, sex, and physical activity, some of which are modifiable and offer ways to manage the OA progression. In recent years, OA has been generally understood to be an inflammatory and biomechanical whole-organ

disease influenced by factors including bone shape, synovitis, diabetes, and age-related inflammation, among others[4]. The whole-organ nature of OA obfuscates the order and causality of tissue changes within the knee, creating a challenge for studying the timeline of OA development. Despite the involvement of all joint tissues in the onset and progression of OA, particular emphasis has been placed in structural and biochemical changes within the articular hyaline cartilage. Articular cartilage is avascular and aneural, resulting in a limited healing ability which makes it particularly vulnerable to degenerative breakdown in OA. Biochemically, the structure of articular cartilage consists of a dense extracellular matrix (ECM) composed of water, collagen, and proteoglycans, with a sparse distribution of chondrocytes. These chondrocytes, together with the ECM, help retain water within the cartilage and contribute to its biomechanical function as a load-distribution and low-friction surface for the knee joint[5]. During early-stage OA there is a disorganization of the ECM structure which leads to excess water infiltrating the cartilage, while late-stage OA includes the disintegration and dehydration of the cartilage which results in appreciable tissue loss. Disruptions to this cartilage structure occur naturally through aging, which coupled with the reduced healing potential of chondrocytes, make OA a highly prevalent disease in older adults above the age of 65[6].

While the full OA pathophysiology is not yet understood, there are coincident pathologies that worsen the disease progression, including subchondral bone sclerosis and cartilage degeneration. The subchondral bone sits beneath the articular cartilage and provides it with mechanical and nutrient support. It consists of two layers, the subchondral bone plate (SBP), a calcified plate directly underneath the cartilage, and subchondral bone trabecula (SBT), a cancellous bone structure which undergoes active remodeling. Early-stage OA changes in the subchondral bone

involve a thinning of the SBP with decreased bone density in the SBT. Late-stage OA is characterized by subchondral bone sclerosis, a thickening of the subchondral bone in response to abnormal loads and results in weakened bone[7]. Furthermore, continuous subchondral bone remodeling occurs both as a regular part of normal joint function and throughout the OA disease process. Considering the close relationship between the subchondral bone and the articular cartilage, there has been a focus on the crosstalk interactions between both tissues during the OA onset and progression. Studies have shown the existence of direct molecular signaling between the cartilage and bone, with growth factors that mediate osteogenesis having a protective effect on chondrocytes[8]. The temporal interactions between the articular cartilage and subchondral bone play an important, yet unclear, role in the onset and development of OA.

1.2 OA imaging

OA is mainly diagnosed in a clinical setting with the Kellgren-Lawrence (KL) radiographic scale, which is composed of five grades ranging from 0 to 4. The main OA features measured using KL grades are tibiofemoral joint space narrowing, osteophytes, subchondral bone sclerosis, and bone deformities[9]. The grading is performed on a posterior-anterior weight-bearing 2D radiograph in order to standardize load-dependent differences in joint spacing between patients. The KL grades determine OA severity based on the previous features and represent no OA (KL=0), minimal/doubtful OA (KL=1), mild OA (KL=2), moderate OA (KL=3), and severe OA (KL=4). While X-rays are a cheap and fast imaging modality to screen for OA, it has two major drawbacks in the form of patient toxicity and poor soft tissue contrast. Ionizing radiation from X-rays is hazardous to patients due to its carcinogenic ability to damage the genetic material in human cells. Additionally, the image contrast in X-rays depends on the level of attenuation of

each tissue, with denser tissues, such as bone, possessing a strong image signal compared to other less dense tissues, such as cartilage. The lack of soft tissue sensitivity of X-ray imaging for cartilage, menisci, and synovium restricts its ability to study and monitor early OA symptoms[10]. In KL scoring, this limitation is lessened by the fact that tibiofemoral joint space narrowing acts as an indirect measure of articular cartilage thickness loss. Nevertheless, radiographic OA changes, like the loss of cartilage resulting in joint space narrowing, and the presence of osteophytes, are endemic of later stages of the OA disease progression, thus limiting X-rays as an imaging tool for early OA diagnosis. Furthermore, there is a reported discordance between radiographic OA features and clinical symptoms such as knee pain[11].

Magnetic resonance imaging (MRI) is a powerful and versatile imaging modality which enables the visualization of anatomical and biochemical features. Unlike X-rays, MRI is safe for patients and relies on the differences in spin frequencies between the protons that make up the molecules in the human body. The MRI image formation starts by first aligning all the protons in the body of the patient along a main magnetic field, applying three perpendicular linear magnetic gradients along each imaging axis, and then perturbing the proton alignment with a set of magnetic pulses, known as a pulse sequence. The linear gradients encode spatial information into the protons in order to locate them within the body during the image construction. The image signal is then measured as the phase and frequency information encoded in the electrical current induced by the magnetic fluctuations from the perturbed protons. This image acquisition allows the tuning of the image contrast to a specific tissue of interest through the design of the pulse sequences. In addition to imaging structural features, MRI has the ability to extract compositional information about the tissues of interest through careful selection of the imaging

parameters. Due to its ability to image all tissues of the joint, MRI is uniquely well-suited for the study of the whole-organ OA disease process. Furthermore, the inflammatory nature of OA requires the imaging of synovium, bone marrow edemas, and joint effusions, which would be missed by X-ray imaging.

The usefulness of MRI as an investigative tool for OA has been recognized through the creation of large knee MRI datasets that track imaging changes on healthy and OA patients. The largest of these imaging studies is the Osteoarthritis Initiative (OAI) dataset, a multi-center longitudinal multi-modality imaging studies in 4,796 patients[12]. The MR imaging was performed on 3T MRI systems with two coronal and three sagittal pulse sequences. The dataset consists of a total of 12 time points ranging from an initial baseline visit to a final 108 month visit with yearly visits in between and a half-year visit for the third and fifth visits. Out of these 12 time points, MRI scans were performed at seven time points, resulting in an imaging span of eight years. Demographic data such as age, body mass index (BMI), and sex was also recorded during each visit. The two sagittal pulse sequences used in the following work are the 3D double echo steady state (3D-DESS) and 2D multi-slice multi-echo (2D-MSME) spin-echo. The 3D-DESS sequence uses water excitation to suppress the bright signal of the fatty tissues within the knee and improve the contrast of the cartilage, menisci, bone, and ligaments. Due to its high spatial resolution, both slice and in-plane, as well as high cartilage contrast, it is used for morphological quantitative analysis for cartilage and bone. The 2D-MSME spin echo sequence on the other hand, possesses comparatively lower spatial resolution but enables compositional imaging of the cartilage through relaxometry analysis. These two sequences together leverage the ability of MRI to investigate both anatomic and physiologic changes associated with OA.

1.3 OA imaging biomarkers

This dissertation will focus on three OA imaging biomarkers acquired from MRI: bone shape, cartilage thickness and cartilage T_2 relaxation time values.

1.3.1 Bone shape

The mechanically-driven nature of OA, coupled with the role of continuous bone remodeling throughout OA development, makes bone shape a promising OA imaging biomarker. Changes in bone shape or geometry alter the load distribution in the joint and may either lead to or exacerbate OA, with well-known changes including osteophytic lipping of the joint periphery, increased tibial plateau size, and subchondral bone attrition. Furthermore, the high rate of changes in the bone compared to cartilage makes it an ideal target for therapies that use bone shape as a measure of drug efficacy[13]. Studies have indeed shown that the shape of the femur, tibia, and patella bones is associated with OA onset and development[14–17]. Neogi et al[14] found that the shape of the femur, tibia, and patella bones predicts the incidence of future OA on knees without any radiographic OA features, suggesting that the intrinsic bone shape in certain patients predisposes them for future OA. This same study showed that bone shapes associated with OA include a widening of the condyles in the femur and tibia, as well as an osteophytic ridge growth around the cartilage plate for all three bones. Additionally, studies have established associations between bone shape and OA-related injuries like Anterior Cruciate Ligament (ACL) tears, highlighting the importance of bone shape for healthy joint function and kinematics[18,19]. In these studies, bone shape is described using statistical shape modelling, a technique which finds the principal modes of variations that describe bone shape in a population. These principal modes

are often projected into a linear discriminant vector that is related to the presence or absence of OA in order to assess the association of certain bone shapes to OA. The problem with such approaches is the lack of supervision of principal component analysis (PCA), which describe shape independently of OA and could miss subtle OA-related bone shape differences. The study detailed in the third chapter of this dissertation addresses these problems by using data-driven, supervised techniques which exploit the bone shape information at the segmentation and classification level.

1.3.2 Cartilage thickness

Cartilage thickness loss is a defining characteristic of OA and has long been used as an OA imaging biomarker. The structural deterioration of the articular cartilage manifests as the joint space narrowing observed in radiographic OA and eventually leads to painful crepitus, a condition where articular cartilage degeneration exposes bone-on-bone friction within the joint. The rate of cartilage degeneration in OA also outpaces the limited regenerative rate of articular cartilage, especially in the case of traumatic injuries, such as ligamentous tears, where large portions of the cartilage are damaged. Studies have shown that different rates of cartilage thickness loss are associated with radiographic OA and, to a lesser degree, pain[20–22]. Generally, cartilage thickness measurements are averaged across anatomical joint regions, defined based on clinical assumptions, in order to simplify the quantitative analysis. Additionally, the location of the cartilage thickness loss matters, with the medial tibiofemoral compartment consistently linked to knee OA, due to the high prevalence of medial OA. The regions commonly used are the medial and lateral compartments, as well as the anterior and posterior aspects for the femoral, tibial, and patellar cartilage. For the medial and lateral femoral compartments, the weight-

bearing region, where the load is distributed across the femorotibial joint, is distinguished due to its role in joint space narrowing[23]. While there has been significant work evaluating the impact of cartilage thickness loss to clinical patient outcomes, only a weak association between cartilage thickness loss and pain has been found[24–27]. This weak association was further evidenced in a recent clinical trial for sprifermin, which found no significant difference in pain for patients with radiographic OA, even after a substantial preservation of 0.05 mm of femorotibial cartilage over a two-year period[28]. The study in the fifth chapter of this dissertation improves the statistical association between cartilage thickness and pain through a data-driven approach for cartilage thickness subregion definition.

1.3.3 Cartilage T_2 relaxation time values

Structural cartilage thickness changes are useful for understanding the OA pathogenesis, but they reflect late-stage manifestations of OA, which are irreversible due to the lack of therapies. The complex organization of the articular cartilage biochemical structure must be maintained for the healthy functioning of the joint. Under normal joint loading, cartilage reversibly deforms to dissipate the weight thanks to the orientation of the collagen fibers coupled with the proteoglycans in the articular cartilage ECM. Disruptions to this microenvironment affect the biomechanical function of articular cartilage as a lubricating and load-distributing tissue in the joint, and are thought to precede morphological cartilage changes[5]. Compositional MRI techniques such as T_2 parametric mapping enable the assessment of the cartilage matrix through the quantitative measurement of water content and collagen fiber organization[29]. The T_2 value of specific tissues depends on the mobility of its protons, with fluid-filled tissues such as cartilage possessing higher T_2 values compared to rigid tissues such as bone. Within healthy articular

cartilage, the ECM bounds the water molecules, thus limiting their mobility and reducing the cartilage T_2 value. During early OA development, damage to the collagen-proteoglycan network generates an imbalance in the osmotic pressure within the ECM and consequently increases the mobility of water within the cartilage. This is reflected in a comparative increase in T_2 values for damaged cartilage over healthy cartilage, observed using parametric fitting of T_2-weighted spin echo sequences. Studies have exploited this physiological effect to establish T_2 relaxation time values as an early-stage OA imaging biomarker which predates any morphological OA changes within the joint[30–33]. Furthermore, elevated cartilage T_2 values between patients without radiographic OA have been linked with knee pain, although more work may be needed to establish a strong association[34]. Given the importance of early detection and intervention in OA, cartilage T_2 values are a promising biomarker for future therapies. The work in the fourth and fifth chapters of this dissertation builds on these studies to explore the contribution of cartilage T_2 relaxation time values on radiographic and symptomatic OA.

Chapter 2: Deep learning applications in OA imaging research

In recent years, deep learning methodologies have revolutionized both the scale and direction of OA imaging research. The advent of convolutional neural network (CNN) and supervised deep learning methods has enabled the large-scale processing and data-driven feature learning of medical imaging data. The three main areas covered in this dissertation where machine learning has impacted the study of OA are: automatic extraction of OA imaging biomarkers, clinical OA diagnosis and incidence prediction, and OA biomarker discovery.

2.1 Automatic OA imaging biomarker extraction

Traditionally, there are two approaches for extracting the three OA imaging biomarkers discussed in Chapter 1.3 from MRI data. The first and most straightforward approach consists of manual segmentation of the tissue of interest by trained users. In practice, a musculoskeletal radiologist trains an inexperienced user for this task, which could take several hours depending on their experience. Additionally, the actual segmentation requires just as much time based on the size and shape of the tissue of interest, as well as the resolution of the MRI scan. As an example, manually segmenting the femur, tibia, and patella bones for a 3D-DESS MRI volume with 160 slices requires upwards of one hour, even for an experienced user using specialized software. The second of these traditional approaches consists of using statistical shape modelling (SSM) methods such as active shape models and active appearance models. These methods are semiautomatic since they require a priori knowledge about the shape of the segmented organ through the definition of matched landmarks across the MRI volumes. The first step in these methods involves the creation of a training set of data with the matched landmarks around the

tissue of interest. PCA of these landmark points generates a point distribution model that describes the mean shape of the target organ. New imaging data can then be segmented by iteratively deforming the points of the mean statistical shape to match the new data based on a least squares minimization[35]. While these methods vastly increase the segmentation throughput over manual approaches, they suffer from the a priori definition of the landmarks, due to the complexity of choosing these regions, and the use of linear methods such as PCA to describe the shape, which would miss subtler nonlinear shapes.

The introduction of CNN methods has dramatically improved the accuracy of computer vision tasks such as object detection, image classification, and image segmentation[36–38]. These methods mimic the natural process through which human neurons learn and process information at different levels of abstraction by using the receptive fields of multiple convolutional filters. The successive application of each convolutional filter to the data generates distinct feature maps that encompass increasing levels of semantic information. This procedure exploits the hierarchical organization of image data, where low-level features, such as edges, can be combined to form high-level features, such as shapes and objects. Before they can be used, CNN models must first be trained on a dataset using an iterative process that updates the parameters of the convolution filters to optimize a specific outcome task. Unseen new data can then be quickly and automatically segmented or classified with a trained CNN. One big advantage of CNN models over classical methods involves the fully data-driven feature selection instead of a priori definitions of the feature importance. The feature learning relies on the inherent variability of the training data, with larger training datasets approximating the population distribution. Given the reliance of CNN approaches on the availability of sizeable training data, their use has been

bolstered by the emergence of large labeled datasets in recent years. Concurrent technological advances in computational resources such as Graphical Processing Units (GPU) have also made the training of large CNN models feasible, yielding breakthroughs in fields ranging from natural language processing[39] to drug discovery[40]. Furthermore, the performance and robustness of individual CNNs can be significantly enhanced through average ensembles of the probabilistic outputs from individual models[41].

In the context of musculoskeletal imaging, advances from CNN approaches are evidenced by the recent body of work on state-of-the-art bone, cartilage, and menisci segmentation from the OAI data[42–46]. The following three chapters in this dissertation leverage ensembled 3D V-Net[47] CNN architectures for the segmentation of bone and cartilage from the 3D-DESS data as a preprocessing step to extract the bone and cartilage biomarkers. The high-quality automatic segmentation of these tissues across the entire OAI is further post-processed to quantify the OA imaging biomarkers from the segmented masks. In this dissertation work, a novel spherical encoding method transforms the segmented bone and cartilage into colocalized 2D articular surface maps that can be used with common CNN architectures. This spherical transformation method has the advantage of reducing the dimensionality of the segmentation maps while preserving the spatial correspondence between the subchondral bone and the articular cartilage. Chapter 4 of this dissertation explores the relationship between the individual OA imaging biomarkers and incident radiographic OA through the use of a spherical encoding biomarker fusion framework.

2.2 Clinical OA diagnosis and incidence prediction

Another prominent application of deep learning in OA imaging research is the creation of predictive models for OA diagnosis and future incidence. Early imaging detection of OA would improve patient outcomes, such as pain and total knee replacement (TKR), through the aggressive management of modifiable risk factors in the disease progression. It would also improve the design of clinical trials for new therapeutics by identifying patients at risk of developing OA and screening them out of healthy control cohorts. Additionally, deterministic predictions from automatic predictive models would address the challenge of inter-reader variability for human readers grading OA[48,49].

Prior to the introduction of CNNs, common methods for OA classification consisted of logistic regression, forest-based classifiers, and supervised clustering classifiers. These methods require an initial phase for feature extraction and engineering from the image data, with commonly defined features including edges, image intensity, textures, and shape descriptors[50–52]. After the feature selection process was completed, the resulting features would then be used to build the classifiers for prediction of existing and future OA symptoms. The main shortcoming of these methods is the reliance on handcrafted features, which may be suboptimal to model OA presence and development. The iterative and open-ended nature of feature engineering also introduces individual bias in the predictive models through the selection of user-specific OA-related features from the image information. CNNs overcome both of these limitation by learning the features directly from the data, leveraging population-wide patterns which may not be obvious to craft even for domain experts. Recent imaging studies have demonstrated the superiority of deep

learning methods for prediction of radiographic OA[53] and TKR[54,55], achieving state-of-the-art area-under-the-curve (AUC) performances over 0.87.

The OAI dataset used in this dissertation is well-suited for CNN approaches, with 50,000 3D-DESS structural and 26,000 2D-MSME compositional scans, each with corresponding clinical OA and demographic variables. The work in this dissertation leverages the large scale of the OAI dataset to train accurate predictive models based on different clinical endpoints for OA. For the third and fourth chapters, KL grades are chosen as the clinical OA endpoint, while for the fifth chapter, patient-reported symptoms of pain are chosen as the clinical OA endpoint.

2.3 Model interpretation and OA biomarker discovery

Despite the superior performance of CNN models over classical methods, they suffer from a costly trade-off between performance and interpretability. The components of linear regression and rule-based expert systems can be teased apart by design to produce highly interpretable predictions. In contrast, CNN models contain millions of parameters and successively apply nonlinear transformations to the data as it moves through the convolutional layers. Once a CNN model is fully trained, understanding the relative importance of each of the parameters for a decision of interest becomes a considerable challenge. This presents a significant roadblock for the adoption of such methods into the clinical workflow, where legal and ethical guidelines demand a rationale for a particular diagnosis. Consequently, shedding light on the decision-making process of CNN models would improve their reliability and translation into healthcare practice. Beyond their use in healthcare settings, the lack of understanding restricts our ability to

learn new associations between the input and output data from the trained models. There is great promise in the use of explainable CNN approaches as a tool for biomarker discovery, given their superhuman performance in disease classification and prediction tasks. In the study of OA, cartilage T_2 and cartilage thickness imaging biomarkers have been summarized through compartment averaging, based on two or more anatomical regions, before their inclusion in predictive models. Averaging across these regions, discussed in Chapter 1.3.2, presupposes an association between relevant clinical outcomes such as pain and these anatomical compartments. The discordance reported between OA imaging biomarkers and clinical OA symptoms[56], such as pain, suggests this compartment averaging might be too simplistic for the complex OA disease process. Interpretable CNN models offer a personalized, data-driven alternative to compartment averaging by averaging patient-specific regions most strongly associated with pain. This approach is the focus of Chapter 5.

The emerging field of explainable deep learning has yielded methods to interpret the predictions of trained CNNs. Different explanatory techniques such as linear proxy models, decision trees, and saliency mapping attempt to understand the CNN model performance by approximating them to linear models, decomposing them into decision trees, or systematically perturbing the inputs to discover the effect on the outputs[57,58]. Unlike other approaches, saliency mapping directly uses the network gradients to generate visualizations of local decision-making importance for a specific input image. This property of saliency mapping allows the interpretability of fully trained CNN models directly, without the need to design a specific architecture or time-consuming perturbation of input images, in the case of occlusion mapping. For large datasets, this efficiency permits the generation of average importance maps which

create a measure of feature importance for the model. Among the saliency mapping strategies, Gradient-weighted Class Activation Mapping (Grad-CAM) has the added benefit being class-discriminative by using the gradient information flowing into the last convolutional layer of the CNN to understand each neuron for a decision of interest[59]. This class-specific saliency map can be overlaid as a heat map of location importance on the input image. Grad-CAM also balances input image regions of high network activation, where neurons fire strongest, and input image regions of high network sensitivity, where changes would most affect the decision. In Chapter 5, a CNN model is trained to classify chronic pain based on the spherical encoding of cartilage thickness. The trained model is then interpreted using Grad-CAM to obtain an average weighting map for cartilage thickness which is most associated with chronic pain.

Chapter 3: Learning osteoarthritis imaging biomarkers from bone surface spherical encoding

3.1 Abstract

The purpose of the study was to learn bone shape features from spherical bone map of knee MRI images using established CNN and use these features to diagnose and predict OA. A bone segmentation model was trained on 25 manually annotated 3D MRI volumes to segment the Femur, Tibia, and Patella from 47,078 3D MRI volumes. Each bone segmentation was converted to a 3D point cloud and transformed into spherical coordinates. Different fusion strategies were performed to merge spherical maps obtained by each bone. A total of 41,822 merged spherical maps with corresponding KL grades for radiographic OA were used to train a CNN classifier model to diagnose OA using bone shape learned features. Several OA Diagnosis models were tested and the weights for each trained model were transferred to the OA Incidence models. The OA incidence task consisted of predicting OA from a healthy scan within a range of eight timepoints, from 1-year to 8-years. The validation performance was compared and the test set performance was reported. The OA Diagnosis model had an AUC of 0.905 on the test set with a sensitivity and specificity of 0.815 and 0.839. The OA Incidence models had an AUC ranging from 0.841 to 0.646 on the test set for the range from 1-year to 8-years. Bone shape was successfully used as a predictive imaging biomarker for OA. This approach is novel in the field of deep learning applications for musculoskeletal imaging and can be expanded to other OA biomarkers.

3.2 Introduction

Osteoarthritis is a degenerative joint disease which affects over 30 million U.S. adults, with the global prevalence of OA approaching 5%[1,60]. Risk factors commonly associated with OA include obesity, aging, and sex[61]. The onset of knee OA is manifested by several changes such as cartilage loss and changes in the meniscus. In addition to degeneration of soft tissues, it has been suggested that changes also occur in the subchondral and trabecular bone. The subchondral bone in particular interacts with the articular cartilage and softens the impact during normal and abnormal mechanical loading of the knee joint[62–64]. Both early-stage and late-stage changes to the subchondral bone are important components of the pathogenesis of OA.

Several investigators have previously proposed bone shape as an OA imaging biomarker, based on anthropometric measures, cross-sectional findings, or shape modeling of knees[14,17,65,66]. Studies based on 2D radiographs have reported sex-based bone shape differences in subjects with lateral and medial OA[67]. The classical approach to represent bone shape has been through SSM, which is a widely used tool to summarize shapes in a comprehensive feature vector. SSM has the ability to not only characterize complex shapes using PCA to reduce the data dimensionality, but also analyze shape differences without *a priori* assumptions, instead of identifying the geometrical features empirically. Furthermore, the 3-dimensional nature of MRI lends itself to SSM approaches and shows great potential in identifying knee OA risk factors and in studying disease pathogenesis; demonstrated in the large body of recent work[14,16,17,68,69]. This technique has also been used to evaluate the contribution of knee shape to ACL tears[18], in order to assess the association between bone shape and the progression of cartilage degeneration[44] as well as altered knee kinematics[19] after ACL reconstruction.

While previous studies show strong evidence of the critical role of the bone shape in the OA development and the ability of MRI and 3D shape modeling to quantify OA features, inferential statistics do not guarantee actual prediction abilities. Additionally, the use of unsupervised linear pattern decompositions as PCA for feature extraction do not guarantee the definition of a feature space that actually captures subtle differences able to characterize OA. The use of supervised feature learning and deep CNN architectures in medical image processing diagnostic tasks show promising results in fully exploiting the image information[70–72]. These techniques have dramatically improved outcomes of challenging problems in a variety of fields such as object detection, classification[73,74], drug discovery and genomics[40]. However, the number of validated applications in MRI and specifically in musculoskeletal imaging research remain limited[42,43,75].

This study aims to fill this gap by developing a knee bone shape feature extraction framework to explore the ability of established CNNs to extract and use knee bone shape features in diagnosing and predicting future incidence of radiographic OA based on Kellgren-Lawrence grade[9].

3.3 Methods

3.3.1 Methods overview

The overall study overview is summarized in **Fig. 3.1**. A bone segmentation model was trained and validated with a dataset of 40 manually segmented MRI volumes to segment the Femur, Tibia, and Patella from 47,078 3D MRI volumes (**Fig. 3.1A)**. Each of the segmented bone masks was converted to a 3D point cloud and rigidly registered to a reference point cloud to account for

rotational variability at scan time (**Fig. 3.1B**). The registered point clouds were then transformed into spherical coordinates and different fusion strategies were performed to merge spherical maps obtained by each bone (**Fig. 3.1C**). A total of 41,822 merged spherical maps with corresponding KL grades were used to train a classifier model to diagnose radiographic OA exclusively using bone shape learned features across all time points. For the OA diagnosis task, several models were tested and their validation performance was compared (**Fig. 3.1D**). The weights for each of these trained models were transferred to the OA Incidence models. The OA incidence task consisted of predicting future OA from the last healthy scan of a patient within a range of eight time points, from 1 year up to 8 years, and was tested on the same models as the OA diagnosis task (**Fig. 3.1E**).

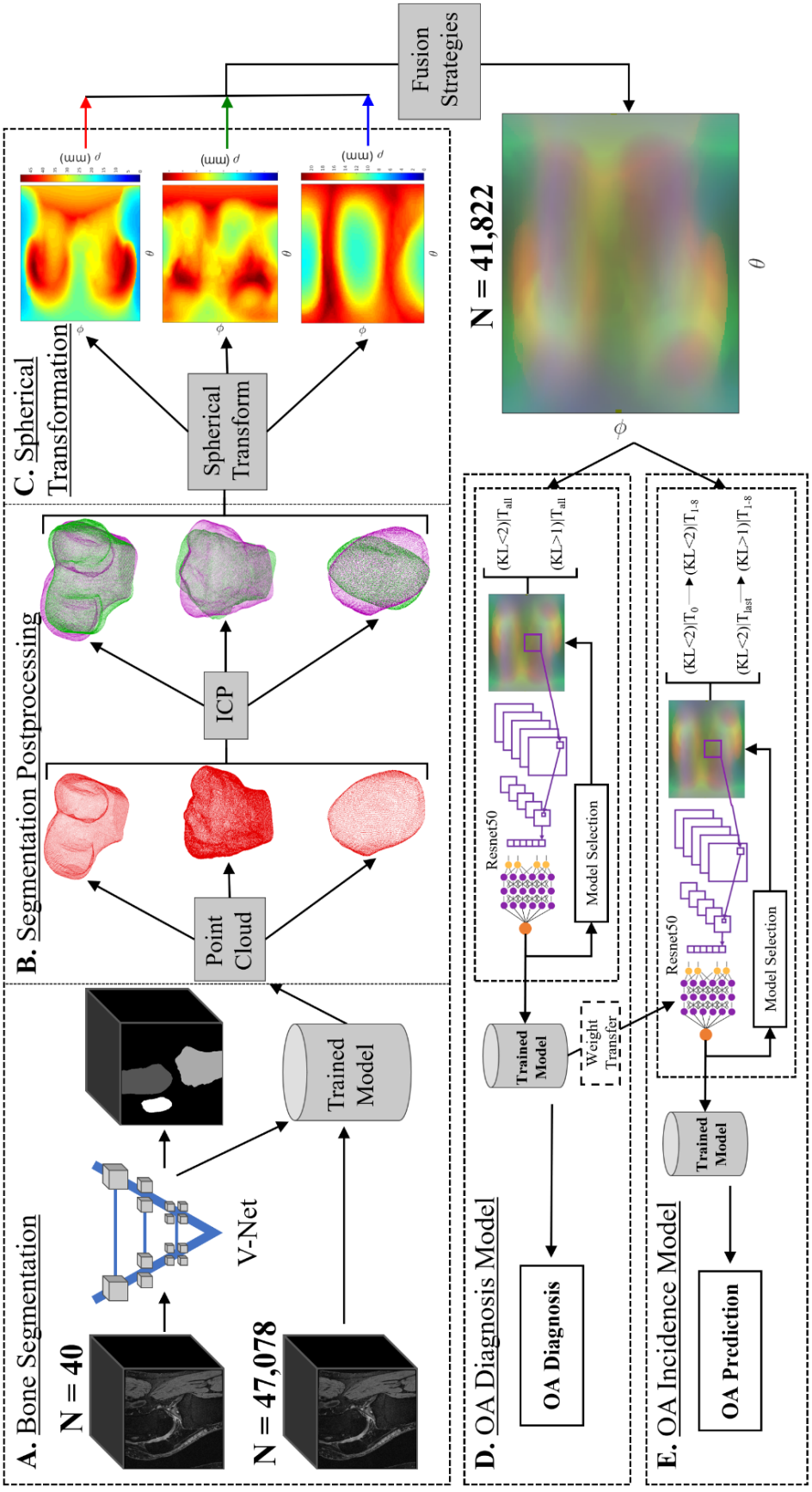

A. Bone Segmentation
N = 40
V-Net
N = 47,078
Trained Model
B. Segmentation Postprocessing
Point Cloud
ICP
C. Spherical Transformation
Spherical Transform
Fusion Strategies
N = 41,822
D. OA Diagnosis Model
Trained Model
Weight Transfer
Resnet50
Model Selection
(KL<2)|T_all
(KL>1)|T_all
OA Diagnosis
E. OA Incidence Model
Trained Model
Resnet50
Model Selection
(KL<2)|T_0 → (KL<2)|T_1-8
(KL<2)|T_last → (KL>1)|T_1-8
OA Prediction

Fig. 3.1 Overview of the study. (A), A V-Net segmentation model was trained and validated with a dataset of 40 3D DESS MRI volumes with the Femur, Tibia, and Patella segmented. The trained model was then used to run inference on 47,078 3D DESS MRI volumes from the OAI dataset. (B), The resulting bone segmentations were rigidly registered using an iterative closest point (ICP) algorithm to account for rotational variability at scan time. (C), The registered point clouds were transformed to spherical coordinates and merged using different fusion strategies. (D), A total of 41,822 spherical bone maps corresponding to patient scans were used to train an OA diagnosis model to classify OA based on bone shape across all time points. Each of the two inputs represents a class in the binary classifier (healthy KL<2 vs. OA KL>1). (E), An OA incidence model, defined as predicting future OA from the last healthy scan of patient within a range of eight time points, from 1 year up to 8 years, was trained using the weights from the OA diagnosis. The first input represents the baseline scans (T_0) from patients that never developed OA on either knee across the following 1 to 8 years ($T_{1\text{-}8}$). The second input represents OA incidence cases, as the last healthy scans (T_{last}) from patients that later developed OA on either knee across the following 1 to 8 years ($T_{1\text{-}8}$). The binary OA Incidence model is therefore represented as: baseline scans from always-healthy patients vs. last healthy scans from future OA patients in 1 to 8 years.

3.3.2 Patient imaging dataset

The imaging data for this study was acquired from the OAI, a multi-center longitudinal multi-modality imaging study in 4,796 patients[12]. This dataset consisted of a total of 12 time points ranging from an initial baseline visit to a final 108 month visit with yearly visits in between and a half-year visit for the third and fifth visits. Demographic data such as age, BMI and sex was recorded during each visit. Out of the 12 time points covered in the OAI, spanning 10 years, only 7 time points had MRI scans performed, which limited the span of the study to 8 years. A total of 41,822 sagittal 3D-DESS volumes from the OAI acquired (3.0T Siemens Trio) were used for this study (FOV = 14 cm; matrix = 384 x 307 x 160; TR/TE = 16.2/4.7 ms; bandwidth = 62.5 kHz; resolution = 0.365 x 0.456 x 0.7 mm). Selected patients had radiographs for both knees to evaluate their KL OA grade. The KL grades represent no OA (KL=0), minimal/doubtful OA (KL=1), mild OA (KL=2), moderate OA (KL=3), and severe OA (KL=4). For the purposes of this study, KL grades of 0 and 1 were determined to be healthy while KL grades of 2, 3, and 4 are considered to be OA.

Out of a total of 47,078 3D DESS volumes, 41,822 had corresponding KL grades and were included in this study. Out of this total, there were 4,506 unique patients, 117 of which only had scans for one of the knees and all the remaining had bilateral knee scan available in the dataset. The KL grade distribution for these 41,822 patients consisted of 16,624 (KL=0), 7,807 (KL=1), 10,240 (KL=2), 5,528 (KL=3) and 1623 (KL=4). The 3D DESS volumes were interpolated by the Siemens reconstruction software (Siemens Healthineers, Erlangen, Germany) from the original 384 x 304 x 160 acquisition resolution to 384 x 384 x 160 for sagittal in-plane isotropic resolution. Each of the 41,822 DESS image volumes used was cropped from 384 x 384 x 160 to 364 x 364 x 140 to remove extra background in the volumes. Each volume was then normalized from 0 to 1 by dividing the volume by its highest intensity.

3.3.3 Bone segmentation

The first step of the study was to accurately segment the bones from the 3D DESS volumes in the OAI dataset. A modified 3D V-Net[76] architecture was used for the Femur, Tibia and Patella bone segmentation (**Fig. 3.1A**). Lateral-medial flipping as well as in-plane rotation data augmentation was performed online to prevent overfitting, when training on a data split of 25 training, 5 validation and 10 testing volumes, for which the manual segmentation was available.

Subsequently, segmented Femur, Tibia, and Patella bones were post-processed to conform to the necessary format for the spherical transformation, such as maintaining the biggest connected component for each bone segmentation followed by morphological closing. The choices of

segmentation post-processing steps were strictly used as a way to sanitize or standardize the data and not to influence the performance of OA classification models. Given the size of the OAI, an additional validation of the bone segmentation accuracy was performed on 60 baseline scans sampled randomly from the OAI. The 60 additional test volumes were representative of the OAI demographic distribution and 30 of the baseline scans were from patients who never developed OA across the entire OAI on both knees. From the remaining scans, 15 were OA Incidence cases and 15 were OA Diagnosis cases. The osteophyte coverage of the bone segmentation network was also assessed. Further details on the architecture selected for the segmentation, adopted training strategies, automatic segmentation post-processing, the additional validation and osteophyte analysis are reported in **Appendix A.1 Bone segmentation and post-processing.**

3.3.4 Spherical transformation

Each post-processed segmented bone was converted to a 3D point cloud and converted to 2D spherical maps centered around the articular surface (**Fig. 3.2**). The transformation from Cartesian coordinates into spherical coordinates was performed by uniformly sampling 224 x 224 points in the point cloud and describing them based on the angle along the x-y plane from the positive x-axis (θ), the elevation angle from the x-y plane (φ) and the distance from the center of the point cloud to the sampled point in the surface (ρ) (**Fig. 3.2**). The angle θ was sampled from $-\pi$ to $+\pi$ while the angle φ was sampled from $-\pi/2$ to $+\pi/2$. Morphological closing was applied to the resulting spherical image to ensure there were no holes. The sampling density of 224 x 224 points, which was required to conform to the ImageNet image size, amounted to 50,000 points. This was an oversampling of the articular surface for each bone, which comprised

30% to 40% of the total points in each point cloud, with the Femur, Tibia, and Patella full point clouds containing on average 20,000, 70,000 and 90,000 points respectively.

Each of the point clouds was also augmented twice by rotating along the distal-proximal axis in a range of -5 to +5 degrees before the spherical transformation.

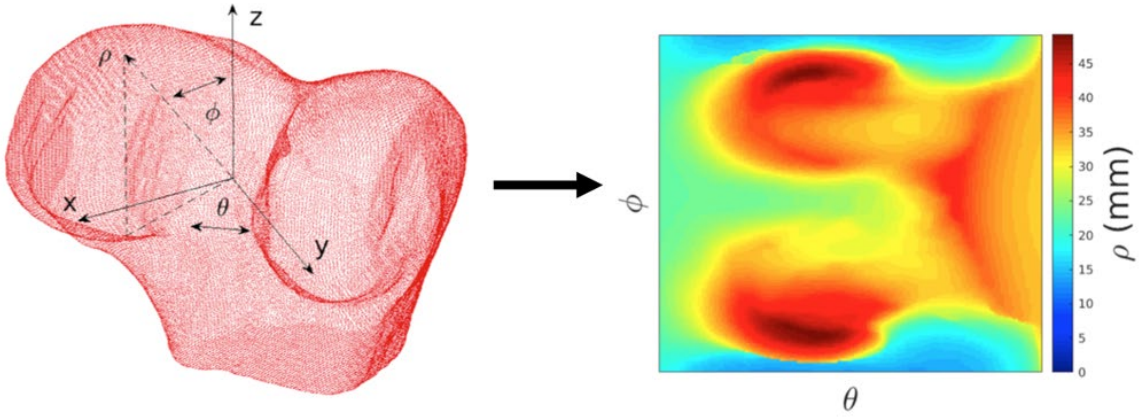


Fig. 3.2 Spherical transformation of the 3D bone point cloud. A femur point cloud is shown with the Cartesian and spherical coordinates. Each point in the surface of the 3D point cloud was transformed into a 2D point in a spherical map where the location was encoded with the two angles (θ, φ) and the distance from the centroid of the point cloud was encoded as the image intensity.

3.3.5 Spherical data formatting

The spherical images for each of the bones were normalized from 0 to 1 by dividing the intensity by the highest intensity for each of the bones. The rescaled spherical images for each patient were merged into an three channel image in the following four combinations: each of the three individual bone spherical maps was replicated three times and converted into a single knee bone

spherical image and the fourth variant was a merged combination of the three bones with the femur spherical image as the first channel, the tibia spherical image as the second channel and the patella spherical image as the third channel. This early fusion model was selected to learn complex features that arise from interactions of bone shape between the different bones in the knee joint. These combinations also allowed the ImageNet pretraining with 3-channel natural images. While the natural images in the ImageNet dataset are spatially correlated, the fourth fusion variant consisted of an artificial construct that contained imperfect spatial relationships between each different bone. The images were then further normalized to have a mean and standard deviation, respectively, of 0.485 and 0.229 for the red channel, 0.456 and 0.224 for the green channel and 0.406 and 0.225 for the blue channel to match the normalization values used for the pre-trained ImageNet[77] weights. This step also removed the bone size information from the spherical bone images, thus avoiding the potentially confounding relationship between bone size and patient sex. The spherical transformation process was validated on the test set used to evaluate the segmentation model by converting the ground truth segmentations into spherical coordinates and then transforming it back to Cartesian coordinates and calculating the distance differences between the closest points in the original. This validation ensured that the bone surface features were accurately represented in the spherical images. This method was iterated identically for the Tibia, Femur and Patella bones.

3.3.6 OA classification model dataset

The 41,822 spherical images were used for a model to diagnose OA and eight OA Incidence models. For the OA Diagnosis model, the dataset was divided into 29,012 training images, 6,365 validation images and 6,445 test images. The healthy controls were patient scans that had no

radiographic OA (KL<2) while the positive cases were patient scans with radiographic OA (KL>1). Both knee scans for each patient were randomly assigned to a single split while controlling for the demographic factors (age, BMI, sex). To test the independence of demographic factors for the positive cases across splits, two different statistical tests were performed. The independence of sex was tested with a Pearson's chi-squared test implemented in scikit-learn[78] using Python (Python Software Foundation, https://www.python.org/). The independence of age and BMI was tested with a one-way Multivariable Analysis of Variance (MANOVA) using a MATLAB implementation. For the OA Incidence models, the healthy controls were baseline patient scans from patients who never developed radiographic OA for both knees across all time points while the positive cases were the last healthy patient scan (KL<2) from patients who later developed radiographic OA. This study looked at eight incidence periods, ranging from one year to eight years for radiographic OA incidence. The training, validation, and test splits were randomized for every OA Incidence period (1-year to 8-year) to balance the classes across splits as well as ensure that the demographic factors were independent across splits. **Table 3.1** summarizes the training, validation and test set splits for all models, along with the P-values of the statistical tests showing independence of demographic factors.

Table 3.1 Training splits information for the bone segmentation, OA Diagnosis, and OA Incidence models. The training, validation, and test set splits were randomly picked into 62.5%, 12.5%, 25% ratios respectively for the bone segmentation and 70%, 15%, 15% ratios respectively for the OA models. The classes were increasingly imbalanced as the OA Incidence period increased due to the lower number of cases in the dataset. Demographic factors were controlled by testing for statistical independence across the splits using a Pearson's chi-squared test (χ^2) for the categorical sex variable and a one-way Multivariate Analysis of Variance for the joint effect of age and BMI. P-values are reported with significance defined as $P<0.05$.

Model	Training (Cases)	Validation (Cases)	Test (Cases)	Cases Ratio	χ^2 Test Correlation (Sex) (P-values)	MANOVA one-way Correlation (Age|BMI) (P-values)
Segmentation	25 (12)	5 (3)	10 (5)	0.500	0.573	0.327
Diagnosis	29012 (12027)	6365 (2753)	6445 (2611)	0.416	0.130	0.105
1-year	2444 (246)	537 (53)	524 (50)	0.101	0.159	0.298
2-year	2495 (297)	548 (64)	537 (63)	0.119	0.206	0.814
3-year	2389 (191)	527 (43)	517 (43)	0.0799	0.516	0.560
4-year	2397 (199)	527 (43)	517 (43)	0.0830	0.220	0.852
5-year	2356 (156)	514 (32)	506 (32)	0.0662	0.860	0.290
6-year	2373 (175)	519 (35)	510 (36)	0.0737	0.591	0.472
7-year	2269 (71)	500 (16)	489 (15)	0.0313	0.559	0.435
8-year	2275 (77)	502 (18)	492 (18)	0.0338	0.998	0.592

3.3.7 OA classification network implementation

Two binary classification models were trained to extract bone shape features from the spherical bone representations and use them to diagnose and predict OA. For the cross sectional OA

diagnosis task (**Fig. 3.1D**), a Resnet[38] architecture with 50 layers (Resnet50) pre-trained with ImageNet weights was implemented in PyTorch[79]. The selection of the Resnet architecture was informed through a CNN architecture grid search that included DenseNet[80], AlexNet[36], SqueezeNet[81] and Resnet. The DenseNet and Resnet architectures outperformed the other architectures and the decision to select the Resnet over the DenseNet was based on the smaller number of training parameters for the Resnet, which allowed a greater batch size. The ImageNet pre-training design choice was validated through a grid search, which included a version of the Resnet50 initialized with a Kaiming normal distribution[82]. The ImageNet pre-trained models achieved faster convergence than the models trained from scratch and consequently allowed for a more comprehensive parameter space search (shown in **Fig. 3.1D-E** as Model Selection) Different layer depths of the Resnet (18-layer, 34-layer, 50-layer, 101-layer, 152-layer) were also investigated with the 50-layer deep model providing the best compromise between accuracy and training speed, important for hyper-parameter optimization. The network architecture uses shortcut residual connections that improve the training performance for deeper models over similar shallower models. The basic structure of the Resnet50 follows the pattern of three convolutional layers with a 1 x 1, 3 x 3, and a 1 x 1 convolutional filter size respectively. Each of these layers is paired with batch normalization and a ReLU activation function. A softmax function was used to activate the last fully connected layer for the positive class.

The OA Diagnosis model was trained first with the following variants: femur, tibia, patella, early fusion, late fusion, logits averaging and majority voting. **Fig. 3.3** shows an overview of the different models used. The femur, tibia, and patella models consisted of three individual Resnet50 trained on each single knee bone spherical image (**Fig. 3.3A-C**). The early fusion model consisted of a Resnet50 trained on the combined spherical images of the femur, tibia, and

patella into a single merged spherical image (**Fig. 3.3D**). The late fusion model was the concatenation of the last 3 layers of the individual Resnet50 trained fused into a fully connected layer and trained end to end (**Fig. 3.3E**). There were two network ensemble methods evaluated: majority voting, where the majority, or median, prediction from all three individual bone network for each patient was used (**Fig. 3.3F**), and logits averaging, where the average of the softmax values outputted by each of the three individual bone networks was used for the prediction (**Fig. 3.3G**).

All OA Diagnosis model variants were initialized with ImageNet weights and fine-tuned using Adam optimizer with a learning rate of 1e-4 and trained end to end using a weighted binary cross entropy loss, based on the class imbalance, with a batch size of 100 in a GeForce GTX Titan 1080 Ti GPU. The OA Incidence models were initialized on the best performing checkpoint from the OA Diagnosis model based on the assumption that there is an overlap between the features for OA Diagnosis and OA Incidence. They were trained using the same parameters as the OA Diagnosis model with the exception of a lower learning rate of 1e-6 for Adam optimizer and a regularization weight decay value of 0.9 (to finetune while preventing overfitting on the training set) and trained for 100 epochs with a batch size of 32.

Network ensemble methods such as logits averaging, and majority voting were used to combine the outputs of the independent bone models. A late fusion model was created by concatenating the output of the last hidden layer of three individual Resnet50 architectures and performing a

global average pooling with a fully connected layer into a one-class softmax (sigmoid) activation function using Keras and a TensorFlow backend.

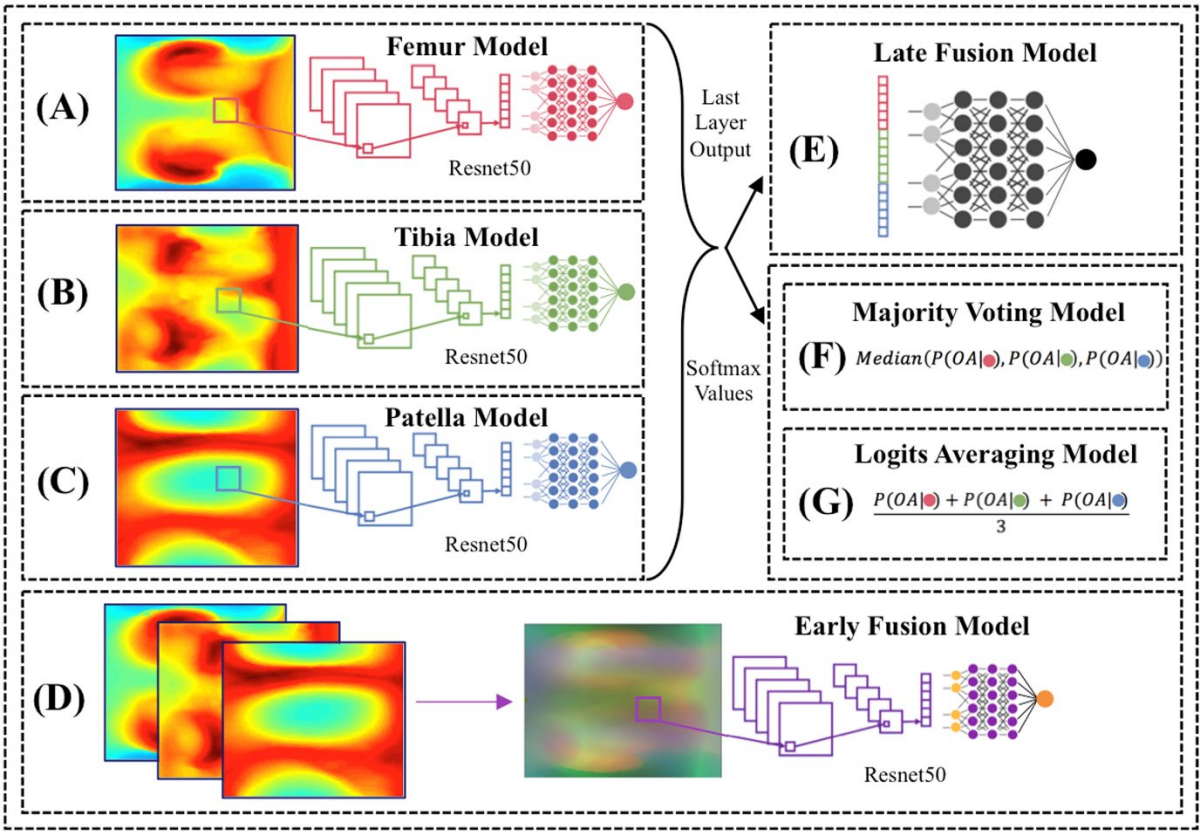


Fig. 3.3 Overview of the model fusion strategies. (A-C), The single bone fusion strategies, with the Femur, Tibia, and Patella shown in order, consisted of replicating the individual spherical bone maps three times and merging them into 3-channel images which were then used as inputs into a Resnet50 classification CNN. (D), The early fusion model merged each of the single bone spherical maps into a 3-channel image, which was then used as input into a Resnet50 classification CNN. (E), The late fusion model concatenated the last layer before the fully connected layer of the individual single bone models and added a fully connected layer that outputs a single softmax prediction for the OA diagnosis and incidence. (F), The first of the ensemble methods consisted of majority voting, where the majority predictions from the individual single bone models, (shown as red, green and blue circles corresponding to the Femur, Tibia, and Patella respectively) was used to determine the final OA diagnosis and prediction. (G), The logits averaging model consisted of averaging the softmax values from the individual single bone models and using the averaged softmax as the OA diagnosis and incidence.

3.3.8 OA classification robustness analysis

The robustness of the OA Diagnosis and first two OA Incidence models to bone atlas choice as well as bone segmentation and spherical transformation errors was evaluated.

The first robustness analysis of the OA classification models consisted of evaluating the impact of bone atlas choice on the performance of the OA Diagnosis and the 2-year and 8-year OA Incidence models. Four patients with different KL grades and demographic information were randomly picked as the bone atlas (for the femur, tibia and patella). The entire framework was rerun on each bone atlas and the OA Diagnosis and the 2-year and 8-year models were retrained using the same splits and hyperparameters as the original framework. The test set accuracy for each model was recorded for each bone atlas.

The second robustness analysis of the OA classification models consisted of evaluating the relationship between the bone segmentation accuracy and the performance of the OA Diagnosis and first two OA Incidence models. A randomly selected set of 30 correct predictions from the test set of the three models was corrupted and the effect of each individual bone corruption on the performance each model was evaluated.

The complete description of the first two analyses can be found in **Appendix A.2 OA Classification Robustness Analysis**.

The third robustness analysis of the OA classification models consisted of evaluating the relationship between the spherical transformation error and the performance of the OA Diagnosis and first two OA Incidence models. For this analysis, 50 correct predictions (25 true positives and 25 true negatives across all models) and 50 false predictions (25 false positives and 25 false negatives for OA Diagnosis, 38 false positives and 12 false negatives for the 1-year OA and 30 false positives and 20 false negatives for the 2-year OA) were selected from the trained OA Diagnosis model and the 1-year and 2-year OA Incidence models. The 1-year and 2-year OA Incidence models were evaluated due to the lack of cases in later year incidences. The distribution of spherical transformation errors measured as MPTS distance errors for the correct and the false predictions was calculated across bones for each model to evaluate the relationship between spherical transformation error and OA classification performance.

3.4 Results

3.4.1 Bone segmentation

The mean post-processed bone segmentation Dice scores for the test set of 10 patients were 97.15% (95% confidence interval = 96.56-97.74%) for the femur, 97.28% (95% confidence interval = 96.64-97.92%) for the tibia, and 95.99% (95% confidence interval = 95.26-96.72%) for the patella. MPTS distance errors were calculated between the manual and automated segmentations for the bone segmentation test set. The MPTS distance errors were 0.45 mm (95% confidence interval = 0.23-0.68 mm) for the Femur, 0.57 mm (95% confidence interval = 0.39-0.74 mm) for the Tibia and 0.51 mm (95% confidence interval = 0.07-0.94 mm) for the Patella, approximately the size of one voxel. **Fig. 3.4** shows representative slices of the 3D bone segmentation results from three different patients along with their respective MR images with the

mean MPTS distance errors over the entire volume. The two types of model error, false positives, where the segmentation misclassified non-bone regions as bone and false negatives, where the model missed the existing bone, are highlighted as cyan and magenta respectively. The complete results of the additional validation are shown in **Supp. Table A.1.** Additionally, the results of the osteophyte analysis are shown in **Supp. Fig. A.2** and **Supp. Fig. A.3**.

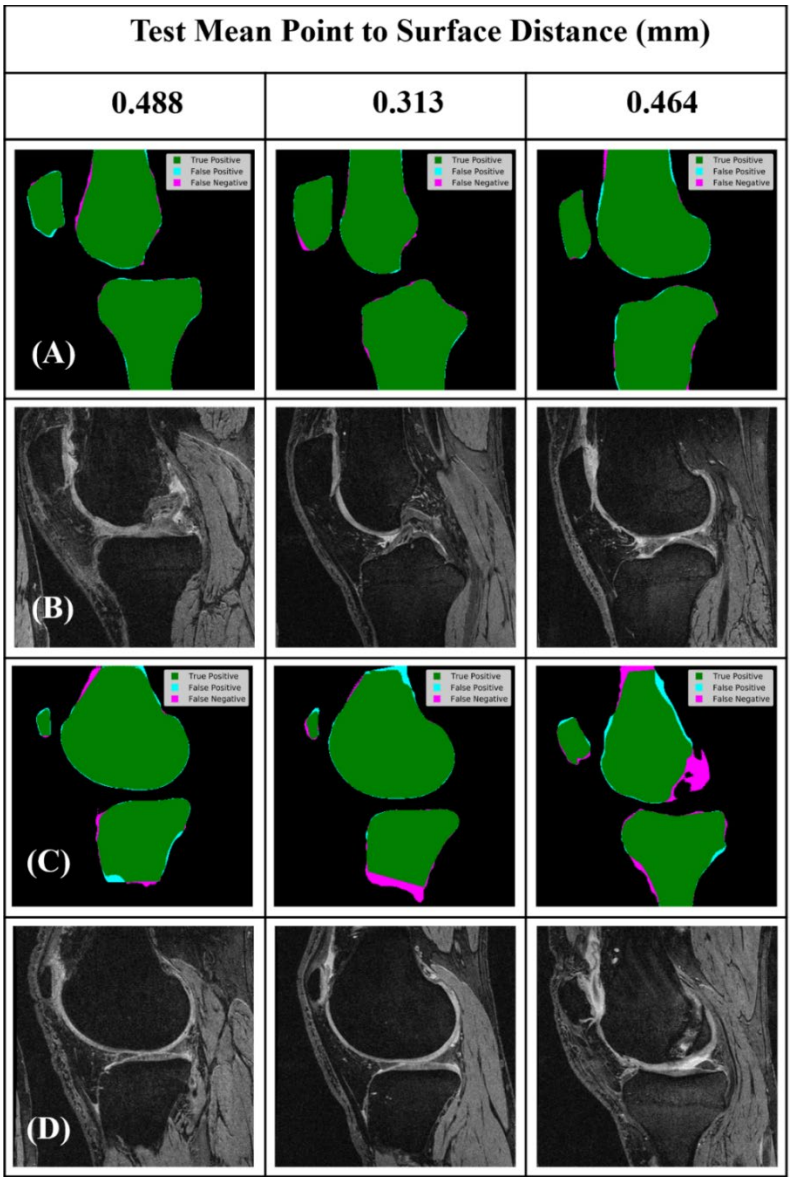


Fig. 3.4 Examples of bone segmentation errors for three scans from the bone segmentation test set with their respective total bone MPTS distance errors. The pixels in agreement between the trained segmentation model inference and the ground truths are labeled as green, representing the true positive cases. The pixels incorrectly classified as bone by the trained segmentation model are labeled as cyan, representing the false positive cases. The pixels missed by the trained segmentation model are labeled as magenta, representing the false negative cases. (A, B), Bone segmentations and corresponding DESS slices respectively for the three patients show minor errors along the bone surface for all three bones. (C, D), Bone segmentations and corresponding DESS slices shown respectively for the same three patients show more severe errors along the tibiofemoral shafts and the femoral intercondylar notch. These errors are likely caused by poor signal as the shaft appears sagittally and partial voluming effects in the intercondylar notch femoral region. The framework cropped the bone shaft and sparsely spherically sampled the intercondylar notch region, thus reducing the effect of these errors on the overall results.

3.4.2 Spherical transformation

The morphologically closed spherical transformation MPTS distance errors for the test set of 10 patients were 0.505 mm (95% confidence interval = 0.534-0.558 mm) for the Femur, 0.272 mm (95% confidence interval = 0.286-0.300 mm) for the Tibia, and 0.129 mm (95% confidence interval = 0.136-0.144 mm) for the Patella. The MPTS distance differences for the 10 patients in the segmentation test set were calculated by transforming the bone point clouds to the spherical coordinates and back to the bone point clouds and calculating the distance differences between the sampled points. The process was accurate at preserving the bone shape at most regions of the bones, except in the intercondylar notch, arguably where the surface curvature changed rapidly.

3.4.3 OA classification models

The validation Receiver Operating Characteristic (ROC) curve results for the binary OA classifier models are summarized in **Table 3.2** and a visual representation is reported in **Fig. 3.5** for the OA Diagnosis and the 1-year and 2-year OA Incidence models. The rest of the OA Incidence models ranging from 3-year to 8-year can be found in the **Supp. Fig. A.1**. For the OA Diagnosis task, the validation AUC for the models ranged from 0.806 to 0.904. The ensemble fusion strategies exhibited the best validation performance for the OA diagnosis task, with the logits averaging model slightly outperforming the majority voting model with a validation AUC of 0.904 and 0.903 respectively. The late and early fusion strategies had the next highest validation performance on average, with a validation AUC of 0.895 and 0.891 respectively. Out of the single bone fusion strategies, the femur model had the best OA diagnostic performance

with a validation AUC of 0.893 closely followed by the tibia model with a validation AUC of 0.887. The patella model had the lowest validation AUC of 0.806. For the OA incidence task, the validation AUC generally decreased with incidence time, however, the validation AUC was above 0.72 for the best fusion strategy across all incidence times, even for the lowest performing 5-year incidence model.

The test set performance of the models was in line with the validation set performance with the exception of the 7-year OA Incidence model, which significantly outperformed in the test set. The models were generally more sensitive to the positive cases with the exception of the 2-year and 8-year model. There was no clear trend in overall performance across each OA Incidence model, with the best performing OA Incidence being the 7-year model. The test set performance, measured in AUC, sensitivity and specificity, is summarized in **Table 3.3**. The OA Diagnosis and the 2-year and 8-year OA Incidence models test set performance for four different bone atlases was also consistent with the original results and is shown in **Supp. Table A.2.**

Table 3.2 Summary of the AUC validation performances from the different model fusion strategies for the OA Diagnosis and OA Incidence tasks.

Model	**AUC (Validation Set)**								
	Diagnosis	**1-year**	**2-year**	**3-year**	**4-year**	**5-year**	**6-year**	**7-year**	**8-year**
Patella	0.806	0.714	0.739	0.624	0.589	0.674	0.640	0.720	0.661
Tibia	0.887	0.756	0.758	0.739	0.694	0.602	0.664	0.639	0.669
Femur	0.893	0.821	0.801	0.738	0.771	0.697	0.729	0.687	0.658
Early Fusion	0.891	0.788	0.751	0.699	0.682	0.683	0.717	0.553	0.654

Late Fusion	0.895	0.778	0.791	0.731	0.760	0.676	0.679	0.698	0.680
Majority Voting	0.903	**0.846**	0.814	0.766	0.748	0.714	**0.746**	0.688	**0.741**
Logits Averaging	**0.904**	0.835	**0.832**	**0.776**	**0.778**	**0.724**	0.740	**0.728**	0.735

Table 3.3 Test set performance for the logits averaging ensemble model for the OA Diagnosis and OA Incidence tasks.

Metric	**Logits Averaging (Test Set)**								
	Diagnosis	1-year	2-year	3-year	4-year	5-year	6-year	7-year	8-year
AUC	0.905	0.818	0.815	0.733	0.764	0.751	0.781	0.841	0.646
Sensitivity	0.815	0.760	0.683	0.721	0.721	0.719	0.694	0.800	0.555
Specificity	0.839	0.751	0.759	0.679	0.696	0.633	0.639	0.656	0.582

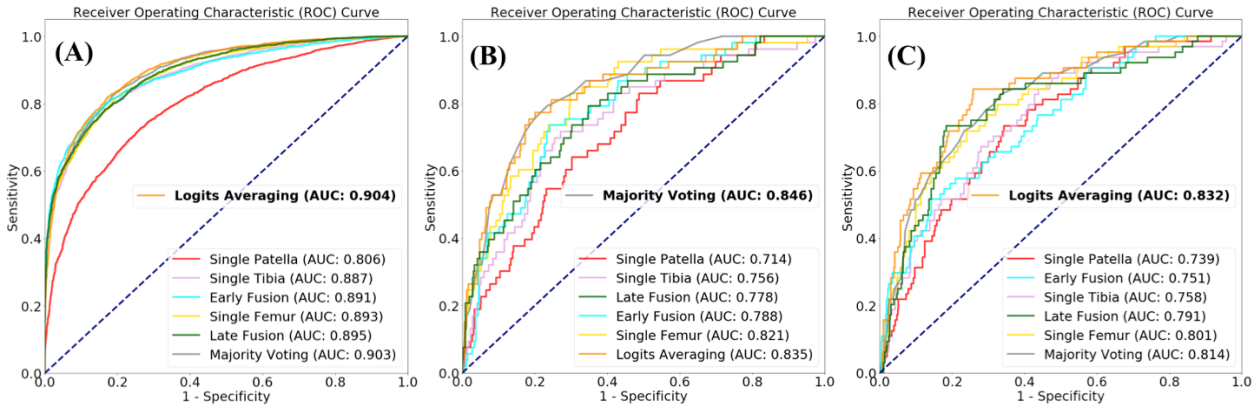


Fig. 3.5 Overview of the validation ROC curve comparisons for the different model fusion strategies. The OA Diagnosis model and the first two OA Incidence models are shown, with the remaining OA Incidence models are shown in the Supp. Fig. A.1. (A), OA Diagnosis model. (B), 1-year OA Incidence model. (C), 2-year OA Incidence model.

3.4.4 OA classification robustness analysis

The OA classification model robustness to bone segmentation accuracy measured in the range of the 95% confidence interval for the test set bone segmentation MPTS distance errors was calculated for the OA Diagnosis model and the 1-year and 2-year OA Incidence models. The OA Diagnosis model MPTS distance errors were 0.582 to 1 for the specificity, 1 to 1 for the sensitivity, and 0.999 to 1 for the AUC across all bones. The 1-year OA Incidence model MPTS distance errors were 0.491 to 0.942 for the specificity, 0.941 to 1 for the sensitivity, and 0.949 to 1 for the AUC across all bones. The 2-year OA Incidence model MPTS distance errors were 0.4 to 0.942 for the specificity, 0.933 to 1 for the sensitivity, and 0.911 to 0.996 for the AUC across all bones. The total MPTS distance errors for the analysis for each bone are shown in **Supp. Fig. A.4**, **Supp. Fig. A.5**, and **Supp. Fig. A.6**.

The complete results of both analyses can be found in the **Appendix A.2 OA classification robustness analysis**.

The OA classification robustness to spherical transformation error, measured in MPTS distance errors, overview is shown in **Fig. 3.6**. The OA Diagnosis model robustness to spherical transformation error is shown in **Fig. 3.6A**. The 1-year OA Incidence model robustness to spherical transformation error is shown in **Fig. 3.6B**. The 2-year OA Incidence model robustness to spherical transformation error is shown in **Fig. 3.6C**. There was no significant increase in spherical transformation MSTP distance error in the false predictions, both positive and negative, compared to the correct predictions.

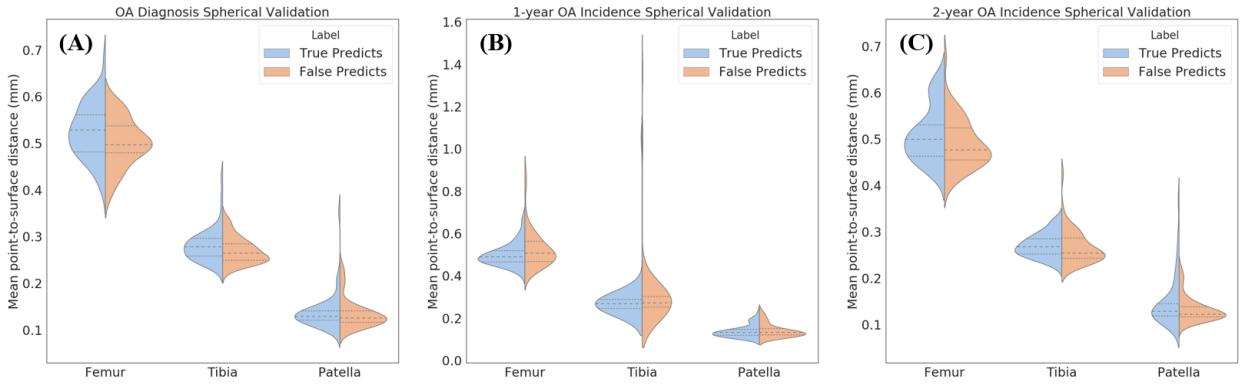


Fig. 3.6 Robustness of the OA classification models to the spherical transformation error measured as mean point-to-surface (MPTS) distance errors from the original point clouds. The average MPTS error, and corresponding 25% quartiles interval, is shown between 50 randomly picked correct predictions from the test set (shown in blue) and 50 randomly picked false predictions from the test set (shown in orange), for both positive and negative cases. There was no significant increase in spherical transformation MPTS distance error in the false predictions, both positive and negative, compared to the correct predictions. (A), OA Diagnosis model. (B), 1-year OA Incidence model. (C), 2-year OA Incidence model.

3.5 Discussion

In this study, we established a model to diagnose and predict knee OA onset within a period ranging from one year to eight years based on extracted bone shape features. The model generates the spherical maps of the Femur, Tibia, and Patella and combines them with a logits averaging network ensemble method to diagnose and predict radiographic knee OA. This model is state-of-the-art for radiographic knee OA diagnosis and OA incidence prediction using solely bone shape.

Classical methods used to represent bone shape based on SSM use PCA to reduce the dimensionality of the bone shape for analysis. This allows each component of the features vector

(mode) to describe a different aspect of the bone shape independent of the other components. The effect of each mode on the average surface can be modeled individually, synthesizing new instances. There are two shortcomings with this approach, the linearity constraint of PCA and the lack of supervision for the feature extraction process. Since PCA is a linear decomposition, the nonlinear relationships within the data are lost and the features described by the different modes may prove too simple to completely capture the bone shape. Furthermore, the unsupervised nature of PCA also means that the features extracted may not necessarily be specific to OA, since the bone shape features may depend on other factors such as demographics. Deep learning approaches address both of these issues by learning representations of data with multiple levels of abstraction, utilizing the fact that many natural image patterns are compositional hierarchies, meaning higher-level features can be decomposed into lower-level feature representations[83]. The hierarchical fashion of deep learning models suggests an improvement upon the established concept of simple data representation using PCA in favor of data-driven representation of relevant information directly from the raw data[83]. Some studies have combined supervised learning techniques such as linear discriminant analysis (LDA) with PCA to link bone shape to OA[14]. LDA best separates two groups (OA and no OA) with a hyperplane in multi-dimensional space, which further reduces the bone shape to a single scalar value representing the distance within the LDA vector for each bone shape. While LDA goes in the direction of adding some supervision to the feature extraction process, the usage of a single vector may be an over simplification of a complex 3D shape, and thus resulting in a robust but potentially less sensitive approach.

The purpose of our study was not to achieve the highest predictive performance in the OA Diagnosis and OA Incidence task, but rather to evaluate the effect of bone shape in the presence and onset of radiographic OA, while accounting for other confounding OA risk factors such as age, sex and BMI. Although the multifactorial nature of OA is well understood, and thus including several of these features together may lead to a more accurate prediction, the study of the single factors individually is also of great interest. This can help identify specific contributions of each factor to better understand the etiology of OA and help define unique OA phenotypes.

While this study brings new insights on the role of deep learning for new imaging OA biomarkers definition, some limitations need to be acknowledged. One of the limitations of the study is the use of radiographic OA based on KL grading as the metric for OA. Radiographic OA measures changes such as tibiofemoral, or joint space, narrowing and osteophyte formation, which occur at more advanced OA stages. This could potentially mean that the last healthy scans considered for the OA Incidence models could already be exhibiting other more subtle OA symptoms, such as loss of cartilage thickness. Another limitation of the study is the small number of OA Incidence cases prevented any further stratification of the OA Incidence models by KL grade increase to better understand the distribution of these OA Incidence subpopulations. The temporal efficacy for the OA Incidence models is also affected by the reshuffling of the splits across incidence periods. A future study could focus on a smaller section of the incident population and follow it across time points. Additionally, since the KL grading is performed on a coronal knee radiograph, only tibiofemoral OA is considered in the diagnosis and the impact of patellofemoral OA is not included in the grading, which could explain the lower performance for

the patella models. Another limitation of this method is the reduced model interpretability when compared to a PCA approach, which could model the modes and understand the relationship between specific bone shapes and OA. The current model would not be able to evaluate the correlation between specific bone shape differences such as tibia slope and the OA diagnosis and OA incidence prediction, but rather assess the general relationship between bone shape and OA. For future studies, using visualization tools, such as Grad-CAM[59], could characterize different bone shape phenotypes for the OA diagnosis and OA incidence tasks. Establishing such a way to phenotype patient bone shape populations could have wide implications in clinical studies for potential treatment of OA as a patient screening tool.

Chapter 4: Spherical Encoding for Multimodal Quantitative MRI OA Biomarker Fusion and Feature Learning

4.1 Abstract

The purpose of the study was to learn features from spherical encoding of multimodal quantitative MRI images using CNNs and use them to diagnose knee OA. Two segmentation model ensembles for bone and cartilage were trained to segment the femur, tibia, and patella bones and cartilage. The trained models were used to segment 21,118 3D-DESS MRI volumes. Bone shape and cartilage thickness maps were obtained from the segmentations. T_2 values were fitted after registering 3D-DESS cartilage masks to matching 2D-MSME spin-echo MRI volumes from a complimentary dataset. Each 3D biomarker map within the cartilage mask was projected onto the articular bone surface and transformed into spherical coordinates. Six different strategies were investigated to merge biomarker spherical maps per bone. The merged spherical maps with corresponding KL grades for radiographic knee OA were used to train a CNN classifier model to diagnose OA. Pairwise McNemar's tests were used to compare the different merging strategies. The single biomarker OA diagnosis models had mean and standard deviation of 86.4±0.1 for test AUC, 70.9±0.2 for sensitivity, 86.0±0.1 for specificity. When considering the biomarkers together, the respective OA diagnosis performance was 87.9±0.1, 73.2±0.2 and 86.1±0.1. Significant performance improvements (p-value=1e-4) were observed when biomarkers were considered simultaneously compared to individually. The performance of each single biomarker and biomarker fusion models generally improved from patella to tibia, and up to femur. The combination of individual OA biomarkers improved the OA diagnosis accuracy over single biomarkers.

4.2 Introduction

Multimodal MRI (MMRI) leverages the ability of MRI to investigate both anatomic and physiologic changes associated with pathologies[84]. Compositional MRI techniques, such as quantitative relaxometry, can discover early signs of a disease that precede subsequent physical manifestations observed with structural imaging techniques[30]. The use of MMRI information can generate robust models of disease onset and progression, with significant applications including: cancer, neuro, and musculoskeletal imaging[85–91].

While MMRI models offer a comprehensive look into a particular pathology, they suffer from challenges such as the spatial colocalization of biomarkers across modalities, due to resolution differences, and increased data dimensionality[92]. While the usage of data from multiple modalities has the potential of improving disease characterization and trajectory prediction, it comes at the cost of reduced model interpretability, since each added modality further obfuscates the relationship between imaging biomarkers and the disease of interest. Common ways to overcome these limitations include reducing the dimensionality of the data by aggregating biomarker values before analysis, such as averaging quantitative mapping values within a clinically relevant anatomic region[93,94]. This strategy exploits *a priori* clinical information about the disease to simplify the fusion of modalities and increases model interpretability, at the expense of data granularity, statistical power, and consequently, model performance.

Morphological features such as shape have been linked to tumor growth[95] and cartilage degeneration[32]. A common dimensionality reduction strategy for these types of features involves

handcrafting geometrical features[96]. However, *a priori* definitions of feature importance are often arbitrarily defined and can overlook relevant information.

The advent of supervised feature learning and deep CNN architectures in medical image diagnostic tasks show promising results in fully exploiting the image information by learning the most relevant data representation for the specific task considered[70–72]. These techniques have dramatically improved outcomes of challenging problems in a variety of fields such as object detection[73], classification[74], drug discovery[97], and genomics[40].

However, the use of deep learning methods often involves a tradeoff between model interpretability and prediction power. Additionally, the risk of model overfitting in applications of MMRI, when the data is considered in its raw form and the number of training examples is relatively low, is often too high and thus hampers the usage of deep learning models. The loss of interpretability can be dampened by occlusion studies that aim to understand the deep learning model performance. Similarly, the risk of model overfitting can be addressed with several regularization techniques at different steps of the data processing pipeline, such as image augmentation and normalization, dropout, and loss penalties, among others. Still, the choice of strategy for MMRI studies between deep learning and more classical feature handcrafting, with all the drawbacks discussed above, is not obvious.

This study aims to address the challenges of model interpretability and overfitting by proposing a spherical encoding method that directly colocalizes *a priori* multimodal imaging biomarkers,

reducing the dimensionality of the data, while allowing for the application of data-driven feature extraction to learn, in a supervised fashion, the best representation of the multimodal data, along with interactions between different biomarkers.

An application to knee OA will explore the ability of deep learning CNNs to learn features from these spherical maps for diagnosing radiographic OA based on KL grade[9]. We hypothesize that the biomarker fusion predictive performance will improve upon the single biomarker predictive performance.

4.3 Methods

4.3.1 Patient imaging dataset

The imaging data for this study was acquired from the OAI, a multi-center longitudinal multimodality imaging study in 4,796 patients[12]. This dataset consisted of a total of 12 time points ranging from an initial baseline visit to a final 108 month visit with yearly visits in between and a half-year visit for the third and fifth visits. Demographic data such as age, BMI, and sex, was recorded during each visit. Out of the 4,796 unique patients, 4,416 had valid T_2, bone, and cartilage biomarker data and were thus included in this study. The T_2 compositional biomarker data was required, in addition to the bone and cartilage morphological biomarker data, in order to investigate both aspects of the OA disease process. The KL grade distribution for the 21,118 3D-DESS volumes consisted of 8,103 (KL=0), 3,972 (KL=1), 5,335 (KL=2), 2,897 (KL=3) and 811 (KL=4). Selected patients had radiographs for right knees to evaluate their KL OA grade. The KL grades represent no OA (KL=0), minimal/doubtful OA (KL=1), mild OA

(KL=2), moderate OA (KL=3), and severe OA (KL=4). For the purposes of this study, KL grades of 0 and 1 were determined to be healthy while KL grades of 2, 3, and 4 are considered to be OA. Out of the 12 time points covered in the OAI, spanning 10 years, only 7 time points had MRI scans performed, which limited the span of the study to 8 years. Furthermore, not all patients had the same number of timepoints available, with some patients having a single timepoint and some having all 7 timepoints.

A total of 21,118 3D-DESS volumes (FOV = 14 cm; matrix = 384 x 307 x 160; TR/TE = 16.2/4.7 ms; bandwidth = 185 Hz/pixel; resolution = 0.456 x 0.3646 x 0.7 mm) and corresponding 2D-MSME spin-echo T_2 volumes (FOV = 12 cm; matrix = 384 x 269 x 21; TR/TE = 2700/10,20,30,40,50,60,70 ms; bandwidth = 250 Hz/pixel; resolution = 0.313 x 0.446 x 3 mm, slice gap = 048 mm, chemical shift = 1.8 pixels) acquired from the OAI (3.0T Siemens Trio) were used for this study.

4.3.2 Methods overview

The overall study overview is summarized in **Fig. 4.1**. A bone and a cartilage segmentation model ensemble were trained on 72 and 148 manually segmented 3D-DESS volumes to segment the femur, tibia, and patella bones and corresponding cartilage. The trained models were used to segment 21,118 3D-DESS volumes (**Fig. 4.1A)**. Bone shape feature and cartilage thickness maps were obtained from the segmented masks. T_2 values were calculated by registering 3D-DESS cartilage masks to the matching 2D-MSME MRI volumes and performing parametric T_2 fitting on the cartilage. Each biomarker was projected onto the articular bone surface (**Fig. 4.1B**) and

transformed into spherical coordinates. Six different strategies were performed to merge spherical maps for each bone (**Fig. 4.1C**). A total of 21,118 merged spherical maps with corresponding KL grades were used to train classifier models to diagnose radiographic OA. A different model was trained and tested for each biomarker model, for a total of 18 models (**Fig. 4.1D**).

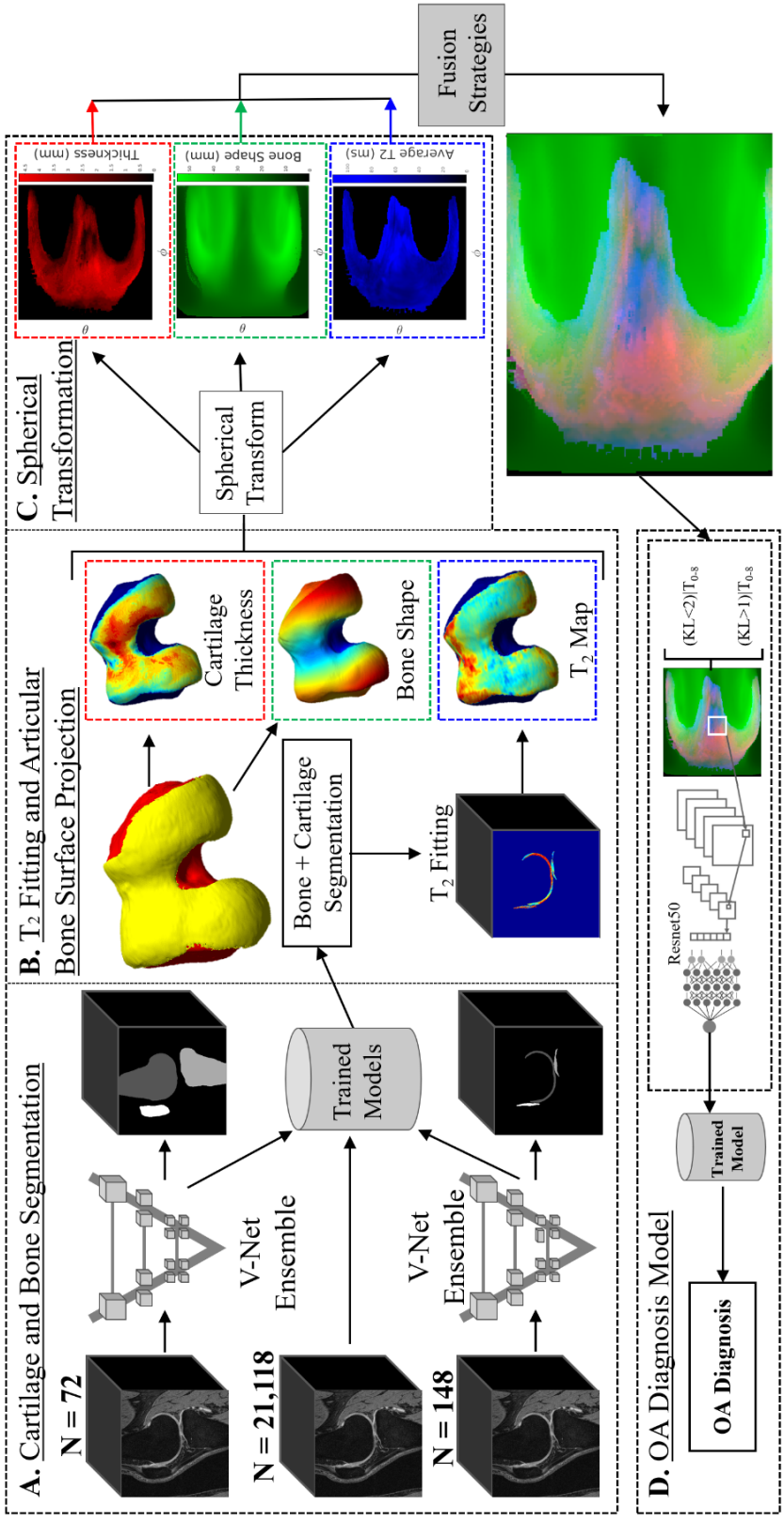

A. Cartilage and Bone Segmentation
N = 72
V-Net Ensemble
Trained Models
N = 21,118
V-Net Ensemble
N = 148
B. T2 Fitting and Articular Bone Surface Projection
Bone + Cartilage Segmentation
T2 Fitting
Cartilage Thickness
Bone Shape
T2 Map
C. Spherical Transformation
Spherical Transform
Thickness (mm)
Bone Shape (mm)
Average T2 (ms)
Fusion Strategies
D. OA Diagnosis Model
Resnet50
(KL<2)|T0-8
(KL>1)|T0-8
Trained Model
OA Diagnosis

Fig. 4.1 Overview of the study. (A) A bone and a cartilage segmentation model ensemble were trained on 72 and 148 manually segmented 3D-DESS volumes to segment the femur, tibia, and patella bones and corresponding cartilage. The trained models were used to segment 21,118 3D-DESS volumes. (B) Bone shape feature and cartilage thickness maps were obtained from the segmented masks. T_2 values were calculated by registering 3D-DESS cartilage masks to the matching 2D-MSME MRI volumes and performing parametric T_2 fitting on the cartilage. Each biomarker was then projected onto the articular bone surface, where each point contained information from each biomarker. (C) The articular bone surface projections were transformed into spherical coordinates. Six different strategies were performed to merge spherical maps per bone. (D) A total of 21,118 merged spherical maps with corresponding KL grades were used to train classifier models to diagnose radiographic OA using the biomarker learned features. A different model was trained and tested for each biomarker strategy per bone, for a total of 18 OA diagnosis models. Each of the two inputs into the OA diagnosis models represents a class in the binary classifier (healthy KL<2 vs. OA KL>1).

4.3.3 Image pre-processing

The 3D-DESS volumes were interpolated by the Siemens reconstruction software (Siemens Healthineers, Erlangen, Germany) from the original 384 x 304 x 160 acquisition resolution to 384 x 384 x 160 for sagittal in-plane isotropic resolution. Similarly, the 2D-MSME volumes were interpolated from 384 x 269 x 21 acquisition resolution to 384 x 384 x 21 for sagittal in-plane isotropic resolution. Each of the 21,118 3D-DESS volumes used was cropped from 384 x 384 x 160 to 364 x 364 x 140 to remove extra background in the volumes. Each volume was then normalized in the [0-1] range by dividing each volume by its 99th percentile highest intensity to remove bright artifacts.

4.3.4 ***Bone and cartilage segmentation***

The first step of the study was to accurately segment the bone and cartilage from the 3D-DESS volumes in the OAI dataset. An ensemble of five 3D V-Net[76] architectures, each trained with different distance-weighted loss functions[98], was used for the femur, tibia and patella bone

segmentation (**Fig. 4.1A**). A full description of the bone segmentation models can be found in **Appendix B.1 Bone segmentation**.

For the cartilage segmentation, an ensemble of three 2D V-Nets and three 3D V-Nets were trained to segment femoral, tibial, and patellar cartilage and menisci (**Fig. 4.1A**). A full description of the cartilage segmentation models can be found in **Appendix B.2 Cartilage segmentation**. This model was also extensively validated in a previous study[20].

4.3.5 Morphometry

The cartilage thickness was calculated for each of the three cartilage masks per sagittal slice using a Euclidean distance transform along the morphological skeleton of each mask. The morphological skeleton was defined as the middle points along the length of each cartilage mask. The distance transform provided the distance from each skeleton point to the edge of the cartilage, which was doubled to obtain the cartilage thickness. For full details of this automatic cartilage thickness method, we refer to a previous study[20]. The bone shape was intrinsically described by the distance from the bone surface of each bone mask to its volumetric centroid. This method was presented and validated in the previous chapter.

4.3.6 Relaxometry

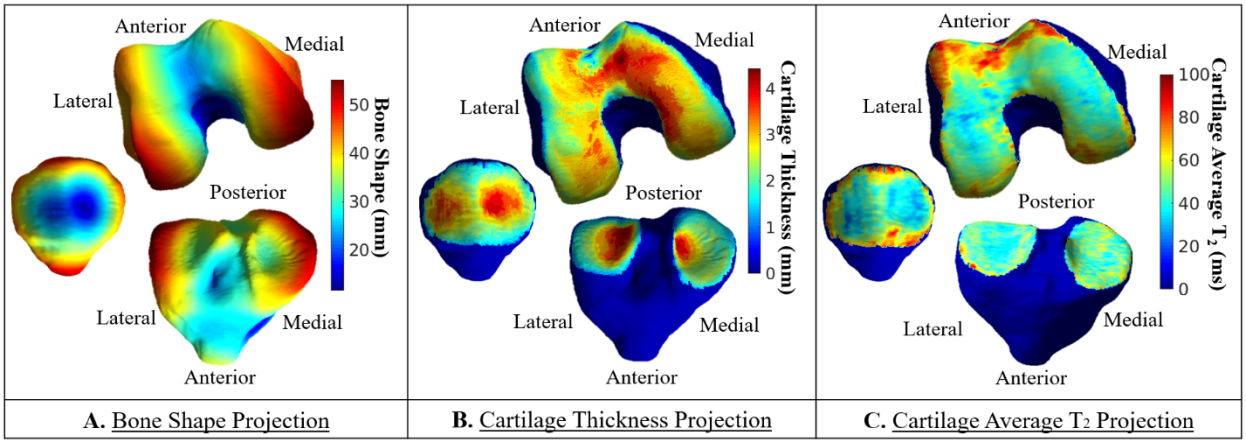
In order to colocalize the three imaging biomarkers considered for this study (bone shape, cartilage thickness and cartilage T_2 relaxation times), the 2D-MSME volumes were rigidly

registered to the 3D-DESS volumes using the Patient Coordinate System (PCS) in the DICOM metadata of both MRI scans. The sagittal in-plane and coronal slice resolution of the 2D-MSME volumes were first matched to the 3D-DESS volumes using bicubic interpolation. The registration was performed using the first echo volume, and the resulting transformation was applied to all echoes. Once the resolutions were matched, the 2D-MSME sagittal slices were spatially shifted to match the 3D-DESS sagittal slices to create MSME-DESS registered volumes. The automatically segmented cartilage mask from the 3D-DESS cartilage segmentation model was then used to isolate the cartilage from the newly registered 2D-MSME. The cartilage T_2 relaxation time values were then computed on the masked 2D-MSME echoes using a three-parameter, Levenberge-Marquardt mono-exponential: $(S(TE) \alpha \exp(-TSL/T_2) + C)$.

4.3.7 Bone surface projection

The tibia and femur bone masks were cropped along the shaft in order to be invariant to the different shaft lengths. The bone and cartilage masks were converted from voxel masks to 3D point clouds, using a marching cubes algorithm implemented in MATLAB, and each 3D biomarker map within the cartilage point cloud was then projected onto the articular bone surface (**Fig. 4.2**). This step mapped each point in the articular surface to a value from each of the three biomarkers: bone shape, cartilage thickness and cartilage T_2 relaxation time values. The bone shape was defined as the distance from the centroid of the bone point cloud to the bone surface (**Fig. 4.2A**). The calculated cartilage thickness of the overlying cartilage was projected to each perpendicular point in the articular bone surface (**Fig. 4.2B**). The superficial, deep, and total average T_2 values for the corresponding section of the cartilage used during the thickness projection were projected to each perpendicular point in the articular bone surface. The

superficial and deep subdivisions of the cartilage used for the T_2 averaging were defined as the respective top and bottom halves of the cartilage, with **Fig. 4.2C** showing the total average T_2 value projection. The projection from the cartilage to the bone surface was calculated using the intersection between the normal vector for each point in the bone surface and the cartilage maps. This normal vector spanning from each point in the bone surface formed a cylinder with a radius of 0.729 mm, empirically set to double the in-plane pixel resolution, that averaged the cartilage thickness and cartilage T_2 values along the cartilage cross-section it covered. This sampling radius addressed the imbalance between the number of bone surface points and cartilage points, and ensured that the bone surface projection was dense.



A. Bone Shape Projection | **B.** Cartilage Thickness Projection | **C.** Cartilage Average T_2 Projection

Fig. 4.2 Articular bone surface biomarker projection. The bone and cartilage masks were converted from voxel masks to 3D point clouds and each 3D biomarker map within the cartilage point cloud was then projected onto the articular bone surface. This step mapped each point in the articular surface to a value from each of the three biomarkers: bone shape, cartilage thickness and T_2 relaxation time values. (A) The bone shape was defined as the distance from the centroid of the bone point cloud to the bone surface. (B) The calculated cartilage thickness of the overlying cartilage was projected to each perpendicular point in the articular bone surface. (C) The total average T_2 values for the corresponding section of the cartilage used during the thickness projection were projected to each perpendicular point in the articular bone surface. The superficial and deep T_2 value projections, not shown here, were calculated using subdivisions of the cartilage used for the T_2 averaging, defined as the respective top and bottom halves of the cartilage.

4.3.8 Spherical transformation

The three biomarkers projected to the articular bone surface were converted to 2D spherical maps (**Fig. 4.3**). The transformation from Cartesian coordinates into spherical coordinates was performed by uniformly sampling 224 x 224 points in the point cloud, to conform to the ImageNet[77] image size for pretraining, and describing them based on the angle along the x-y plane from the positive x-axis (θ), the elevation angle from the x-y plane (φ) and the distance from the center of the point cloud to the sampled point in the surface (ρ) (**Fig. 4.3A**). The angle θ was sampled from $-\pi$ to $+\pi$ for all bones while the angle φ was sampled from $-\pi/2$ to $+\pi/8$ for the femur and tibia and from $-\pi/2$ to $+\pi/8$ for the patella. Bicubic interpolation was performed between the sampled points to create densely sampled spherical maps. The sampling was designed to be centered around the articular surface to ensure the cartilage would be centered for each bone (**Fig. 4.3B-D**). The sampling density of 224 x 224 points was an oversampling of the articular surface for each bone, which comprised 30% to 40% of the total points in each bone point cloud, with the femur, tibia, and patella full bone point clouds containing on average 20,000, 70,000 and 90,000 points respectively.

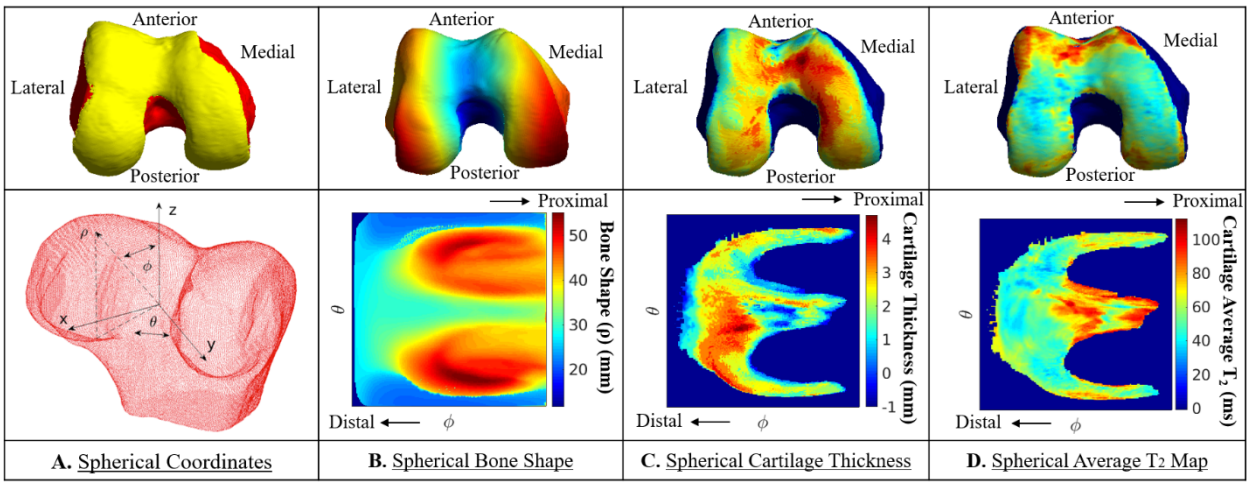


Fig. 4.3 Biomarker 2D spherical maps. The three biomarkers projected to the articular bone surface were converted to 2D spherical maps. (A) The transformation from Cartesian coordinates into spherical coordinates was performed by uniformly sampling 224 x 224 points in the point cloud and describing them based on the angle along the x-y plane from the positive x-axis (θ), the elevation angle from the x-y plane (φ) and the distance from the center of the point cloud to the sampled point in the surface (ρ). The angle θ was sampled from -π to +π for all bones while the angle φ was sampled from -π/2 to +π/8 for the femur and tibia and from -π/2 to +π/8 for the patella. The sampling was designed to be centered around the articular surface to ensure the cartilage would be centered for each bone. (B) Bone shape 2D spherical map. (C) Cartilage thickness 2D spherical map. (D) Cartilage average T_2 value 2D spherical map.

4.3.9 Spherical data formatting

The spherical images were group normalized by the minimum and maximum biomarker value from each bone for all the patients. The normalized spherical images for each patient were merged into three-channel 8-bit images, with the six strategies shown for the femur in **Fig. 4.4**. The choice of three-channel images leveraged pre-trained CNN models on the ImageNet dataset and was motivated by a previous study[99].

The first three strategies consisted of the single biomarkers: cartilage thickness, bone shape, and cartilage T_2. The *cartilage thickness* strategy consisted of the cartilage thickness spherical maps

replicated three times into a spherical image (**Fig. 4.4A**). The *bone shape* strategy consisted of the bone shape spherical maps replicated three times into a spherical image (**Fig. 4.4B**). The *cartilage T_2* strategy consisted of the deep, superficial, and average T_2 spherical maps as the first, second, and third channels respectively, (**Fig. 4.4C**). The last three strategies consisted of the biomarker fusions: morphological cartilage and bone fusion, morphological and compositional cartilage fusion and all biomarkers fusion. The *morphological cartilage and bone fusion* consisted of the cartilage thickness and bone shape spherical maps as the first and second channels respectively, with the last channel empty (**Fig. 4.4D**). The *morphological and compositional cartilage fusion* consisted of the deep and superficial T_2 spherical maps as the first and second channels respectively with the third channel consisting of the cartilage thickness spherical map (**Fig. 4.4E**). The *all biomarkers fusion* consisted of the cartilage thickness, bone shape, and average T_2 spherical map as the first, second and third channels respectively (**Fig. 4.4F**).

The spherical maps were directly colocalized for each bone, with each point describing the same geometric location in the articular surface. This colocalization allowed the model to learn local features that arise from interactions between the different biomarkers across the same bone. Each channel was normalized separately. To illustrate for the morphological and compositional cartilage fusion (**Fig. 4.4E**), a pixel in the spherical image with elevated T_2 values for both the deep and superficial cartilage layers as well as cartilage thinning could have a 3-channel value of (204, 204, 26), which would be a dark yellow. Another pixel in the same spherical image with elevated T_2 values for the superficial cartilage layer with average cartilage thickness and T_2

values for the superficial cartilage layer could have a 3-channel value of (128, 204, 128), which would be a dark green.

The images were then further normalized to have a mean and standard deviation, respectively, of 0.485 and 0.229 for the red channel, 0.456 and 0.224 for the green channel and 0.406 and 0.225 for the blue channel to match the normalization values used for the pre-trained ImageNet weights.

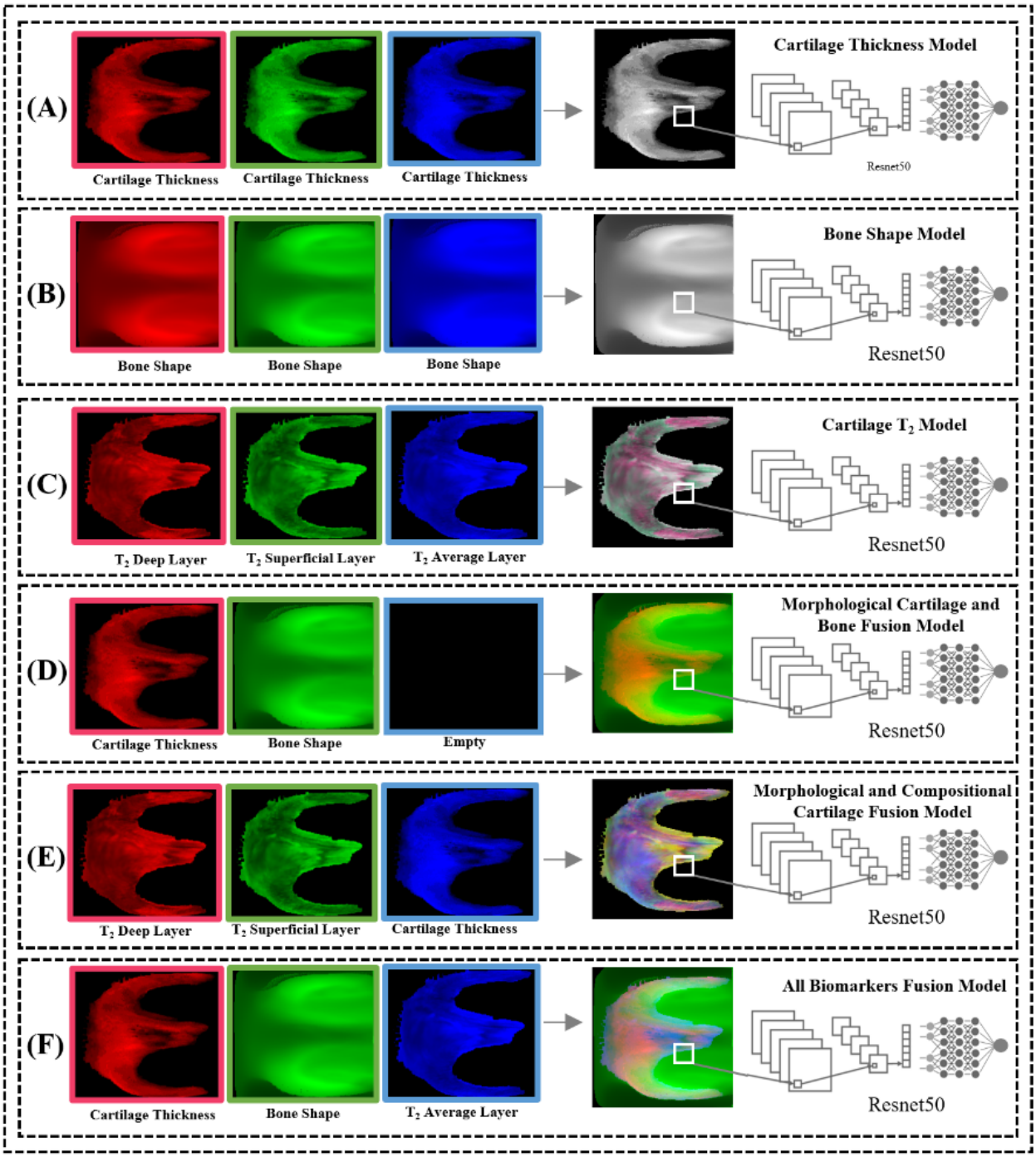


Fig. 4.4 Overview of the biomarker model strategies, shown for the femur. The normalized spherical images for each patient were merged into a three-channel 8-bit image. (A-C) The first three strategies consisted of the single biomarkers: cartilage thickness, bone shape, and cartilage T_2. (A) The cartilage thickness strategy consisted of the cartilage thickness spherical maps replicated three times into a spherical image. (B) The bone shape strategy consisted of the bone shape spherical maps replicated three times into a spherical image. (C) The cartilage T_2 strategy consisted of the deep, superficial, and average T_2 spherical maps as the first, second, and third

channels respectively. (D-F) The last three fusion strategies consisted of the biomarker fusions: morphological cartilage and bone fusion, morphological and compositional cartilage fusion and all biomarkers fusion. (D) The morphological cartilage and bone fusion consisted of the cartilage thickness and bone shape spherical maps as the first and second channels respectively, with the last channel empty. (E) The morphological and compositional cartilage fusion consisted of the deep and superficial T_2 spherical maps as the first and second channels respectively with the third channel consisting of the cartilage thickness spherical map. (F) The all biomarkers fusion consisted of the cartilage thickness, bone shape, and average T_2 spherical map as the first, second and third channels respectively.

4.3.10 OA classification model dataset

The 21,118 spherical images were used to train a model to diagnose OA. The dataset was divided into 12,634 training images, 2,558 validation images and 5,926 test images, with no patient overlap across splits. The healthy controls were patient scans that had no radiographic OA (KL<2) while the positive cases were patient scans with radiographic OA (KL>1). Right knee scans for each patient were randomly assigned to a single split while controlling for the demographic factors (age, BMI, sex). To test the independence of demographic factors for the OA cases across splits, two different statistical tests were performed. The independence of sex was tested with a Pearson's $\chi 2$ test implemented in scikit-learn[78] using Python (Python Software Foundation, https://www.python.org/). The independence of age and BMI was tested with a one-way MANOVA using a MATLAB implementation. **Table 4.1** summarizes the training, validation and test set splits for the bone segmentation and OA diagnosis models, along with the p-values of the statistical tests showing independence of demographic factors.

Table 4.1. Training, validation, and test splits information for the bone segmentation and OA diagnosis models. The training, validation, and test set splits were randomly picked into 55%, 15%, 30% ratios respectively for the bone segmentation, and 58%, 12%, 28% ratios respectively for the OA diagnosis model. Demographic factors were controlled by testing for statistical independence across the splits using a Pearson's $\chi 2$ test for the categorical sex variable and a one-way MANOVA for the joint effect of age and BMI. Bold p-values are significant (p-value < 0.05).

Task	Model	Training (Cases)	Validation (Cases)	Test (Cases)	Cases Ratio	χ^2 Test Correlation (Sex) (p-values)	MANOVA one-way Correlation (Age|BMI) (p-values)
Segmentation	Bone	57 (29)	15 (8)	30 (16)	0.52	0.75	0.41
Classification	OA Diagnosis	12,634 (5,402)	2,558 (1,111)	5,926 (2,530)	0.43	0.12	0.19

4.3.11 OA classification network implementation

A total of 18 binary classification models, one for each biomarker strategy per bone, were trained to extract biomarker features from the spherical biomarker representations and use them to diagnose OA (**Fig. 4.1D**). A Resnet[38] architecture with 50 layers (Resnet50) pre-trained with ImageNet weights was implemented in PyTorch[79]. The choice of architecture and hyperparameters was informed by our previous study on the relationship between bone shape and radiographic OA[99]. The Resnet50 network architecture uses shortcut residual connections that improve the training performance for deeper models over similar shallower models. The basic structure of the Resnet50 follows the pattern of three convolutional layers with a 1 x 1, 3 x 3, and a 1 x 1 convolutional filter size respectively. Each of these layers is paired with batch normalization and a rectified linear unit activation function. Additionally, a dropout rate of 0.15 was used to improve generalizability of the model during training, randomly turning off activations at a rate of 15%.

All OA diagnosis model variants were initialized with ImageNet weights and fine-tuned using Adam optimizer with a learning rate of 1e-5 with a regularization weight decay value of 0.9, in order to finetune while preventing overfitting on the training set. The training was performed for 100 epochs with an early stopping 15-epoch patience for validation loss non-improvement over the best validation loss reached. The models were also trained end to end using a weighted binary cross entropy loss, based on the class imbalance, with a batch size of 300 in a Tesla V100 32GB GPU.

The OA diagnosis models were trained using the different biomarker strategies outlined in **Fig. 4.4**. The OA diagnosis models for each biomarker strategy were ensembled across the bones by averaging the softmax values outputted by each network. Therefore, each of the six biomarker models had a total of five predictive values: for the patella, for the tibia, for the femur, for the averaged predictive values of the tibia and femur, and for the average predictive values of all three bones. For the averaged ensembles, each anatomical region contributes equally to the final prediction.

4.4 Results

4.4.1 Bone and cartilage segmentation

The mean segmentation dice scores and their corresponding 95% confidence intervals (CI95) for the bone and cartilage respective test sets of 30 and 28 patients are shown in **Table 4.2**. The MPTS distance errors were also calculated between the manual and automated segmentations for both test sets, shown in **Table 4.2**.

Table 4.2. Summary of the bone and cartilage segmentation test set performances, shown both as Dice scores and MPTS distance errors, with their corresponding 95% confidence intervals.

Segmentation Model (n = test #)	Class	Dice Scores (Mean ± CI95)	MPTS (mm) (Mean ± CI95)
Bone (n = 30)	Femur	98.0% ± 0.32%	0.406 ± 0.051
	Tibia	98.0% ± 0.26%	0.390 ± 0.047
	Patella	96.4% ± 0.70%	0.370 ± 0.055
Cartilage (n = 28)	Femoral	90.0% ± 0.74%	0.247 ± 0.021
	Tibial	88.6% ± 1.3%	0.223 ± 0.036
	Patellar	85.7% ± 2.5%	0.555 ± 0.194

The bone segmentation mean test dice scores with corresponding CI95 were 98.0% ± 0.32%, 98.0% ± 0.26%, and 96.4% ± 0.70% for the femur, tibia, and patella respectively. These were a 0.8%, 0.7%, and 0.4% improvement over the respective mean test dice scores in our previous study[99]. The bone segmentation mean test MSTP distance errors with corresponding CI95 were 0.406 ± 0.051 mm, 0.390 ± 0.047 mm, and 0.370 ± 0.055 mm for the femur, tibia, and patella respectively. These were a 0.044 mm, 0.18 mm, and 0.14 mm improvement over the respective mean MPTS distance errors in our previous study[99].

The cartilage segmentation mean test dice scores with corresponding CI95 were 90.0% ± 0.74%, 88.6% ± 1.3%, and 85.7% ± 2.5% for the femoral, tibial, and patellar cartilage respectively. The cartilage segmentation mean test MPTS distance errors with corresponding CI95 were 0.247 ± 0.021 mm, 0.223 ± 0.036 mm, and 0.555 ± 0.194 mm for the femoral, tibial, and patellar cartilage respectively.

The cartilage segmentation results were further validated in a previous study[20] beyond the 28 patients in the test set by calculating the cartilage thickness of 4,129 patients with corresponding manual cartilage thickness measurements for the femur and tibia. **Fig. 4.5** shows representative slices of the 3D bone and cartilage segmentation results from three different patients along with their respective MR images with the mean MPTS distance errors over the entire volume. The pixels in agreement between the trained segmentation model inference and the ground truths are labeled as green, representing the true positive cases. The two types of model error, false positives, where the segmentation misclassified non-bone or non-cartilage regions and false negatives, where the model missed the existing bone or cartilage, are highlighted as cyan and magenta respectively.

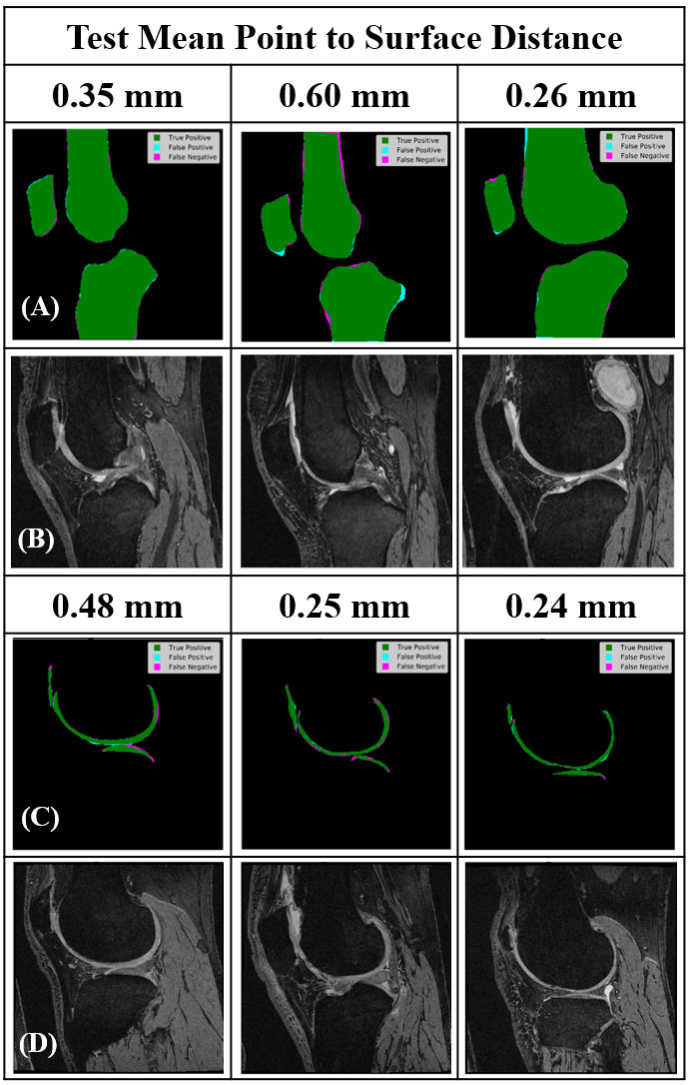


Fig. 4.5. Examples of bone and cartilage segmentation errors for three patients from the respective bone and cartilage segmentation test sets. Representative slices of the 3D bone and cartilage segmentation are shown along with their respective 3D-DESS images with the mean MPTS distance errors over the entire volume. The pixels in agreement between the trained segmentation model inference and the ground truths are labeled as green, representing the true positive cases. The two types of model error, false positives, where the segmentation misclassified non-bone or non-cartilage regions and false negatives, where the model missed the existing bone or cartilage, are highlighted as cyan and magenta respectively. (A, B) Bone segmentations and corresponding 3D-DESS slices for the three patients show minor errors along the articular bone surface for all three bones. The errors present can be observed along the femoral and tibial shaft, as well as the distal facet of the patella. (C, D) Cartilage segmentations and corresponding 3D-DESS slices shown for three different patients shows diffuse segmentation errors along the cartilage. Both of these errors are likely caused by signal heterogeneity and partial voluming effects. Only the articular bone surface was sampled during the spherical transformation, reducing the effect of certain bone segmentation errors along the shaft and intercondylar notch on the overall results.

4.4.2 Spherical transformation validation

The spherical transformation method was validated over the dataset for both the average cartilage thickness and the average cartilage T_2 time values for each bone. **Fig. 4.6** shows Bland-Altman plots comparing the original average values of cartilage thickness and cartilage T_2 values to the spherically transformed average values for each bone. The differences between the average biomarker values were calculated using the original average values as a reference, by subtracting the original average values from the average spherical values for each biomarker. Most differences for the average cartilage thickness ranged from -0.051 to 0.216 mm for the femoral cartilage, from 0.039 to 0.410 mm for the tibial cartilage, and from -0.14 to 0.35 mm for the patellar cartilage. These average cartilage thickness deviations between the original and spherically transformed average data are within the in-plane pixel resolution for the 3D-DESS volumes. Most differences for the average cartilage T_2 values ranged from -5.37 to 4.67 ms for the femoral cartilage, from -2.79 to 3.95 ms for the tibial cartilage, and from -6.27 to 3.27 ms for the patellar cartilage. Overall, the spherical transformation was accurate at preserving the biomarkers at most regions of the articular surface of each bone.

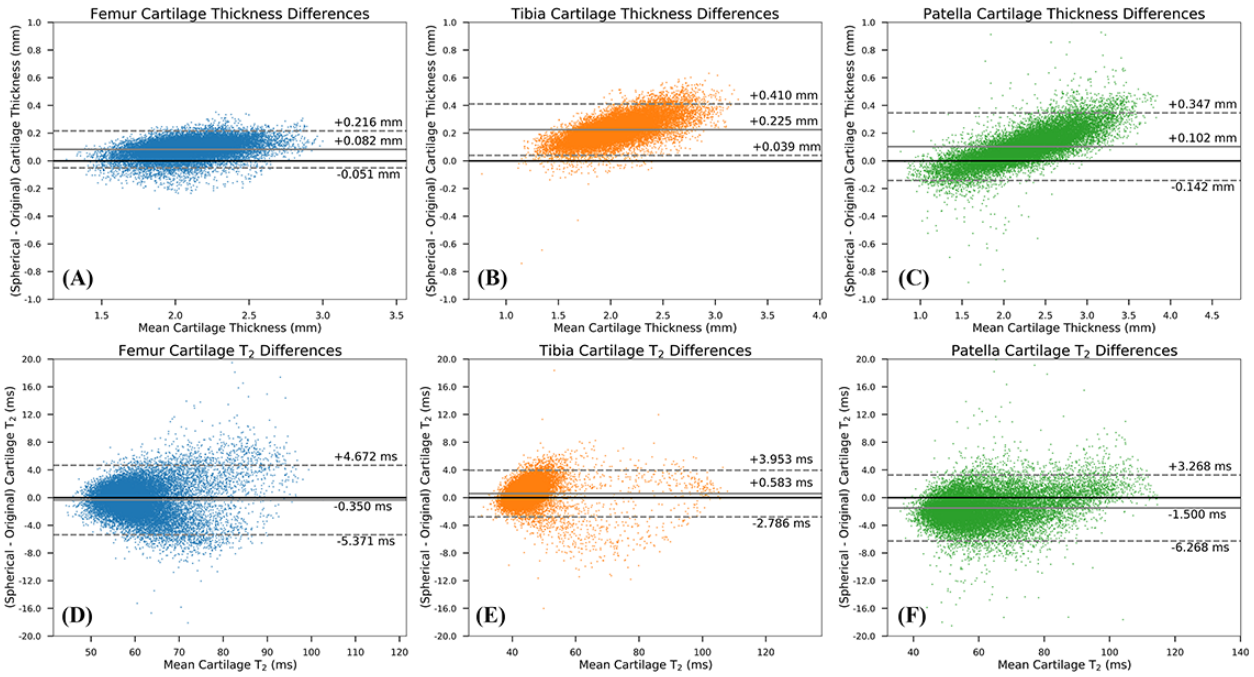


Fig. 4.6 Bland-Altman plots comparing the original average values of cartilage thickness and cartilage T_2 to the spherically transformed average values for each bone. The differences between the average biomarker values were calculated using the original average values as a reference, by subtracting the original average values from the average spherical values for each biomarker. The solid black line represents the zero difference. The solid gray line represents the mean difference and the dashed gray lines represent two standard deviations above or below the mean. (A) Differences between average spherical cartilage thickness and average original cartilage thickness for the femur. (B) Differences between average spherical cartilage thickness and average original cartilage thickness for the tibia. (C) Differences between average spherical cartilage thickness and average original cartilage thickness for the patella. (D) Differences between average cartilage T_2 values and average original cartilage T_2 values for the femur. (E) Differences between average cartilage T_2 values and average original cartilage T_2 values for the tibia. (F) Differences between average cartilage T_2 values and average original cartilage T_2 values for the patella.

4.4.3 OA diagnosis models

The test ROC curve results, defined as the sensitivity, the specificity, and AUC for the binary OA diagnosis models, along with their respective CI95, are summarized in **Table 4.3**. The ROC metrics are given for each single biomarker and biomarker fusion OA diagnosis models for each bone, as well as the softmax averaging ensembled results across the tibia and femur (TF), and all three bones (PTF). The single biomarker OA diagnosis models had an average test AUC with

standard deviation of 86.4 ± 0.09, with a sensitivity and specificity of 70.9 ± 0.2 and 86.0 ± 0.1 respectively. The biomarker fusion OA diagnosis models had a test AUC with standard deviation of 87.9 ± 0.1, with a sensitivity and specificity of 73.2 ± 0.2 and 86.1 ± 0.1 respectively. On average, the sensitivity, specificity and AUC improved from the single biomarker models to the biomarker fusion models.

Table 4.3 Bootstrapped (n=100) test set OA diagnosis ROC performance for all six biomarker models per bone, as well as two different ensembles across the bones. Sensitivity, specificity, and AUC values are shown respectively, along with their corresponding 95% confidence intervals. The best performances per bone and ensembling strategy are bolded. PTF = Patella + Tibia + Femur ensemble. TF = Tibia + Femur ensemble

Biomarker Type	Biomarker Model	Test Set ROC (Sensitivity/Specificity/AUC) (Mean ± CI95)				
		Patella	**Tibia**	**Femur**	**PTF**	**TF**
Single	Cartilage T_2	67.5 ± 0.18 73.9 ± 0.16 77.6 ± 0.12	70.0 ± 0.20 85.3 ± 0.12 86.0 ± 0.10	75.5 ± 0.16 81.5 ± 0.14 86.0 ± 0.10	77.2 ± 0.15 87.5 ± 0.12 89.9 ± 0.08	75.6 ± 0.17 86.5 ± 0.12 89.2 ± 0.08
	Cartilage Thickness	68.1 ± 0.17 72.7 ± 0.16 77.0 ± 0.12	68.5 ± 0.20 86.7 ± 0.12 85.5 ± 0.10	69.4 ± 0.19 90.9 ± 0.09 89.0 ± 0.08	73.7 ± 0.17 90.8 ± 0.10 90.6 ± 0.08	72.3 ± 0.19 91.2 ± 0.09 90.3 ± 0.07
	Bone Shape	62.2 ± 0.20 81.2 ± 0.13 78.3 ± 0.11	67.0 ± 0.19 91.6 ± 0.09 87.9 ± 0.09	73.1 ± 0.16 86.3 ± 0.10 88.5 ± 0.08	71.2 ± 0.17 91.9 ± 0.10 89.9 ± 0.08	72.4 ± 0.16 91.5 ± 0.10 90.4 ± 0.08
Fusion	Morphological Cartilage and Bone Fusion	55.3 ± 0.19 88.0 ± 0.11 80.8 ± 0.10	71.7 ± 0.17 89.6 ± 0.09 89.6 ± 0.08	72.5 ± 0.17 90.0 ± 0.10 90.1 ± 0.07	72.9 ± 0.17 93.1 ± 0.09 91.7 ± 0.07	74.3 ± 0.16 92.6 ± 0.09 91.8 ± 0.07
	Morphological and Compositional Cartilage Fusion	67.0 ± 0.18 76.7 ± 0.14 78.5 ± 0.12	78.0 ± 0.16 76.8 ± 0.14 86.1 ± 0.09	**75.0 ± 0.17** **83.6 ± 0.12** **87.7 ± 0.08**	78.6 ± 0.17 85.4 ± 0.11 89.5 ± 0.09	78.6 ± 0.16 83.2 ± 0.13 89.5 ± 0.08
	All Biomarkers Fusion	**64.3 ± 0.17** **83.0 ± 0.13** **81.0 ± 0.11**	**76.4 ± 0.16** **86.0 ± 0.12** **89.8 ± 0.07**	76.3 ± 0.13 85.5 ± 0.12 89.2 ± 0.09	**78.2 ± 0.16** **89.6 ± 0.10** **91.7 ± 0.07**	**78.8 ± 0.16** **87.7 ± 0.12** **91.7 ± 0.07**

The results of the biomarker fusion OA diagnosis models were compared to the results of the single biomarker OA diagnosis models using pairwise McNemar's tests. The McNemar's test calculates the probability that the performance of two binary classifiers is different, based on the proportion of misclassification errors, both false positives and false negatives. The pairwise

McNemar's tests were performed for every combination of single biomarker and biomarker fusion models, as well as their ensembled performance, as shown in **Table 4.4**. For the *morphological cartilage and bone fusion*, consisting of the bone shape and cartilage thickness biomarkers, all the biomarker fusion OA diagnosis models per bone, as well as the ensembled models across bones, were different from the single biomarker OA diagnosis models, with the exception of the *cartilage T_2* and *bone shape* models on the tibia. For the *morphological and compositional cartilage fusion*, consisting of the superficial and deep cartilage T_2 and the cartilage thickness biomarkers, there was a difference for the *cartilage T_2* models on the tibia, femur, and TF models. There was also a difference for the *cartilage thickness* model on the TF model. Finally, there was a difference observed for the *bone shape* model on the tibia model. For the *all biomarkers fusion*, consisting of the cartilage T_2, cartilage thickness and bone shape biomarkers, the biomarker fusion OA diagnosis models per bone, as well as the ensembled models, were all different from the single biomarker OA diagnosis models, with the exception of the *cartilage T_2* and *bone shape* models on the tibia, the *cartilage thickness* and *bone shape* models on the femur, and the *bone shape* model on the TF model.

Table 4.4 McNemar's test p-values between the single biomarker and the biomarker fusion strategies based on the test set ROC performance. Bold p-values are significant (p-values < 0.05). T_2 = Cartilage T_2 biomarker model. Thk = Cartilage Thickness biomarker model. Bone = Bone Shape biomarker model. PTF = Patella + Tibia + Femur ensemble. TF = Tibia + Femur ensemble

Biomarker Fusion Model	Patella (p-values)			Tibia (p-values)			Femur (p-values)			PTF (p-values)			TF (p-values)		
	T_2	Thk	Bone	T_2	Thk	Bone	T_2	Thk	Bone	T_2	Thk	Bone	T_2	Thk	Bone
Morphological Cartilage and Bone Fusion	**1e-4**	**4e-3**	**1e-4**	0.06	**1e-4**	0.98	**1e-4**	**1e-4**	**1e-3**	**1e-4**	**1e-4**	**1e-4**	**1e-4**	**1e-4**	**7e-3**
Morphological and Compositional Cartilage Fusion	0.23	0.12	0.44	**0.04**	0.40	**1e-4**	**1e-3**	0.94	0.87	0.21	0.61	0.78	**4e-3**	**0.02**	0.90
All Biomarkers Fusion	**1e-4**	**4e-3**	**1e-4**	0.08	**1e-4**	0.98	**0.03**	0.36	0.23	**1e-4**	**1e-4**	**1e-4**	0.01	0.05	1.0

4.5 Discussion

In this study, we established a biomarker fusion spherical encoding method and applied it to diagnosing knee OA. The model generates spherical maps of OA imaging biomarkers for the femur, tibia, and patella to diagnose radiographic knee OA. This model is novel for radiographic knee OA diagnosis using solely a combination of three clinically relevant biomarkers, bone shape, and cartilage thickness and cartilage T_2 relaxation time values.

Classical approaches used to analyze MMRI data generally either reduce the multimodal data dimensionality by aggregating *a priori* imaging biomarkers[100], or register the original multimodal data to a reference atlas prior to voxel-based statistical analysis[101]. The first approach improves the model interpretability while reducing the data granularity, which limits its potential findings due to the loss of more subtle, local biomarker interactions. The latter approach suffers from the challenge inherent in registering different modalities, such as nonrigid geometric deformation and data interpolation. Furthermore, both approaches typically rely on linear regression methods for the data analysis, which miss the more complex nonlinear relationships within the data.

Deep learning approaches address both these issues by learning nonlinear representations of raw data with multiple levels of abstraction, utilizing the fact that many natural image patterns are compositional hierarchies, meaning higher-level features can be decomposed into lower-level feature representations[83]. However, deep learning-based, fully data-driven approaches to MMRI can suffer from model overfitting, due to the large number of parameters needed to describe each

modality, as well as the common lack of labeled training data in MMRI studies. Our proposed framework, combines the benefits of classical feature handcrafting and deep learning approaches, while avoiding their respective shortcomings of loss of data granularity and reduced interpretability. By exploiting *a priori* clinical information to define the imaging biomarkers, the multimodal data dimensionality can be drastically reduced compared to a fully data-driven approach while still retaining enough data for the analysis. Furthermore, the ability of the proposed framework to directly colocalize MMRI biomarkers within a single image representation can leverage the power of CNNs to extract local semantic features arising from interactions between each biomarker.

It is worth nothing that the purpose of the framework was not to achieve the highest predictive performance diagnosing OA, but rather to evaluate the individual and combined effect of the biomarkers in the presence of radiographic OA. Given the multifactorial nature of OA, the study of biomarker crosstalk at different stages of the disease, such as the longitudinal relationship between cartilage thickness and subchondral bone shape, is of great clinical interest. The proposed framework can help identify individual as well as combined contributions of each biomarker to better understand the etiology of OA and help define unique OA phenotypes.

While this spherical encoding method is highlighted in musculoskeletal imaging with knee OA, it can be extended to other clinical challenges in cancer, cardio and neuroimaging. The spherical transformation is based on the assumption that structures of interest are spheroidal, which is the case for the femur, tibia, and patella in the knee. There are also numerous spheroidal structures in

the human body such as the brain, liver, heart, and even tumors, for which this technique can be applied. In the particular case of neuroimaging, morphological biomarkers like the cortical thickness and subcortical volume of the brain have been linked to neurodegenerative disorders such as Alzheimer's disease[102,103]. The method proposed by this study could be adapted to such an application, provided there is a cortical brain segmentation, and combined with other functional imaging biomarkers in the brain cortex from modalities like fMRI. Such an application could yield interesting studies looking at the spatiotemporal relationship between these morphological and functional imaging biomarkers throughout the disease onset and development.

Although this study brings new insights on the role of deep learning for MMRI biomarker fusion, some limitations need to acknowledged. One of the limitations of the study is the use of radiographic OA based on KL grading as the clinical definition for OA. Radiographic OA measures changes such as tibiofemoral, or joint space, narrowing and osteophyte formation, which occur at more advanced OA stages. This affects the efficacy of using compositional biomarkers such as cartilage T_2 relaxation times which measure early changes that precede the onset of radiographic OA. Additionally, since the KL grading is performed on a frontal knee radiograph, only tibiofemoral OA is considered in the diagnosis and the impact of patellofemoral OA is not included in the grading, which is reflected in the comparatively lower performance for the patella OA diagnosis models. Finally, the computational time required to process the large-scale multimodal data into the spherical maps is another limiting factor, with the full processing of the dataset lasting a week. For future directions, the use of visualization tools, such as Grad-CAM[59], could characterize different biomarker phenotypes for OA diagnosis. An occlusion study

using Grad-CAMs to understand which anatomical regions in each bone contribute the most to a future OA diagnosis could shed light on the complex etiology of OA.

Chapter 5: Uncovering Associations Between Data-Driven Learned qMRI Biomarkers and Chronic Pain

5.1 Abstract

Knee pain is the most common and debilitating symptom of knee OA. While there is a perceived association between OA imaging biomarkers and pain, there are weak or conflicting findings for this relationship. This study uses Deep Learning (DL) models to elucidate associations between bone shape, cartilage thickness and T_2 relaxation times extracted from MRI and chronic knee pain. Grad-CAM applied on the trained chronic pain DL models are used to evaluate the locations of features associated with presence and absence of pain. For the cartilage thickness biomarker, the presence of features sensitive for pain presence were generally located in the medial side, while the features specific for pain absence were generally located in the anterior lateral side. This suggests that the association of cartilage thickness and pain varies, requiring a more personalized averaging strategy. We propose a novel DL-guided definition for cartilage thickness spatial averaging based on Grad-CAM weights. We showed a significant improvement modeling chronic knee pain with the inclusion of the novel biomarker definition: likelihood ratio test p-values of 7.01×10^{-33} and 1.93×10^{-14} for DL-guided cartilage thickness averaging for the femur and tibia, respectively, compared to the cartilage thickness compartment averaging.

5.2 Introduction

Knee pain is the most prominent and debilitating symptom of knee OA, a degenerative joint disease which affects over 13% of U.S. adults[1]. Notably, knee pain affects up to 7.3% of the total US population over 25 years of age, and the costs of medical care and loss of productivity are

rising[104]. The development of OA involves all joint tissues and is characterized by changes in the cartilage and bone. Given the lack of noninvasive treatment options to reverse the progression of structural joint degeneration, the medical care of OA has shifted to symptomatic pain management in a clinical setting[105,106]. While there is a widely perceived association of structural joint change with pain, previous studies linking OA imaging biomarkers to the presence of knee pain have not yet verified a strong correlation[11,56,107,108].

The sources of OA-related knee pain are not yet fully understood, with tissues such as bone and cartilage implicated through direct and indirect mechanisms. In particular, the aneural nature of cartilage obfuscates its involvement in the pain process, with surrounding tissue interactions being proposed as the source of pain[109]. Structurally, OA pathogenesis is marked by progressive degradation of the cartilage extracellular matrix, with early-stage changes including cartilage hydration, proteoglycan loss, and disruption of collagen. This process can be observed using quantitative Magnetic Resonance Imaging (qMRI) through imaging biomarkers such as T_2 relaxation time[101]. Late-stage OA is characterized by cartilage dehydration and structural breakdown, which results in measurable cartilage thickness loss on high resolution 3D MRI[21]. Alongside these cartilage changes, remodeling also occurs in the trabecular and subchondral bone, which can be observed with MRI-derived bone shape measurements[14]. Some early bony changes such as bone marrow lesions (BML) can predate cartilage degeneration, while presence of large osteophytes can act as a measure of advanced OA severity[110].

These imaging biomarkers (cartilage T_2, cartilage thickness and bone shape) have been classically extracted through compartment averaging, with femur, tibia, and patella divided into two or more functional regions[94,111]. This is an intuitive approach, given the prevalence of medial OA observed in patient populations, and there is particular emphasis placed in the medial compartment when conducting quantitative analysis of these biomarkers. While predictive models built with these imaging biomarker definitions tend to be interpretable, they suffer from decreased data granularity and statistical power. Furthermore, the discordance between OA-related imaging biomarkers and knee pain suggests that this methodology could be too reductive for a complex and multifactorial disease such as OA.

The advent of supervised feature learning and deep CNN architectures in medical image diagnostic tasks shows promising results in fully exploiting the image information by learning the most relevant data representation for the specific task considered[70–72]. However, the use of deep learning (DL) methods involve a tradeoff between model interpretability and performance, with classical rule-based expert systems[112] and regression models being highly interpretable but not as accurate. In the last few years, a renewed focus on DL model interpretability has produced explanatory techniques such as linear proxy models, decision trees, and saliency mapping[57,58]. These approaches attempt to understand the DL model performance by approximating CNNs to linear models, decomposing CNNs into decision trees, or systematically perturbing the inputs to discover the effect on the outputs. Saliency mapping in particular, has the benefit of being scalable by directly probing the gradients in a neural network to generate visualizations of local decision-making importance for a specific input image. Among these, Grad-CAM has the added benefit being class-discriminative by using the gradient information flowing into the last

convolutional layer of the CNN to understand each neuron for a decision of interest[59]. The resulting class-specific saliency map can be visualized as a heat map of location importance overlaid on the input image. Grad-CAM strikes a balance between emphasizing input image regions of high network activation, where neurons fire strongest, and input image regions of high network sensitivity, where changes would most affect the decision.

This study aims to uncover latent relationships between chronic knee pain and three MRI-based OA imaging biomarkers; cartilage T_2, cartilage thickness and bone shape by explaining CNN decisions using Grad-CAM. As a secondary aim, we propose a novel DL-guided and personalized definition of cartilage thickness compartment averaging based on Grad-CAM activations. We hypothesize these DL-guided imaging biomarkers will better explain chronic knee pain over classically extracted image biomarkers through *a priori* defined compartment averaging.

5.3 Methods

5.3.1 Aim and study overview

This study uses three known OA quantitative MR imaging biomarkers: bone shape, cartilage thickness and T_2 relaxation times, to train OA-related chronic knee pain classification models. It then leverages the trained models to determine the spatial averaging weights for each biomarker that are most correlated to chronic knee pain classification. In the next paragraph we present an overall study overview, with all the steps explained in detail in the subsequent sections.

First, the biomarkers are extracted from the knee MRI dataset by using two automatic segmentation models for the femur, tibia, and patella bones and corresponding cartilage. The cartilage thickness and T_2 relaxation times are then calculated from the cartilage segmentations while the bone shape is calculated from the bone segmentations. The three biomarkers are projected into the surface of the femur, tibia, and patella bones and transformed into spherical coordinates to obtain 2D images. Six different strategies were performed to merge biomarker spherical maps for each bone. Each of the six strategies for each bone was used to train individual chronic knee pain classification models, which were pretrained to classify radiographic OA, for a total of 18 models. Grad-CAM interpretation spherical maps of the entire hold out test set for all chronic knee pain models were inverted to the original bone surfaces and harmonized to a single atlas. Local group analysis of the two true predictive groups, true positives and true negatives, were compared to assess the local spatial difference in pain features for each group using a statistical parametric mapping technique. Two cartilage thickness averages were obtained using classically identified clinical compartments and using the Grad-CAM for each patient as a local weighting factor of the averaging (DL-guided). Logistic regression models were then used to compare the associations of DL-guided OA quantitative imaging biomarkers and a priori clinical compartments average biomarkers to chronic knee pain.

5.3.2 Imaging dataset

The details of the patient imaging dataset can be found on section 3.3.2.

5.3.3 Clinical outcome definition

Chronic pain labels were defined using clinical data from the OAI available for a subset of the patients. The chronic pain label was defined as patient timepoints which reported a knee pain, aching, or stiffness more than half of the days of a month for more than six months of the past 12 months. The no chronic pain label was defined as patient timepoints which did not report any knee pain, aching, or stiffness in the past 12 months. To control for nonspecific sources of pain outside of the knee, we excluded patients showing the presence of wide-spread pain syndrome, defined as reported pain concurrently in above-waist joints (shoulder, elbow, wrist, hand), below-waist joints (hip, knee, ankle, and foot), and axial joints (back and neck) for more than half of the days in the previous 30 days[107]. This localized definition of chronic pain focuses on pain symptoms lasting for months compared to shorter term clinical pain definitions such as the Western Ontario and McMaster Universities Osteoarthritis Index[113] (WOMAC) scores and the Knee injury and Osteoarthritis Outcome Score[114] (KOOS), which focus on the previous seven days. OA and its detectable imaging features may be more likely in patients who consistently reported pain within a yearlong period[107,115].

5.3.4 Patient inclusion

The three main criteria for inclusion of a knee image volume from a specific patient timepoint in this cross-sectional study were the existence of a KL grade, a chronic pain label, and matching 3D-DESS and 2D-MSME image volumes. Starting with a total of 47,078 3D-DESS image volumes, 261 image volumes were excluded due to poor inference quality from the bone and cartilage segmentation models (defined as a segmentation volume outside of three standard

deviations from the mean training segmentation), 22,464 image volumes from left patient knees were excluded due to absence of 2D-MSME for left knee image volumes, 3,235 image volumes were excluded due to missing KL grades for the visit, and 13,681 image volumes were excluded following exclusion criteria of the chronic pain definition described above. This selection resulted in 7,437 cross-sectional timepoints from 3,067 unique patients. The patient selection flowchart is summarized in **Fig. 5.1**.

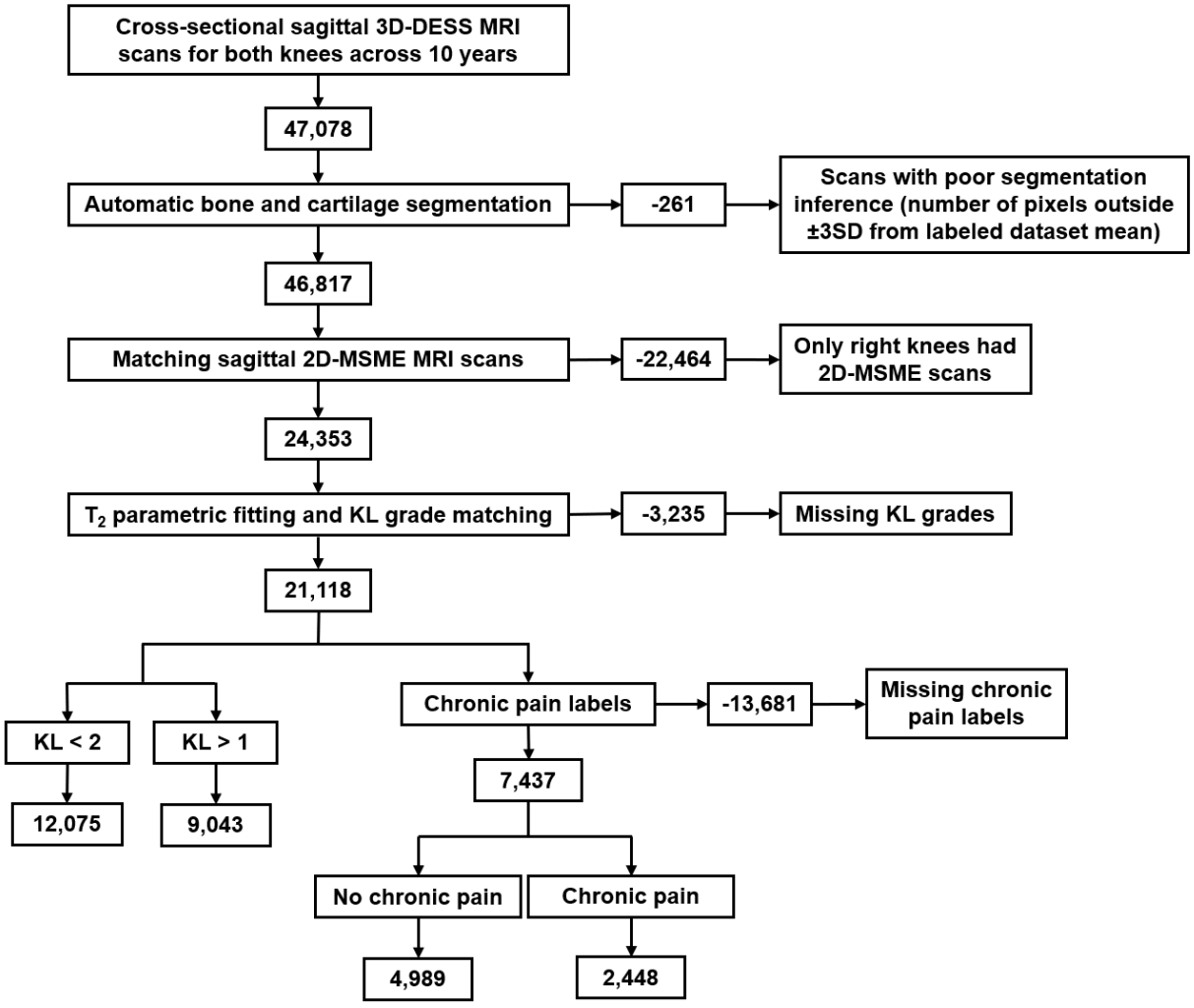


Fig. 5.1 The inclusion criteria for a knee image volume from a specific patient timepoint in this cross-sectional study. The three main criteria were the existence of a KL grade, a chronic pain label, and matching 3D-DESS and 2D-MSME image volumes, which resulted in 7,437 cross-sectional timepoints from 3,067 unique patients.

5.3.5 Bone and cartilage segmentation

The details of the bone and cartilage segmentation can be found on section 4.3.4.

5.3.6 Morphometry

The details of the morphometry can be found on section 4.3.5.

5.3.7 Relaxometry

The details of the relaxometry can be found on section 4.3.6.

5.3.8 Bone surface projection

The details of the bone surface projection can be found on section 4.3.7.

5.3.9 Spherical transformation

The details of the spherical transformation can be found on section 4.3.8.

5.3.10 Chronic pain model training

A total of 18 binary classification models, one for each biomarker strategy per bone, were trained to extract biomarker features from the spherical biomarker representations and use them to

predict chronic knee pain (**Supp. Fig. C.1**). Each chronic pain model was trained using 7,437 spherical images divided into 4,029 training images, 1,257 validation images and 2,151 test images, with no patient overlap across splits. To test the independence of demographic factors (sex, age, BMI) for the chronic pain cases across splits, two different statistical tests were performed. The independence of sex was tested with a Pearson's $\chi 2$ test while the independence of age and BMI was tested with a one-way MANOVA. **Table 5.1** summarizes the training, validation and test set splits for the segmentation and classification models, along with the p-values of the statistical tests showing independence of demographic factors.

The chronic pain prediction models were pretrained on an OA classification task. There were 21,118 cross-sectional timepoints from 4,416 unique patients. The KL grade distribution consisted of 8,103 (KL=0), 3,972 (KL=1), 5,335 (KL=2), 2,897 (KL=3) and 811 (KL=4). The KL grades represent no OA (KL=0), minimal/doubtful OA (KL=1), mild OA (KL=2), moderate OA (KL=3), and severe OA (KL=4). For the purposes of this study, KL grades of 0 and 1 were determined to be healthy while KL grades of 2, 3, and 4 are considered to be OA.

This study evaluated three types of Resnet[38] architectures with 18, 34, and 50 layers (Resnet18, Resnet34, Resnet50) with a binary class output. The Resnet18 and Resnet34 architecture consists of stacked building blocks of two convolutional layers with a 3x3 convolutional filter size, while the Resnet50 architecture follows the pattern of three convolutional layers with a 1x1, 3x3, and a 1x1 convolutional filter size respectively. For all architectures, each convolutional layer is paired with batch normalization and a rectified linear unit activation function.

Model training optimization for all 18 models was performed using the training and validation splits with two different learning rates ($1x10^{-4}$ and $1x10^{-5}$), three types of Resnet (Resnet18, Resnet34, Resnet50), three initialization strategies (Random[82], ImageNet, OA), and four variants of layer freezing during training (first layer, first two layers, all layers, no layers), for a total of 612 combinations. The model optimization was performed with Adam optimizer for 100 epochs with an early stopping 15-epoch patience for validation loss non-improvement over the best validation loss reached. The models were trained end to end using a class-weighted binary cross entropy loss, based on the class imbalance, with a batch size of 300. The test set was held out for each model during training optimization and the test performance was evaluated just once for the optimal 18 models.

Table 5.1 Training, validation, and test splits information for the segmentation and classification models. Demographic factors were controlled by testing for statistical independence across the splits using a Pearson's $\chi 2$ test for the categorical sex variable and a one-way MANOVA for the joint effect of age and BMI. Bold p-values are significant (p-value < 0.05).

Task	Model	Training (cases)	Validation (cases)	Test (cases)	Cases ratio	χ^2 test correlation (sex) (p-values)	MANOVA one-way correlation (age|BMI) (p-values)
Segmentation	Bone	57 (29)	15 (8)	30 (16)	0.520	0.745	0.413
	Cartilage	118 (114)	28 (28)	28 (28)	0.977	0.156	**$1x10^{-4}$**
Classification	OA	12,634 (5,402)	2,558 (1,111)	5,926 (2,530)	0.428	0.121	0.190
	Chronic Pain	4,029 (1,324)	1,257 (411)	2,151 (713)	0.329	0.179	0.0848

5.3.11 Grad-CAM model interpretation for imaging biomarker discovery

The overreaching goal of this study is to uncover associations between quantitative MR imaging biomarkers and chronic knee pain. We used the Grad-CAM model interpretation technique to obtain a class discriminative localization map for each prediction. We first compute the gradient of the class of interest (before the softmax function) with respect to feature maps of the last convolutional layer in the Resnet. These gradients flowing back are global average-pooled to obtain the neuron importance weights for the target class. A heat map of location importance is then up sampled to match the image size and overlaid on the input image.

We leveraged the invertible property of our spherical transformation method to generate articular surface importance heat maps for model interpretation for each bone and for each single biomarker. This process was performed on every patient in the hold out test set (n = 2,151) and is illustrated for the femur on **Fig. 5.2**.

The vertices of a reference bone surface, selected to match the average demographic distribution of the test set, were mapped on all the bone surfaces in the test set using a fully automatic landmark-matching algorithm. The strategy used in this study was based on the one proposed by Lombaert, H. et al[116]. The maximum and minimum local curvatures were used for coupling homologous points on two surfaces. Both these features were locally defined on the surfaces and used to identify the landmark matching solved using Coherent Point Drift[117]. After the landmark matching procedure, with the heat maps in the same reference space, localized group analysis was performed to compare the true positive (TP_{Pain}) and true negative (TN_{NoPain}) model

predictions for each single biomarker. Local Statistical Parametric Mapping (SPM) was performed on these two groups to assess differences in location of important features significant for presence of pain (TP_{Pain}) or specific for absence of pain (TN_{NoPain}). Point-by-point SPM was performed using ANOVA group comparison considering age, sex and BMI as confounding factors.

An ad-hoc analysis was then performed to compare the ability to explain chronic knee pain between cartilage thickness imaging biomarkers averaged using clinical compartments and a novel DL-guided definition based on weight averaging of the cartilage thickness with the scaled values of Grad-CAM as weights. Two logistic regression models were built to predict chronic knee pain, both with age, BMI, sex, and clinical compartment cartilage thickness averages, and one with DL-guided cartilage thickness averages. The performance of the nested models was compared using a likelihood ratio χ^2 test to determine the significance of the improvement of adding the DL-guided cartilage thickness averages. The linearity of the regression models and simplification of the analysis was used to compare the associations with pain of the classical and DL-guided biomarkers, instead of identifying nonlinear associations between the biomarkers and pain.

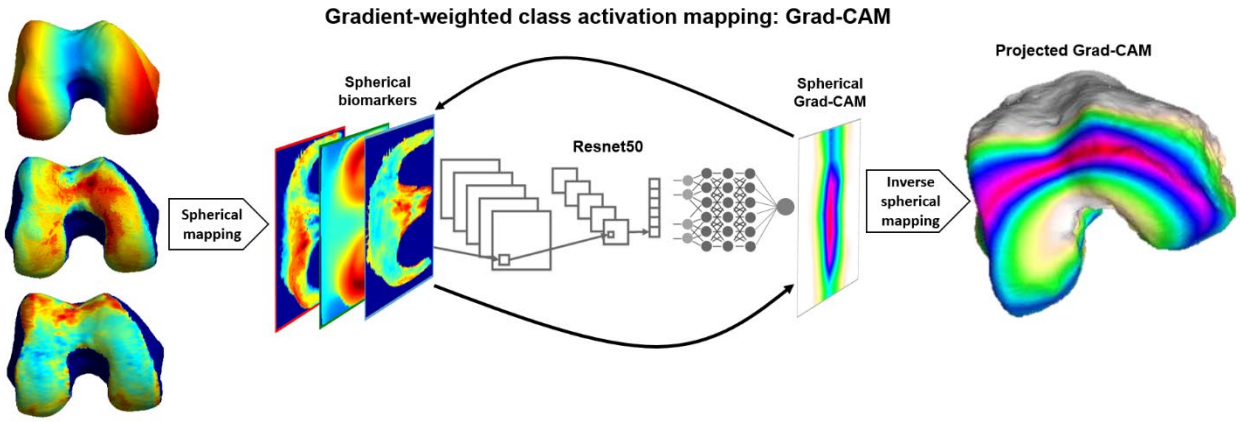


Fig. 5.2 We used the Grad-CAM model interpretation technique to obtain a class discriminative localization map for each prediction. We first computed the gradient of the class of interest (before the softmax function) with respect to feature maps of the last convolutional layer in the Resnet. These gradients flowing back are global average-pooled to obtain the neuron importance weights for the target class. A heat map of location importance is then up sampled to match the image size and overlaid on the input image. We then leveraged the invertible property of our spherical transformation method to generated articular surface importance heat maps for model interpretation for each bone and for each single biomarker. This process was performed on every patient in the hold out test set (n = 2,151) and is illustrated for the femur.

5.4 Results

5.4.1 Chronic pain model performance

The results of the model optimization were evaluated using the validation sensitivity, specificity, and AUC as well as the coefficient of variation of the validation AUC, as a measure of training smoothness. **Supp. Fig. C.2** summarizes the optimization results for the best performing models for each initialization strategy. The OA pretrained Resnet50 models consistently outperformed the randomly initialized models and exhibited smoother validation AUC than the ImageNet pretrained models. The model optimization informed the global selection of a Resnet50 pretrained to predict OA and fine-tuned to predict chronic pain for all 18 models, with the individual selection of the optimal learning rate and layer freezing for each model.

The test ROC curve results, defined as the sensitivity, the specificity, and AUC for the binary pretraining OA diagnosis task models, along with their respective 95% CI, are summarized in **Table 4.3**. The ROC metrics are given for each single biomarker and biomarker fusion pretraining OA diagnosis task models for each bone, as well as the ensembled averaged performance across all bones. The test sensitivity, specificity, and AUC respectively, ranged from 67.5 ± 0.2, 73.9 ± 0.2, and 77.6 ± 0.1 to 72.5 ± 0.2, 90.0 ± 0.1, and 90.1 ± 0.1. The bone shape model was the best performing single biomarker model for all bones. The femur biomarkers were the best performing models, followed by the tibia and the patella biomarker models.

For the chronic knee pain models, based on the results of the model optimization, the best model combination consisted of Resnet50 with OA pretraining, which were used for the test results. The test sensitivity, specificity, and AUC respectively, ranged from 57.9 ± 0.3, 70.6 ± 0.3, and 68.0 ± 0.2 to 53.0 ± 0.4, 83.8 ± 0.2, and 74.1 ± 0.2. The test performance followed a similar trend to the OA pretraining task, with the bone shape models outperforming the other single biomarker models for all bones. The performance across each bone also followed the decreasing trend of femur to tibia to patella. The cartilage T_2 models had a more balanced performance and higher sensitivity compared to the bone shape and cartilage thickness models, which tended to be more specific to chronic pain. Most models tended to be more specific than sensitive to chronic pain, and biomarker fusion models showed increased performance compared to the single biomarker models. The full test ROC results, defined as the sensitivity, the specificity, and AUC for the binary chronic pain models, along with their respective 95% CI, are summarized in **Table 5.2**.

The ROC metrics are given for each single biomarker and biomarker fusion chronic pain models for each bone, as well as the ensembled averaged performance across all bones.

Table 5.2 Bootstrapped (n=100) test set chronic pain ROC performance for all six biomarker models per bone, as well as an average ensemble across all bones. Sensitivity, specificity, and AUC values are shown respectively, along with their corresponding 95% confidence intervals. The best performances per bone and ensemble are bolded. PTF = Patella + Tibia + Femur ensemble.

Biomarker type	Biomarker model	Test set ROC (sensitivity/specificity/AUC) (mean ± 95% CI)			
		Patella	Tibia	Femur	PTF
Single	Cartilage T_2	60.0 ± 0.354 69.5 ± 0.246 69.7 ± 0.238	51.4 ± 0.380 78.4 ± 0.219 71.1 ± 0.241	64.8 ± 0.345 66.9 ± 0.246 72.4 ± 0.250	62.6 ± 0.317 74.1 ± 0.206 74.7 ± 0.216
	Cartilage thickness	57.9 ± 0.324 70.6 ± 0.256 68.0 ± 0.232	51.7 ± 0.366 80.4 ± 0.208 71.9 ± 0.246	57.1 ± 0.398 77.4 ± 0.206 72.9 ± 0.236	56.3 ± 0.360 79.1 ± 0.185 73.8 ± 0.241
	Bone shape	54.4 ± 0.361 77.3 ± 0.214 69.3 ± 0.248	52.2 ± 0.369 82.3 ± 0.198 73.1 ± 0.225	57.7 ± 0.389 78.3 ± 0.221 73.5 ± 0.240	56.9 ± 0.339 81.7 ± 0.195 74.3 ± 0.206
Fusion	Morphological bone and cartilage fusion	**63.3 ± 0.365** **69.7 ± 0.229** **71.8 ± 0.235**	**53.0 ± 0.389** **83.8 ± 0.170** **74.1 ± 0.223**	51.4 ± 0.333 81.1 ± 0.205 72.9 ± 0.218	**55.4 ± 0.335** **82.5 ± 0.192** **75.4 ± 0.228**
	Morphological and compositional cartilage fusion	59.9 ± 0.381 65.2 ± 0.265 68.9 ± 0.265	48.4 ± 0.367 81.2 ± 0.202 70.8 ± 0.237	**52.3 ± 0.355** **83.0 ± 0.191** **74.0 ± 0.220**	55.8 ± 0.330 79.5 ± 0.207 74.7 ± 0.194
	All biomarkers fusion	52.1 ± 0.423 77.2 ± 0.206 69.8 ± 0.234	50.8 ± 0.356 83.5 ± 0.186 74.1 ± 0.202	53.6 ± 0.387 80.8 ± 0.208 72.9 ± 0.258	51.6 ± 0.345 83.7 ± 0.193 74.4 ± 0.223

5.4.2 Grad-CAM model interpretation for imaging biomarker discovery

From the whole test set of 2,151 patients, a total of 137 TP_{Pain} cases and 379 TN_{NoPain} cases were selected, which consisted of the intersection of the correctly classified cases for all 18 models. This intersection, despite the reduction in number of samples, was selected over choosing

different sets for each model in an attempt to perform an analysis that could provide a direct comparison between the different biomarkers. For the TP_{Pain} group, the average and standard deviation for the age and BMI was 63.7 ± 8.5 and 31.2 ± 4.9 respectively, with 61 male and 76 female patients. For the TN_{NoPain} group the average and standard deviation for the age and BMI was 60.7 ± 9.7 and 25.5 ± 4.0 respectively, with 164 male and 215 female patients. Additionally, the race distribution of the TP_{Pain} group consisted of 30 Black or African American patients, 103 white patients and 4 patients with unreported race, while for the TN_{NoPain} group, the race distribution consisted of 15 African American patients, 362 white patients and 1 patient with unreported race.

Fig. 5.3 shows the results of the Grad-CAM statistical parametric mapping group analysis for each single biomarker for all three bones. After landmark matching, average Grad-CAM surfaces were generated for each biomarker for the two groups. The first two columns of each subfigure show the TP_{Pain} and TN_{NoPain} group average maps. In the third column, the results of the local SPM analysis are shown as a p-value surface. **Fig. 5.3A** shows the results of the femur bone. For the bone shape feature, similar patterns of elevations were observed in TP_{Pain} and TN_{NoPain}. In both groups, the majority of the Grad-CAM elevation was co-localized in the anterior medial femoral area. High values of these maps are indicative of common patterns in the whole group, since Grad-CAM elevations distributed in different locations for each patient would be averaged out over the group. Similar patterns in two groups, as it is observed for the femur bone shape feature, are indicative of similar location of features being exploited by the model for the assessment of both pain presence and absence.

In cartilage thickness and T_2 imaging biomarkers, the locations of features that were sensitive for the presence of chronic pain are distinct from the locations of features that were specific for absence of chronic pain. Features sensitive for pain presence are located in the medial femoral condyle, while features that are specific for pain absence are located in the anterior femoral area, particularly in the trochlea.

Similar relationships were observed for the tibia (**Fig. 5.3B**), where the location of important bone shape features was similar in the two groups. For cartilage thickness, the medial plateau was almost exclusively observed as significant for the TP_{Pain} group while both the medial and lateral plateaus showed importance for the TN_{NoPain} group. The T_2 biomarker in the tibia showed weak elevations in the group Grad-CAM, which demonstrates scattered peaks on the individual maps of patients.

Results on the patella bone and cartilage are shown in **Fig. 5.3C**. Bone shape biomarker features sensitive to the pain were located in the lateral facet, while features specific for absence of pain were located in the most inferior aspect of the patella bone. A similar pattern was observed for cartilage thickness, with the pattern seemingly inverted for cartilage T_2.

Table 5.3 shows the results of the chronic pain logistic regression using demographic factors, such as age, sex, and BMI, and the standard cartilage compartment averages compared with the same model with the addition of the DL-guided thickness averages. For the femur and tibia, the

DL-guided biomarker is a significantly better predictor of the chronic pain outcome, with likelihood ratio test p-values of 7.01×10^{-33} and $1.93 \times 10^{-14,}$ respectively.

Table 5.3 Logistic regression results for the cartilage thickness biomarker for all bones. The demographic factors, such as age, BMI, and sex, are included to the logistic regression models as well as the different cartilage thickness averaging methods. The results are shown for the two definitions for OA imaging biomarkers, clinical compartment average and DL-guided weighted average for the femur, tibia, and patella. LF = Lateral Femur, MF = Medial Femur, MT = Medial Tibia, LT = Lateral Tibia, M = Medial, L= Lateral.

Biomarker	Bone	Method	Variable	Estimate	Standard error	P-value	Likelihood ratio p-value
Cartilage thickness (n = 2,151)	**Femur**	**Classical:** clinical compartment average	Intercept	-3.59	0.621	7.56×10^{-9}	**7.01×10^{-33}**
			Age	-0.011	5.1×10^{-3}	3.06×10^{-2}	
			BMI	0.077	1.01×10^{-2}	1.84×10^{-14}	
			Sex	-0.193	0.114	9.05×10^{-2}	
			LF Thickness	-0.582	0.264	2.77×10^{-2}	
			MF Thickness	1.289	0.246	1.77×10^{-7}	
		Proposed: DL-guided weighted average	Intercept	-2.52	0.645	9.16×10^{-5}	
			Age	-1.97×10^{-2}	5.36×10^{-3}	2.33×10^{-4}	
			BMI	5.14×10^{-2}	1.06×10^{-2}	1.13×10^{-6}	
			Sex	-8.78×10^{-2}	0.118	0.455	
			LF Thickness	2.25	0.364	6.55×10^{-10}	
			MF Thickness	2.29	0.268	1.57×10^{-17}	
			DL-Thickness	-3.66	0.32	2.02×10^{-30}	
	Tibia	**Classical:** clinical compartment average	Intercept	0.496	0.638	0.437	**1.93×10^{-14}**
			Age	-1.81×10^{-2}	5.17×10^{-3}	4.5×10^{-4}	
			BMI	0.078	0.01	7.27×10^{-15}	
			Sex	-0.356	0.106	8.0×10^{-4}	
			LT Thickness	-0.445	0.182	1.47×10^{-2}	
			MT Thickness	-0.537	0.15	3.42×10^{-4}	
		Proposed: DL-guided	Intercept	0.81	0.65	0.213	
			Age	-2.11×10^{-2}	5.27×10^{-3}	6.06×10^{-5}	

Biomarker	Bone	Method	Variable	Estimate	Standard error	P-value	Likelihood ratio p-value
		weighted average	BMI	$7.03x10^{-2}$	$1.02x10^{-2}$	$4.86x10^{-12}$	
			Sex	-0.387	0.107	$3.12x10^{-4}$	
			LT Thickness	0.289	0.208	0.165	
			MT Thickness	0.108	0.173	0.533	
			DL-Thickness	-1.37	0.184	$9.6x10^{-14}$	
	Patella	**Classical:** clinical compartment average	Intercept	1.21	0.644	$6.05x10^{-2}$	0.851
			Age	$-2.38x10^{-2}$	$5.33x10^{-3}$	$8.09x10^{-6}$	
			BMI	$6.64x10^{-2}$	$1.03x10^{-2}$	$1.06x10^{-10}$	
			Sex	-0.389	0.105	$2.28x10^{-4}$	
			L Thickness	-0.398	0.118	$7.12x10^{-4}$	
			M Thickness	-0.424	0.12	$3.97x10^{-4}$	
		Proposed: DL-guided weighted average	Intercept	1.215	0.645	$5.95x10^{-2}$	
			Age	$-2.38x10^{-2}$	$5.33x10^{-3}$	$7.96x10^{-6}$	
			BMI	$6.63x10^{-2}$	$1.03x10^{-2}$	$1.20x10^{-10}$	
			Sex	-0.39	0.106	$2.24x10^{-4}$	
			L Thickness	-0.376	0.166	$2.39x10^{-2}$	
			M Thickness	-0.401	0.173	$2.06x10^{-2}$	
			DL-Thickness	$-4.57x10^{-2}$	0.243	0.851	

(a)
(b)
(c)
Grad-CAM
0
0.8
P-value
1
10^{-20}
ANCOVA group difference
TP_{Pain} (n=137)
$TN_{NonPain}$ (n=379)
Bone Shape
Cartilage Thickness
Cartilage T_2
A
P
M
L

Fig. 5.3 The vertices of a reference bone surface, selected to match the average demographic distribution of the test set, were mapped on all the bone surfaces in the test set using a fully automatic landmark-matching algorithm. The maximum and minimum local curvatures were used for coupling homologous points on two surfaces. Both these features were locally defined on the surfaces and used to identify the landmark matching. After the landmark matching procedure, with the heat maps in the same reference space, localized group analysis was performed to compare the true positive (TP_{pain}) and true negative (TN_{Nopain}) model predictions for each single biomarker. Local Statistical Parametric Mapping (SPM) was performed on these two groups to assess differences in location of important features significant for presence of pain (TP_{pain}) or specific for absence of pain (TN_{Nopain}). Point-by-point SPM was performed using ANOVA group comparison considering age, sex and BMI as confounding factors.

5.5 Discussion

In this study, we propose a DL-guided definition for OA quantitative imaging biomarkers which is more strongly associated to chronic knee pain than the clinical compartment average definition. We report likelihood ratio test significant p-values of 7.01×10^{-33} and 1.93×10^{-14} for DL-guided cartilage thickness averaging for the femur and tibia, respectively, compared to the cartilage thickness compartment averaging, for predicting chronic pain. The difference is reported even with the inclusion of demographic factors such as age, BMI, and sex to the regression models, which have been linked to pain[118]. This method for quantitative imaging biomarker discovery is specific to each patient, instead of being predefined based on clinical assumptions, which suggests there are personalized changes not reflected by known OA-related regions.

The average Grad-CAM saliency maps for the TP_{Pain} and TN_{NoPain} groups revealed an interesting heterogeneity in the localization of the features sensitive to pain and specific to no pain. This observation of distinct locations for pain specific and non-pain specific features for the cartilage thickness biomarker was surprising and previously unreported, to the best of our knowledge. The

activation regions for the cartilage thickness across all bones showed pain specific features generally located in the medial side, while the non-pain specific features were generally located in the lateral side. This finding generated the hypothesis that the weak association between cartilage thickness and clinically relevant outcomes, such as pain, could be partly attributed to patient-specific heterogenous importance in the locations of cartilage thickness variation. Furthermore, this process might explain why the use of averages across the entire compartment would produce a weak association or even a discordance between the imaging biomarkers and pain. This selectivity between pain and non-pain specific features could be indicative of local regulatory behavior for knee pain, where areas that produce the pain could be mediated by areas associated with a lack of pain.

A recent study by Bacon, K. et al[25] found a weak association between medial femorotibial cartilage thickness loss and knee pain, reporting a significant 0.32 ± 0.11 mean change in WOMAC pain scores resulting from a 0.1 mm cartilage thickness loss over a 24 month period. This correlation, while statistically significant, did not surpass the minimally clinical importance difference for WOMAC pain scores[119]. Similarly, a reduction in central medial femorotibial compartment cartilage thickness was reported to be weakly associated with pain progression with an odds ratio of 1.3 ± 0.2[24]. Our work has two key differences with these studies, the definition of chronic knee pain, instead of pain defined by the WOMAC scale, and the use of DL-guided cartilage thickness averaging, instead of compartment averaging. Our use of chronic knee pain as a clinical outcome has the advantage of focusing on persistent pain experienced over the course of a year, which is likelier to capture meaningful changes in cartilage thickness compared to the week-long recall period for WOMAC pain scores. The DL-guided approach serves as a

personalized approach for region of interest definition, which allows for the extraction of an imaging biomarker more associated to pain.

The bone shape biomarker has generally been described in previous works using statistical shape modelling to compare different shape variations between case groups[16,120]. Unlike cartilage thickness and cartilage T_2 biomarkers, there is no obvious way to apply the Grad-CAM saliencies to the bone shape maps, since averaging bone shape values may not be appropriate. For cartilage T_2, we did not find a difference in the association between classical compartment averaging and the DL-guided weight averaging to chronic pain. While cartilage T_2 times have been shown to be associated with pain[34], we did not find an improvement in the inclusion of the DL-guided weight averaging to the classical compartment averaging in the regression models. This suggests that the nature of the behavior for cartilage thickness and cartilage T_2 may be different, with the latter exhibiting a weaker pain feature heterogeneity. Compartment averaging for T_2 relaxation times may be sufficient in explaining chronic pain.

Although this study brings new insights on the role of deep learning for quantitative imaging biomarker discovery, several limitations need to be acknowledged. One of the limitations of the study is the focus on structural changes, which omits the impact of inflammatory changes that have been consistently linked to pain. Bone marrow lesions and synovitis, in particular, have been reported to play a role in the pain process and are not directly reflected by our biomarkers[121]. Additionally, the pain performance improvement of the biomarker fusion models over the single biomarker models suggests that there are some added pain-related interactions

between biomarkers. These were not further inspected due to the reduced interpretability of combining the biomarkers at the input level. The use of the intersection of all 18 models limited the findings to the set of imaging features that are most persistently associated with chronic pain. This could result in the loss of more nuanced patient-specific relationships to pain. The OAI is also a limited dataset and findings based on it may not be generalizable to the general population. It is also worth noting that the purpose of the study was not to achieve the highest predictive performance for chronic pain, but rather to understand local associations between the biomarkers and chronic pain.

The findings of this work could improve the imaging biomarker definition for clinical trials, with patient-specific imaging biomarkers that are more strongly correlated to clinical outcomes such as pain. A recent clinical trial for the disease-modifying osteoarthritis drug sprifermin showed a protective effect for femorotibial average cartilage thickness loss of 0.1 mm over a period of 2 years[28]. The same trial found no significant effect for this substantial cartilage preservation on the WOMAC pain scores, which highlights the importance of stronger predictors for pain. Our proposed DL-guided cartilage thickness averaging could be used to evaluate the effect of such cartilage-preserving treatments on pain, tailoring the imaging biomarker to the clinical outcome.

References

1. Cisternas MG, Murphy L, Sacks JJ, Solomon DH, Pasta DJ, Helmick CG. Alternative Methods for Defining Osteoarthritis and the Impact on Estimating Prevalence in a US Population-Based Survey. *Arthritis Care Res (Hoboken)*. 2016;68(5):574-580. doi:10.1002/acr.22721

2. Murphy LB, Cisternas MG, Pasta DJ, Helmick CG, Yelin EH. Medical Expenditures and Earnings Losses Among US Adults With Arthritis in 2013. *Arthritis Care & Research*. 2018;70(6):869-876. doi:10.1002/acr.23425

3. Charlesworth J, Fitzpatrick J, Perera NKP, Orchard J. Osteoarthritis- a systematic review of long-term safety implications for osteoarthritis of the knee. *BMC Musculoskeletal Disorders*. 2019;20(1):151. doi:10.1186/s12891-019-2525-0

4. Mobasheri A, Batt M. An update on the pathophysiology of osteoarthritis. *Annals of Physical and Rehabilitation Medicine*. 2016;59(5):333-339. doi:10.1016/j.rehab.2016.07.004

5. Sophia Fox AJ, Bedi A, Rodeo SA. The Basic Science of Articular Cartilage. *Sports Health*. 2009;1(6):461-468. doi:10.1177/1941738109350438

6. The Burden of Musculoskeletal Diseases in the United States, Fourth Edition. BMUS: The Burden of Musculoskeletal Diseases in the United States. Accessed May 5, 2021. https://www.boneandjointburden.org/fourth-edition

7. Donell S. Subchondral bone remodelling in osteoarthritis. *EFORT Open Rev*. 2019;4(6):221-229. doi:10.1302/2058-5241.4.180102

8. Hu Y, Chen X, Wang S, Jing Y, Su J. Subchondral bone microenvironment in osteoarthritis and pain. *Bone Research*. 2021;9(1):1-13. doi:10.1038/s41413-021-00147-z

9. Kohn MD, Sassoon AA, Fernando ND. Classifications in Brief: Kellgren-Lawrence Classification of Osteoarthritis. *Clin Orthop Relat Res*. 2016;474(8):1886-1893. doi:10.1007/s11999-016-4732-4

10. Chan WP, Lang P, Stevens MP, et al. Osteoarthritis of the knee: comparison of radiography, CT, and MR imaging to assess extent and severity. *American Journal of Roentgenology*. 1991;157(4):799-806. doi:10.2214/ajr.157.4.1892040

11. Bedson J, Croft PR. The discordance between clinical and radiographic knee osteoarthritis: a systematic search and summary of the literature. *BMC Musculoskelet Disord*. 2008;9:116. doi:10.1186/1471-2474-9-116

12. Peterfy CG, Schneider E, Nevitt M. The osteoarthritis initiative: report on the design rationale for the magnetic resonance imaging protocol for the knee. *Osteoarthritis Cartilage*. 2008;16(12):1433-1441. doi:10.1016/j.joca.2008.06.016

13. Neogi T, Felson DT. Bone as an imaging biomarker and treatment target in OA. *Nature Reviews Rheumatology*. 2016;12(9):503-504. doi:10.1038/nrrheum.2016.113

14. Neogi T, Bowes MA, Niu J, et al. Magnetic Resonance Imaging-Based Three-Dimensional Bone Shape of the Knee Predicts Onset of Knee Osteoarthritis: Data From the Osteoarthritis Initiative: 3-D Bone Shape Predicts Incident Knee OA. *Arthritis & Rheumatism*. 2013;65(8):2048-2058. doi:10.1002/art.37987

15. Barr AJ, Dube B, Hensor EMA, et al. The relationship between three-dimensional knee MRI bone shape and total knee replacement-a case control study: data from the Osteoarthritis Initiative. *Rheumatology (Oxford)*. 2016;55(9):1585-1593. doi:10.1093/rheumatology/kew191

16. Hunter D, Nevitt M, Lynch J, et al. Longitudinal validation of periarticular bone area and 3D shape as biomarkers for knee OA progression? Data from the FNIH OA Biomarkers Consortium. *Ann Rheum Dis*. 2016;75(9):1607-1614. doi:10.1136/annrheumdis-2015-207602

17. Bredbenner TL, Eliason TD, Potter RS, Mason RL, Havill LM, Nicolella DP. Statistical shape modeling describes variation in tibia and femur surface geometry between Control and Incidence groups from the Osteoarthritis Initiative database. *J Biomech*. 2010;43(9):1780-1786. doi:10.1016/j.jbiomech.2010.02.015

18. Pedoia V, Lansdown DA, Zaid M, et al. Three-Dimensional MRI-Based Statistical Shape Model and Application to a Cohort of Knees with Acute ACL Injury. *Osteoarthritis Cartilage*. 2015;23(10):1695-1703. doi:10.1016/j.joca.2015.05.027

19. Lansdown DA, Pedoia V, Zaid M, et al. Variations in Knee Kinematics After ACL Injury and After Reconstruction Are Correlated With Bone Shape Differences. *Clinical Orthopaedics and Related Research*. 2017;475(10):2427. doi:10.1007/s11999-017-5368-8

20. Iriondo C, Liu F, Calivà F, Kamat S, Majumdar S, Pedoia V. Towards Understanding Mechanistic Subgroups of Osteoarthritis: 8 Year Cartilage Thickness Trajectory Analysis. *Journal of Orthopaedic Research*. n/a(n/a). doi:10.1002/jor.24849

21. Reichenbach S, Yang M, Eckstein F, et al. Does cartilage volume or thickness distinguish knees with and without mild radiographic osteoarthritis? The Framingham Study. *Ann Rheum Dis*. 2010;69(1):143-149. doi:10.1136/ard.2008.099200

22. Eckstein F, Cotofana S, Wirth W, et al. Painful Knees have Greater Rates of Cartilage Loss than Painless Knees After Adjusting for Radiographic Disease Stage: Data from the OA Initiative. *Arthritis Rheum*. 2011;63(8):2257-2267. doi:10.1002/art.30414

23. Wirth W, Le Graverand M-PH, Wyman BT, et al. Regional Analysis of Femorotibial Cartilage Loss in a Subsample from the Osteoarthritis Initiative Progression Subcohort. *Osteoarthritis Cartilage*. 2009;17(3):291-297. doi:10.1016/j.joca.2008.07.008

24. Eckstein F, Collins JE, Nevitt MC, et al. CARTILAGE THICKNESS CHANGE AS AN IMAGING BIOMARKER OF KNEE OSTEOARTHRITIS PROGRESSION – DATA FROM THE FNIH OA BIOMARKERS CONSORTIUM. *Arthritis Rheumatol*. 2015;67(12):3184-3189. doi:10.1002/art.39324

25. Bacon K, LaValley MP, Jafarzadeh SR, Felson D. Does cartilage loss cause pain in osteoarthritis and if so, how much? *Annals of the Rheumatic Diseases*. 2020;79(8):1105-1110. doi:10.1136/annrheumdis-2020-217363

26. McAlindon TE, LaValley MP, Harvey WF, et al. Effect of Intra-articular Triamcinolone vs Saline on Knee Cartilage Volume and Pain in Patients With Knee Osteoarthritis: A Randomized Clinical Trial. *JAMA*. 2017;317(19):1967-1975. doi:10.1001/jama.2017.5283

27. Wluka AE, Wolfe R, Stuckey S, Cicuttini FM. How does tibial cartilage volume relate to symptoms in subjects with knee osteoarthritis? *Ann Rheum Dis*. 2004;63(3):264-268. doi:10.1136/ard/2003.007666

28. Hochberg MC, Guermazi A, Guehring H, et al. Effect of Intra-Articular Sprifermin vs Placebo on Femorotibial Joint Cartilage Thickness in Patients With Osteoarthritis: The FORWARD Randomized Clinical Trial. *JAMA*. 2019;322(14):1360-1370. doi:10.1001/jama.2019.14735

29. Guermazi A, Alizai H, Crema MD, Trattnig S, Regatte RR, Roemer FW. Compositional MRI techniques for evaluation of cartilage degeneration in osteoarthritis. *Osteoarthritis and Cartilage*. 2015;23(10):1639-1653. doi:10.1016/j.joca.2015.05.026

30. Prasad AP, Nardo L, Schooler J, Joseph GB, Link TM. T1ρ and T2 relaxation times predict progression of knee osteoarthritis. *Osteoarthritis Cartilage*. 2013;21(1):69-76. doi:10.1016/j.joca.2012.09.011

31. David-Vaudey E, Ghosh S, Ries M, Majumdar S. T2 relaxation time measurements in osteoarthritis. *Magnetic Resonance Imaging*. 2004;22(5):673-682. doi:10.1016/j.mri.2004.01.071

32. Pedoia V, Su F, Amano K, et al. Analysis of the articular cartilage T1ρ and T2 relaxation times changes after ACL reconstruction in injured and contralateral knees and relationships with bone shape. *Journal of Orthopaedic Research*. 2017;35(3):707-717. doi:10.1002/jor.23398

33. Razmjoo A, Caliva F, Lee J, et al. T2 analysis of the entire osteoarthritis initiative dataset. *Journal of Orthopaedic Research*. n/a(n/a). doi:https://doi.org/10.1002/jor.24811

34. Baum T, Joseph GB, Arulanandan A, et al. Association of MRI-based knee cartilage T2 measurements and focal knee lesions with knee pain - data from the Osteoarthritis Initiative. *Arthritis Care Res (Hoboken)*. 2012;64(2):248-255. doi:10.1002/acr.20672

35. Cootes TF, Edwards GJ, Taylor CJ. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2001;23(6):681-685. doi:10.1109/34.927467

36. Krizhevsky A, Sutskever I, Hinton GE. ImageNet Classification with Deep Convolutional Neural Networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, eds. *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc.; 2012:1097-1105. Accessed September 26, 2019.

http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

37. Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift für Medizinische Physik*. 2019;29(2):102-127. doi:10.1016/j.zemedi.2018.11.002

38. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. *arXiv:151203385 [cs]*. Published online December 10, 2015. Accessed November 6, 2018. http://arxiv.org/abs/1512.03385

39. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:181004805 [cs]*. Published online May 24, 2019. Accessed May 9, 2021. http://arxiv.org/abs/1810.04805

40. Leung MKK, Delong A, Alipanahi B, Frey BJ. Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets. *Proceedings of the IEEE*. 2016;104(1):176-197. doi:10.1109/JPROC.2015.2494198

41. Ganaie MA, Hu M, Tanveer* M, Suganthan* PN. Ensemble deep learning: A review. *arXiv:210402395 [cs]*. Published online April 6, 2021. Accessed May 9, 2021. http://arxiv.org/abs/2104.02395

42. Norman B, Pedoia V, Majumdar S. Use of 2D U-Net Convolutional Neural Networks for Automated Cartilage and Meniscus Segmentation of Knee MR Imaging Data to Determine Relaxometry and Morphometry. *Radiology*. 2018;288(1):177-185. doi:10.1148/radiol.2018172322

43. Liu F, Zhou Z, Jang H, Samsonov A, Zhao G, Kijowski R. Deep Convolutional Neural Network and 3D Deformable Approach for Tissue Segmentation in Musculoskeletal Magnetic Resonance Imaging. *Magn Reson Med*. 2018;79(4):2379-2391. doi:10.1002/mrm.26841

44. Pedoia V, Li X, Su F, Calixto N, Majumdar S. Fully Automatic Analysis of the Knee Articular Cartilage T1ρ relaxation time using Voxel Based Relaxometry. *J Magn Reson Imaging*. 2016;43(4):970-980. doi:10.1002/jmri.25065

45. Peake EJ, Chevasson R, Pszczolkowski S, Auer DP, Arthofer C. *Ensemble Learning for Robust Knee Cartilage Segmentation: Data from the Osteoarthritis Initiative*. Pathology; 2020. doi:10.1101/2020.09.01.267872

46. Desai AD, Caliva F, Iriondo C, et al. The International Workshop on Osteoarthritis Imaging Knee MRI Segmentation Challenge: A Multi-Institute Evaluation and Analysis Framework on a Standardized Dataset. *arXiv:200414003 [cs, eess]*. Published online May 26, 2020. Accessed May 9, 2021. http://arxiv.org/abs/2004.14003

47. Milletari F, Navab N, Ahmadi S-A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In: *2016 Fourth International Conference on 3D Vision (3DV)*. ; 2016:565-571. doi:10.1109/3DV.2016.79

48. Sheehy L, Culham E, McLean L, et al. Validity and sensitivity to change of three scales for the radiographic assessment of knee osteoarthritis using images from the Multicenter Osteoarthritis Study (MOST). *Osteoarthritis Cartilage*. 2015;23(9):1491-1498. doi:10.1016/j.joca.2015.05.003

49. Marinetti A, Tessarolo F, Ventura L, et al. Morphological MRI of knee cartilage: repeatability and reproducibility of damage evaluation and correlation with gross pathology examination. *Eur Radiol*. 2020;30(6):3226-3235. doi:10.1007/s00330-019-06627-5

50. Shamir L, Ling SM, Scott WW, et al. Knee x-ray image analysis method for automated detection of osteoarthritis. *IEEE Trans Biomed Eng*. 2009;56(2):407-415. doi:10.1109/TBME.2008.2006025

51. Gornale SS, Patravali PU, Manza RR. Detection of Osteoarthritis using Knee X-Ray Image Analyses: A Machine Vision based Approach. *International Journal of Computer Applications*. 2016;145(1):20-26.

52. Hladůvka J, Phuong BTM, Ljuhar R, et al. Femoral ROIs and Entropy for Texture-based Detection of Osteoarthritis from High-Resolution Knee Radiographs. *arXiv:170309296 [cs]*. Published online March 27, 2017. Accessed May 9, 2021. http://arxiv.org/abs/1703.09296

53. Tiulpin A, Thevenot J, Rahtu E, Lehenkari P, Saarakkala S. Automatic Knee Osteoarthritis Diagnosis from Plain Radiographs: A Deep Learning-Based Approach. *Scientific Reports*. 2018;8(1):1727. doi:10.1038/s41598-018-20132-7

54. Leung K, Zhang B, Tan J, et al. Prediction of Total Knee Replacement and Diagnosis of Osteoarthritis by Using Deep Learning on Knee Radiographs: Data from the Osteoarthritis Initiative. *Radiology*. 2020;296(3):584-593. doi:10.1148/radiol.2020192091

55. Tolpadi AA, Lee JJ, Pedoia V, Majumdar S. Deep Learning Predicts Total Knee Replacement from Magnetic Resonance Images. *Scientific Reports*. 2020;10(1):6371. doi:10.1038/s41598-020-63395-9

56. Yusuf E, Kortekaas MC, Watt I, Huizinga TWJ, Kloppenburg M. Do knee abnormalities visualised on MRI explain knee pain in knee osteoarthritis? A systematic review. *Ann Rheum Dis*. 2011;70(1):60-67. doi:10.1136/ard.2010.131904

57. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining Explanations: An Overview of Interpretability of Machine Learning. *arXiv:180600069 [cs, stat]*. Published online February 3, 2019. Accessed April 7, 2021. http://arxiv.org/abs/1806.00069

58. Zhang Y, Tiňo P, Leonardis A, Tang K. A Survey on Neural Network Interpretability. *arXiv:201214261 [cs]*. Published online March 3, 2021. Accessed April 7, 2021. http://arxiv.org/abs/2012.14261

59. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *arXiv:161002391 [cs]*. Published online October 7, 2016. Accessed May 13, 2019. http://arxiv.org/abs/1610.02391

60. Cross M, Smith E, Hoy D, et al. The global burden of hip and knee osteoarthritis: estimates from the Global Burden of Disease 2010 study. *Ann Rheum Dis*. 2014;73(7):1323-1330. doi:10.1136/annrheumdis-2013-204763

61. Hootman JM, Helmick CG. Projections of US prevalence of arthritis and associated activity limitations. *Arthritis & Rheumatism*. 2006;54(1):226-229. doi:10.1002/art.21562

62. Müller-Gerbl M, Griebl S, Putz R, Goldmann A, Kuhr M, Taeger K. Assessment of subchondral bone density distribution patterns in patients subjected to correction osteotomy. *Trans Orth Soc*. 1994;19:574.

63. Müller-Gerbl M, Putz R, Hodapp N, Schulte E, Wimmer B. Computed tomography-osteoabsorptiometry for assessing the density distribution of subchondral bone as a measure of long-term mechanical adaptation in individual joints. *Skeletal Radiol*. 1989;18(7):507-512. doi:10.1007/BF00351749

64. Pauwels F. *Biomechanics of the Locomotor Apparatus: Contributions on the Functional Anatomy of the Locomotor Apparatus*. Springer-Verlag; 1980. Accessed June 12, 2019. https://www.springer.com/us/book/9783642671401

65. Lynch JA, Parimi N, Chaganti RK, Nevitt MC, Lane NE. The association of proximal femoral shape and incident radiographic hip OA in elderly women. *Osteoarthritis Cartilage*. 2009;17(10):1313-1318. doi:10.1016/j.joca.2009.04.011

66. Baker-LePain JC, Lynch JA, Parimi N, et al. Variant Alleles of the WNT Antagonist FRZB Are Determinants of Hip Shape and Modify the Relationship between Hip Shape and Osteoarthritis. *Arthritis Rheum*. 2012;64(5):1457-1465. doi:10.1002/art.34526

67. Wise BL, Kritikos L, Lynch JA, et al. Proximal Femur Shape Differs Between Subjects with Lateral and Medial Knee Osteoarthritis and Controls: The Osteoarthritis Initiative. *Osteoarthritis Cartilage*. 2014;22(12):2067-2073. doi:10.1016/j.joca.2014.08.013

68. Chan EF, Farnsworth CL, Koziol JA, Hosalkar HS, Sah RL. Statistical shape modeling of proximal femoral shape deformities in Legg–Calvé–Perthes disease and slipped capital femoral epiphysis. *Osteoarthritis and Cartilage*. 2013;21(3):443-449. doi:10.1016/j.joca.2012.12.007

69. Bowes MA, Vincent GR, Wolstenholme CB, Conaghan PG. A novel method for bone area measurement provides new insights into osteoarthritis and its progression. *Ann Rheum Dis*. 2015;74(3):519-525. doi:10.1136/annrheumdis-2013-204052

70. Lee H, Tajmir S, Lee J, et al. Fully Automated Deep Learning System for Bone Age Assessment. *J Digit Imaging*. 2017;30(4):427-441. doi:10.1007/s10278-017-9955-8

71. Becker AS, Blüthgen C, Phi van VD, et al. Detection of tuberculosis patterns in digital photographs of chest X-ray images using Deep Learning: feasibility study. Published online March 1, 2018. doi:info:doi/10.5588/ijtld.17.0520

72. Ribli D, Horváth A, Unger Z, Pollner P, Csabai I. Detecting and classifying lesions in mammograms with Deep Learning. *Sci Rep*. 2018;8(1):4165. doi:10.1038/s41598-018-22437-z

73. Lee H, Grosse R, Ranganath R, Ng AY. Unsupervised learning of hierarchical representations with convolutional deep belief networks. *Commun ACM*. 2011;54(10):95. doi:10.1145/2001269.2001295

74. Kallenberg M, Petersen K, Nielsen M, et al. Unsupervised Deep Learning Applied to Breast Density Segmentation and Mammographic Risk Scoring. *IEEE Transactions on Medical Imaging*. 2016;35(5):1322-1331. doi:10.1109/TMI.2016.2532122

75. Chaudhari AS, Fang Z, Kogan F, et al. Super-resolution musculoskeletal MRI using deep learning. *Magnetic Resonance in Medicine*. 2018;80(5):2139-2154. doi:10.1002/mrm.27178

76. Milletari F, Navab N, Ahmadi S-A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. Published online June 15, 2016. Accessed October 22, 2018. https://arxiv.org/abs/1606.04797

77. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: A Large-Scale Hierarchical Image Database. :8.

78. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*.:6.

79. Paszke A, Gross S, Chintala S, et al. Automatic differentiation in PyTorch. In: ; :4.

80. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks. Published online August 25, 2016. Accessed November 6, 2018. https://arxiv.org/abs/1608.06993

81. Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. Published online February 24, 2016. Accessed September 26, 2019. https://arxiv.org/abs/1602.07360v4

82. He K, Zhang X, Ren S, Sun J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *arXiv:150201852 [cs]*. Published online February 6, 2015. Accessed September 26, 2019. http://arxiv.org/abs/1502.01852

83. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444. doi:10.1038/nature14539

84. Uludağ K, Roebroeck A. General overview on the merits of multimodal neuroimaging data fusion. *NeuroImage*. 2014;102:3-10. doi:10.1016/j.neuroimage.2014.05.018

85. Hasan KM, Walimuni IS, Abid H, Wolinsky JS, Narayana PA. Multi-modal Quantitative MRI Investigation of Brain Tissue Neurodegeneration in Multiple Sclerosis. *J Magn Reson Imaging*. 2012;35(6):1300-1311. doi:10.1002/jmri.23539

86. Scheidhauer K, Walter C, Seemann MD. FDG PET and other imaging modalities in the primary diagnosis of suspicious breast lesions. *Eur J Nucl Med Mol Imaging*. 2004;31 Suppl 1:S70-79. doi:10.1007/s00259-004-1528-7

87. Pietrzyk U, Herholz K, Schuster A, Stockhausen H-M v., Lucht H, Heiss W-D. Clinical applications of registration and fusion of multimodality brain images from PET, SPECT, CT, and MRI. *European Journal of Radiology*. 1996;21(3):174-182. doi:10.1016/0720-048X(95)00713-Z

88. Calhoun VD, Sui J. Multimodal fusion of brain imaging data: A key to finding the missing link(s) in complex mental illness. *Biol Psychiatry Cogn Neurosci Neuroimaging*. 2016;1(3):230-244. doi:10.1016/j.bpsc.2015.12.005

89. Liu S, Cai W, Liu S, et al. Multimodal neuroimaging computing: a review of the applications in neuropsychiatric disorders. *Brain Inform*. 2015;2(3):167-180. doi:10.1007/s40708-015-0019-x

90. Hasan KM, Walimuni IS, Abid H, Datta S, Wolinsky JS, Narayana PA. Human Brain Atlas-based Multimodal MRI Analysis of Volumetry, Diffusimetry, Relaxometry and Lesion Distribution in Multiple Sclerosis Patients and Healthy Adult Controls: Implications for understanding the Pathogenesis of Multiple Sclerosis and Consolidation of Quantitative MRI Results in MS. *J Neurol Sci*. 2012;313(1-2):99-109. doi:10.1016/j.jns.2011.09.015

91. Tempany C, Jayender J, Kapur T, et al. Multimodal Imaging for Improved Diagnosis and Treatment of Cancers. *Cancer*. 2015;121(6):817-827. doi:10.1002/cncr.29012

92. Hao X, Xu D, Bansal R, et al. Multimodal magnetic resonance imaging: The coordinated use of multiple, mutually informative probes to understand brain structure and function. *Hum Brain Mapp*. 2011;34(2):253-271. doi:10.1002/hbm.21440

93. Gracien R-M, Petrov F, Hok P, et al. Multimodal Quantitative MRI Reveals No Evidence for Tissue Pathology in Idiopathic Cervical Dystonia. *Front Neurol*. 2019;10. doi:10.3389/fneur.2019.00914

94. Souza RB, Feeley BT, Zarins ZA, Link TM, Li X, Majumdar S. T1rho MRI Relaxation in Knee OA Subjects with Varying Sizes of Cartilage Lesions. *Knee*. 2013;20(2):113-119. doi:10.1016/j.knee.2012.10.018

95. Limkin EJ, Reuzé S, Carré A, et al. The complexity of tumor shape, spiculatedness, correlates with tumor radiomic shape features. *Scientific Reports*. 2019;9(1):4329. doi:10.1038/s41598-019-40437-5

96. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer*. 2012;48(4):441-446. doi:10.1016/j.ejca.2011.11.036

97. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. *Drug Discovery Today*. 2018;23(6):1241-1250. doi:10.1016/j.drudis.2018.01.039

98. Caliva F, Iriondo C, Martinez AM, Majumdar S, Pedoia V. Distance Map Loss Penalty Term for Semantic Segmentation. *arXiv:190803679 [cs, eess]*. Published online August 9, 2019. Accessed October 28, 2020. http://arxiv.org/abs/1908.03679

99. Martinez AM, Caliva F, Flament I, et al. Learning osteoarthritis imaging biomarkers from bone surface spherical encoding. *Magnetic Resonance in Medicine*. n/a(n/a). doi:10.1002/mrm.28251

100. Li X, Kuo D, Theologis A, et al. Cartilage in Anterior Cruciate Ligament–Reconstructed Knees: MR Imaging T1ρ and T2—Initial Experience with 1-year Follow-up. *Radiology*. 2011;258(2):505-514. doi:10.1148/radiol.10101006

101. Pedoia V, Gallo MC, Souza RB, Majumdar S. A longitudinal Study using voxel-based relaxometry: association between cartilage T1ρ and T2 and patient reported outcome changes in hip osteoarthritis. *J Magn Reson Imaging*. 2017;45(5):1523-1533. doi:10.1002/jmri.25458

102. Busovaca E, Zimmerman ME, Meier IB, et al. Is the Alzheimer's disease cortical thickness signature a biological marker for memory? *Brain Imaging Behav*. 2016;10(2):517-523. doi:10.1007/s11682-015-9413-5

103. Racine AM, Brickhouse M, Wolk DA, Dickerson BC. The personalized Alzheimer's disease cortical thickness index predicts likely pathology and clinical progression in mild cognitive impairment. *Alzheimers Dement (Amst)*. 2018;10:301-310. doi:10.1016/j.dadm.2018.02.007

104. Deshpande BR, Katz JN, Solomon DH, et al. The number of persons with symptomatic knee osteoarthritis in the United States: Impact of race/ethnicity, age, sex, and obesity. *Arthritis Care Res (Hoboken)*. 2016;68(12):1743-1750. doi:10.1002/acr.22897

105. Bhosale AM, Richardson JB. Articular cartilage: structure, injuries and review of management. *British Medical Bulletin*. 2008;87(1):77-95. doi:10.1093/bmb/ldn025

106. Goodwin DW, Dunn JF. High-Resolution Magnetic Resonance Imaging of Articular Cartilage: Correlation with Histology and Pathology. *Topics in Magnetic Resonance Imaging*. 1998;9(6):337.

107. Minciullo L, Parkes MJ, Felson DT, Cootes TF. Comparing image analysis approaches versus expert readers: the relation of knee radiograph features to knee pain. *Ann Rheum Dis*. 2018;77(11):1606-1609. doi:10.1136/annrheumdis-2018-213492

108. Neogi T, Frey-Law L, Scholz J, et al. Sensitivity and sensitisation in relation to pain severity in knee osteoarthritis: trait or state? *Ann Rheum Dis*. 2015;74(4):682-688. doi:10.1136/annrheumdis-2013-204191

109. Felson DT. Imaging abnormalities that correlate with joint pain. *Br J Sports Med*. 2011;45(4):289-291. doi:10.1136/bjsm.2010.081398

110. Neogi T. Clinical significance of bone changes in osteoarthritis. *Ther Adv Musculoskelet Dis*. 2012;4(4):259-267. doi:10.1177/1759720X12437354

111. Eckstein F, Wirth W. Quantitative Cartilage Imaging in Knee Osteoarthritis. *Arthritis*. 2011;2011. doi:10.1155/2011/475684

112. Grosan C, Abraham A. Rule-Based Expert Systems. In: Grosan C, Abraham A, eds. *Intelligent Systems: A Modern Approach*. Intelligent Systems Reference Library. Springer; 2011:149-185. doi:10.1007/978-3-642-21004-4_7

113. Riddle DL, Perera RA. The WOMAC Pain Scale and Crosstalk From Co-occurring Pain Sites in People With Knee Pain: A Causal Modeling Study. *Physical Therapy*. 2020;100(10):1872-1881. doi:10.1093/ptj/pzaa098

114. Roos EM, Lohmander LS. The Knee injury and Osteoarthritis Outcome Score (KOOS): from joint injury to osteoarthritis. *Health Qual Life Outcomes*. 2003;1:64. doi:10.1186/1477-7525-1-64

115. Davis KD, Flor H, Greely HT, et al. Brain imaging tests for chronic pain: medical, legal and ethical issues and recommendations. *Nature Reviews Neurology*. 2017;13(10):624-638. doi:10.1038/nrneurol.2017.122

116. Lombaert H, Grady L, Polimeni JR, Cheriet F. FOCUSR: feature oriented correspondence using spectral regularization--a method for precise surface matching. *IEEE Trans Pattern Anal Mach Intell*. 2013;35(9):2143-2160. doi:10.1109/TPAMI.2012.276

117. Besl PJ, McKay ND. A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1992;14(2):239-256. doi:10.1109/34.121791

118. Rogers MW, Wilder FV. The association of BMI and knee pain among persons with radiographic knee osteoarthritis: A cross-sectional study. *BMC Musculoskelet Disord*. 2008;9:163. doi:10.1186/1471-2474-9-163

119. Olsen MF, Bjerre E, Hansen MD, Tendal B, Hilden J, Hróbjartsson A. Minimum clinically important differences in chronic pain vary considerably by baseline pain and methodological factors: systematic review of empirical studies. *J Clin Epidemiol*. 2018;101:87-106.e2. doi:10.1016/j.jclinepi.2018.05.007

120. Bowes MA, Kacena K, Alabas OA, et al. Machine-learning, MRI bone shape and important clinical outcomes in osteoarthritis: data from the Osteoarthritis Initiative. *Annals of the Rheumatic Diseases*. 2021;80(4):502-508. doi:10.1136/annrheumdis-2020-217160

121. Neogi T. The Epidemiology and Impact of Pain in Osteoarthritis. *Osteoarthritis Cartilage*. 2013;21(9):1145-1153. doi:10.1016/j.joca.2013.03.018

122. Drozdzal M, Vorontsov E, Chartrand G, Kadoury S, Pal C. The Importance of Skip Connections in Biomedical Image Segmentation. *arXiv:160804117 [cs]*. Published online August 14, 2016. Accessed June 7, 2019. http://arxiv.org/abs/1608.04117

123. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. *arXiv:14126980 [cs]*. Published online December 22, 2014. Accessed June 7, 2019. http://arxiv.org/abs/1412.6980

124. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. :8.

125. Kamnitsas K, Ledig C, Newcombe VFJ, et al. Efficient Multi-Scale 3D CNN with Fully Connected CRF for Accurate Brain Lesion Segmentation. Published online March 18, 2016. doi:10.1016/j.media.2016.10.004

126. Myronenko A. 3D MRI brain tumor segmentation using autoencoder regularization. *arXiv:181011654 [cs, q-bio]*. Published online October 27, 2018. Accessed September 27, 2019. http://arxiv.org/abs/1810.11654

127. Wu Y, He K. Group Normalization. *arXiv:180308494 [cs]*. Published online March 22, 2018. Accessed September 27, 2019. http://arxiv.org/abs/1803.08494

128. Pereyra G, Tucker G, Chorowski J, Kaiser Ł, Hinton G. Regularizing Neural Networks by Penalizing Confident Output Distributions. *arXiv:170106548 [cs]*. Published online January 23, 2017. Accessed November 25, 2020. http://arxiv.org/abs/1701.06548

Appendix A: Supplementary Material to Chapter 3

A.1 Bone segmentation and post-processing

A.1.1 Bone segmentation network implementation

The first step of the study was to accurately segment the bones from the 3D DESS volumes in the OAI dataset. A CNN architecture was trained and used to segment the bone from the entire OAI dataset (**Fig. 3.1A**).

A modified 3D V-Net[76] architecture was adapted from an existing TensorFlow (Google, Mountain View, CA) implementation (https://github.com/MiguelMonteiro/VNet-Tensorflow) for the Femur, Tibia and Patella bone segmentation. The architecture consisted of an encoder-decoder network with the encoder network compressing the most relevant features for the segmentation task while the decoder network decompresses these features to reconstruct the labeled segmented volume. The decoder network has five levels, with each level doubling the number of convolutional filters and using short shortcut connections between each layer input and output in the form of element-wise addition. The network also uses long shortcut connections between each mirroring level by concatenating the layer output of each encoder layer to the layer input of its corresponding mirrored decoder layer. These connections have been shown to improve the uniform update of weights for deeper CNNs and improve gradient stability[122]. The activation function used after each convolution was a parametric leaky rectified linear unit (pReLU), trained on the last dimension of the input, and the last fully connected layer was activated with a softmax function for all the classes (femur, tibia, patella, background). A soft Dice loss function was used for the training.

A.1.2 Bone segmentation network training

The number of feature channels outputted at the very first encoding convolution, which determines the overall number of learned features in the V-Net, was empirically set to 8 and a batch size of one sample per feed-forward was used, which was the memory limit of the GPU. The network was trained using Adam optimizer[123] with a learning rate of 1e-4 using TensorFlow in a GeForce GTX Titan 1080 Ti GPU (NVIDIA, Santa Clara, CA). All the weights for the 3D convolutional layers were randomly initialized with a Xavier uniform distribution[124]. The training was performed for a total of 215 epochs and stopped early at 185 epochs after a 30-epoch patience for validation Dice non-improvement over the best validation Dice was reached. Data augmentation was performed online with an independent 50% chance of flipping the input volume along the lateral-medial dimension and an independent 50% chance to randomly rotate the sagittal plane in a range of -5 to +5 degrees in 1-degree increments. The labels were truncated to the integer part after the 2D sagittal affine rotation to ensure there were no artificial partial volume effects introduced by the augmentation.

The bone segmentation dataset consisted of 40 3D DESS volumes that were carefully annotated by a trained user. These 40 patients were selected from the greater OAI patient population as 20 matched pairs of patients that exhibited a 2-year OA incidence from a healthy baseline initial scan matched with healthy controls that did not exhibit 2-year OA incidence from a healthy baseline time point. The selected scans included in the segmentation dataset were the healthy baselines for both the healthy controls and the 2-year OA incidence cases. The age, BMI, and sex

were balanced for the OA incidence cases and the healthy controls with a mean age of 58.4±6.2 and 58.4±6.2 respectively. The BMI for the cases and healthy controls were 26.2±3.0 and 26.4±2.9 respectively. The sex split for the cases and healthy controls were 12 males and 8 females for both. The segmented patient MRI volumes for both the OA incidence and the healthy controls were from a healthy baseline time point. The network training was performed with 25 patients with 5 patients used for training validation. The model was evaluated using a test set with 10 unseen patient volumes. **Table 3.1** summarizes the distribution of OA incidence cases and healthy controls for the bone segmentation dataset as well as the statistical independence tests for confounding demographic factors across splits.

A.1.3 Bone segmentation post-processing

The trained V-Net segmentation model was then used to segment the Femur, Tibia, and Patella from a total of 47,078 3D DESS volumes. The inference was performed in 8 batches of 6,000 volumes and each batch lasted 3 hours. The inferred knee bone segmentations were further processed in MATLAB (MathWorks, Natick, MA) to conform to the necessary format for the spherical transformation. Biggest connected component analysis was performed on each bone segmentation to select the largest component and remove spurious artifacts followed by morphological closing performed over the entire volume. As a quality control measure, any resulting post-processed mask with a bone volume outside of three standard deviations of the training annotations was excluded from the study. There were 420 volumes excluded by this quality control measure. Each mask was converted to a smoothed point cloud, with the femur segmentation mask cropped along the femoral shaft before point cloud conversion in order to be invariant to the different femoral shaft lengths when sampling. The bone point clouds were

centered at the origin of the new coordinate system, with each point representing the actual distance in millimeters from this centroid to the surface of the bone. Each point cloud from a right knee was flipped along the lateral-medial dimension (to match the right knee scan orientation) and rigidly registered using an ICP algorithm[117] to a reference point cloud, randomly selected from a patient in the segmentation dataset, for each bone type to account for rotational variability at scan time (**Fig. 3.1B**). The orientation of the reference point cloud ensured that the articular surfaces for each bone were facing the same direction and no positional information was included in our shape model. The computation time for this post-processing step was a total of 18 hours split across 8 batches of 6,000 volumes.

A.1.4 Bone segmentation validation

The bone segmentation model was further validated with an additional test set of 60 manually segmented baseline scans randomly selected to represent the demographic distribution of the OAI population. Out of the 60 volumes, 30 were healthy controls from baseline scans of patients that never developed OA in both knees across all time points, 15 were from baseline scans of OA incidence cases with KL2> and 15 were baseline scans from OA diagnosis cases with KL1<, all selected randomly from the OAI baseline scan population. Out of the OA incidence cases, there were six 1-year cases, one 2-year case, one 3-year case, one 4-year, one 6-year case and four 8-year cases. Out of the OA diagnosis cases, there were nine KL2 cases, four KL3 cases and two KL4 cases. The scans were manually segmented by six authors of the study after 5 hours training with an experienced annotator. The manual segmentation of each case took 1 hour. Each user segmented an equal distribution of OA, control and incidence cases and were blinded to the labels. After segmenting the volumes, the performance of the automatic bone segmentation

model was evaluated on the three cohorts in terms of both Dice and MPTS distance errors. Each cohort was then tested for statistical significance to determine whether there was a bias introduced by the bone segmentation model for specific cohorts. There was no significant difference found across each cohort for each bone and for both metrics evaluated. The complete results of this validation are summarized in the **Supp. Table A.1**.

A.1.5 Osteophyte analysis

An analysis of the osteophyte coverage of the bone segmentation network was performed by a musculoskeletal radiologist on 20 randomly picked patients from the OAI baseline with osteophytes based on MOAKS grading. There was a total of 97 osteophytes (22 patellar, 33 tibial, and 42 femoral) across the 20 selected patients of representatively distributed MOAKS grades of 1 (small osteophytes), 2 (medium osteophytes) and 3 (large osteophytes). The radiologist identification was based on a 3D evaluation of the osteophyte volume coverage by the bone segmentation model on the sagittal DESS. There were four osteophyte identification categories ranging from not identified (<50%) to fully identified (>90%), with varying levels of identification (50-70% and 70-90%), as shown in **Supp. Fig. A.2**. The results of the osteophyte analysis are summarized in **Supp. Fig. A.3**. The bone segmentation network generally demonstrated correct identification of osteophytes, with at least partial coverage (>50%) of osteophytes on 80% (n = 78) of the total osteophytes observed in the analysis (n = 97). The patellar osteophytes were the least captured subtype of the total osteophytes, with only 60% (n = 13) of the total patellar osteophytes (n = 22) at least partially identified (>50%), which could potentially explain the lower performance of the Patella Diagnosis model.

A.1.6 3D patch-based approach comparison

A 3D patch-based approach for the bone segmentation was evaluated for the bone segmentation dataset. Four sets of experiments using patches of two different sizes, 48×48×40 and 96×96×40, were performed. For each patch size, a binary segmentation and a multiclass semantic segmentation model were trained. The network was a 2-level V-Net architecture with 16 channels after the first convolution. The output for the binary network was activated by a sigmoid function and the multiclass model by a softmax function. The batch size was set to 8 for both cases to maximize computational availability. Adam optimizer was also used with an initial learning rate of 5e-5. The training was performed for 100 epochs and early stopping was set to a 15-epoch patience for validation Dice non-improvement over the best validation Dice. No data augmentation was performed with a 95% keep probability dropout set as the regularization measure. At inference time, patches of each volume were re-arranged to form the volume and compute the volumetric dice over the whole volume. The best performing model was the binary segmentation model with a patch size of 96x96x40, with a validation and test Dice of 88.3% and 89.3% respectively.

Generally, the performance of the models increased with patch size, which then leads to a reduced batch size as the patch size approaches the dimensions of the full volume. The binary patch-based models outperformed the multiclass patch-based models, which introduced the added non-trivial challenge of separating the binary classification into the different bones during the post-processing steps. This challenge, paired with the lower performance when compared to

a full 3D approach reinforces our selection of a fully 3D volumetric approach, even at the expense of a limited batch size of volume. The possibility of having batches with different distribution as inputs, was reduced by normalizing our inputs in order to increase the similarity among different batches.

A.1.7 Bone segmentation discussion

On a further important note, the extraction of bone shape highly relies on a precise segmentation of the structures of interest. Previous studies have relied on advanced methods that segmented 3D structures using a patch-based approach[125] or a slice-based approach[43]. Patch-based approached have the drawback of limiting the spatial context, which could in turn affect the segmentation capability of a network, as shown in previous studies and our empirical findings, where context appeared to be crucial in order to achieve a high quality segmentation[126]. A slice-based approach has the benefit of being less memory demanding than a full volume 3D approach while still preserving more spatial context than a patch-based approach. Ultimately, the use of a small batch size for this study was encouraged by the results achieved by our empirical findings and previous studies[126,127] for full volume 3D segmentation approaches. The modified V-Net architecture of choice was an optimized version of the V-Net originally proposed customized to our specific application. A future direction of this study could investigate the underlying effect of different segmentation approaches on the performance of the OA classification tasks. It is also worth noting that while the segmentation dataset was limited to 40 3D volumes, the OA bone shape feature extraction portion of the proposed framework leverages 41,822 samples, thus better exploiting the scale of a large dataset such as the OAI.

One of the limitations acknowledged for the bone segmentation network was the comparatively poor osteophyte coverage for the patellar osteophytes with respect to the femoral and tibial osteophytes. While this could have potentially impacted the performance of our patella diagnosis model, it is worth noting that the radiographic KL grades used for the labels of the OA Diagnosis models are based on coronal radiographs that focus on tibiofemoral osteophytes and omit patellar osteophytes[9]. Furthermore, the presence of tibiofemoral osteophytes is strongly correlated with KL grades ranging from 2 to 4, which would impact the performance of the femoral and tibial OA Diagnosis models. For the OA Incidence models, only KL grades of 0 or 1 were selected and therefore osteophytes are not expected to similarly impact the performance of these models.

Instead of training with a bigger dataset, a more careful evaluation of the bone segmentation model was prioritized to avoid erroneous observations on the ability of the downstream classifiers to diagnose and predict future incidence of OA. Evenly distributed bone segmentation errors between OA and control classes as the ones observed in this ad-hoc experiment might decrease the classification performance by obscuring some relevant features and cause suboptimal performance. However, high sensitivity and specificity of the OA classification models show that the bone segmentation, even if imperfect, is able to capture sufficient bone shape features in order to diagnose and predict future incidence of OA. Better bone segmentation models in future studies might report improvements over this first experience with deep learning-based bone shape OA imaging biomarker extraction.

A.2 OA classification robustness analysis

A.2.1 Choice of bone atlas

The robustness of the OA Diagnosis and the 1-year and 2-year OA Incidence classification models to the choice of bone atlas was evaluated. Four patients with different KL grades and demographic information were randomly picked as the bone atlas (for the femur, tibia and patella). The entire framework was rerun on each bone atlas and the OA Diagnosis model and the 2-year and 8-year OA Incidence models were retrained using the same splits and hyperparameters as the original framework. The total computational time to rerun the entire framework was 55 hours, with 18 hours for the atlas registration and spherical transformation, 25 hours to retrain the femur, tibia, and patella OA Diagnosis models (needed for the logits averaging model ensemble) as well as 12 hours to retrain the femur, tibia, and patella 2-year and 8-year OA Incidence models (needed for the logits averaging model ensemble). Due to the large size of the dataset, each framework run generated a terabyte of storage, which paired with the computational time limited this analysis to only four atlases with varied KL grades and demographic distribution. The test set accuracy for each model was recorded for each bone atlas. The test ROC values for each atlas framework run were consistent with the original atlas test ROC values. The complete results of this analysis are shown in **Supp. Table A.2**.

A.2.2 Bone segmentation errors

Another additional validation ad hoc analysis of our framework evaluated the OA classification robustness to segmentation errors. The relationship between the accuracy of the segmentation model and the performance of the OA Diagnosis and first two OA Incidence models was

assessed. A total of 30 cases with a correct prediction, 15 true positives and 15 true negatives, for the test set of the OA Diagnosis model and the 1-year and 2-year OA Incidence models was randomly selected. The bone point clouds were corrupted by moving the coordinates of the points from 0mm to 3.2mm in 0.16mm increments by adding increasing amounts of Gaussian noise. The overall corruption amount was computed using MPTS distance errors. The altered point clouds were converted into spherical coordinates and inferred on the trained single-bone OA Diagnosis model and the trained single-bone 1-year and 2-year OA Incidence models. The effect of each individual bone corruption on the logits averaging ensemble performance was evaluated. Sensitivity, specificity and AUC of the 15 true positives and 15 true negatives were calculated as a function of MPTS distance. An overview of the results for each bone is shown in **Supp. Fig. A.4** for the Femur, **Supp. Fig. A.5** for the Tibia and **Supp. Fig. A.6** for the Patella. The average MPTS distance error for the segmentation test set along with the corresponding 95% confidence interval is shown as green and red vertical dotted lines respectively. The analysis showed that the OA Diagnosis and 1-year and 2-year OA Incidence models were sensitive to perturbations on each bone at different levels. The Tibia appeared to have the largest impact on the overall classification accuracy across all models, followed by the Femur. The Patella seemed to have the least impact in the overall classification accuracy, which is supported by the fact that radiographic KL grades evaluate tibiofemoral OA and largely ignore patellofemoral OA. The specificity of the models was the most sensitive ROC metric to the perturbation, suggesting that the models tend to over-predict anomalies as positive cases for both OA Diagnosis and OA Incidence. Furthermore, this analysis highlights the value of using a network ensemble method such as logits averaging instead of a single-bone model due to the added robustness to these segmentation errors.

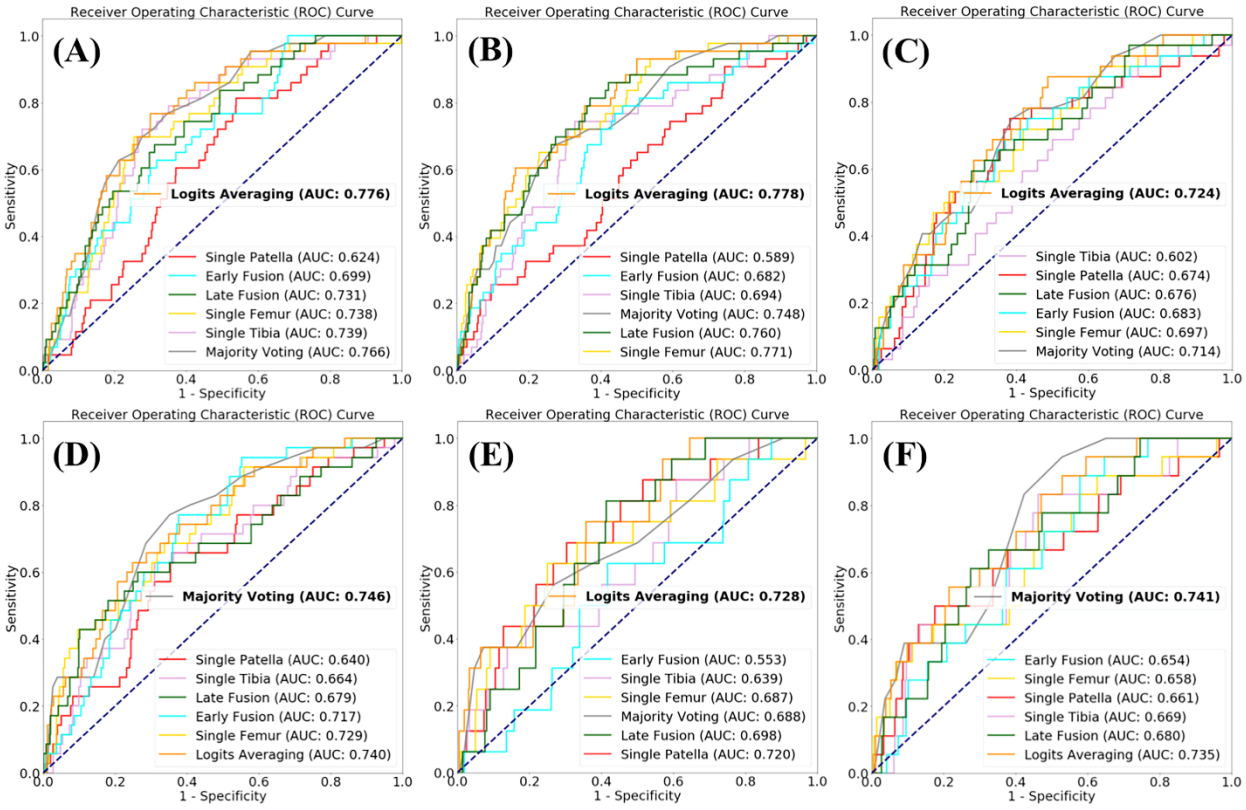
Supp. Table A.1 Overview of the additional validation of the bone segmentation model. Each of the three cohorts was tested for statistical significance to determine whether there was a bias introduced by the bone segmentation model for specific cohorts. There was no significant difference found across each cohort for each bone and for both metrics evaluated.

Results (Mean ± SD)	Bilateral No OA (N = 30)	OA Incidence (N = 15)	OA (N = 15)	Total (N =60)	OAI Baseline (N = 9592)
OA Cohort Distribution	All KL<2 for all time points	KL<2 at baseline with future incidence	All KL>1 at baseline	-	-
Age	58.5 ± 9.37	57.27 ± 8.57	60.4 ± 8.98	58.7 ± 9.15	61.16 ± 9.19
BMI (kg/m^2)	26.85 ± 4.89	30.35 ± 5.63	28.73 ± 6.19	28.2 ± 5.62	28.62 ± 4.84
Sex (Male/Female)	13/17	5/10	6/9	24/36	1992/2804
Femur MPTS (mm)	0.487 ± 0.117	0.545 ± 0.173	0.536 ± 0.159	0.514 ± 0.142	-
Femur Dice	96.7 ± 0.69	96.2 ± 1.15	96.1 ± 1.40	96.5 ± 1.07	-
Tibia MPTS (mm)	0.607 ± 0.235	0.603 ± 0.152	0.587 ± 0.132	0.601 ± 0.190	-
Tibia Dice	95.9 ± 1.04	95.9 ± 0.698	95.8 ± 0.894	95.9 ± 0.933	-
Patella MPTS (mm)	0.405 ± 0.0714	0.411 ± 0.0978	0.506 ± 0.286	0.431 ± 0.158	-
Patella Dice	94.6 ± 1.01	93.9 ± 1.38	93.8 ± 2.46	94.2 ± 1.63	-

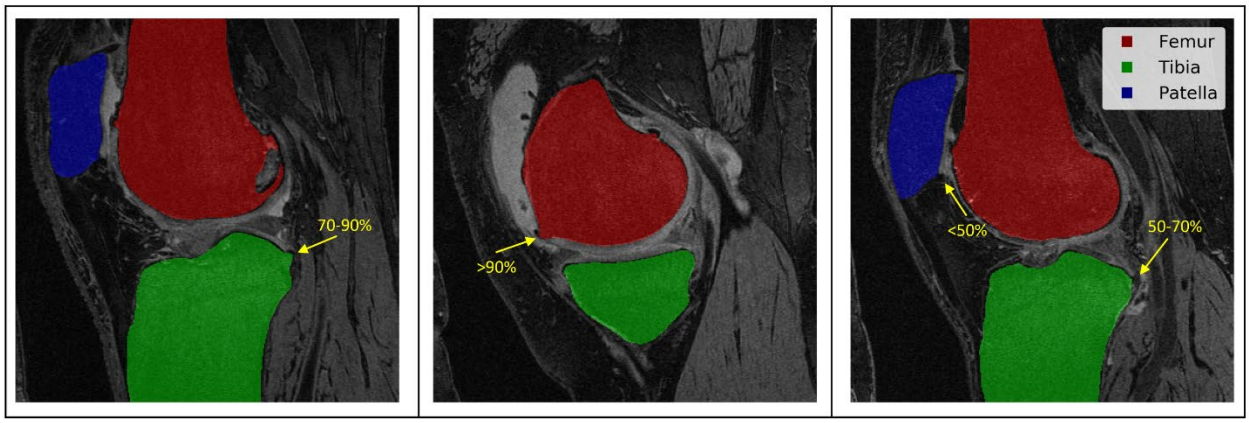
Supp. Table A.2 Overview of the OA classification robustness to the choice of bone atlas. The entire framework was rerun, with three of the OA Diagnosis and the 2-year and 8-year OA Incidence models retrained on the same splits, on four different bone atlases. The test ROC values for each atlas framework run were consistent with the original atlas test ROC values.

Results	Original Atlas	Atlas #1	Atlas #2	Atlas #3	Atlas #4
KL Grade	0	3	2	3	4
Age	67	64	72	67	58
BMI (kg/m^2)	25.5	23.0	25.6	38.7	26.1
Sex	Female	Female	Male	Male	Female
OA Diagnosis Test (Sensitivity/Specificity/AUC)	0.815 0.839 0.905	0.792 0.859 0.900	0.796 0.852 0.901	0.805 0.844 0.901	0.805 0.849 0.906
OA 2-year Incidence Test (Sensitivity/Specificity/AUC)	0.683 0.759	0.635 0.840	0.619 0.804	0.698 0.734	0.714 0.778

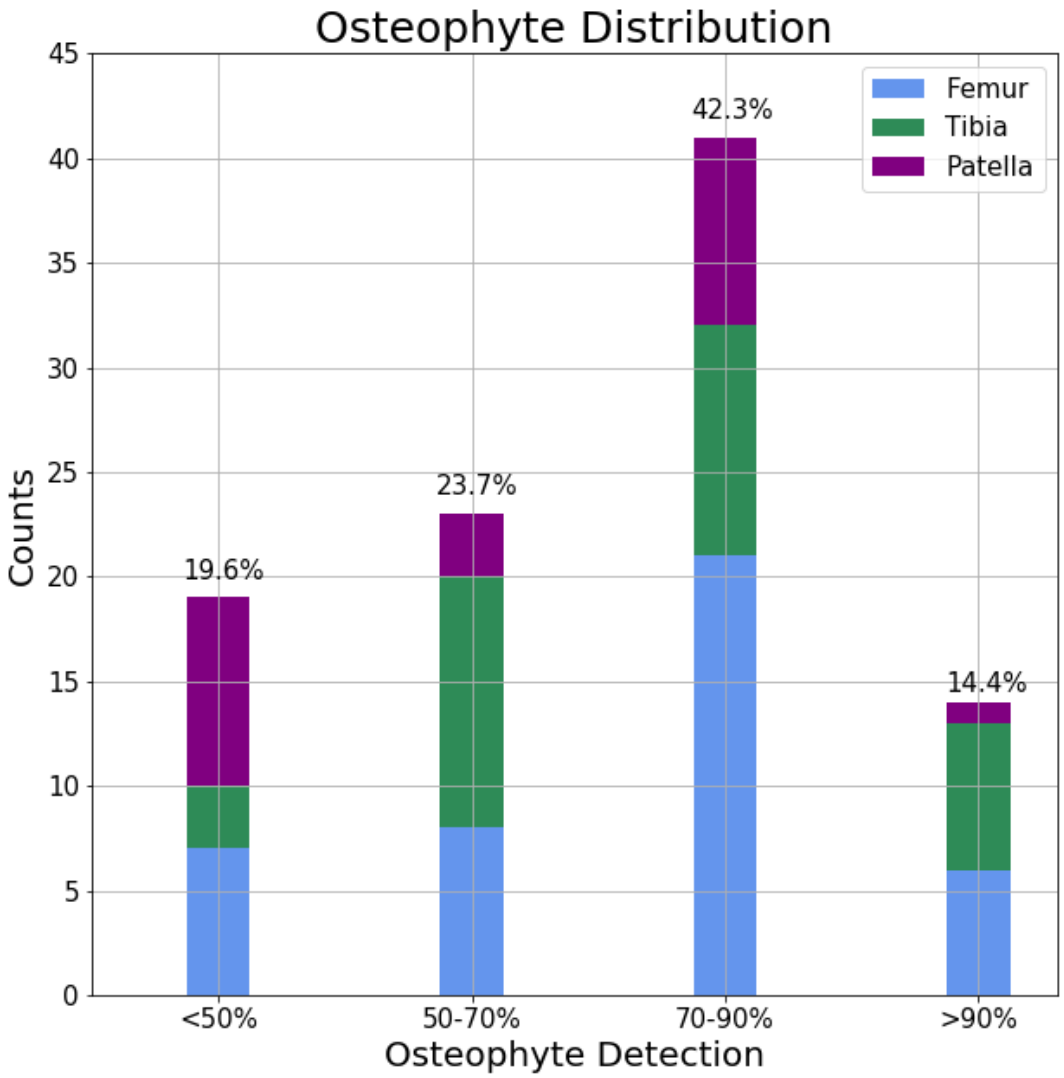
Results	Original Atlas	Atlas #1	Atlas #2	Atlas #3	Atlas #4
	0.815	0.838	0.814	0.804	0.830
OA 8-year Incidence Test (Sensitivity/Specificity/AUC)	0.555	0.611	0.667	0.667	0.611
	0.582	0.694	0.614	0.635	0.597
	0.646	0.696	0.647	0.692	0.647



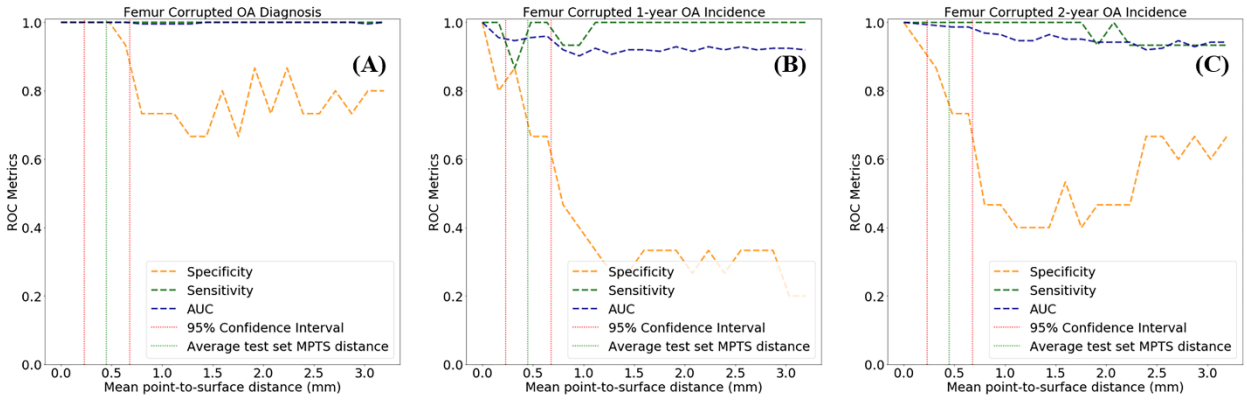
Supp. Fig. A.1 Overview of the validation ROC curve comparisons for the different model fusion strategies. (A-F), 3-year to 8-year OA Incidence models, shown in order.



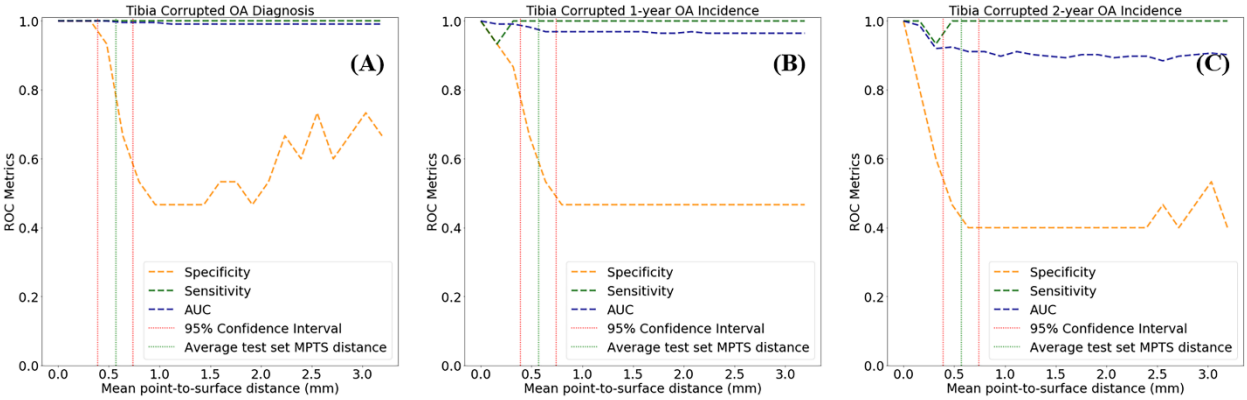
Supp. Fig. A.2 A representative slice from three patients out of the 20 randomly picked patients from the osteophyte analysis. The radiologist identification was based on a 3D evaluation of the osteophyte volume coverage by the bone segmentation model on the sagittal DESS. All four osteophyte identification categories are shown, ranging from not identified (<50%) to fully identified (>90%), with varying levels of identification (50-70% and 70-90%).



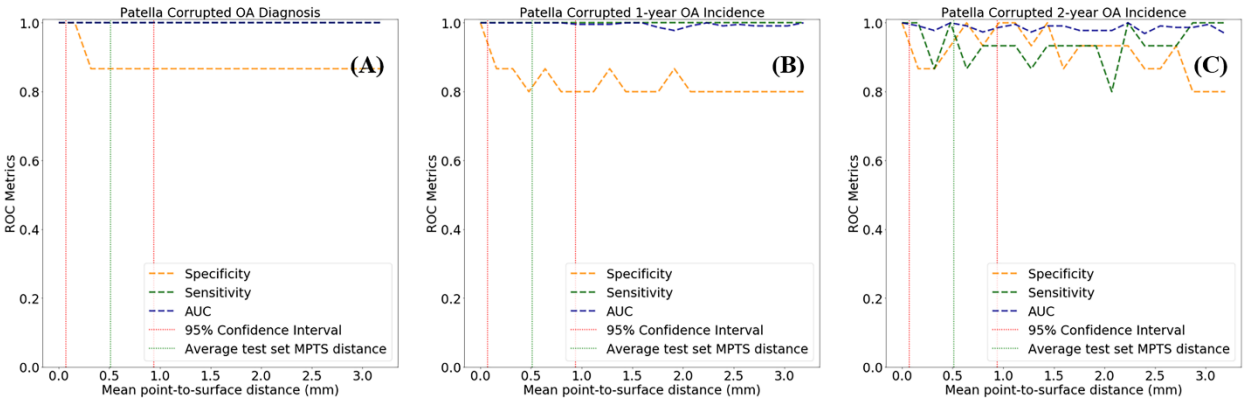
Supp. Fig. A.3 The results of the osteophyte analysis of the bone segmentation network. The analysis was performed by a musculoskeletal radiologist on 20 randomly selected patients from the OAI baseline with osteophytes based on MOAKS grading. There was a total of 97 osteophytes (22 patellar, 33 tibial, and 42 femoral). The osteophyte analysis generally demonstrated correct identification of osteophytes by the bone segmentation network, with at least partial coverage (>50%) of osteophytes on 80% (n = 78) of the total osteophytes observed in the analysis (n = 97). The patellar osteophytes were the least captured subtype of the total osteophytes, which could potentially explain the lower performance of the Patella Diagnosis model.



Supp. Fig. A.4 Robustness of the OA classification models to the Femur segmentation error measured as mean point-to-surface distance errors on the automatic segmentations. The segmentation error was simulated by adding increasingly more Gaussian noise to the Femur point cloud coordinates before the spherical transformation. The average MPTS distance error and corresponding 95% confidence interval between the automatic Femur segmentations in the segmentation test set and the manual segmentations is included. (A), OA Diagnosis model. (B), 1-year OA Incidence model. (C), 2-year OA Incidence model.



Supp. Fig. A.5 Robustness of the OA classification models to the Tibia segmentation error measured as mean point-to-surface distance errors on the automatic segmentations. The segmentation error was simulated by adding increasingly more Gaussian noise to the Tibia point cloud coordinates before the spherical transformation. The average MPTS distance error and corresponding 95% confidence interval between the automatic Tibia segmentations in the segmentation test set and the manual segmentations is included. (A), OA Diagnosis model. (B), 1-year OA Incidence model. (C), 2-year OA Incidence model.



Supp. Fig. A.6 Robustness of the OA classification models to the Patella segmentation error measured as mean point-to-surface distance errors on the automatic segmentations. The segmentation error was simulated by adding increasingly more Gaussian noise to the Patella point cloud before the spherical transformation. The average MPTS distance error and corresponding 95% confidence interval between the automatic Patella segmentations in the segmentation test set and the manual segmentations is included. (A), OA Diagnosis model. (B), 1-year OA Incidence model. (C), 2-year OA Incidence model.

Appendix B: Supplementary Material to Chapter 4

B.1 Bone segmentation

B.1.1 Bone segmentation network implementation

The first step of the study was to accurately segment the bones from the 3D-DESS volumes in the OAI dataset. An ensemble of five 3D V-Net[76] architectures were trained and tested on 72 and 30 3D-DESS volumes, respectively, and used to segment the bone from the entire OAI dataset (**Fig. 3.1A**).

A modified V-Net architecture was adapted from an existing TensorFlow 1.0 (Google, Mountain View, CA) implementation (https://github.com/MiguelMonteiro/VNet-Tensorflow) for the femur, tibia and patella bone segmentation. The 3D V-Net The architecture consisted of an encoder-decoder network with the encoder network compressing the most relevant features for the segmentation task while the decoder network decompresses these features to reconstruct the labeled segmented volume. The decoder network has five levels, with each level doubling the number of convolutional filters and using short shortcut connections between each layer input and output in the form of element-wise addition. The network also uses long shortcut connections between each mirroring level by concatenating the layer output of each encoder layer to the layer input of its corresponding mirrored decoder layer. These connections have been shown to improve the uniform update of weights for deeper CNNs and improve gradient stability[122]. The activation function used after each convolution was a ReLU, trained on the last dimension of the input, and the last fully connected layer was activated with a softmax function for all the classes (femur, tibia, patella, background). Additionally, a dropout rate of 0.05 was

used to improve generalizability of the model during training, randomly turning off activations at a rate of 5%.

Each of the five V-Net models was trained with a with different distance-weighted loss functions[98]. The distance weighting was an added penalty to ensure that the segmentation accuracy was prioritized along the surface of the bone and cartilage. This ensured that the articular bone surface was as accurate as possible prior to the biomarker projection. Additionally, given the class imbalance between the different bones, with the femur being much larger than the patella, class weights were added to four of the losses to ensure that the learning process was balanced. The distance-weighted loss functions were: class-weighted dice loss, class-weighted cross-entropy loss, mixed weighted cross-entropy and class-weighted dice loss (with the weighting factor for the cross-entropy loss equal to 0.1), class-weighted penalized confident output cross-entropy loss[128], and regular dice loss.

B.1.2 Bone segmentation network training

A batch size of one sample per feed-forward was used, which was the memory limit of the GPU. The network was trained using Adam optimizer[123] with a learning rate of 5e-4 using TensorFlow 1.10 in a Titan 1080 Ti 12GB GPU (NVIDIA, Santa Clara, CA). All the weights for the 3D convolutional layers were randomly initialized with a Xavier uniform distribution[124]. The training was performed for a total of 500 epochs and stopped early after a 30-epoch patience for validation loss non-improvement over the best validation loss reached. MRI volumes were cropped from Data augmentation was performed online with an independent 50% chance of

flipping the input volume along the lateral-medial dimension and an independent 50% chance to randomly rotate the sagittal plane in a range of -5 to +5 degrees in 1-degree increments. The labels were truncated to the integer part after the 2D sagittal affine rotation to ensure there were no artificial partial volume effects introduced by the augmentation.

The bone segmentation training consisted of 102 3D-DESS volumes that were carefully annotated by trained users. The age and BMI for the training split with the respective standard deviation was 57.2 ± 7.4 and 27.5 ± 5.2 respectively. The age and BMI for the validation split with the respective standard deviation was 60.9 ± 10.6 and 28.9 ± 4.2 respectively. The age and BMI for the test split with the respective standard deviation was 59.4 ± 7.6 and 27.2 ± 4.7 respectively. The sex split for training, validation and test splits was 31 males/26 females, 7 males/8 females, and 11 males/19 females respectively. The network training was performed with 57 patients with 15 patients used for training validation. The model was evaluated using a test set with 30 unseen patient volumes. **Table 3.1** summarizes the distribution of OA cases and healthy controls for the bone segmentation dataset as well as the statistical independence tests for confounding demographic factors across splits.

B.1.3 Bone segmentation inference and ensembling

The trained V-Net bone ensemble segmentation model was then used to segment the femur, tibia, and patella from a total of 47,078 3D-DESS volumes in the OAI. The inference was performed in 8 batches of 6,000 volumes and each batch lasted 3 hours. The inferred bone segmentation

masks for all five models were then subsequently ensembled by averaging the softmax values for each bone across all models.

B.2 Cartilage segmentation

B.2.1 Cartilage segmentation network implementation

A cartilage and menisci segmentation model ensemble was trained on 148 3D-DESS volumes and tested on 28 3D-DESS volumes[20]. The trained ensemble consisted of three 2D V-Net and three 3D V-Net architectures and was used to segment the cartilage and menisci in the OAI dataset (**Fig. 3.1A**).

The same 3D V-Net architecture as the bone segmentation V-Net was implemented in Tensorflow 1.10. The 2D V-Net architectures were derived from the 3D V-Net, where the convolution kernels are modified to accommodate 2D data. The 2D V-Nets were 2 levels deep with 4 convolutions at each level, and 4 convolutions at the bottom level, all activated with ReLU functions. At the output layers, a sigmoid activation produced the tissue segmentations. A dropout rate of 0.05 was used to improve generalizability of the model during training, randomly turning off activations at a rate of 5%.

B.2.2 Cartilage segmentation network training

The network was trained using Adam optimizer with a learning rate of 1e-4 using TensorFlow in a Titan 1080 Ti 12GB GPU or V100 32GB GPU. All the weights for the convolutional layers

were randomly initialized with a Xavier uniform distribution. The training was performed with an early stopping patience criterion of 30 epochs, when validation loss non-improvement over the best validation loss was reached. Training volumes were augmented offline using a random combination of geometric and intensity-based transforms, chosen to simulate 3D variations in patient positioning, bone shape, cartilage thickness, and MR imaging artifacts. Pooled training/validation data totaled 2812 3D-DESS volumes: 148 original volumes plus 2664 augmented volumes. Volumes were also flipped to medial-first orientation, center-cropped to 344x344x140 and normalized to their 85th intensity percentile. 2D models were trained using slices of the original dataset, while 3D models were trained using the augmented dataset to prevent overfitting.

The cartilage and menisci segmentation dataset consisted of 176 3D-DESS volumes that were provided by IMorphics. The age and BMI for the training-validation split with the respective standard deviation was 59.9 ± 1.6 and 30.9 ± 0.7 respectively. The age and BMI for the test split with the respective standard deviation was 71.4 ± 2.9 and 30.8 ± 1.6 respectively. The sex split for training-validation and test splits was 72 males/76 females and 18 males/10 females respectively. Each of the six segmentation models was trained on an independent data split of 50 training and 98 validation volumes, with the same 28 testing volumes, for which the manual segmentation was available.

B.2.3 Cartilage segmentation inference and ensembling

The trained V-Net cartilage and menisci ensemble segmentation model was then used to segment the femoral, tibial, and patellar cartilage, as well as the menisci, from a total of 47,078 3D-DESS volumes in the OAI. The inference was performed in 8 batches of 6,000 volumes and each batch lasted 3 hours. Softmax prediction values from the 3D and 2D models from each of the independent splits were ensembled to produce the final probability maps. Since the OAI only collected matching T_2 MSME, needed for the compositional T_2 spherical maps, MRI scans for the right knee of each patient, a subset of 21,118 out of the 47,078 segmented volumes were selected for this study.

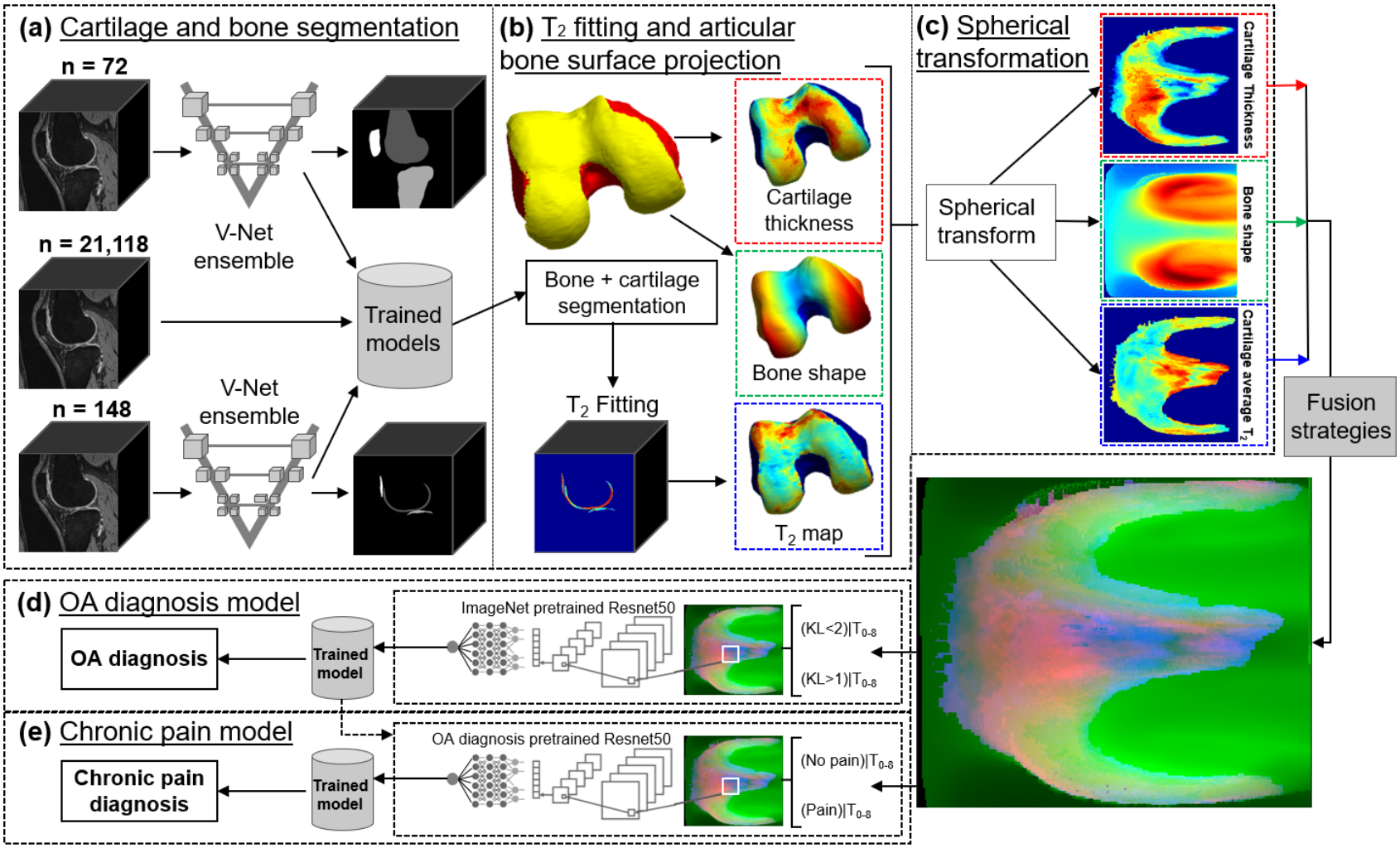
Appendix C: Supplementary Material to Chapter 5

C.1 OA diagnosis network implementation

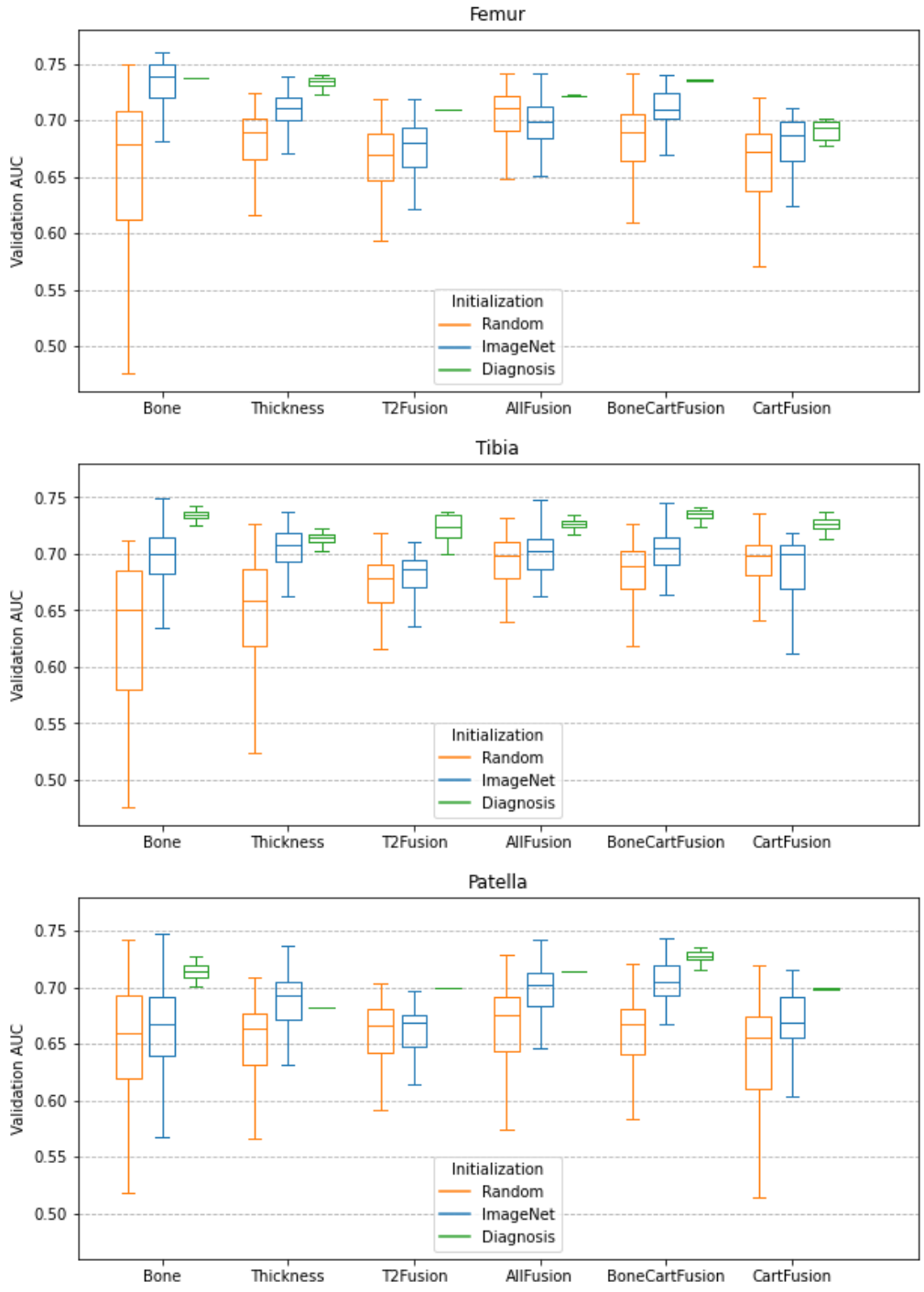
A total of 18 binary classification models, one for each biomarker strategy per bone, were trained to extract biomarker features from the spherical biomarker representations and use them to diagnose OA (**Supp. Fig. C.1**). A Resnet[38] architecture with 50 layers (Resnet50) pre-trained with ImageNet weights was implemented in PyTorch[79]. The choice of architecture and hyperparameters was informed by our previous study on the relationship between bone shape and radiographic OA[99]. The Resnet50 network architecture uses shortcut residual connections that improve the training performance for deeper models over similar shallower models. The basic structure of the Resnet50 follows the pattern of three convolutional layers with a 1 x 1, 3 x 3, and a 1 x 1 convolutional filter size respectively. Each of these layers is paired with batch normalization and a ReLU activation function.

All OA diagnosis model variants were initialized with ImageNet weights and fine-tuned using Adam optimizer with a learning rate of 1e-5 with a regularization weight decay value of 0.9, in order to finetune while preventing overfitting on the training set. The training was performed for 100 epochs with an early stopping 15-epoch patience for validation loss non-improvement over the best validation loss reached. The models were also trained end-to-end using a weighted binary cross entropy loss, based on the class imbalance, with a batch size of 300 in a Tesla V100 32GB GPU.

The OA diagnosis models were trained using the different biomarker strategies outlined in **Fig. 3**. The OA diagnosis models for each biomarker strategy were ensembled across the bones by averaging the softmax values outputted by each network. Therefore, each of the six biomarker models had a total of five predictive values: for the patella, for the tibia, for the femur, for the averaged predictive values of the tibia and femur, and for the average predictive values of all three bones. For the averaged ensembles, each anatomical region contributes equally to the final prediction.


(a) Cartilage and bone segmentation
n = 72
n = 21,118
n = 148
V-Net ensemble
V-Net ensemble
Trained models
(b) T2 fitting and articular bone surface projection
Bone + cartilage segmentation
T2 Fitting
Cartilage thickness
Bone shape
T2 map
(c) Spherical transformation
Spherical transform
Cartilage Thickness
Bone shape
Cartilage average T2
Fusion strategies
(d) OA diagnosis model
OA diagnosis
Trained model
ImageNet pretrained Resnet50
(KL<2)|T0-8
(KL>1)|T0-8
(e) Chronic pain model
Chronic pain diagnosis
Trained model
OA diagnosis pretrained Resnet50
(No pain)|T0-8
(Pain)|T0-8

Supp. Fig. C.1 (a) A bone and a cartilage segmentation model ensemble were trained on 72 and 148 manually segmented 3D-DESS volumes to segment the femur, tibia, and patella bones and corresponding cartilage. The trained models were used to segment 21,118 3D-DESS volumes. (b) Bone shape feature and cartilage thickness maps were obtained from the segmented masks. T2 values were calculated by registering 3D-DESS cartilage masks to the matching MSME MRI volumes and performing parametric T2 fitting on the cartilage. Each biomarker was then projected onto the articular bone surface, where each point contained information from each biomarker. (c) The articular bone surface projections were transformed into spherical coordinates. Six different strategies were performed to merge spherical maps per bone. (d) A total of 21,118 merged spherical maps with corresponding KL grades were used to train classifier models to diagnose radiographic OA using the biomarker learned features. A different model was trained and tested for each biomarker strategy per bone, for a total of 18 OA diagnosis models. Each of the two inputs into the OA diagnosis models represents a class in the binary classifier (healthy KL<2 vs. OA KL>1). (e) A total of 7,437 merged spherical maps with corresponding chronic pain labels were used to train classifier models pretrained on its corresponding OA diagnosis model to predict chronic pain. A different model trained and tested for each biomarker strategy per bone, for a total of 18 OA diagnosis models. Each of the two inputs into the chronic pain models represents a class in the binary classifier (chronic pain vs. no chronic pain).



Supp. Fig. C.2 Model training optimization results shown for all 18 models using the training and validation splits with two different learning rates (1x10-4 and 1x10-5), three types of Resnet (Resnet18, Resnet34, Resnet50), three initialization strategies (Random, ImageNet, OA), and four variants of layer freezing during training (first layer, first two layers, all layers, no layers), for a total of 612 combinations. The best performing models for each initialization strategy are shown with the validation AUC for each biomarker and bone.

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:
493E28E867594A2...
Author Signature

5/11/2021
Date