This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2018.2840082, IEEE Transactions on NanoBioscience

A Novel Method to Predict Knee Osteoarthritis Progression on MRI Using Machine Learning Methods

Yaodong Du, Rania Almajalid, Juan Shan
Department of Computer Science
Seidenberg School of CSIS, Pace University
New York City, NY, USA
yd67578n@pace.edu, ra56319p@pace.edu, jshan@pace.edu

Ming Zhang
Division of Rheumatology
Tufts Medical Center
Boston, MA, USA
MZhang@tuftsmedicalcenter.org

***Abstract*—This study explored the hidden biomedical information from knee MR images for osteoarthritis (OA) prediction. We have computed the Cartilage Damage Index (CDI) information from 36 informative locations on tibiofemoral cartilage compartment from 3D MR imaging and used PCA analysis to process the feature set. Four machine learning methods (artificial neural network (ANN), support vector machine (SVM), random forest and naïve Bayes) were employed to predict the progression of OA, which was measured by change of Kellgren and Lawrence (KL) grade, Joint Space Narrowing on Medial compartment (JSM) grade and Joint Space Narrowing on Lateral compartment (JSL) grade. To examine the different effect of medial and lateral informative locations, we have divided the 36-dimensional feature set into 18-dimensional medial feature set and 18-dimensional lateral feature set and run the experiment on four classifiers separately. Experiment results showed that the medial feature set generated better prediction performance than the lateral feature set, while using the total 36-dimensional feature set generated the best. PCA analysis is helpful in feature space reduction and performance improvement. For KL grade prediction, the best performance was achieved by ANN with AUC = 0.761 and F-measure = 0.714. For JSM grade prediction, the best performance was achieved by random forest with AUC = 0.785 and F-measure = 0.743, while for JSL grade prediction, the best performance was achieved by the ANN with AUC = 0.695 and F-measure = 0.796. As experiment results showing that the informative locations on medial compartment provide more distinguishing features than informative locations on lateral compartment, it could be considered to select more points from the medial compartment while reduce the number of points from the lateral compartment to improve clinical CDI design.**

Keywords—knee osteoarthritis; cartilage damage index; informative locations; feature representation; machine learning

I. Introduction

Knee osteoarthritis (OA) is a disease that increases in incidence and prevalence with advancing age, such that in those over the age of 60, about 10% of men and 13% of women have symptomatic knee OA [1]. With the aging of the population, the number of people in the age range with the greatest severity of OA continues to increase [2]. Furthermore, OA is a leading cause of morbidity and disability, and thus carries high socioeconomic costs. In 2004, arthritis was estimated to cost the United States $336 billion, or 3% of the gross domestic product, with OA as the most common form of arthritis [3, 4].

The pathology of OA disease is still unclear and there are no interventions that effectively modify the OA disease process [5]. In clinical studies, OA is mainly diagnosed through medical images. Measurement of cartilage change is a primary assessment of structural progression of OA and is used to evaluate the effectiveness of new treatments. Magnetic resonance (MR) imaging is a noninvasive technology that can generate 3-dimensional images of intra-articular soft-tissue structures, including cartilage. However, obtaining accurate and reproducible quantitative measurements from MRI scans is burdensome due to the structure and morphology of the knee as well as the nature of MR imaging [6]. It may take up to six hours for a reader to manually segment each series of 3-dimensional (3D) knee MRI. Furthermore, operators who use cartilage segmentation software often need extensive training [7] which further contributes to the time and cost.

Over the past decade, researchers have developed different approaches to reduce the burden of measuring knee cartilage on MR images. These includes segmenting alternate MR slices or confining measurements to partial regions of cartilage [8-10]. Computer-aided algorithms (e.g., active contours, B-splines) have also been developed to assist with cartilage segmentation for MR images [26-28]. Unfortunately, these methods lack sufficient accuracy and reliability to detect small cartilage changes [10, 25]. Thus, there remains a need among researchers for a quantification method that can be rapidly computed and has good reproducibility, validity, and sensitivity to change.

In recent studies, a novel and sensitive cartilage biomarker, called cartilage damage index (CDI), was developed and validated by Zhang et al. [11, 12]. The CDI demonstrated increased measurement efficiency and scale responsiveness when measurements of cartilage thickness were confined to points on the cartilage surface. Using CDI was able to detect up to a 14.3% annual cartilage change compared to 2~3% annual cartilage change detected by traditional methods [29].

This work is inspired by the observation that in the process of computing CDI score for a knee joint, each of the 60 index locations was measured separately, but only the summation of all these locations or a subgroup of these locations were used to compute CDI. In this work, we focus on using data mining and machine learning methods to fully explore the information contained in each index location. We treat CDI information from each location as an individual feature dimension and use principle component analysis to find the optimum feature representation. The optimum feature set serves as the input for machine learning methods to learn the mapping function

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2018.2840082, IEEE Transactions on NanoBioscience

between cartilage change at CDI locations and OA severity grade change. We used Kellgren and Lawrence (KL) grade, Joint Space Narrowing on Medial compartment (JSM) grade and Joint Space Narrowing on Lateral compartment (JSL) grade as OA severity grade in this study.

The rest of the paper is organized as follows. In Section II, we described the materials and methods used in this research, including OAI database, definition and measurement of CDI locations, feature analysis, machine learning methods and evaluation metrics. In Section III, we presented and analyzed the results from the experiments described in Section II. Finally, in Section IV, we drew conclusion and discussed future work.

II. Materials and Methods

A. Data

All the data and MR images used in this study were selected from the Osteoarthritis Initiative (OAI). OAI was initiated to promote the evaluation of OA biomarkers as potential surrogate endpoints [14]. It includes four clinical centers that recruited approximately 4800 men and women (ages 45-79 years) with or at risk for knee OA. The participants underwent annual knee radiography and MR scans during the first four years and then biannually for the subsequent 4 years. The radiographic assessments are available from the OAI. The OAI cohort has excellent follow-up retention rates at 4 years (90%) and 6 years (~88%). The full set of publicly-available OAI data can be viewed on the OAI website. In this study, we selected a convenience sample of 100 pairs of knee (both baseline and 24-month MR scans) that had complete data (i.e., clinical, static knee alignment, semi-quantitative radiographic grading, and joint space width). We selected our samples to represent the range of radiographic OA severity (KL scores 0 to 4) enriched with knees that showed radiographic worsening over time (KL scores changes between baseline and 24-month follow-up).

B. Cartilage Damage Index

Cartilage Damage Index (CDI) is a novel osteoarthritis cartilage damage quantification method that utilizes informative locations on knee MR images [11-13]. The CDI quantifies cartilage thickness by measuring certain informative locations on the reconstructed cartilage layer instead of measuring cartilage on all MR slides. The informative locations are selected based on the statistical analysis that certain articular cartilage locations are more susceptible to occurrence of OA damage and thus may be more informative in the measurement of OA progression. To measure CDI, totally 60 locations on the cartilage layer are selected, including 18 locations from medial tibiofemoral compartment, 18 locations from lateral tibiofemoral, and 24 locations from patella compartment (Fig. 1). CDI has been validated using images from Osteoarthritis Initiative (OAI) database and successfully applied to clinical trials [23]. Statistical studies show that CDI is associated with commonly used OA severity measures including Joint Space Narrowing (JSN) grade, Kellgren and Lawrence (KL) score, Joint Space Width (JSW), and knee alignment with p-values <0.05 [11, 12].

These informative locations are selected from regions on articular surface where cartilage denudation frequently happen. In the study to find the most informative locations, a 3D articular surface of the distal femur and proximal tibia is constructed using sequence of 2D MR slides, as shown in Fig. 1. Then the 3D surface is projected to a 2D rectangular coordinate systems to represent the articular surface of the distal femur and proximal tibia.

18 informative locations are selected within medial and lateral femur compartments (yellow dots in Fig. 1), 18 informative locations are selected within medial and lateral tibia, and 24 informative locations are selected within patella. In specific, 9 locations are selected within the region of the most commonly denuded areas on the medial femur, medial tibia, lateral femur and lateral tibia, and 12 locations within medial patellofemoral and lateral patellofemoral respectively. In this paper, we used 36 informative locations from medial and lateral tibiofemoral compartments to do the analysis because they are more related with OA progression.

To measure the CDI information for a new set of MR images (one knee), first step is to indicate the most medial and lateral MR image slices within the knee. These images designate the minimum and maximum values of the medial-to-lateral axis on the 2D coordinate system. Next, the software automatically determines the MR image slices that contain the informative locations. On each of these slices the bone-cartilage boundary need to be manually traced by an experienced expert using predefined segmentation rules. The software then translates the length of the bone-cartilage boundary to a standardized anterior-to-posterior axis and indicates the predefined informative locations so that the expert could measure the cartilage thickness at those points (Fig. 2). The software then computes the CDI score by summing the products of cartilage thickness, cartilage length (anterior-posterior), and voxel size from each informative location.

C. Feature Analysis

Unlike the definition of CDI score which computes the summation of thickness information from all CDI locations, we used the thickness information from each informative location as an individual feature. For each informative location, the thickness change over two years (subtracting baseline data from 24-month data) is computed as one input feature. Therefore, 36 informative locations generate a 36-dimensional feature set. The corresponding class label is the change of KL grades (change or no-change). These 36 features are further divided into two groups, 18 features from medial tibiofemoral compartment (including medial tibia and medial femur) and 18 features from lateral tibiofemoral compartment (including lateral tibia and lateral femur). We plan to analyze the 36 features as well as the two subgroups (medial and lateral) separately as research showed that medial OA is more common than lateral OA [15, 16].

We analyzed the feature space by principal component analysis (PCA) [17], with the purpose to find the most representative optimum feature set. PCA projects data onto a new space in which consecutive dimensions contain less and less of the variance of the original dataspace and compresses the most important information onto a subspace with lower dimensionality than the original space. Before running PCA, as a preprocessing step, we normalized data into range [0, 1] for each dimension. We tested the feature space with 5-100% of the projected subspace using 10-fold cross-validation to establish how many principal components needed to be included to reach full performance.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2018.2840082, IEEE Transactions on NanoBioscience

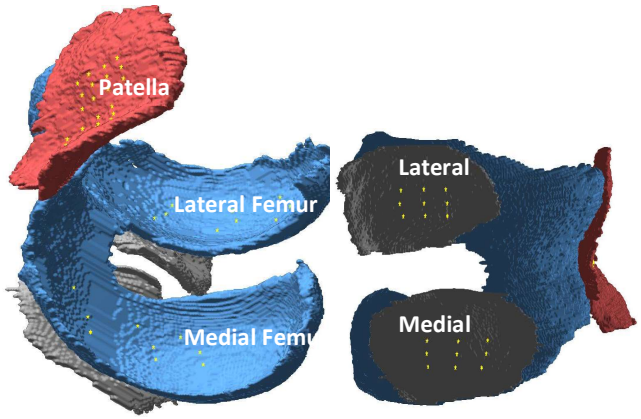

Fig. 1. Informative locations (yellow points) on 3D cartilage layer [13].

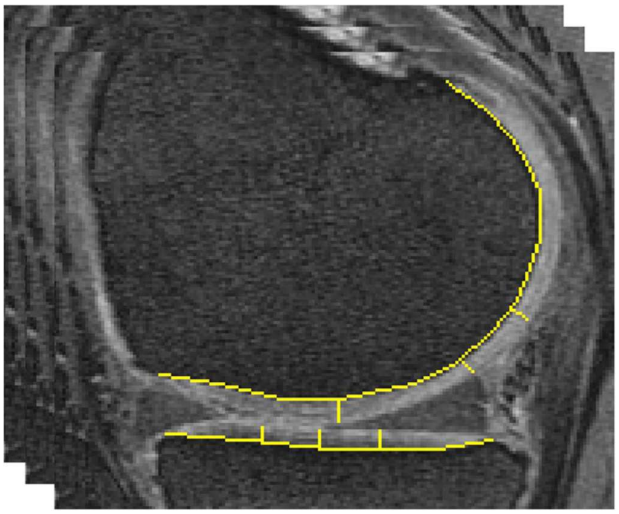

Fig. 2. The thickness measurement of six CDI locations on one MR slide of the medial tibiofemoral compartment [12].

D. Machine Learning Methods

We explored the use of four machine learning methods to learn the mapping function between the CDI feature space and OA severity denoted by KL grades. The four machine learning methods are artificial neural network (ANN), support vector machine (SVM), random forest and naive Bayes.

ANNs are powerful classifiers that are based on the structure and functions of biological neural networks [18]. An ANN is composed of an input layer, an output layer and one or more hidden layers. In this work, a single hidden layer with n neurons was employed as the network structure, where n is computed as (# of attributes + # of classes)/2. The backpropagation algorithm is used to update the weights of neurons.

A SVM constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space to separate data [19]. It uses a kernel function to map data into the higher dimensionality to obtain a better distribution and therefore a better classification result. SVMs are have been reported to be a superior method in many classification problems. In this work, the radial basis function (RBF) was adopted as the kernel function.

A random forest is an ensemble learning method that constructs a multitude of decision trees at training time and outputs the class that is the overall prediction of the individual trees [20]. Random forests correct the overfitting problem of decision trees and are commonly used for classification, regression and other tasks with an efficient performance on large scale data bases.

Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features [21]. It is a popular method for text categorization problem and finds application on automatic medical diagnosis. In other classification fields, naïve Bayes is usually not as competitive as other more advanced machine learning methods such as SVM and ANN, but in this work, we found that it often achieved good performance in the experiments of OA severity prediction.

All the classifiers were implemented in Weka software package [22], which was used to run experiments in this work.

E. Evaluation

10-fold cross-validation was used for training and testing procedure for all four classifiers, in which data were divided into 10 equal groups, and for each iteration, one was held out for testing while the remaining nine groups were used for training, until all the data had been used as testing data once.

We used several metrics to evaluate performance of our classifiers: precision (also called positive predictive value (PPV)), recall (also called sensitivity), F-measure, Matthew's correlation coefficient (MCC), and the area under the receiver operating characteristic (ROC) curve (AUC). ROC curves provide an indication of the tradeoff between classification sensitivity and specificity as the classifier confidence threshold increases or decreases. The F-measure, provides an indication of overall classification accuracy as a weighted average of precision and recall for a specified confidence threshold. MCC is a powerful accuracy evaluation criterion of machine learning methods. Especially, when the number of negative samples and positive samples are obviously unbalanced, MCC gives a better evaluation than overall accuracy. The formulas of the evaluation metrics are provided below.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{2}$$

$$\text{F} - \text{Measure} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision+Re}} \tag{3}$$

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP+FP})(\text{TP+FN})(\text{TN+FP})(\text{TN+FN})}} \tag{4}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives. In this work, positive class is defined as KL grade is changed after 24-month follow-up and negative class is defined as KL grade has no-change after 24-month follow-up. Same strategy was used for JSM and JSL grades.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2018.2840082, IEEE Transactions on NanoBioscience

III. Experiment and Results

A. Experiments for KL grade prediction

1) Experiment 1: Predict KL grade change using 18 medial tibiofemoral informative locations

Experiment 1 used 18 information locations on medial tibiofemoral compartment to predict the change of KL grade. For each informative location, the product of cartilage thickness, cartilage length (anterior-posterior), and voxel size was computed from both baseline and 24-month data, the change of the two products was used to represent CDI information of each informative location. The 18-dimensional feature set was normalized first and then processed by PCA. We tested the performance of all four machine learning methods with different PCA component percentages using 10-fold cross-validation. The ROC performance of each of the four machine learning methods was plotted in Figs. 3-6.

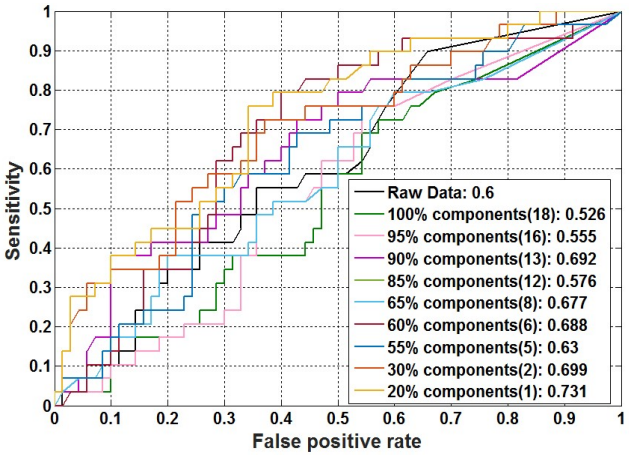

Fig. 3. ROC curves of ANN classifier with different percentages of PCA components obtained from 18 medial features.

For ANN, the best performance was achieved with AUC 0.731 and F-measure 0.708 when using the top 1 PCA component which covered 20% of PCA variance. For SVM, the best performance was achieved with AUC 0.691 and F-measure 0.586 using the top 10 PCA components which covered 70% PCA variance. It should be noted that the MCC was 0, which indicated that the SVM classified all samples into one class. For random forest, using top 65% PCA achieved its best performance but the performance was weaker than the best performance of ANN. Surprisingly, among all the four classifiers, the best performance was achieved by naïve Bayes with raw data, i.e., AUC 0.742 and F-measure 0.700. The result indicated that PCA analysis did not help improving the performance of naïve Bayes classifier using the 18-dimensional medial feature set, but did help the other three classifiers improve the performance using this feature set. The best performance of each classifier was summarized in Table I with different evaluation metrics.

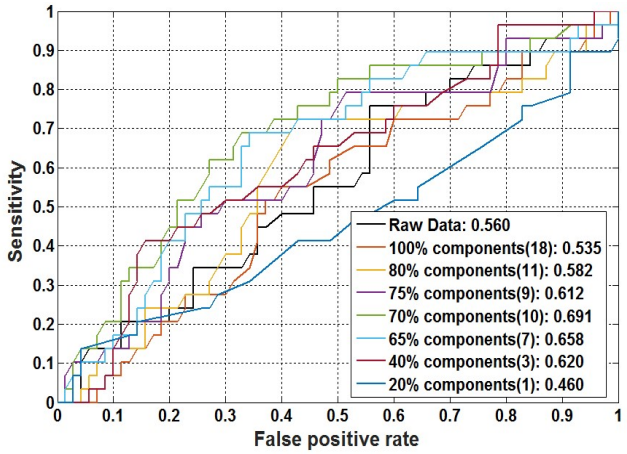

Fig. 4. ROC curves of SVM classifier with different percentages of PCA components obtained from 18 medial features.

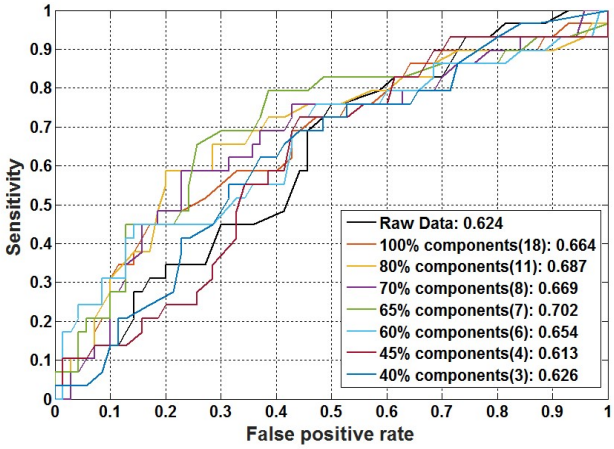

Fig. 5. ROC curves of random forest classifier with different percentages of PCA components obtained from 18 medial features.

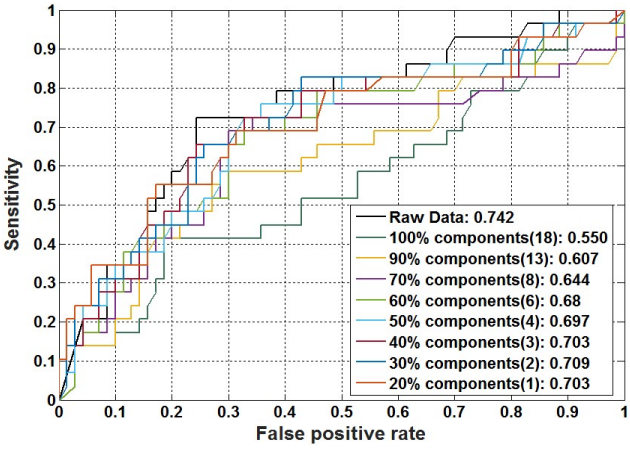

Fig. 6. ROC curves of naïve Bayes classifier with different percentages of PCA components obtained from 18 medial features.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2018.2840082, IEEE Transactions on NanoBioscience

TABLE I. BEST PERFORMANCE OF EACH OF THE FOUR CLASSIFIERS ON 18 MEDIAL FEATURES

Classifier	PCA variance	Precision	Recall	F-Measure	MCC	ROC area
ANN	Top 20%	0.714	0.737	0.708	0.285	0.731
SVM	Top 70%	0.5	0.707	0.586	0	0.691
Random Forest	Top 65%	0.653	0.697	0.655	0.144	0.702
Naive Bayes	Raw data	0.744	0.687	0.700	0.362	0.742

2) Experiment 2: Predict KL grade change using 18 lateral tibiofemoral informative locations

As research showed that cartilage damage is more likely to happen on medial tibiofemoral compartment than lateral tibiofemoral compartment [15, 16], we decided to analyze the informative locations from the two compartments separately. The similar experiments were conducted as described in Experiment 1, by replacing the 18 medial informative locations with 18 lateral informative locations. Figs. 7-10 plotted the performance of the four machine learning methods with different PCA component percentages using 10-fold cross-validation.

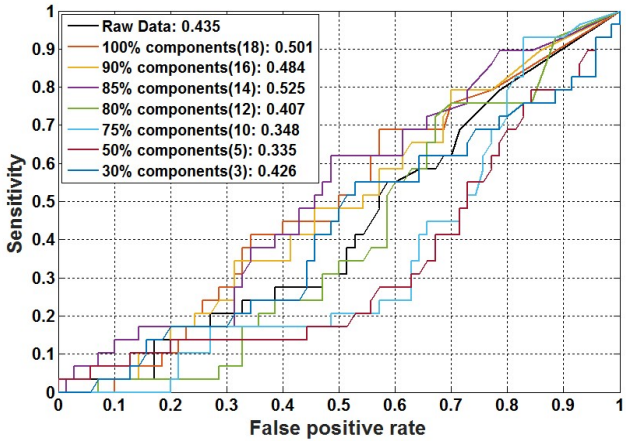

Fig. 7. ROC curves of ANN classifier with different percentages of PCA components obtained from 18 lateral features.

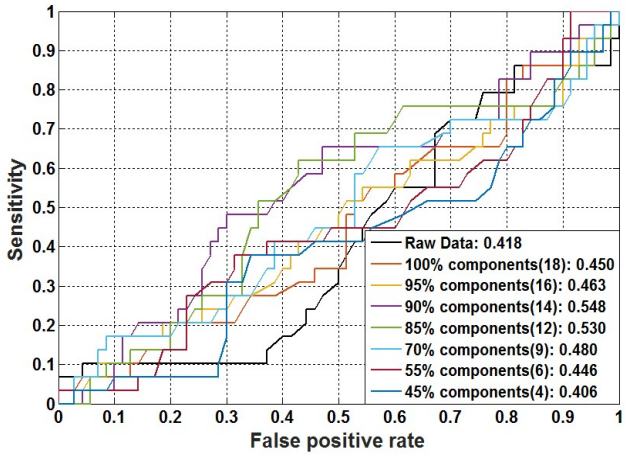

Fig. 8. ROC curves of SVM classifier with different percentages of PCA components obtained from 18 lateral features.

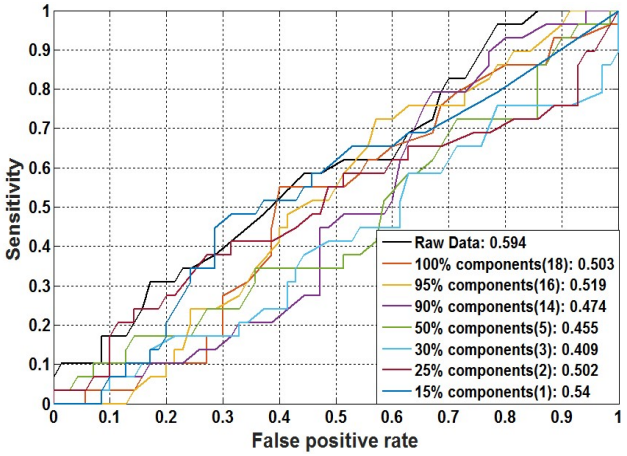

Fig. 9. ROC curves of random forest classifier with different percentages of PCA components obtained from 18 lateral features.

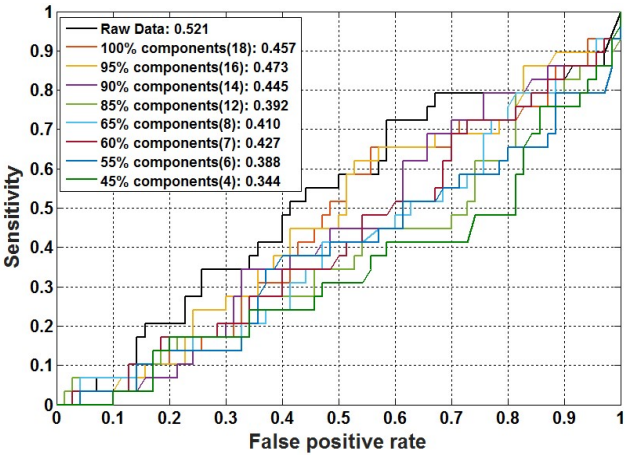

Fig. 10. ROC curves of naïve Bayes classifier with different percentages of PCA components obtained from 18 lateral features.

Using lateral feature set, we can see that the performance of all four classifiers dropped compared with using medial feature set (see Figs. 7-10). The best performance was achieved by random forest with AUC 0.594 and F-measure 0.612. The experiment results indicated that medial informative locations contain more important and distinguishing information than lateral informative locations, for KL grade change prediction. Table II summarized the best performance of each method using the 18-dimensional lateral feature set.

TABLE II. BEST PERFORMANCE OF EACH OF THE FOUR CLASSIFIERS USING PCA ANALYSIS ON 18 LATERAL FEATURES

Classifier	PCA variance	Precision (PPV)	Recall (Sensitivity)	F-Measure	MCC	ROC Area
ANN	85%	0.556	0.556	0.556	-0.073	0.525
SVM	90%	0.5	0.707	0.586	0	0.548
Random forest	Raw data	0.6	0.677	0.612	0.028	0.594
Naive Bayes	Raw data	0.612	0.657	0.625	0.06	0.521

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2018.2840082, IEEE Transactions on NanoBioscience

3) Experiment 3: Predict KL grade change using 36 medial and lateral tibiofemoral informative locations

In this experiment, we combined both medial and lateral features to form the 36-dimensional feature set. We ran PCA analysis and machine learning methods on this feature set similar as described in Experiment 1 and Experiment 2. Figs. 11-14 plotted the performance of the four machine learning methods with different PCA component percentages using 10-fold cross-validation.

When using the features from all 36 informative locations, the performance of ANN and SVM improved as compared with using medial or lateral features separately, while the performance of the random forest and naïve Bayes was about the same. The best performance of the four classifiers was achieved by ANN using top 55% PCA, with AUC 0.761 and F-measure 0.714. This is also the best performance among all classifiers using three different feature sets. Table III summarized the best performance of each method using the 36-dimensional feature set.

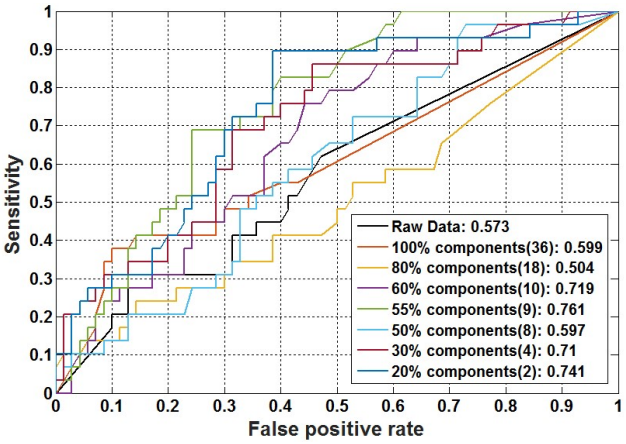

Fig. 11. ROC curves of ANN classifier with different percentages of PCA components obtained from 36 features.

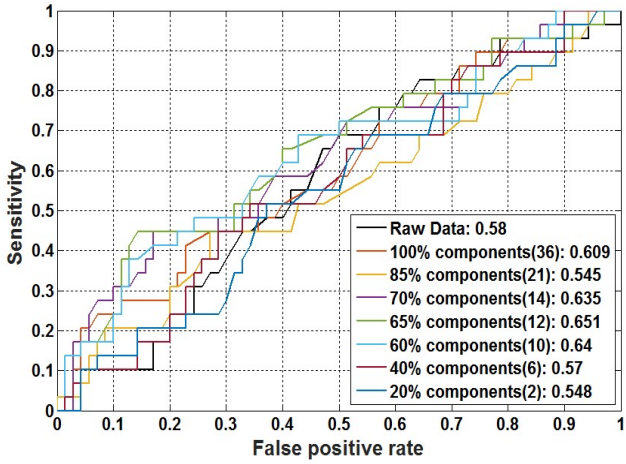

Fig. 12. ROC curves of SVM classifier with different percentages of PCA components obtained from 36 features.

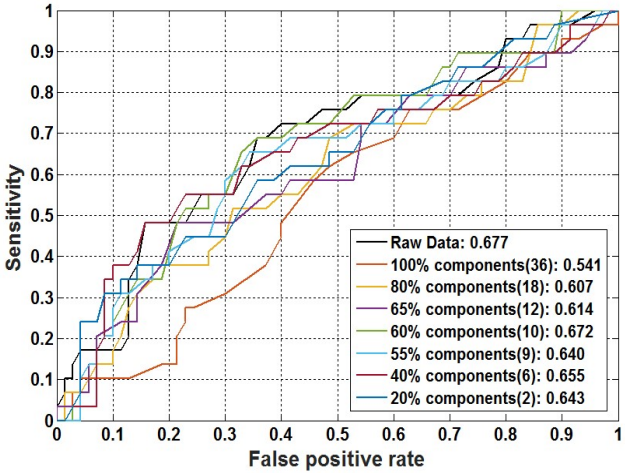

Fig. 13. ROC curves of random forest classifier with different percentages of PCA components obtained from 36 features.

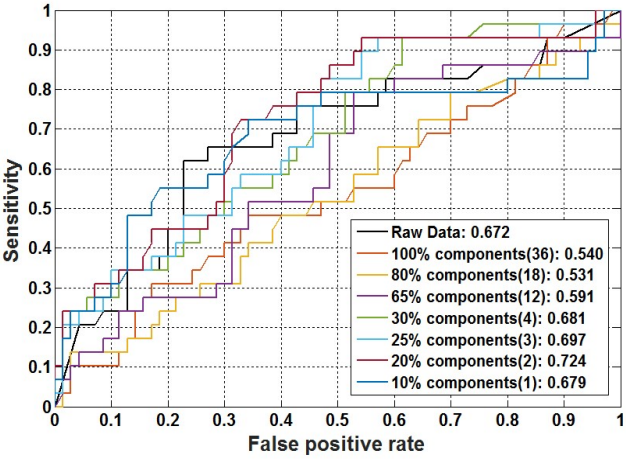

Fig. 14. ROC curves of naïve Bayes classifier with different percentages of PCA components obtained from 36 features.

TABLE III. BEST PERFORMANCE OF EACH OF THE FOUR CLASSIFIERS USING PCA ANALYSIS ON BOTH MEDIAL AND LATERAL FEATURES

Classifier	PCA variance	Precision (PPV)	Recall (Sensitivity)	F-Measure	MCC	ROC Area
ANN	Top 55%	0.712	0.717	0.714	0.304	0.761
SVM	Top 65%	0.703	0.717	0.624	0.145	0.651
Random forest	Raw data	0.681	0.717	0.660	0.182	0.677
Naive Bayes	Top 20%	0.699	0.727	0.685	0.237	0.724

B. Experiments for JSM and JSL grades prediction

JSM and JSL grades refer to the medial and lateral compartments for the Joint Space Narrowing (JSN) measurement, respectively. Besides KL grade, JSM and JSL are often used as measures for OA severity. In previous section, we evaluated the relation between CDI feature points and KL grade; in this section, we extended the experiment to evaluate the relation between CDI feature points and the two new

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2018.2840082, IEEE Transactions on NanoBioscience

measures, JSL and JSM grades. The goal is to validate if the similar pattern could be detected between CDI and different OA severity measures.

1) Experiment 1: Predict the change of JSM grade.

JSM grade is measured based on the joint space on the medial side of a knee joint. To predict the JSM grade change, we first tried the 18 information locations on medial compartment. Besides, we also tried using the CDI on the whole tibiofemoral compartment, which comprises of both lateral and medial, since experiment in section III.A showed that all 36 CDI points generate best prediction result for KL grade. The four methods of the machine learning and PCA analysis on this feature set were run similarly as described in section III.A. Tables IV and V provide a summary of the best performance of each machine learning method, using the 18-dimensional feature set and the 36-dimensional feature set, respectively, for JSM prediction.

TABLE IV. BEST PREDICTION PERFORMANCE OF EACH MACHINE LEARNING METHOD FOR JSM GRADE USING 18 MEDIAL FEATURES

Classifier	PCA variance	Precision (PPV)	Recall (Sensitivity)	F-Measure	MCC	ROC Area
ANN	Top 30%	0.688	0.707	0.691	0.268	0.694
SVM	Top 20%	0.737	0.747	0.720	0.354	0.705
Random forest	Top 100%	0.695	0.717	0.689	0.270	0.704
Naive Bayes	Raw Data	0.742	0.707	0.716	0.39	0.738

TABLE V. BEST PREDICTION PERFORMANCE OF EACH MACHINE LEARNING METHOD FOR JSM GRADE USING 36 FEATURES

Classifier	PCA variance	Precision (PPV)	Recall (Sensitivity)	F-Measure	MCC	ROC Area
ANN	Top 25%	0.709	0.727	0.703	0.303	0.717
SVM	Top 75%	0.465	0.657	0.544	-0.119	0.681
Random forest	Top 55%	0.746	0.758	0.743	0.396	0.785
Naive Bayes	Raw Data	0.715	0.657	0.669	0.318	0.742

As Tables IV and V showed, using 36-dimensional feature set generated better prediction results in terms of different evaluation metrics, than using 18 medial features. The best AUC 0.785 was achieved by random forest classifier with top 55% of PCA components using 36 CDI features. The performance of three machine learning methods (ANN, naïve Bayes, and random forest) improved after adopting 36 features as compared to using medial features only, while the SVM's performance recorded a slight drop.

For each feature set, we selected the classifier that achieved the best performance and depicted their performance in Figs. 15-16. When using 18 medial features, the best performance (AUC = 0.738) was from naïve Bayes utilizing raw data; when using 36 features, the best performance (AUC = 0.785) was from random forest utilizing top 55% of PCA components.

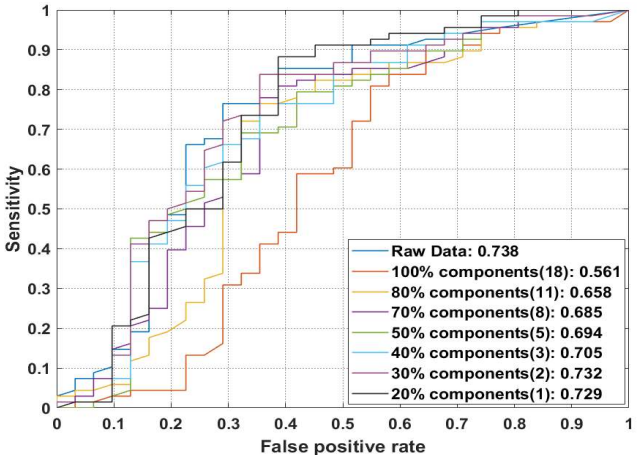

Fig. 15. ROC curves of naïve Bayes with different percentages of PCA components obtained from 18 medial features for JSM prediction.

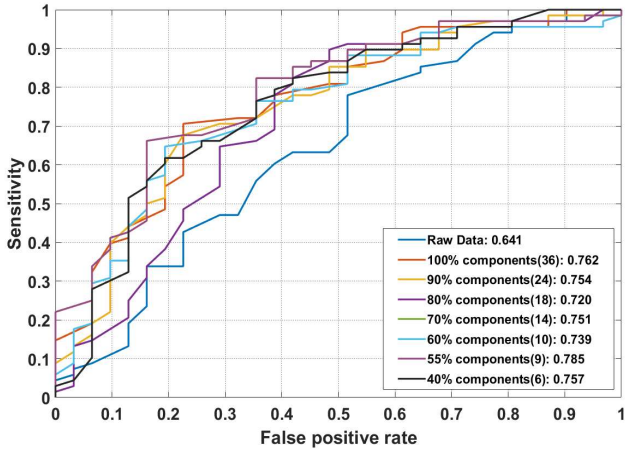

Fig. 16. ROC curves of random forest with different percentages of PCA components obtained from 36 features for JSM prediction.

2) Experiment 2: Predict the change of JSL grade.

JSL grade is measured from the lateral side of joint space narrowing. Therefore, we used the 18 information locations on the lateral compartment, as well as the 36 information locations on both lateral and medial compartments. The interesting finding is similar to what we observed from the experiment for JSM grade prediction, that using 36 CDI points achieved better prediction accuracy than using 18 lateral CDI points, although JSL is measure based on lateral compartment only.

The best performance exhibited by each clarifier was summarized in Tables VI and VII, for 18-dimensional feature set and 36-dimensional feature set, respectively. Figs. 17-18 displayed the ROC performance associated with the best machine learning method for each feature set (best performance using 18 features was from naïve Bayes with AUC 0.657 while best performance using 36 features was from ANN with AUC 0.695).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNB.2018.2840082, IEEE Transactions on NanoBioscience

TABLE VI. BEST PREDICTION PERFORMANCE OF EACH MACHINE LEARNING METHOD FOR JSN GRADE USING 18 LATERAL FEATURES

Classifier	PCA variance	Precision (PPV)	Recall (Sensitivity)	F-Measure	MCC	ROC Area
ANN	Top 55%	0.855	0.869	0.86	0.313	0.639
SVM	Top 98%	0.771	0.869	0.817	-0.038	0.618
Random forest	Raw Data	0.771	0.869	0.817	-0.038	0.677
Naive Bayes	Top 10%	0.769	0.848	0.807	-0.066	0.657

TABLE VII. BEST PREDICTION PERFORMANCE OF EACH MACHINE LEARNING METHOD FOR JSN GRADE USING 36 FEATURES

Classifier	PCA variance	Precision (PPV)	Recall (Sensitivity)	F-Measure	MCC	ROC Area
ANN	Top 100%	0.785	0.808	0.796	-0.01	0.695
SVM	Top 97%	0.815	0.808	0.811	0.131	0.648
Random forest	Raw Data	0.771	0.869	0.817	-0.038	0.662
Naive Bayes	Top 65%	0.901	0.889	0.845	0.272	0.606

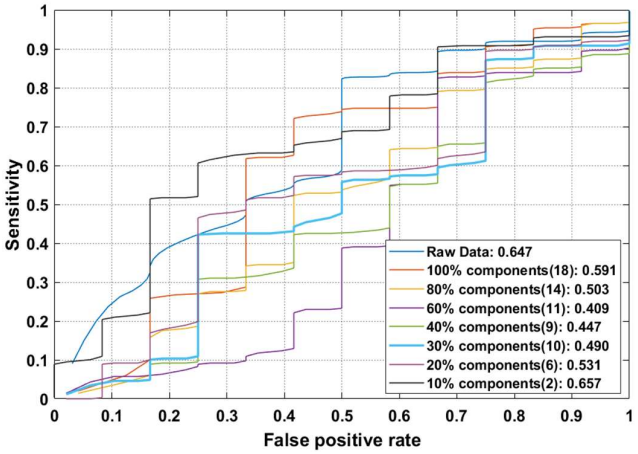

Fig. 17. ROC curves of naïve Bayes classifier with different percentages of PCA components obtained from 18 lateral features for JSN prediction.

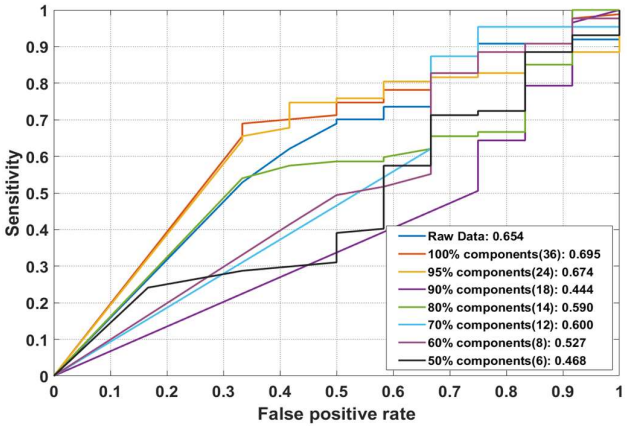

Fig. 18. ROC curves of ANN classifier with different percentages of PCA components obtained from 36 features for JSN prediction.

From the experiment results, it is clear that the correlation between the JSL grade change and the CDI feature sets is weaker than the correlation between the JSM grade and CDI feature sets. When all 36 informative locations were used, prediction performance improved for both JSM and JSL grades, though accuracy of JSL prediction is still lower than the accuracy of JSM prediction (AUC 0.695 vs. AUC 0.785).

From the experiment results we can find that the medial CDI locations provide more distinguishing and informative features than the lateral CDI locations when predicting the change of joint space narrowing. This observation is consistent with the findings from clinical studies which showed the degenerative changes in the knee for OA disease are more prone to affect the medial compartment than the lateral compartments [24].

IV. DISCUSSION AND CONCLUSION

In this paper, we used machine learning methods to explore the hidden biomedical information contained in the clinically used Cartilage Damage Index (CDI), to predict the change of KL, JSM, and JSL grades, respectively. These measurements provide different perspective to measure the progression of knee osteoarthritis disease. We computed the CDI information from each of the 36 informative locations on tibiofemoral compartment from 3D knee MR imaging and used PCA analysis as a feature selection method. The processed feature set and original raw feature set were served as input to four machine learning methods (ANN, SVM, random forest and naïve Bayes). In particular, to examine the possible different effect of medial and lateral informative locations, we have divided the 36-dimensional feature set into 18-dimensional medial feature set and 18-dimensional lateral feature set and run the experiment on all four classifiers separately with 10-fold cross-validation.

Several interesting findings are observed. First, for KL grade prediction, using medial feature set generated better prediction performance than using lateral feature set, while the total 36-dimensional feature set generated the best. Similar finding was observed for JSM and JSL grade prediction, which indicates that CDI points from medial compartment contains more valuable information for OA progression prediction. Therefore, clinical application of CDI could consider to select more points from the medial tibiofemoral compartment while reduce the number of points selected from the lateral tibiofemoral compartment.

Second, PCA analysis is helpful in feature space reduction and performance improvement, for OA severity grade change prediction. The best performance of KL grade prediction was achieved by ANN using top 55% of PCA components on the 36-dimensional feature set. The best performance of JSM prediction was achieved by random forest also using top 55% of PCA components, while the best performance of JSL prediction was achieved by ANN using 100% of PCA components, both on 36-dimensional feature set.

Third, though JSM is defined and measured based on the medial part of knee joint, it is found that using CDI locations on both medial and lateral parts can better predict the change of JSM, than using CDI locations from medial part only. Consistent finding is observed for JSL grade prediction. Our rationale behind this is that medial and lateral compartments are correlated closely with each other, so using CDI on both compartments could provide more complete information for

machine learning methods to predict the change of the severity grade.

In the future, we are going to incorporate cartilage information from patella into the analysis, i.e., another 24 informative locations defined by CDI. Patella compartment was usually paid less attention than femur and tibia. We will analyze the informative locations from patella and test the classifiers using combined feature set with medial and lateral CDI locations. Besides, we will enlarge our dataset by selecting more cases from OAI database. The current dataset size is limited and prevents us from applying techniques such as deep learning strategies which require large amount of training samples.

REFERENCES

[1] Zhang Y, Jordan JM. Epidemiology of osteoarthritis. Clin Geriatr Med. 2010;26(3):355-69.

[2] Brandt KD, Doherty M, Lohmander LS. Osteoarthritis. 2nd ed. Oxford ; New York: *Oxford University Press*; 2003. xiii, 598 p.

[3] Rosemont IL. United States Bone and Joint Decade. The Burden of Musculoskeletal Diseases in the United States. *American Academy of Orthopaedic Surgeons*, 2008.

[4] Chu CR, Williams AA, Coyle CH, Bowers ME. Early diagnosis to enable early treatment of pre-osteoarthritis. *Arthritis research & therapy*. 2012;14(3):212.

[5] Bhatia D, Bejarano T, Novo M. Current interventions in the management of knee osteoarthritis. *Journal of pharmacy & bioallied sciences*. 2013;5(1):30-8.

[6] Shim H, Chang S, Tao C, Wang JH, Kwoh CK, Bae KT. Knee cartilage: efficient and reproducible segmentation on high-spatial-resolution MR images with the semiautomated graph-cut algorithm method. *Radiology*. 2009;251(2):548-56.

[7] J.L. Jaremko, R.W.T. Cheng, R.G.W. Lambert, A.F. Habib, J.L. Ronsky, "Reliability of an efficient MRI-based method for estimation of knee cartilage volume using surface registration," Osteoarthritis Cartilage, vol. 14, pp. 914–922, 2006

[8] Y. Yin, X. Zhang, R. Williams, X. Wu, D. Anderson, M. Sonka, "LOGISMOS – Layered Optimal Graph Image Segmentation of Multiple Objects and Surfaces: cartilage segmentation in the knee joint," IEEE Trans Med Imaging, vol. 29, pp. 2023–2037, 2010.

[9] J. Fripp, S. Crozier, S. Warfield, S. Ourselin, "Automatic segmentation and quantitative analysis of the articular cartilages from magnetic resonance images of the knee," IEEE Trans Med Imaging, vol. 29, pp. 55–64, 2010.

[10] F. Eckstein, W. Wirth, "Quantitative cartilage imaging in knee osteoarthritis," Arthritis, vol. 2011, pp. 1–19, 2011.

[11] M. Zhang, J. B. Driban, L. Lyn Price, G. H. Lo, E. Miller, and T. E. McAlindon, "Development of a Rapid Cartilage Damage Quantification Method for the Lateral Tibiofemoral Compartment Using Magnetic Images: Data from the Osteoarthriis Initiative," Hindawi Publishing Corporation, 2015.

[12] M. Zhang, J. B. Driban, L. Lyn Price, D. Harper, G. H. Lo, E. Miller, R. J. Warld, and T. E. McAlindon, "Development of a Rapid Knee Cartilage Damage Quantification Method Using Magnetic Resonance Images," BMC Musculokeletal Disorders, vol. 15, pp. 264, 2014.

[13] M. Zhang, L. Lyn Price, A. R. Canavatchel, J. B. Driban, P. Yuan, G. H. Lo, T. E. McAlindon, "Cartilage Loss Primarily Occurs in the Most Affected Tibiofemoral Compartment with no Evidence of a Ceiling Effect among Advanced-Stage Disease: A Two-year Longtitudinal Study of Data from the Osteoarthritis," 2016 ACR/ARHP Annual Meeting, 2016.

[14] F. Eckstein, M. Hudelmaier, W. Wirth, B. Kiefer, R. Jackson, J. Yu, C.B. Eaton, E. Schneider, "Double echo steady state magnetic resonance imaging of knee articular cartilage at 3 Tesla: a pilot study for the Osteoarthritis Initiative," Ann Rheum Dis, vol. 65, pp. 433–441, 2006.

[15] L.D. Bennett, J.C. Buckland-Wright, "Meniscal and articular cartilage changes in knee osteoarthritis: a cross-sectional double-contrast macroradiographic study," Rheumatology, vol. 41, pp. 917–923, 2002.

[16] L. Sharma, J. Song, D. Dunlop, D. Felson, C.E. Lewis, N. Segal, J. Torner, T.D. Cooke, J. Hietpas, J. Lynch, M. Nevitt, "Varus and valgus alignment and incident and progressive knee osteoarthritis," Ann Rheum Dis, vol. 69, pp. 1940–1945, 2010.

[17] J. Shlens, "A Tutorial on Principal Component Analysis," ArXiv, pp. 1–13, 2014.

[18] D. Rumelhart, G. Hinton, and R. Williams, "Learning internal representations by error propagation," 1985.

[19] C. Cortes and V. Vapnik, "Support-Vector Networks," Mach. Learn., vol. 297, pp. 273–297, 1995.

[20] L. Breiman, "Random Forests," Mach. Learn., vol. 4, pp. 5–32, 2001.

[21] D. J. Hand, K. Yu, "Idiot's Bayes — not so stupid after all?", International Statistical Review, vol. 69 (3), pp. 385–399, 2011.

[22] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, "The WEKA data mining software: An update," SIGKDD Explorations, vol. 11(1), pp. 10-18, 2009.T. E. McAlindon, M. LaValley, W. F. Harvey,

[23] L. Lyn Price, J. B. Driban, M. Zhang, R. Ward, "Effects of Intra-Articular Triamcinolone vs Saline on Knee Cartilage Volume and Pain in Patients With Knee Osteoarthritis: A Randomized Clinical Trial," JAMA, vol. 317(19), pp. 1967-1975, 2017.

[24] F. Cicuttini, A. Wluka, J. Hankin, Y. Wang"Longitudinal study of the relationship between knee angle and tibiofemoral cartilage volume in subjects with knee osteoarthritis." Rheumatology, vol. 43(3), pp. 321-324, 2004.

[25] Pedoia V, Majumdar S, Link TM. Segmentation of joint and musculoskeletal tissue in the study of arthritis. *Magma*. 2016;29(2):207-21.

[26] P.M.M. Cashman, R.I. Kitney, M.A. Gariba, M.E. Carter, "Automated techniques for visualization and mapping of articular cartilage in MR images of the osteoarthritic knee: a base technique for the assessment of microdamage and submicro damage," Trans NanoBioscience 2002, vol. 1, pp. 42–51, 2002.

[27] Z.T. Hussain, S.S. Usha, "Automated image processing and analysis of cartilage MRI: enabling technology for data mining applied to osteoarthritis," In AIP Conference Proceedings, pp. 262–276, 2007.

[28] G. Vincent, C. Wolstenholme, I. Scott, M. Bowes, "Fully Automatic Segmentation of the Knee Joint using Active Appearance Models," In Medical Image Analysis for the Clinic: A Grand Challenge, 2010.

[29] Zhang M, Price LL, Canavatchel A, Driban JB, Yuan P, Lo GH, McAlindon TE. Cartilage Loss Primarily Occurs in the Most Affected Tibiofemoral Compartment with no Evidence of a Ceiling Effect among Advanced-Stage Disease: A Two-year Longitudinal Study of Data from the Osteoarthritis. *Arthritis Rheum*., 2016.