Title
Learning osteoarthritis imaging biomarkers from bone surface spherical encoding

Journal
Magnetic Resonance in Medicine, 84(4)

ISSN
0740-3194

Authors
Martinez, Alejandro Morales
Caliva, Francesco
Flament, Io
et al.

Publication Date
2020-10-01

DOI
10.1002/mrm.28251

Peer reviewed

HHS Public Access

Author manuscript

Magn Reson Med. Author manuscript; available in PMC 2020 October 01.

Published in final edited form as:

Magn Reson Med. 2020 October ; 84(4): 2190–2203. doi:10.1002/mrm.28251.

Learning osteoarthritis imaging biomarkers from bone surface spherical encoding

Alejandro Morales Martinez[1,2,3], **Francesco Caliva**[1], **Io Flament**[1], **Felix Liu**[4], **Jinhee Lee**[1], **Peng Cao**[5], **Rutwik Shah**[1], **Sharmila Majumdar**[1,2,3], **Valentina Pedoia**[1,2,3,6]

[1]Department of Radiology and Biomedical Imaging, University of California, San Francisco, California

[2]Graduate Program in Bioengineering, University of California, San Francisco, California

[3]Graduate Program in Bioengineering, University of California, Berkeley, California

[4]Department of Epidemiology and Biostatistics, University of California, San Francisco, California

[5]Department of Diagnostic Radiology, The Hong Kong University, Hong Kong, China

[6]Center for Digital Health Innovation (CDHI), University of California, San Francisco, California

Abstract

Purpose: To learn bone shape features from spherical bone map of knee MRI images using established convolutional neural networks (CNN) and use these features to diagnose and predict osteoarthritis (OA).

Methods: 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 Kellgren-Lawrence 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 time points, from 1 y to 8 y. The validation performance was compared and the test set performance was reported.

Results: The OA Diagnosis model had an area-under-the-curve (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 y to 8 y.

Conclusion: 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.

Correspondence Valentina Pedoia, Department of Radiology and Biomedical Imaging, University of California, San Francisco, 1700 4th St., Suite 201C, CA 94143. valentina.pedoia@ucsf.edu.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

Keywords

bone shape; deep learning; musculoskeletal MRI; osteoarthritis

1 | INTRODUCTION

Osteoarthritis (OA) is a degenerative joint disease which affects over 30 million U.S. adults, with the global prevalence of OA approaching 5%.[1,2] Risk factors commonly associated with OA include obesity, aging, and sex.[3] 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.[4–6] 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.[7–10] Studies based on 2D radiographs have reported sex-based bone shape differences in subjects with lateral and medial OA.[11] The classical approach to represent bone shape has been through Statistical Shape Modeling (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 principal component analysis (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.[8,10,12–14] This technique has also been used to evaluate the contribution of knee shape to anterior cruciate ligament (ACL) tears,[15] in order to assess the association between bone shape and the progression of cartilage degeneration[16] as well as altered knee kinematics[17] 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, such 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 convolutional neural networks (CNN) architectures in medical image processing diagnostic tasks show promising results in fully exploiting the image information.[18–20] These techniques have dramatically improved outcomes of challenging problems in a variety of fields such as object detection, classification,[21,22] drug discovery and genomics.[23] However, the number of validated applications in MRI, and specifically in musculoskeletal imaging research, remain limited.[24–26]

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 (KL) grade.[27]

2 | METHODS

2.1 | Methods overview

The overall study overview is summarized in Figure 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 (Figure 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 (Figure 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 (Figure 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 (Figure 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 y up to 8 y, and was tested on the same models as the OA diagnosis task (Figure 1E).

2.2 | Patient imaging dataset

The imaging data for this study was acquired from the Osteoarthritis Initiative (OAI), a multi-center longitudinal multi-modality imaging study in 4796 patients.[28] 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, body mass index (BMI), and sex, were recorded during each visit. Out of the 12 time points covered in the OAI, spanning 10 y, only 7 time points had MRI scans performed, which limited the span of the study to 8 y. A total of 41 822 3D sagittal double echo steady-state (DESS) volumes from the OAI acquired (3.0T Siemens Trio) were used for this study (field of view = 14 cm; matrix = 384 × 307 × 160; repetition time/echo time = 16.2/4.7 ms; bandwidth = 62.5 kHz; resolution = 0.365 × 0.456 × 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. The human studies were conducted with the approval of the Institutional Review Board.

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 4506 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),

7807 (KL = 1), 10,240 (KL = 2), 5528 (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 × 304 × 160 acquisition resolution to 384 × 384 × 160 for sagittal in-plane isotropic resolution. Each of the 41,822 DESS image volumes used was cropped from 384 × 384 × 160 to 364 × 364 × 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.

2.3 | 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[29] architecture was used for the femur, tibia, and patella bone segmentation (Figure 1A). Lateral-medial flipping as well as in-plane rotation data augmentation were 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 Supporting Information Section 1: Bone Segmentation and Post-processing, which is available online.

2.4 | 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 (Figure 2). The transformation from Cartesian coordinates into spherical coordinates was performed by uniformly sampling 224 × 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) (Figure 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 × 224 points, which was required to conform to the ImageNet image size, amounted to approximately 50 000 points. This was an over-sampling 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.

2.5 | 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 three-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[30] 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.

2.6 | 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, 6365 validation images, and 6445 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[31] 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. 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 1 y to 8 y for radiographic OA incidence. The training, validation, and test splits were randomized for every OA Incidence period (1 y to 8 y) to balance the classes across splits as well as ensure that the demographic factors were independent across splits. Table 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.

2.7 | 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 (Figure 1D), a Resnet[32] architecture with 50 layers (Resnet50) pre-trained with ImageNet weights was implemented in PyTorch.[33] The selection of the Resnet architecture was informed through a CNN architecture grid search that included DenseNet,[34] AlexNet,[35] SqueezeNet,[36] 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.[37] 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 Figure 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 hyperparameter 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 1×1, 3×3, and a 1×1 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. Figure 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 (Figure 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 (Figure 3D). The late fusion model was the concatenation of the last three layers of the individual Resnet50 trained fused into a fully connected layer and trained end to end (Figure 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 (Figure 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 (Figure 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 fine tune 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.

2.8 | 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 y and 8 y 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 y and 8 y 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 Supporting Information Section 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-y, and 30 false positives and 20 false negatives for the 2-y OA) were selected from the trained OA Diagnosis model and the 1-y and 2-y OA Incidence models. The 1-y and 2-y OA Incidence models were evaluated due to the lack of cases in later year incidences. The

distribution of spherical transformation errors measured as mean point-to-surface (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 | RESULTS

3.1 | 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. Figure 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 Supporting Information Table S1. Additionally, the results of the osteophyte analysis are shown in Supporting Information Figures S2 and S3.

3.2 | 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.

3.3 | OA classification models

The validation receiver operating characteristic (ROC) curve results for the binary OA classifier models are summarized in Table 2 and a visual representation is reported in Figure 5 for the OA Diagnosis and the 1-y and 2-y OA Incidence models. The rest of the OA Incidence models ranging from 3 y to 8 y can be found in the Supporting Information Figure S1. 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-y incidence model.

The test set performance of the models was in line with the validation set performance with the exception of the 7-y 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-y and 8-y model. There was no clear trend in overall performance across each OA Incidence model, with the best performing OA Incidence being the 7-y model. The test set performance, measured in AUC, sensitivity, and specificity, is summarized in Table 3. The OA Diagnosis and the 2-y and 8-y OA Incidence models test set performance for four different bone atlases was also consistent with the original results and is shown in Supporting Information Table S2.

3.4 | 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-y and 2-y 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-y 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-y 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 Supporting Information Figure S4, Supporting Information Figure S5, and Supporting Information Figure S6.

The complete results of both analyses can be found in the Supporting Information Section 2: OA Classification Robustness Analysis.

The OA classification robustness to spherical transformation error, measured in MPTS distance errors, overview is shown in Figure 6. The OA Diagnosis model robustness to spherical transformation error is shown in Figure 6A. The 1-y OA Incidence model robustness to spherical transformation error is shown in Figure 6B. The 2-y OA Incidence model robustness to spherical transformation error is shown in Figure 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.

4 | DISCUSSION

In this study, we established a model to diagnose and predict knee OA onset within a period ranging from 1 y to 8 y 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.

Classic 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, using the fact that many natural image patterns are compositional hierarchies, meaning higher-level features can be decomposed into lower-level feature representations.[38] 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.[38] Some studies have combined supervised learning techniques such as linear discriminant analysis (LDA) with PCA to link bone shape to OA.[10] LDA best separates two groups (OA and no OA) with a hyper-plane 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 use of a single vector may be an over simplification of a complex 3D shape, 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 gradient-weighted class activation mapping (Grad-CAM[39]), 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.

5 | CONCLUSIONS

In this study, we demonstrated that bone shape can be used as a predictive biomarker for OA through the use of a spherical transformation and coupled with a classification CNN. This approach is novel in the field of deep learning applications for musculoskeletal imaging and shows promise for future expansions to study other OA biomarkers such as cartilage thickness and T2 relaxation times values.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding information

National Institute of Arthritis and Musculoskeletal and Skin Diseases, Grant/Award Number: R00AR070902 and R61AR073552

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. 2016;68:574–580.
2. 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;73:1323–1330. [PubMed: 24553908]
3. Hootman JM, Helmick CG. Projections of US prevalence of arthritis and associated activity limitations. Arthritis Rheum. 2006;54:226–229. [PubMed: 16385518]
4. 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.

5. 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;18:507–512. [PubMed: 2588028]
6. Pauwels F Biomechanics of the Locomotor Apparatus: Contributions on the Functional Anatomy of the Locomotor Apparatus. Berlin, Heidelberg: Springer-Verlag; 1980 https://www.springer.com/us/book/9783642671401. Accessed June 12, 2019.
7. Lynch JA, Parimi N, Chaganti RK, Nevitt MC, Lane NE. The association of proximal femoral shape and incident radiographic hip OA in elderly women. Osteoarthr Cartil OARS Osteoarthr Res Soc. 2009;17:1313–1318.
8. 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;43:1780–1786. [PubMed: 20227696]
9. 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;64:1457–1465. [PubMed: 22544526]
10. 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 Rheum. 2013;65:2048–2058. [PubMed: 23650083]
11. 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. Osteoarthr Cartil OARS Osteoarthr Res Soc. 2014;22:2067–2073.
12. 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 Cartilage. 2013;21:443–449. [PubMed: 23274103]
13. 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;74:519–525. [PubMed: 24306109]
14. 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;75:1607–1614. [PubMed: 26483253]
15. 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. Osteoarthr Cartil OARS Osteoarthr Res Soc. 2015;23:1695–1703.
16. 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 JMRI. 2016;43:970–980. [PubMed: 26443990]
17. Lansdown DA, Pedoia V, Zaid M, et al. Variations in knee kinematics after ACL injury and after reconstruction are correlated with bone shape differences. Clin Orthop. 2017;475: 2427–2435. [PubMed: 28451863]
18. Lee H, Tajmir S, Lee J, et al. Fully automated deep learning system for bone age assessment. J Digit Imaging. 2017;30:427–441. [PubMed: 28275919]
19. Becker AS, Blüthgen C, Phi van VD, et al. Detection of tuber-culosis patterns in digital photographs of chest X-ray images using deep learning: Feasibility study. Int J Tuberc Lung Dis. 2018;22:328–335. [PubMed: 29471912]
20. Ribli D, Horváth A, Unger Z, Pollner P, Csabai I. Detecting and classifying lesions in mammograms with deep learning. Sci Rep. 2018;8:4165. [PubMed: 29545529]
21. Lee H, Grosse R, Ranganath R, Ng AY. Unsupervised learning of hierarchical representations with convolutional deep belief networks. Commun ACM. 2011;54:95.
22. Kallenberg M, Petersen K, Nielsen M, et al. Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. IEEE Trans Med Imaging. 2016;35:1322–1331. [PubMed: 26915120]

23. Leung MKK, Delong A, Alipanahi B, Frey BJ. Machine learning in genomic medicine: A review of computational problems and data sets. Proc IEEE. 2016;104:176–197.
24. Chaudhari AS, Fang Z, Kogan F, et al. Super-resolution musculoskeletal MRI using deep learning. Magn Reson Med. 2018;80:2139–2154. [PubMed: 29582464]
25. 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;79:2379–2391. [PubMed: 28733975]
26. 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;288:177–185. [PubMed: 29584598]
27. Kohn MD, Sassoon AA, Fernando ND. Classifications in brief: Kellgren-Lawrence classification of osteoarthritis. Clin Orthop. 2016;474:1886–1893. [PubMed: 26872913]
28. Peterfy CG, Schneider E, Nevitt M. The osteoarthritis initiative: Report on the design rationale for the magnetic resonance imaging protocol for the knee. Osteoarthr Cartil OARS Osteoarthr Res Soc. 2008;16:1433–1441.
29. 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), Stanford, CA, 6 2016 pp. 565–571. https://arxiv.org/abs/1606.04797. Accessed October 22, 2018.
30. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database In 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, 2009 pp. 248–255.
31. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–2830.
32. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. ArXiv151203385 Cs. 12 2015 http://arxiv.org/abs/1512.03385. Accessed November 6, 2018.
33. Paszke A, Gross S, Chintala S, et al. Automatic differentiation in PyTorch. In 31st Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, 2017.
34. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. 8 2016 https://arxiv.org/abs/1608.06993. Accessed November 6, 2018.
35. 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. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf. Accessed September 26, 2019.
36. 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 2 2016 https://arxiv.org/abs/1602.07360v4. Accessed September 26, 2019.
37. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. ArXiv150201852 Cs. 2 2015 http://arxiv.org/abs/1502.01852. Accessed September 26, 2019.
38. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–444. [PubMed: 26017442]
39. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. ArXiv161002391 Cs. 10 2016 https://arxiv.org/abs/1610.02391. Accessed May 13, 2019.

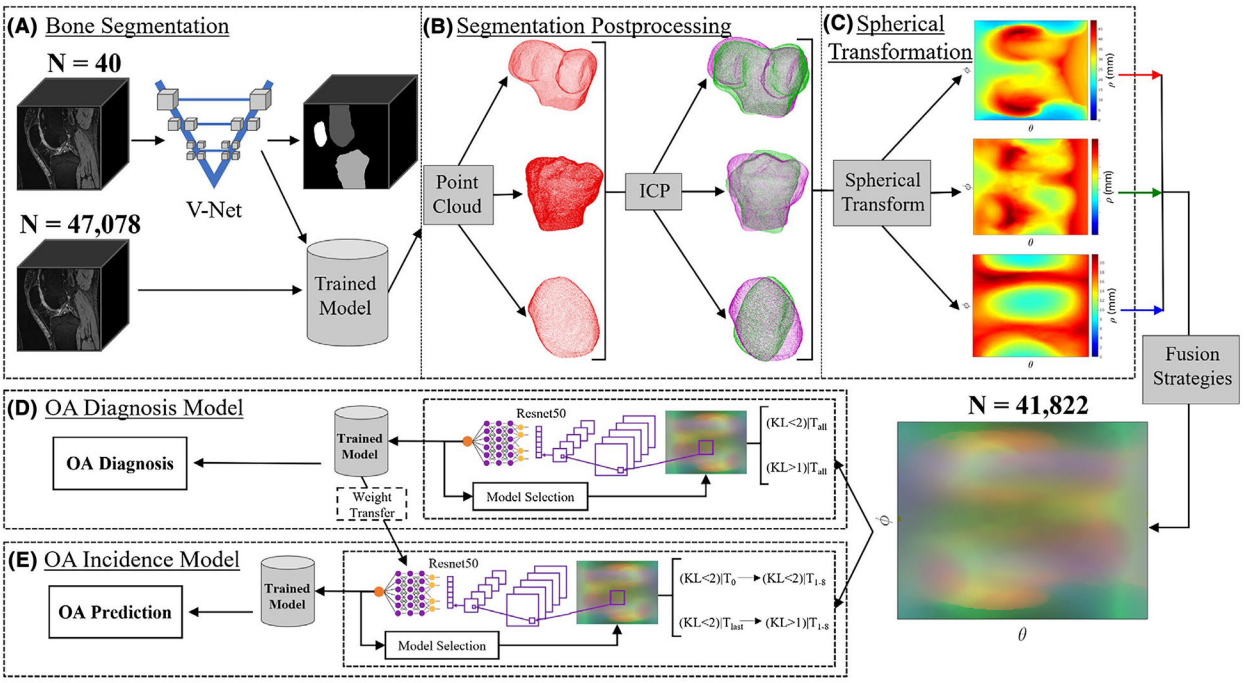

FIGURE 1.

Overview of the study. (A), A V-Net segmentation model was trained and validated with a dataset of 40 3D double echo steady-state (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 Osteoarthritis Initiative (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 osteoarthritis (OA) based on bone shape across all time points. Each of the two inputs represents a class in the binary classifier (healthy Kellgren-Lawrence [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 y up to 8 y, 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 y ($T_{1–8}$). 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 y ($T_{1–8}$). 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 y

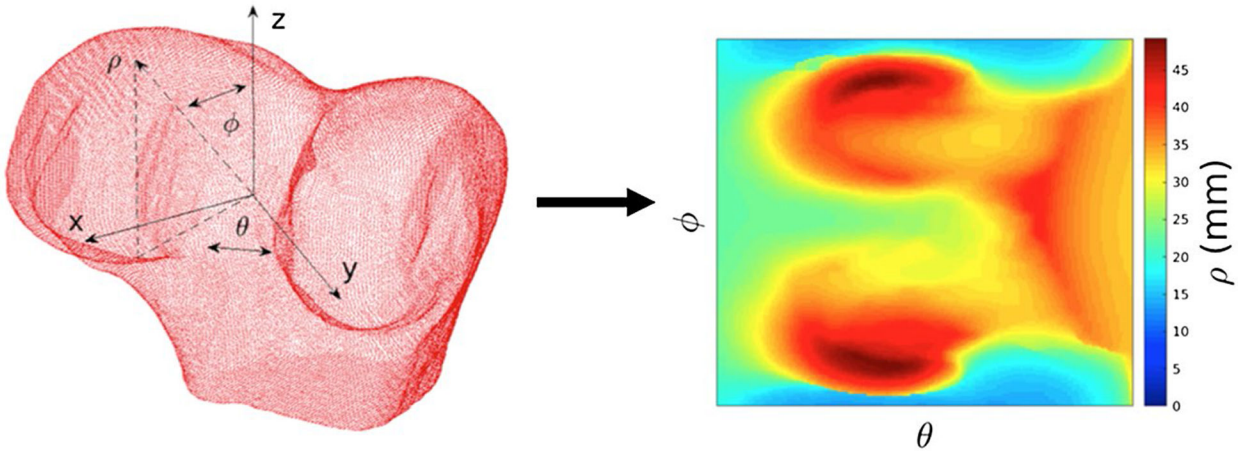

FIGURE 2.

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

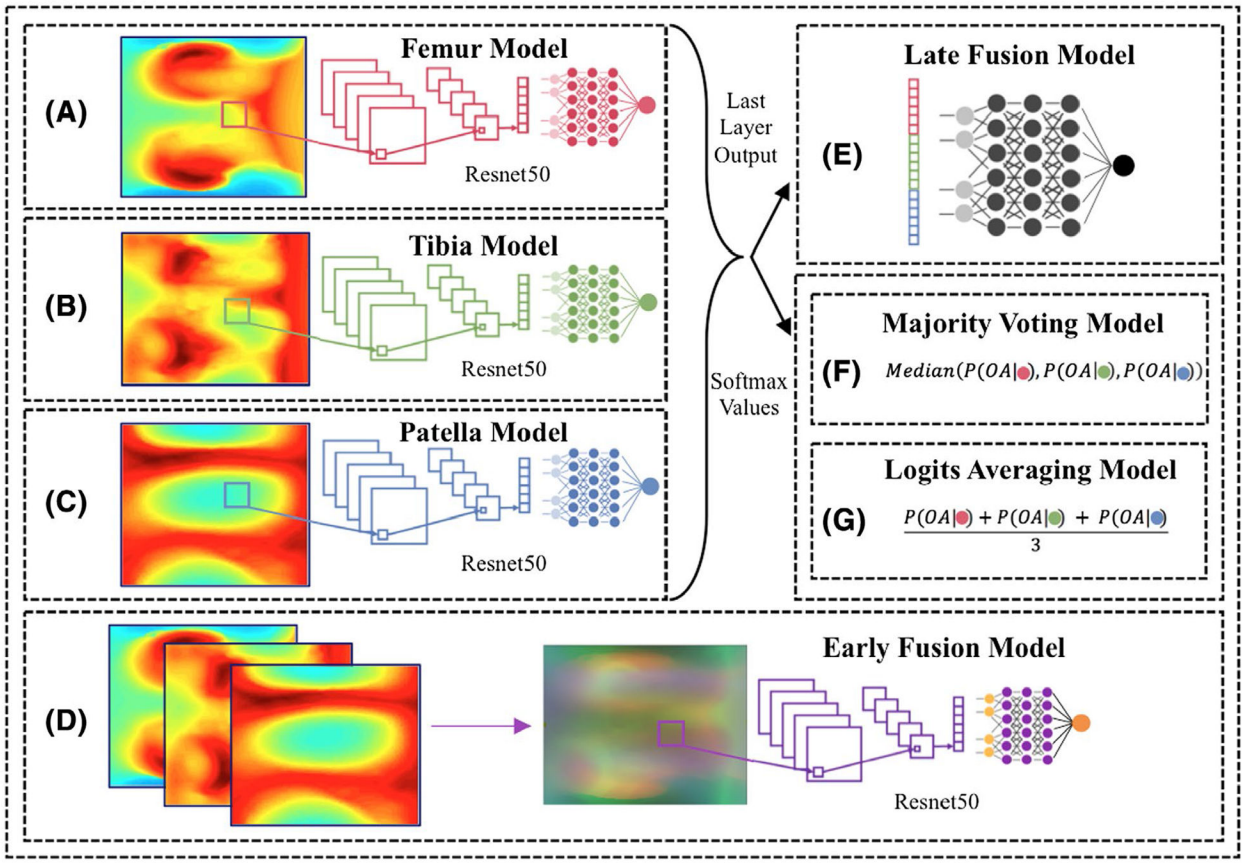

FIGURE 3.

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 three-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 three-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 osteoarthritis (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

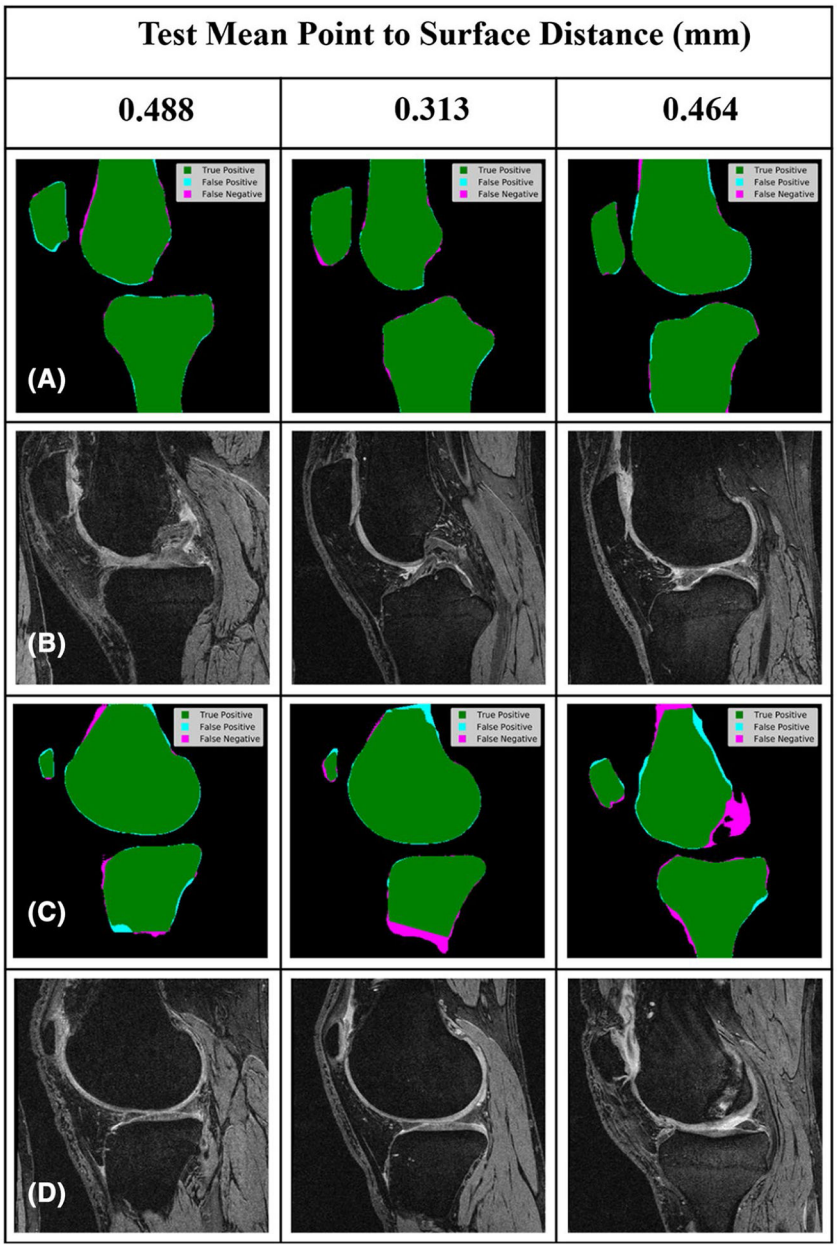

FIGURE 4.

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 double echo steady-state (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

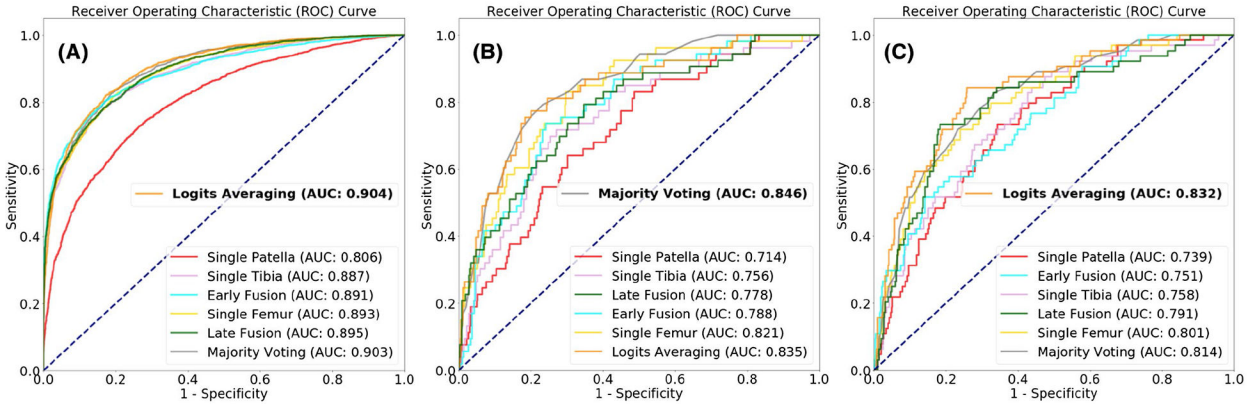

FIGURE 5.

Overview of the validation receiver operating characteristics (ROC) curve comparisons for the different model fusion strategies. The osteoarthritis (OA) Diagnosis model and the first two OA Incidence models are shown, with the remaining OA Incidence models are shown in the Supporting Information Figure S1. (A), OA Diagnosis model. (B), 1-y OA Incidence model. (C), 2-y OA Incidence model

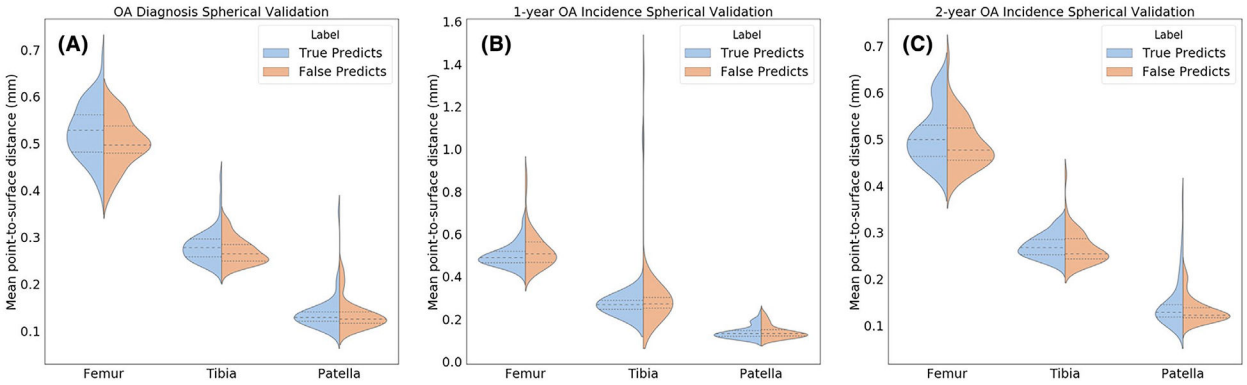

FIGURE 6.

Robustness of the osteoarthritis (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-y OA Incidence model. (C), 2-y OA Incidence model

TABLE 1

Training splits information for the bone segmentation, OA Diagnosis, and OA Incidence models

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-y	2444 (246)	537 (53)	524 (50)	0.101	0.159	0.298
2-y	2495 (297)	548 (64)	537 (63)	0.119	0.206	0.814
3-y	2389 (191)	527 (43)	517 (43)	0.0799	0.516	0.560
4-y	2397 (199)	527 (43)	517 (43)	0.0830	0.220	0.852
5-y	2356 (156)	514 (32)	506 (32)	0.0662	0.860	0.290
6-y	2373 (175)	519 (35)	510(36)	0.0737	0.591	0.472
7-y	2269 (71)	500 (16)	489 (15)	0.0313	0.559	0.435
8-y	2275 (77)	502 (18)	492 (18)	0.0338	0.998	0.592

Note: 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 $P < .05$.

TABLE 2

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

	AUC (Validation Set)								
Model	**Diagnosis**	**1-y**	**2-y**	**3-y**	**4-y**	**5-y**	**6-y**	**7-y**	**8-y**
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

Note: The bold values represent the best performing fusion strategy for each model.

TABLE 3

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

	Logits Averaging (Test Set)								
Metric	**Diagnosis**	**1-y**	**2-y**	**3-y**	**4-y**	**5-y**	**6-y**	**7-y**	**8-y**
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