Cartilage-X Automated Anomaly Detection in Knee MRIs, Final Report

C Iriondo, D Jain, R Muhamedrahimov, V Papanikolaou, K Trotskovsky, L Sun

Introduction

Over 25% of US adults suffer from knee pain, creating an urgent need for an efficient way to assess cartilage and bone lesions to match patients to effective treatments. Patients undergo Magnetic Resonance Imaging (MRI) scans across the country, but the radiology expertise and time required to interpret these images is scarce. Cartilage-X uses AI-guided lesion detection to predict the location and severity of lesions, providing a clinically feasible method to quantitatively assess knee lesions and decrease the time and expertise needed.

Background

When assessing a knee MRI, physicians look for two key indicators of knee health: cartilage lesions and bone marrow edema. Cartilage lesions are thickness defects on the cartilage surface, where defect span and depth correspond to lesion severity. Radiologists assign healthy cartilage a 0 or 1, and lesions a 2, 2.5, 3, 4, 5, or 6 from least severe to most severe. Bone marrow edema (BME) is the presence of hyper-intense zones in the subchondral bone caused by the accumulation of fluid. Radiologists assign healthy bone a 0 while edema is scored 1, 2, or 3 according to extent of hyper-intensity. These two signals, while present in the same MRI image, required different preprocessing strategies and were ultimately classified by different algorithms.

Data Preprocessing

The data consisted of 1,800 MR images and segmentations (binary masks containing ones at the locations of the cartilage and zeros otherwise). The dataset of 3D MR images was cropped to the boundaries of existing cartilage determined by the 3D segmentations. This resulted in a ~50-fold reduction in image size (512x512x200 to 200x150x30), reducing the computational load. After cropping, image intensities were normalized. While helpful, cropping was not sufficient. First, the images were still large, leading to memory issues when defining tensors. Second, the C-shape cartilage occupied only a small fraction of the cropped volume, increasing the number of irrelevant features provided to the network. Further preprocessing was performed for each classification problem.

Cartilage Lesions: Each 2D slice of cartilage was flattened from C-shape to a rectangular shape by: (1) identifying cartilage edges using segmentation boundaries (2) creating a Laplacian potential map between the edges (3) finding streamlines connecting the top edge and the bottom edge (4) performing a Forneffet transform to straighten all curves in the top edge while preserving length of streamlines. The cartilage thickness (which is important for lesion classification) was preserved. Before flattening, the 3D segmentations were cleaned (combining pieces, dilating, morphological closing of holes etc.) since segmentations were sourced from a previous machine learning algorithm (UNet). The flattening pipeline was implemented in Matlab using built-in functions and previous UCSF work. Flattening resulted in a ~6-fold decrease in image size (200x150x30 to 250x20x20) reducing the required memory and removing redundant features.

Bone Marrow Edema (BME): Cropped 3D volumes were projected into 2D images by taking the mean of the 3D volume along the z-axis. Projection images were resized to a fixed 120x120. Eliminating z-axis slices drastically reduced the size of each image, allowing the whole dataset to be loaded into memory for rapid algorithm testing.

Goal 1: Classifying Cartilage Lesions

The smallest detectable cartilage lesions appear as a few mm change in cartilage intensity and can be present within one or multiple slices of the MRI from multiple views. As such, features of interest are the relative (not

absolute) pixel intensity values. A convolutional neural network with three-dimensional filters was used to take into account correlations between adjacent pixels in all dimensions for cartilage lesion classification. A binary classifier assigned a 0–healthy or 1–presence of a cartilage lesion to each image.

Classification

The model was built using TensorFlow. First, to accommodate the large size of the input data, an input pipeline was created to load training batches into memory on each training iteration, while allowing for data of various input shapes and file formats. Additionally, to account for class imbalance, the option to sample positive and negative training examples in different proportions was implemented. For this case, the positive class accounted for approximately 14% of the full dataset. Given the relatively small size of the dataset of ~3000 samples, oversampling was tested as a method of improving on performance with respect to class imbalance.

To build a model architecture, the model was initially trained on a small subset of the training data to ensure it could be trained and overfit. The architecture was then scaled up, using the training set performance to determine the approximate limits of sample and model complexity, from which point the optimal model was determined, under constraints of time and computational memory. Ultimately, a model architecture consisting of 4 convolutions with 2 pooling layers resulted in the best model performance, with a validation accuracy of 0.75, precision of 0.28, and recall of 0.61, including regularization and oversampling of positive samples to approximate balanced classes.

Goal 2: Classifying Bone Marrow Edema

Bone marrow edema appears as bright, fuzzy spots of at least half a cm in diameter in the subchondral bone, and span across multiple slices along the z-axis. The mean z-axis projection images captured variations in signal intensity across the subchondral bone. A binary classifier assigned a 0–healthy or 1–presence of bone marrow edema to each image.

Classification

As a first attempt at BME classification, the 120x120 images were unraveled into a single vector and tested on simple classification models. These inputs were fed into a Logistic Regression model where performance was poor (57% accuracy). Support Vector Machine performed the best (59% accuracy) while Random Forest models performed the worst (52% accuracy). Hyperparameter tuning and efforts to reduce class imbalance did not improve results. Although BME features were visible to a non-radiologist, the 2D spatial information lost during the unraveling of the image provided to be valuable in the classification task.

A 2D convolutional neural net a 3x3 filter size and 2 convolutional layers was built using Keras, a high-level TensorFlow wrapper. The network was trained for 15 epochs on the original dataset, achieving 94% and 96% test and validation accuracy, respectively. High accuracy proved deceiving as precision and recall values showed the network had learned to classify all cases as negative (no lesion) due to class imbalance. Performing data augmentation on the training set's minority class greatly improved the neural net's performance, described in detail below.

Data Augmentation

There was significant class imbalance in the dataset (majority class ~85% of dataset), which arises from the multi-class pathology, rarity of positive cases in the dataset, and the highly correlated images (MRIs from the same patient under different loading conditions, at different time points, etc.). Accuracy was high, but precision and recall values were low. To counter the class imbalance, minority classes were combined into a single positive class for

binary classification (lesion/no lesion). In addition, the correlated data was kept in the same test/train/validate group.

Data augmentation by oversampling the minority was not generalizable for detecting minority classes, whereas undersampling the majority resulted in a loss of data and loss in accuracy of prediction of majority class, which was undesirable. Within the training set, the images in the minority class were cropped, shifted, sheared, and brightness/contrast filtered. Including the augmented training data in the 2D CNN significantly improved performance. The recall value reached 0.69 (for affine transformation) and 0.61 for intensity variations. After training the net for 30 epochs with the augmented dataset (affine), test accuracy reached 85%, but recall remained 0.45. The validation results were remarkably improved, with an accuracy of 100% and a recall of 1.

Future Improvements

Data Preprocessing	• Establish quality guidelines for images/segmentations to run flattening and projection algorithms • Extract cartilage features directly from the flattened images and try a simpler and faster classifier • Create multi-plane projection images (sagittal, coronal, axial)
Cartilage Lesions	• Optimize 3D-CNN, iterate through more model architectures • Online batch sampling to tackle class imbalance and improve consistency of training • Transfer learning: pre-train model on balanced dataset, initialize new model using trained weights • Alternative models, including recurrent convolutional neural networks • Introduce demographic features (age, BMI, gender) in last layer of network
Bone Marrow Edema	• Optimize 2D-CNN architecture (number of layers, filter size, weighted sampling) • Train the net with combined data augmentations (affine + filter, affine + intensity, etc) • Implement multi-class classification with BME score • Implement LIME to visualize activation map of CNN and lesion location
Data Augmentation	• Conduct principal component analysis (PCA) on the images to identify the components that are indicative of cartilage or bone lesions, alter non-significant components to generate synthetic instances • Apply data augmentation on volumes for 3D CNN
Overall	• Test generalizability with a fresh dataset • Define product strategy (deployment, regulatory, etc)

Team Effort

All members contributed significantly towards the project throughout the semester with particular expertise in: CI: study design, data sourcing/cleaning, preprocessing, DJ: data augmentation, RM: data cleaning, cartilage lesion detection, VP: bone marrow edema detection, KT: preprocessing, LS: bone marrow edema detection

Collaboration with Mentors

We were lucky to collaborate with multiple mentors. UCSF mentor Valentina Pedoia and the MQIR group helped troubleshoot data access rights, server issues, as well as specific strategies regarding preprocessing and network architecture. We met with Kevin Li on a few occasions, and he brought in a great industry perspective as well some clever strategies for improving our model performance (many of his suggestions are included in our future steps).