Rytky, S. J.O.; Tiulpin, Aleksei; Frondelius, T.; Finnilä, M. A.J.; Karhula, S. S.; Leino, Janina; Pritzker, K. P.H.; Valkealahti, M.; Lehenkari, P.; Joukainen, Antti; Kröger, H.; Nieminen, H. J.; Saarakkala, S.

Automating three-dimensional osteoarthritis histopathological grading of human osteochondral tissue using machine learning on contrast-enhanced micro-computed tomography

Published in:
Osteoarthritis and Cartilage

DOI:
10.1016/j.joca.2020.05.002

Published: 01/08/2020

Document Version
Publisher's PDF, also known as Version of record

Published under the following license:
CC BY

Please cite the original version:
Rytky, S. J. O., Tiulpin, A., Frondelius, T., Finnilä, M. A. J., Karhula, S. S., Leino, J., Pritzker, K. P. H., Valkealahti, M., Lehenkari, P., Joukainen, A., Kröger, H., Nieminen, H. J., & Saarakkala, S. (2020). Automating three-dimensional osteoarthritis histopathological grading of human osteochondral tissue using machine learning on contrast-enhanced micro-computed tomography. *Osteoarthritis and Cartilage*, *28*(8), 1133-1144. https://doi.org/10.1016/j.joca.2020.05.002

Osteoarthritis and Cartilage 28 (2020) 1133–1144

Osteoarthritis and Cartilage

Automating three-dimensional osteoarthritis histopathological grading of human osteochondral tissue using machine learning on contrast-enhanced micro-computed tomography

S.J.O. Rytky [†] [*], A. Tiulpin [†] [‡], T. Frondelius [†], M.A.J. Finnilä [†] [§], S.S. Karhula [†] [‡], J. Leino [†], K.P.H. Pritzker [||] [¶], M. Valkealahti [#], P. Lehenkari [§] [#] [††], A. Joukainen [‡‡], H. Kröger [‡‡], H.J. Nieminen [§§] [†], S. Saarakkala [†] [‡]

[†] *Research Unit of Medical Imaging, Physics and Technology, University of Oulu, Oulu, Finland*
[‡] *Department of Diagnostic Radiology, Oulu University Hospital, Oulu, Finland*
[§] *Medical Research Center, University of Oulu, Oulu, Finland*
[||] *Department of Laboratory Medicine and Pathobiology, Surgery University of Toronto, Toronto, Ontario, Canada*
[¶] *Mount Sinai Hospital, Toronto, Ontario, Canada*
[#] *Department of Surgery and Intensive Care, Oulu University Hospital, Oulu, Finland*
[††] *Cancer and Translational Medical Research Unit, Faculty of Medicine, University of Oulu, Oulu, Finland*
[‡‡] *Department of Orthopaedics, Traumatology and Hand Surgery, Kuopio University Hospital, Kuopio, Finland*
[§§] *Dept. of Neuroscience and Biomedical Engineering, Aalto University, Espoo, Finland*

ARTICLE INFO

Article history:
Received 22 August 2019
Accepted 1 May 2020

Keywords:
Osteoarthritis
3D histopathological grading
Contrast-enhanced micro-computed tomography
Machine learning
Cartilage
Textural analysis

SUMMARY

Objective: To develop and validate a machine learning (ML) approach for automatic three-dimensional (3D) histopathological grading of osteochondral samples imaged with contrast-enhanced micro-computed tomography (CEμCT).

Design: A total of 79 osteochondral cores from 24 total knee arthroplasty patients and two asymptomatic donors were imaged using CEμCT with phosphotungstic acid -staining. Volumes-of-interest (VOI) in surface (SZ), deep (DZ) and calcified (CZ) zones were extracted depth-wise and subjected to dimensionally reduced Local Binary Pattern -textural feature analysis. Regularized linear and logistic regression (LR) models were trained zone-wise against the manually assessed semi-quantitative histopathological CEμCT grades (diameter = 2 mm samples). Models were validated using nested leave-one-out cross-validation and an independent test set (4 mm samples). The performance was primarily assessed using Mean Squared Error (MSE) and Average Precision (AP, confidence intervals are given in square brackets).

Results: Highest performance on cross-validation was observed for SZ, both on linear regression (MSE = 0.49, 0.69 and 0.71 for SZ, DZ and CZ, respectively) and LR (AP = 0.9 [0.77–0.99], 0.46 [0.28–0.67] and 0.65 [0.41–0.85] for SZ, DZ and CZ, respectively). The test set evaluations yielded increased MSE on all zones. For LR, the performance was also best for the SZ (AP = 0.85 [0.73–0.93], 0.82 [0.70–0.92] and 0.8 [0.67–0.9], for SZ, DZ and CZ, respectively).

Conclusion: We present the first ML-based automatic 3D histopathological osteoarthritis (OA) grading method which also adequately perform on grading unseen data, especially in SZ. After further development, the method could potentially be applied by OA researchers since the grading software and all source codes are publicly available.

[*] Address correspondence and reprint requests to: S.J.O. Rytky, Research Unit of Medical Imaging, Physics and Technology, University of Oulu, Oulu, Finland, POB 5000, FI-90014, Oulu, Finland. Tel.: 358-45-672-6367.

E-mail addresses: santeri.rytky@oulu.fi (S.J.O. Rytky), aleksei.tiulpin@oulu.fi (A. Tiulpin), juha.frondelius@student.oulu.fi (T. Frondelius), mikko.finnila@oulu.fi (M.A.J. Finnilä), sakari.karhula@oulu.fi (S.S. Karhula), janina.leino@student.oulu.fi (J. Leino), kenpritzker@gmail.com (K.P.H. Pritzker), maarit.valkealahti@oulu.fi (M. Valkealahti), petri.lehenkari@oulu.fi (P. Lehenkari), antti.joukainen@kuh.fi (A. Joukainen), heikki.kroger@kuh.fi (H. Kröger), heikki.j.nieminen@aalto.fi (H.J. Nieminen), simo.saarakkala@oulu.fi (S. Saarakkala).

https://doi.org/10.1016/j.joca.2020.05.002
1063-4584/

Introduction

Conventional microscopic histopathological grading of osteochondral tissue is the gold standard for assessment of osteoarthritis (OA) severity *ex vivo*. The most commonly used OA grading methods are Osteoarthritis Research Society International (OARSI)[1] and Mankin[2] scoring systems[3]. Mankin scoring system was developed based on late-stage OA samples, having limitations for assessment of early OA[4] and disease extent[5]. Consequently, the OARSI grading system was introduced later to address these issues, offering more sensitivity to the mild and moderate progressive changes in articular cartilage (AC), as well as functional information on cartilage properties[6]. Generally, histopathological grading methods sensitive to early changes are highly valuable for drug development and basic OA research[7]. Furthermore, sensitive grading methods might potentially be utilized in developing biomarkers, which are essential when developing prevention of the late-stage disease or non-surgical disease-modifying treatments[8,9].

The conventional histopathological methods are complex, destructive and time consuming[4], and also unable to capture all of the OA-induced changes within the full sample volume. Recently, methods combining multiple thin sections into 3D volume through image registration have been proposed[10,11]. However, such approaches can only avert partly the problem of two-dimensionality with the expense of a more laborious protocol.

Multiple 3D histopathological grading methods for different tissues have been proposed in the literature, based on magnetic resonance imaging (MRI)[12–15], optical imaging[16], ultrasound[17], and atomic force microscopy[18]. 3D grading methods could possibly serve as a reference for clinical 3D modalities, as well as higher resolution 3D techniques. Contrast-enhanced micro-computed tomography (CEμCT) has shown potential in fast quantitation of osteochondral features while preserving the sample and reducing user bias[19]. We recently introduced a protocol for contrast-enhanced micro-computed tomography (CEμCT) using phosphotungstic acid (PTA) as a collagen-specific contrast agent[20,21], and consequently, developed a 3D OA grading system to assess each AC zone separately[22]. However, the current 3D μCT grading system still requires manual assessment, thus, having a risk for user-dependent bias. The automation of this process could provide more objective evaluations.

Recently, methods for the quantitative 3D analysis of AC surface[23,24], calcified cartilage[25] and full cartilage tissue[19] degeneration, as well as chondrocyte organization[26,27] with CEμCT, have been reported. However, most of the current methods are either limited to a single osteochondral zone[23–25] or not validated via independent testing[19]. The current implementations could be improved by developing more generalizable methods applicable to analyze multiple different osteochondral zones while utilizing more advanced validation techniques that show their feasibility on unseen data.

The development of machine learning (ML) techniques has enabled a data-driven approach in pattern recognition and decision making without the need for explicit programming. Several grading methods have been proposed in fields outside OA research[28,29]. ML has been applied in clinical OA research in several domains, such as the prediction of OA severity[30–33] and progression[15,34,35] using X-ray radiographs[30,31,33,34] or MRI analysis[15,32,35]. However, little attention has been paid to ML in pre-clinical OA research[26,36,37].

In this study, we aim to automate the recently proposed histopathological grading[22] of CEμCT imaged osteochondral samples using ML. The feasibility of performing the automatic grading in different cartilage zones, and the robustness of the developed models to a sample acquisition protocol change, are assessed with an independent test set.

Materials and methods

Sample preparation

Osteochondral cores were harvested from tibial plateaus and femoral weight-bearing areas of human knee joints (Supplementary Fig. 1). A total of 90 cores were extracted from 24 total knee arthroplasty (TKA) patients and two asymptomatic donors. Subsequently, 79 samples that contained both the cartilage and bone layers were included in the study, and split into two datasets based on the core diameter Ø:

- Cross-validation set; 19 patients, $n = 34$, Ø = 2 mm, ethical approval PPSHP 78/2013, Ethical committee of Northern Ostrobothnia's Hospital District
- Test set; seven patients, $n = 45$, Ø = 4 mm, ethics approval PPSHP 78/2013; PSSHP 58/2013 & 134/2015, Research Ethics Committee of the Northern Savo Hospital District

Detailed sample and patient distributions are given in Supplementary Table 1. After the core extraction, all the samples were kept frozen at −80°C. Before the imaging, the samples were thawed and then fixed in 10% neutral-buffered formalin for 5 days. Fixation was followed by a minimum of 8h wash in 70% ethanol and minimum 48h immersion in 70% ethanol, 1% w/v PTA solution[20,21]. PTA is negatively charged and can bind to collagen ionically since collagen has a positive net charge in low pH solution. Electromagnetic repulsion of the negative proteoglycans might hinder the effect of PTA binding, and sufficient time for the passive diffusion of the contrast agent should be allowed[21]. To prevent sample drying during μCT imaging, each sample was wrapped in Parafilm (Parafilm M, Bemis Company Inc, Neenah, WI, USA) and orthodontic wax (Orthodontic Wax, Ortomat Hepola, Turku, Finland).

Imaging

The imaging was conducted right after the PTA immersion was completed. Samples were imaged using a desktop μCT setup

Dataset	Zone	Grade 0	Grade 1	Grade 2	Grade 3
Cross-validation	S	7	11	13	3
	D	8	16	8	2
	C	8	16	7	3
Test	S	2	19	9	14
	D	0	16	15	13
	C	0	24	11	9

S = Surface zone, D = Deep zone, C = Calcified zone.

Table I

Distribution of μCT grades assessed from the reconstructions (used as ground truth). The cross-validation set contained only a small number of samples from grade three and a reduced number of healthy samples, while almost no healthy samples were found in the test set. Otherwise, samples were distributed relatively evenly

(Skyscan 1,272; Bruker microCT, Kontich, Belgium; Scanning parameters: 45 kV, 222 µA, 3.2 µm voxel side length, 3,050 ms, two frames/projection, 1,200 projections, additional 0.25 mm aluminum filter).

During the imaging of the test set, we improved the data acquisition protocol by checking the sample voids – areas of deep cartilage with no PTA accumulation (Supplementary Video 2 in *Nieminen et al.*[20]). We observed that the voids appeared due to the insufficient diffusion time, especially in samples with a very thick AC layer. In the new protocol, upon detection of a void in the µCT scan, the sample was re-immersed in PTA to allow full diffusion to deep AC.

3D histopathological grading

We used reconstructed data to determine the semi-quantitative 3D histopathological grades for each sample, corresponding to the analyzed zones[22]. J. Leino conducted the grading according to the published method[22], assessing each sample independently. In this study we used the following grades:

- Surface continuity: Smooth and continuous = 0; Slightly discontinuous = 1; Moderately discontinuous = 2; Severely discontinuous = 3,
- Deep cartilage (zone 3, DZ) extracellular matrix (ECM) disorganization: Normal = 0; Slightly disorganized = 1; Moderately disorganized = 2; Severely disorganized = 3
- Calcified cartilage (zone 4, CZ) ECM disorganization: Normal = 0; Slightly disorganized = 1; Moderately disorganized = 2; Severely disorganized = 3

Grade distribution is presented in Table I and graphically in Supplementary Fig. 2. Besides the multiclass grades, we also used dichotomized grades and split them into intact/mild VOI degeneration and moderate/severe VOI degeneration groups (Grades zero and one were grouped against two and three). In our analyses, we excluded the grades from the middle zone (zone 2), since we

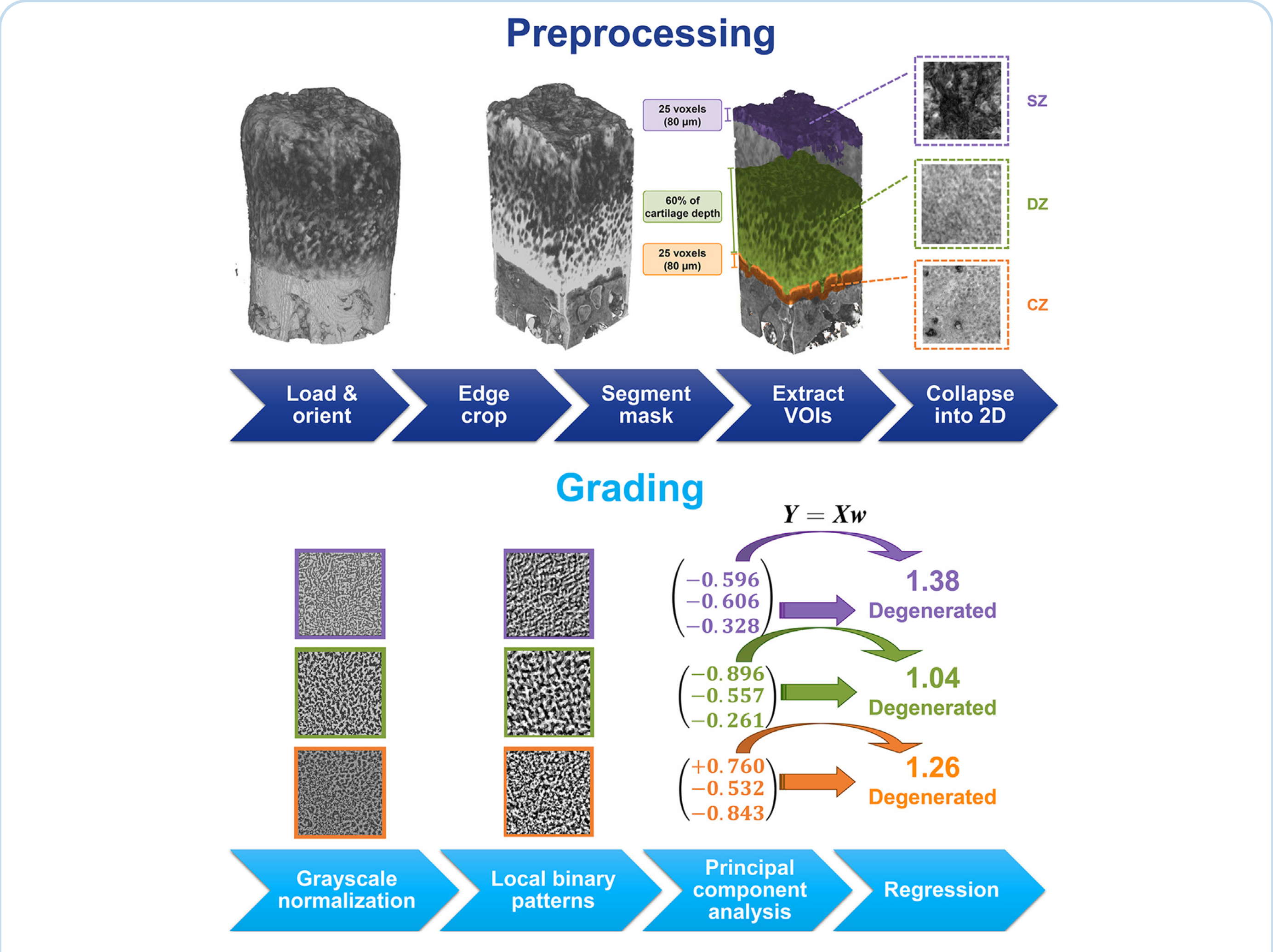

Fig. 1 The workflow of the analysis methods used for CEµCT imaged samples. SZ = surface zone, DZ = deep zone, CZ = calcified zone.

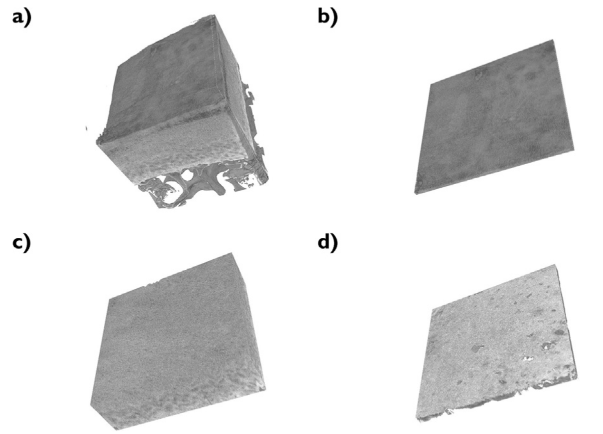

Fig. 2 **a)** Oriented and edge-cropped VOI from a healthy/mildly degenerated osteochondral sample in the test set (harvested from an asymptomatic donor), **b)** Sub-VOI from the cartilage surface, **c)** deep cartilage, and **d)** calcified tissue. A smooth and continuous surface is visible. Deep and calcified ECM are well organized.

believe that the automatic selection of transitional zone is error-prone, especially in the late-OA samples with a delaminated surface layer (ECM loss extending to the transitional zone).

Basic data pre-processing

A python *ad hoc* software was developed to preprocess the image stacks and train the regression and classification models. The method workflow is illustrated in Fig. 1. The reconstructed samples were loaded and oriented using an optimization algorithm. The center of the sample in the XY plane was detected by finding the center of mass of the image stack summed along the *Z*-axis (Z – sample's depth dimension). Edges of the sample were cropped using detected center and pre-defined VOI size (1300 μm · 1300 μm · Z for Ø = 2 mm, 2600 μm · 2600 μm · Z for Ø = 4 mm). Orientation and edge cropping processes are further described in Supplementary Fig. 3.

Calcified cartilage segmentation and VOI extraction

After cropping the sample edges, the calcified cartilage interface (tidemark) was segmented in a slice-by-slice-manner. We adapted the method from[38] and utilized U-Net[39] – a Deep convolutional neural network (CNN). The proposed modification included the use of transfer learning, for which we utilized a ResNet-34 encoder pretrained on the ImageNet dataset[40] (~14 million natural images labeled into 1,000 generic classes). The segmentation model was trained on the cross-validation set (Ø = 2 mm). To enrich the training data, we used such data augmentations as random translations, rotations, scaling, flips, added noise, brightness and contrast modifications. During inference, the model predictions were averaged along coronal and sagittal planes, and the segmentation probability map was thresholded with 0.5 (cross-validation set) or 0.2 (test set). The lower thresholding was applied for the test set due to images having a lower signal (see Supplementary Fig. 4).

Once the calcified tissue mask was acquired, the average depth of AC was calculated using the mask and the surface coordinates of the samples. The depth for DZ was set as 60% of AC depth to ensure that the full zone was included also on the samples with the delaminated surface layer. The lower limit for DZ was set to 30 μm above the segmentation mask to ensure that the interface and calcified tissues were not included in DZ. The surface was detected using the Otsu threshold, and surface zone (SZ) was set extending 80 μm below (25 slices). CZ was set as 80 μm thick volume immediately below DZ. Here, we used small zone thickness values to focus on the detailed surface features and account for samples with thin CZ. Extracted volumes (Figs. 2 and 3) were collapsed into two-dimensional (2D) texture images summing their mean and the standard deviation depth-wise.

Finally, all the Ø = 4 mm samples included in the test set were split into nine smaller sub-images (with dimensions half to the original image) to increase prediction reliability. This was also done to make sure that the textural features of the large image have similar relative size and impact on the resulting feature descriptor used to predict the 3D grades of the sample, compared to the features trained on cross-validation.

Feature extraction

Prior to the feature extraction, possible misalignment artifacts that appeared during preprocessing were automatically cropped out. In the algorithm, possible defects on the image corners were detected using adaptive thresholding and cropped. Subsequently, we performed a local normalization by subtracting from each pixel of its neighborhood's weighted intensity. Here, we used a gaussian kernel for intensity weighing. The kernel parameters were optimized independently for each sample zone (Supplementary Table 2).

To extract the features related to cartilage degeneration, Median Robust Extended Local Binary Patterns (MRELBP) were calculated according to *Liu et al*[41]. In this case, the texture analysis is conducted by comparing median filtered patches in the 2D images on multiple scales. A total of 32 features were extracted. Two features were obtained by thresholding the image patch by the full images' mean intensity. Ten features from small, large and radial local binary pattern (LBP) images are collected, comparing the center patch to eight neighboring patches using rotation-invariant uniform mapping (nine uniform- and one non-uniform patterns). The combined histogram was normalized. Features that did not have any occurrences were excluded, resulting in 28 features. Finally, we mean-centered the data.

After the data centering, a principal component analysis (PCA) -based whitening was used, and consequently, the dimensionality of the extracted feature vectors was also reduced. Here, 90% of the explained variance was set as a threshold for finding the optimal number of principal components. Eventually, three components were automatically selected for all the cartilage zones.

Automatic grading

After the PCA, we used the obtained features to train two regression models on cross-validation. In particular, we used leave-one-patient-out (LOPO) cross-validation, using samples from each individual patient as a validation set, against a model trained on the rest of the patients in the dataset. The cross-validation set had two samples per patient (Supplementary Table 1). Firstly, a linear regression model was trained against the ground truth μCT grades. Here, we used L2 (ridge) regularization with a coefficient of 0.1 and assumed a continuous outcome. Secondly, a binary logistic regression (LR) model (also with L2 regularization) was trained to assess the sample's degeneration.

For the test set images, the developed models were evaluated for all the nine sub-images separately and the average of their predictions was finally used. The models trained with the best hyperparameters from the cross-validation set were selected. To also estimate the effect of switching the cross-validation- and test sets, separate models were subsequently trained for Ø = 4 mm samples using LOPO cross-validation (Replication experiment, see Supplementary Table 3).

Model interpretability

To assess the interpretability of the model's decisions, we used the SHapley Additive exPlanations (SHAP) method[42]. Briefly, it is a game-theoretic approach capable of assessing the impact of individual MRELBP features used via analytical calculation of the Shapley values. The results describe the user if low or high values of the feature contribute to a higher or lower predicted grade.

Parameter optimization

To tune the hyperparameters for MRELBP and grayscale normalization, we used the Bayesian hyperparameter optimization algorithm from the Hyperopt package[43]. To avoid overfitting, we performed a “nested leave-one-out” cross-validation (Fig. 4). In particular, during the leave-one-out (LOO), we used a hyperparameter search on the N-1 (33 out of 34) samples using another, nested LOPO cross-validation split. A regression model was trained for each optimization batch of 33 samples. Optimization was conducted on the cross-validation set evaluating a maximum of 100 parameter sets per iteration. The algorithm converged to the same solution on most of the iterations (30/34 for SZ, 34/34 for DZ and 18/34 for CZ) and we used the most frequent solution as the hyperparameter selection for each zone. Optimized sets of parameters are listed in Supplementary Table 2.

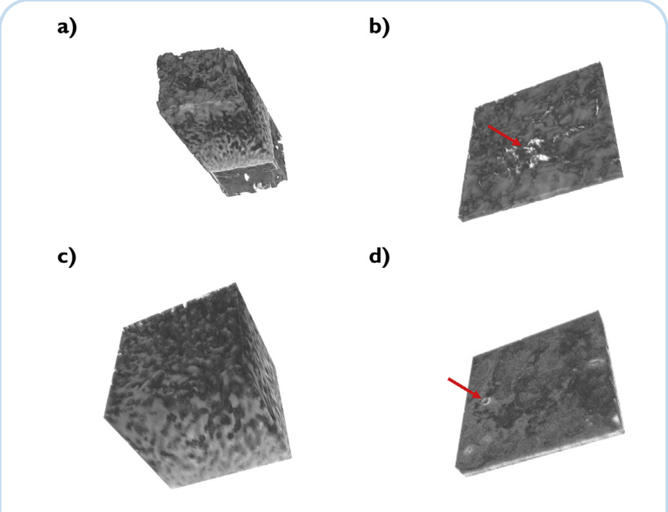

Fig. 3

a) Oriented and edge-cropped VOI from a degenerated osteochondral sample in the cross-validation set (harvested from a TKA patient), **b)** Sub-VOI from the cartilage surface, **c)** deep cartilage, and **d)** calcified tissue. Surface discontinuities, as well as deep and calcified ECM disorganization, are clearly visible. Vascular infiltration and surface discontinuities are shown with a red arrow.

Statistical analyses

Predictions of the linear regression models were assessed using the mean squared error (MSE) and Spearman's correlation analysis. For the LR models, receiver operating characteristic (ROC) curves and precision–recall curves (PRC) were calculated. We evaluated the area under the ROC curve (AUC) and the average precision (AP) of PRC. The 95% confidence intervals were estimated via stratified bootstrapping with 2000 iterations. To further analyze the performance of the binary classification models, we calculated the precision, recall and F1 scores under the threshold of 0.5.

Results

Detection of sample degeneration

The main results of the experiments are presented in Table II and Figs. 5–6. For the cross-validation set, we obtained the AUCs of 0.92 (0.80, 0.99), 0.62 (0.41, 0.81) 0.71 (0.48, 0.90) for SZ, DZ and CZ, respectively. Here, the parentheses indicate 95% confidence intervals. Having the threshold of 0.5 for LR's predictions, the precision (positive predictive value) of the model was found to be high on SZ (0.83), while it remained weak and moderate on DZ and CZ (0.35 and 0.50, respectively). The recall was found to be very high on SZ (0.94) and strong for DZ and CZ (0.70 and 0.60, respectively). F1 scores were 0.88, 0.47 and 0.55 for SZ, DZ and CZ respectively. APs from PRC curves were 0.89 (0.77, 0.99), 0.46 (0.28, 0.67) and 0.65 (0.41, 0.85) for SZ, DZ and CZ, respectively.

For the test set, we obtained the AUCs of 0.81 (0.68, 0.92), 0.68 (0.51, 0.83) and 0.77 (0.62, 0.89) for SZ, DZ and CZ, respectively. Precisions were 0.77, 0.86 and 0.62 for SZ, DZ, and CZ, respectively. The recall was 0.74 on SZ, 0.66 for DZ and 0.62 for CZ. F1 scores were 0.76, 0.75, and 0.62 for Z, DZ and CZ, respectively. APs from PRC curves were 0.85 (0.73, 0.93), 0.82 (0.70, 0.92) and 0.80 (0.67, 0.90) for SZ, DZ and CZ, respectively. Comparable detection accuracy was found for SZ compared to the cross-validation set, while a performance increase was seen on DZ and CZ. The AP of the DZ and CZ models increased by 0.36 and 0.15 compared to the cross-validation set.

ROC and PRC curves (Fig. 5) show that the model for SZ is performing best compared to all zones. On the cross-validation set, ROC curves show that CZ performs slightly better compared to DZ, but the difference is even more obvious in the PRC plot. Similar results can be seen on the test set.

Automatic grading

The performances of all the linear regression models are summarized in Table II and Fig. 6. In particular, the linear regression model yielded MSEs of 0.49, 0.69 and 0.71 for SZ, DZ and CZ, respectively. Strong Spearman's correlation was observed for SZ ($\rho = 0.68$), while weak correlations were observed for DZ ($\rho = 0.24$) and CZ ($\rho = 0.18$) compared to the manual grades.

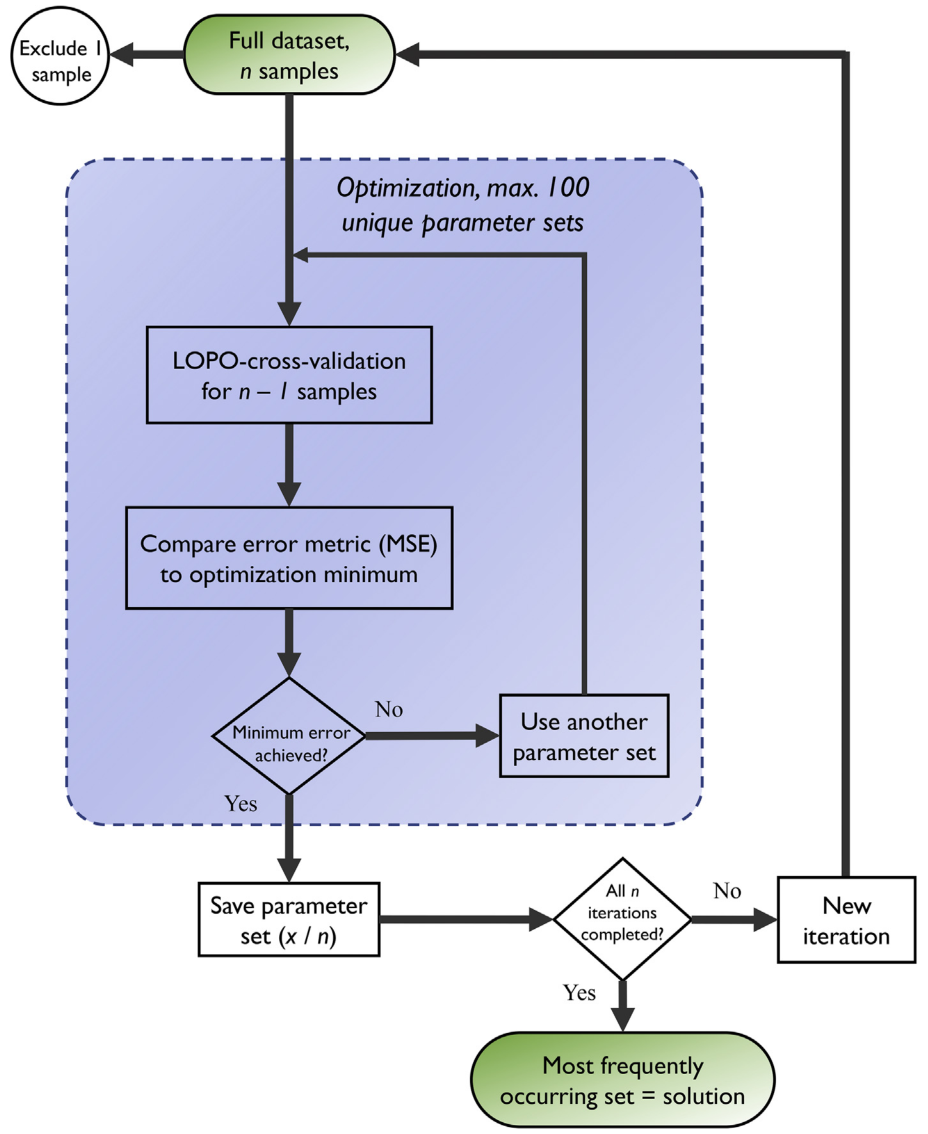
Exclude 1 sample
Full dataset, n samples
Optimization, max. 100 unique parameter sets
LOPO-cross-validation for n – 1 samples
Compare error metric (MSE) to optimization minimum
Minimum error achieved?
No
Use another parameter set
Yes
Save parameter set (x / n)
All n iterations completed?
No
New iteration
Yes
Most frequently occurring set = solution

Fig. 4 Flowchart describing the nested cross-validation method used in the parameter optimization. First, LOO is performed resulting in ***n* - *1*** samples in the optimizations. A maximum of 100 parameter sets are evaluated in the optimization algorithm, where regression is performed with the LOPO split. Initial LOO results in 34 optimization results, and the most frequent parameter set is used as a final solution.

For the test set, we evaluated the predictions using the models that were saved during the training of the cross-validation set. The test set yielded MSEs of 0.87, 1.33 and 1.01 for SZ, DZ and CZ, respectively. Spearman's correlation was moderate ($\rho = 0.46, 0.59$) on SZ and CZ, and weak ($\rho = 0.28$) on DZ.

Interpretability analysis

The results of the interpretability analysis are shown in Supplementary Figs. 6–7. The most significant features from the image are the center (threshold by image mean), non-uniform features (Large and Small radius) as well as uniform patterns consisting of three or four connected neighbors on the Small and Large radius (U-3 and U-4). High sample degeneration is associated with a high amount of image patches below mean intensity (high Center -, low Center +), low amount of non-uniform patterns and a high number of uniform patterns.

Replication experiment

The replication experiment was performed to assess the transferability of the developed texture-based volume analysis technique. The results from the model trained separately for the test set with LOPO cross-validation are shown in Supplementary Table 3. Ridge regression showed improvement in MSE (0.87 → 0.82, 1.33 → 0.81, 0.73 → 0.55, for SZ, DZ and CZ, respectively) but not in Spearman's correlation. LR yielded worse results using ROC/AUC and PRC analysis based on AP (0.85 → 0.74, 0.82 → 0.77 and 0.80 → 0.69, for SZ and CZ, respectively). However, additional parameters show that recall and F1 score are improved in SZ, when

Dataset	Zone	Linear Regression			Logistic Regression				
		MSE	SC	*p*-value	AUC (95% CI)	AP (95% CI)	Prec.	Recall	F1
Cross-validation	S	0.49	0.68	<0.0001	0.92 (0.80–0.99)	0.89 (0.77–0.99)	0.83	0.94	0.88
	D	0.69	0.24	0.16	0.62 (0.41–0.81)	0.46 (0.28, 0.67)	0.35	0.70	0.47
	C	0.71	0.18	0.30	0.71 (0.48, 0.90)	0.65 (0.41–0.85)	0.50	0.60	0.55
Test	S	0.87	0.46	0.002	0.81 (0.68–0.92)	0.85 (0.73–0.93)	0.77	0.74	0.76
	D	1.33	0.28	0.06	0.68 (0.51–0.83)	0.82 (0.70, 0.92)	0.86	0.66	0.75
	C	0.73	0.59	<0.0001	0.77 (0.62–0.89)	0.80 (0.67–0.90)	0.62	0.62	0.62

S = Surface zone, D = Deep zone, C = Calcified zone, SC = Spearman's correlation, Prec. = Precision, CI = Confidence interval.

Table II Performance of trained linear (ridge) and logistic regression models. Confidence intervals for 95% are given in parentheses. Statistical variables for linear regression are on the left side of the table and variables for logistic regression are on the right side

using the threshold of 0.5 for the LR model (recall: 0.74 → 0.83, F1 score: 0.76 → 0.79).

Software prototype

We implemented the developed automatic 3D grading method in an open-source software package for Windows OS (Supplementary Video). Currently, the models trained using a python script are exported into an intermediate format and loaded by the software to predict the degeneration of unseen samples. Additional features of the software are manual tools for artefact cropping and also the advanced visualization pipeline. The source code of the software is available on GitHub: https://github.com/MIPT-Oulu/3DHistoGrading.

Discussion

In this paper, we investigated the feasibility of automation of the 3D μCT grading system for osteochondral human samples. We developed a method based on ML to predict the grades of degeneration for AC surface, deep and calcified cartilage zones in an automatic manner. The trained models were evaluated in two settings – via cross-validation and on a completely independent dataset. This allowed the assessment of generalization of the developed method to the unseen data, as well as its robustness and applicability to new data acquisition settings.

From the experiments, we found that our models are more suited for the detection of the presence of overall degeneration in the analyzed VOI, instead of fine-grained grading. This is probably due to the limited number of training samples. The results showed that the surface degeneration can be detected reliably (AUC of 0.92 and AP of 0.89). Detection of CZ disorganization yielded moderate performance (AUC of 0.71 and AP of 0.65). The lowest performance was seen for the DZ (AUC of 0.62 and AP of 0.46). It could be that the subtle changes in deep ECM organization are too difficult to distinguish from the PTA contrast.

On the other hand, the results are highly generalizable to different data acquisition settings as shown in our experiments. On the test set, an AP increase of 0.15 was observed for the CZ model and 0.36 for the DZ model. The findings suggest that besides the SZ, the precise segmentation pipeline could allow grading the CZ. To further increase the reliability of the presented models, novel data augmentation and semi-supervised grading techniques, e.g., domain adaptation[44,45], could be utilized in the future.

To assess the interpretability of the trained models, we used the SHAP method (Supplementary Figs. 6 and 7). We observed that the SZ and CZ models (that had superior performance to the DZ model) benefitted highly from the two center features. Especially, a higher degeneration was predicted for high negative center values, i.e. in the case of a large proportion of dark areas on the image. This suggests that these zones have prevalent changes in the tissue 3D-structure such as fissures and vascular infiltration. Also, even though the non-uniform features and some uniform features (U-3 and U-4) were impactful, they likely provide only supplementary information and perform poorly without the Center feature.

The replication experiment was conducted to study the effect of switching the training and evaluation datasets. The lower performance when training the models on the test set could be explained by two phenomena: First, compared to the test set, the cross-validation set had a higher number of patients and more heterogeneous grade distribution (Table I, Supplementary Fig. 2). Second, the μCT imaging parameters were optimized for Ø = 2 mm. We analyzed this both visually and quantitatively, comparing the images with the filtered data (Supplementary Figs. 4 and 5). For the test set, MSE against the filtered data was higher (mean MSE = 29.6) compared to the cross-validation set (mean MSE = 5.8). Both peak signal-to-noise ratio and structural similarity index were higher in the cross-validation set (mean values 40.2 and 0.84 compared to 33.3 and 0.71). All three metrics suggest higher image quality in the cross-validation set.

The concept of generalization is crucial in ML. If the training process is not designed carefully, the models can overfit[46] to the training data, memorizing the individual samples instead of learning broad and meaningful associations. In such cases, the method works seemingly well on the training and validation sets and fails to generalize to new samples outside the training process. The risk of overfitting is elevated with a low number of training samples. To facilitate the generalization of our method, we used multiple techniques: MRELBP histogram normalization, PCA-based dimensionality reduction (lower number of features are more unlikely to result in overfitting), L2 regularization, as well as choosing simple models (linear and logistic regression) that do not overfit as easily as more complex ones (such as CNN, random forest or support vector machine). We also used nested LOO, where a Bayesian hyperparameter search was performed at each iteration of cross-validation. Finally, we used an independent test set to prove the generalization of the method to samples with lower image quality and different grade distribution.

Besides the robust validation scheme, we also tackled the issue of a thorough evaluation of the results. When making a binary classification, ROC curves are often reported[47]. They are easily

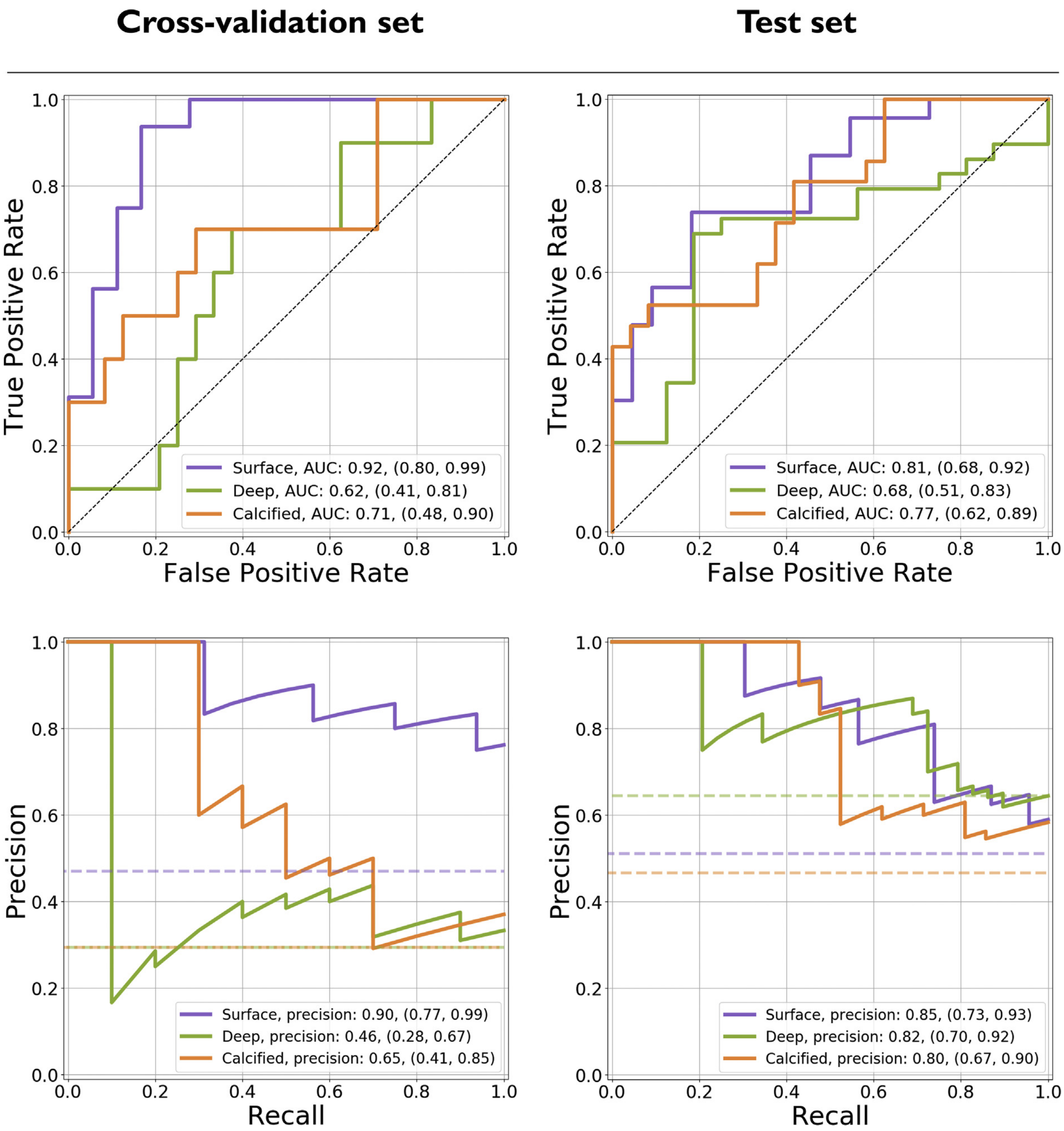

Fig. 5 Receiver operating characteristic (ROC) and precision–recall curves (PRC) for each dataset. Values for bootstrapped AUCs and APs with 95% confidence intervals are shown. From both curves, it can be clearly seen that especially the surface models are performing well compared to the baseline.

understood and allow assess performance well on evenly distributed datasets. However, the PRCs are more descriptive on imbalanced datasets and provide information on the positive predictive value of the models[48,49]. The use of the ROC curve analysis can even lead to false conclusions on classifier reliability when using imbalanced data due to wrong interpretations of the true positive rate[48]. We consider the use of a different metric for classification models to be one of the core strengths of this study.

Our group has previously utilized a novel method for quantitative surface morphology assessment. Similarly to the handcrafted surface features presented by *Ylitalo et al.*[23], our ML approach here showed the highest sensitivity for SZ for detecting intact samples. This highlights the importance of surface features, although the presented ML method can provide a comprehensive description of pathological changes of other cartilage zones as well. These studies are not otherwise directly comparable either since a different split (grades 0–1 against 2–3, instead of 0 against $\geq$ 1) was used here to better balance the grade distributions of the different groups (class distribution in *Ylitalo et al.*[23] was seven against 29 for the surface). Further, in the current study, we conducted a more thorough validation with nested LOO, PRC and interpretability analysis, as well as independent testing.

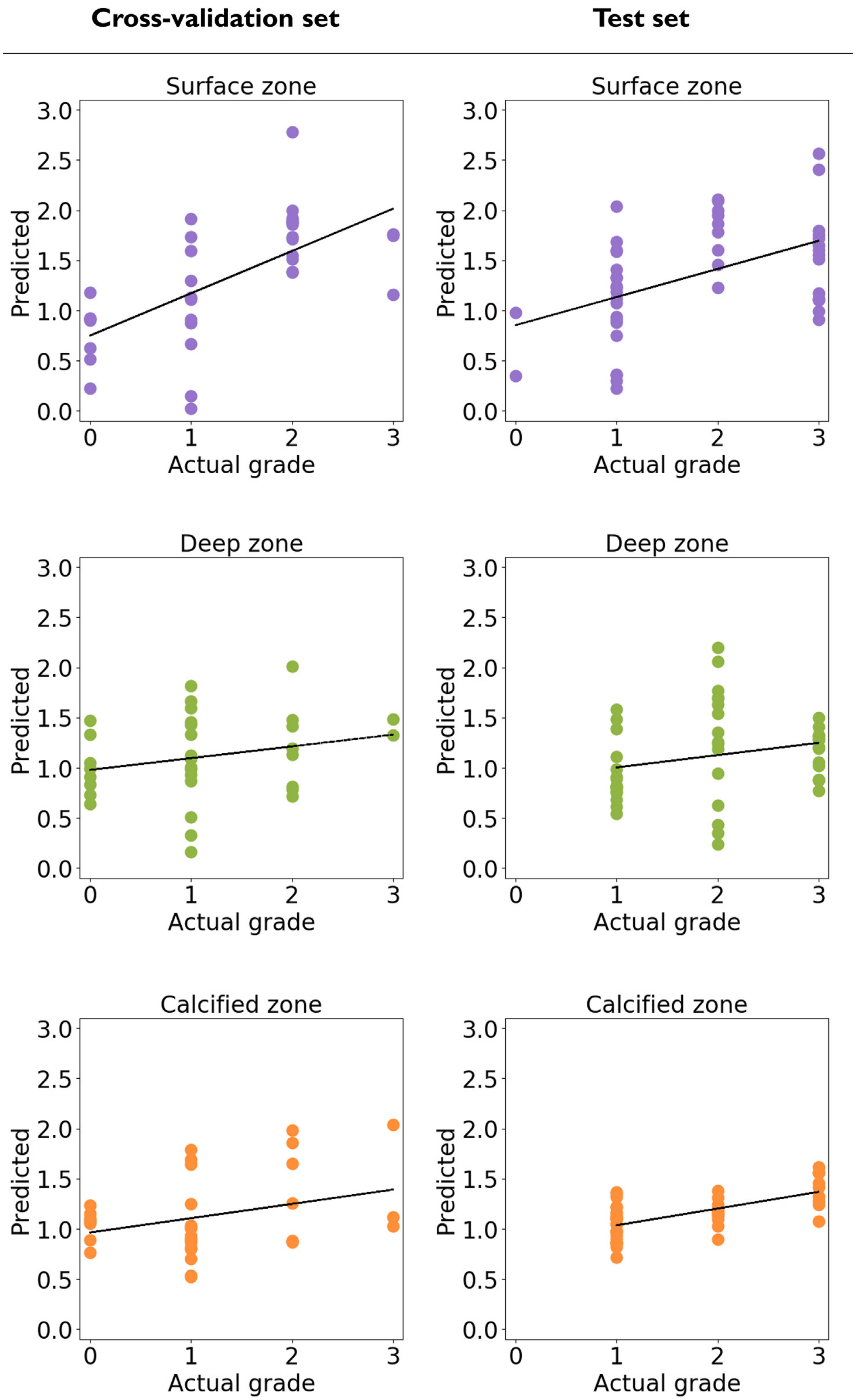

Fig. 6 Predictions obtained from the Ridge regression models on the cross-validation (left column) and test sets (right column). Predictions in most models are very close to grade 1, showing that ridge regression has little power to distinguish individual grades in this case. On the cross-validation set, predictions for SZ and CZ as well as for test set SZ, low and high grades can be visually separated from each other.

This study has several important limitations. First and foremost, a very reliable and accurate model might require hundreds or thousands of samples from different subjects, and the current model was created based only on 34 samples from 19 TKA patients. Secondly, we had to include one freeze–thaw cycle for the samples due to practical reasons. Thirdly, datasets used in the study were very heterogeneous due to different core diameters, causing decreased image quality in the test set. Fourthly, the distribution of μCT grades was also different in the test set, which could be due to lower patient count or the lack of multiple graders. Fifthly, DZ model performance might increase if a smaller depth of cartilage was used (e.g., 30–40% instead of 60% of cartilage depth[50]), better avoiding inclusion of the transitional zone. Finally, we would like to point out, that the chosen methods assume sample independency. While our samples are not independent, we note that their visual appearance may vary drastically even when the osteochondral plugs are extracted from similar anatomical locations. Therefore, we believe that the impact of not strictly fulfilling the assumption of independent and identically distributed data is very minor, and does not affect our results.

As a conclusion, this study shows that automatic 3D histopathological grading of osteochondral samples is feasible from CEμCT with minimal user input. Our model could potentially be used to provide a second opinion for OA researchers requiring a reliable assessment of OA *ex-vivo* severity, mainly at SZ and CZ. However, the method and the software are not fully ready for practical use and further development, including the acquisition of a bigger training dataset and external validation, is highly recommended. This would likely increase the reliability of the analysis also for zones other than the cartilage surface. To the best of our knowledge, this is the first report presenting an ML-based 3D histopathologic OA grading model, which also adequately generalizes to unseen data. All codes used, and the software prototype developed during this study are available on the project's GitHub page (https://github.com/MIPT-Oulu/3DHistoGrading).

Author Contributions

Conception and design: SJOR, AT, MAJF, SSK, KPHP HJN, SS.

Data analysis, development of the pipeline and the software prototype: SJOR, TF, AT.

Data acquisition: SJOR, MAJF, SSK, JL, MV, PL, AJ, HK.

Drafting the manuscript: SJOR, AT.

Critical revision for important intellectual content and approval of the manuscript: all authors.

Conflict of interest

HJN has received Academy of Finland grant, has several patent publications (University of Oulu, University of Helsinki, Philips Healthcare, Photono Oy, SWAN Cytologics, Revenio), and also receives royalties from them. SS has received grants from European Research Council, Academy of Finland and Sigrid Juselius Foundation. AT was supported by KAUTE foundation.

Other authors report no conflicts of interest.

Funding sources

Funding sources are not associated with the scientific contents of the study.

Acknowledgements

The financial support of the Academy of Finland (grants no. 268378, and 303786); Sigrid Juselius Foundation; European Research Council under the European Union's Seventh Framework Programme (FP/2007–2013) / ERC Grant Agreement no. 336267; Kaute Foundation; and the strategic funding of University of Oulu are acknowledged.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.joca.2020.05.002.

References

1. Pritzker KP, Gay S, Jimenez SA, Ostergaard K, Pelletier JP, Revell PA, *et al.* Osteoarthritis cartilage histopathology: grading and staging. Osteoarthritis Cartilage 2006;14(1): 13–29, https://doi.org/10.1016/j.joca.2005.07.014.
2. Mankin HJ, Dorfman H, Lippiello Zarins A. Biochemical and metabolic abnormalities in articular cartilage from osteoarthritic human hips. II. Correlation of morphology with biochemical and metabolic data. J Bone Joint Surg Am 1971 Apr;53(3):523–37.
3. Rutgers M, van Pelt MJP, Dhert WJA, Creemers LB, Saris DBF. Evaluation of histological scoring systems for tissue-engineered, repaired and osteoarthritic cartilage. Osteoarthritis Cartilage 2010;18(1):12–23, https://doi.org/10.1016/j.joca.2009.08.009.
4. Pauli C, Whiteside R, Heras FL, Nesic D, Koziol J, Grogan SP, *et al.* Comparison of cartilage histopathology assessment systems on human knee joints at all stages of osteoarthritis development. Osteoarthritis Cartilage 2012;20(6):476–85, https://doi.org/10.1016/j.joca.2011.12.018.
5. Custers RJH, Creemers LB, Verbout AJ, van Rijen MHP, Dhert WJA, Saris DBF. Reliability, reproducibility and variability of the traditional histologic/histochemical grading system vs the new OARSI osteoarthritis cartilage histopathology assessment system. Osteoarthritis Cartilage 2007;15(11): 1241–8, https://doi.org/10.1016/j.joca.2007.04.017.
6. Waldstein W, Perino G, Gilbert SL, Maher SA, Windhager R, Boettner F. OARSI osteoarthritis cartilage histopathology assessment system: a biomechanical evaluation in the human knee. J Orthop Res 2016;34(1):135–40, https://doi.org/10.1002/jor.23010.
7. Pollard TCB, Gwilym SE, Carr AJ. The assessment of early osteoarthritis. Bone Joint Lett J 2008;90-B(4):411–21, https://doi.org/10.1302/0301-620X.90B4.20284.
8. Mobasheri A, Henrotin Y. Biomarkers of (osteo)arthritis. Biomarkers 2015;20(8):513–8, https://doi.org/10.3109/1354750X.2016.1140930.
9. Chu CR, Williams AA, Coyle CH, Bowers ME. Early diagnosis to enable early treatment of pre-osteoarthritis. Arthritis Res Ther 2012;14(3):212, https://doi.org/10.1186/ar3845.
10. Song Y, Treanor D, Bulpitt A, Magee D. 3D reconstruction of multiple stained histology images. J Pathol Inf 2013;4(2):7, https://doi.org/10.4103/2153-3539.109864.
11. Alic L, Haeck JC, Bol K, Klein S, van Tiel ST, Wielepolski PA, *et al.* Facilitating tumor functional assessment by spatially relating 3D tumor histology and in vivo MRI: image registration approach. PloS One 2011;6(8), e22835, https://doi.org/10.1371/journal.pone.0022835.
12. Dou T, Zhou W. 2D and 3D convolutional neural network fusion for predicting the histological grade of hepatocellular carcinoma. In: 24th International Conference on Pattern Recognition (ICPR). Beijing, China: IEEE; 2018, https://doi.org/10.1109/ICPR.2018.8545806.
13. Fetit AE, Novak J, Peet AC, Arvanitis TN. 3D texture analysis of MR images to improve classification of paediatric brain

tumours: a preliminary study. Stud Health Technol Inf 2014, https://doi.org/10.3233/978-1-61499-423-7-213.

14. Liu Y, Zhang Y, Cheng R, Liu S, Qu F, Yin X, *et al.* Radiomics analysis of apparent diffusion coefficient in cervical cancer: a preliminary tudy on histological grade evaluation. J Magn Reson Imag 2019;49(1):280–90, https://doi.org/10.1002/jmri.26192.
15. Ashinsky BG, Bouhrara M, Coletta CE, Lehallier B, Urish KL, Lin P, *et al.* Predicting early symptomatic osteoarthritis in the human knee using machine learning classification of magnetic resonance images from the osteoarthritis initiative. J Orthop Res 2017;35(10):2243–50, https://doi.org/10.1002/jor.23519.
16. Lee J, Shih Y, Wei M, Sun C, Chiang B. Classification of established atopic dermatitis in children with the in vivo imaging methods. J Biophot 2018, e201800148, https://doi.org/10.1002/jbio.201800148. 0(0).
17. Schone M, Mannicke N, Somerson JS, Marquass B, Henkelmann R, Mochida J, *et al.* 3D ultrasound biomicroscopy for assessment of cartilage repair tissue: volumetric characterisation and correlation to established classification systems. Eur Cell Mater 2016 Feb 8;31:119–35, https://doi.org/10.22203/eCM.v031a09.
18. Peng Z, Wang M. Three dimensional surface characterization of human cartilages at a micron and nanometre scale. Wear 2013;301(1):210–7, https://doi.org/10.1016/j.wear.2012.11.056.
19. Kerckhofs G, Sainz J, Maréchal M, Wevers M, Van de Putte T, Geris L, *et al.* Contrast-enhanced nanofocus X-ray computed tomography allows virtual three-dimensional histopathology and morphometric analysis of osteoarthritis in small animal models. Cartilage 2014;5(1):55–65, https://doi.org/10.1177/1947603513501175.
20. Nieminen HJ, Ylitalo T, Karhula S, Suuronen J, Kauppinen S, Serimaa R, *et al.* Determining collagen distribution in articular cartilage using contrast-enhanced micro-computed tomography. Osteoarthritis Cartilage 2015;23(9):1613–21, https://doi.org/10.1016/j.joca.2015.05.004.
21. Karhula SS, Finnilä MA, Lammi MJ, Ylärinne JH, Kauppinen S, Rieppo L, *et al.* Effects of articular cartilage constituents on phosphotungstic acid enhanced micro-computed tomography. PloS One 2017;12(1), e0171075, https://doi.org/10.1371/journal.pone.0171075.
22. Nieminen HJ, Gahunia HK, Pritzker KPH, Ylitalo T, Rieppo L, Karhula SS, *et al.* 3D histopathological grading of osteochondral tissue using contrast-enhanced micro-coputed tomography. Osteoarthritis Cartilage 2017;25(10):1680–9, https://doi.org/10.1016/j.joca.2017.05.021.
23. Ylitalo T, Finnilä MAJ, Gahunia HK, Karhula SS, Suhonen H, Valkealahti M, *et al.* Quantifying complex micro-topography of degenerated articular cartilage surface by contrast-enhanced micro-computed tomography and parametric analyses. J Orthop Res 2019, https://doi.org/10.1002/jor.24245. 0.
24. Maerz T, Newton MD, Matthew HWT, Baker KC. Surface roughness and thickness analysis of contrast-enhanced articular cartilage using mesh parameterization. Osteoarthritis Cartilage 2016;24(2):290–8, https://doi.org/10.1016/j.joca.2015.09.006.
25. Kauppinen S, Karhula SS, Thevenot J, Ylitalo T, Rieppo L, Kestilä I, *et al.* 3D morphometric analysis of calcified cartilage properties using micro-computed tomography. Osteoarthritis Cartilage 2019;27(1):172–80, https://doi.org/10.1016/j.joca.2018.09.009.
26. Nagarajan MB, Coan P, Huber MB, Diemoz PC, Glaser C, Wismüller A. Computer-aided diagnosis in phase contrast imaging x-ray computed tomography for quantitative characterization of ex vivo human patellar cartilage. IEEE Trans Biomed Eng 2013;60(10):2896–903, https://doi.org/10.1109/TBME.2013.2266325.
27. Nagarajan MB, Coan P, Huber MB, Diemoz PC, Glaser C, Wismüller A. Computer-Aided diagnosis for phase-contrast X-ray computed tomography: quantitative characterization of human patellar cartilage with high-dimensional geometric features. J Digit Imag 2014;27(1):98–107, https://doi.org/10.1007/s10278-013-9634-3.
28. Litjens G, Sánchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, *et al.* Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep 2016;6(1):26286, https://doi.org/10.1038/srep26286.
29. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, *et al.* Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542(7639):115–8, https://doi.org/10.1038/nature21056.
30. Tiulpin A, Thevenot J, Rahtu E, Lehenkari P, Saarakkala S. Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. Sci Rep 2018;8(1):1727, https://doi.org/10.1038/s41598-018-20132-7.
31. Xue Y, Zhang R, Deng Y, Chen K, Jiang T. A preliminary examination of the diagnostic value of deep learning in hip osteoarthritis. PloS One 2017;12(6), https://doi.org/10.1371/journal.pone.0178992.
32. Madelin G, Poidevin F, Makrymallis A, Regatte RR. Classification of sodium MRI data of cartilage using machine learning. Magn Reson Med 2015;74(5):1435–48, https://doi.org/10.1002/mrm.25515.
33. Tiulpin A, Saarakkala S. Automatic Grading of Individual Knee Osteoarthritis Features in Plain Radiographs Using Deep Convolutional Neural Networks 2019. arXiv 2019 doi.org/1907.vol.8020.
34. Tiulpin A, Klein S, Bierma-Zeinstra SMA, Thevenot J, Rahtu E, van Meurs J, *et al.* Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data. Sci Rep 2019;9(1):20038, https://doi.org/10.1038/s41598-019-56527-3.
35. Pedoia V, Haefeli J, Morioka K, Teng H-, Nardo L, Souza RB, *et al.* MRI and biomechanics multidimensional data analysis reveals R2-R1ρ as an early predictor of cartilage lesion progression in knee osteoarthritis. J Magn Reson Imag 2018;47(1):78–90, https://doi.org/10.1002/jmri.25750.
36. Swan AL, Stekel DJ, Hodgman C, Allaway D, Alqahtani MH, Mobasheri A, *et al.* A machine learning heuristic to identify biologically relevant and minimal biomarker panels from omics data. BMC Genom 2015;16(1), https://doi.org/10.1186/1471-2164-16-S1-S2.
37. Ashinsky BG, Coletta CE, Bouhrara M, Lukas VA, Boyle JM, Reiter DA, *et al.* Machine learning classification of OARSI-scored human articular cartilage using magnetic resonance imaging. Osteoarthritis Cartilage 2015;23(10):1704–12, https://doi.org/10.1016/j.joca.2015.05.028.
38. Tiulpin A, Finnilä M, Lehenkari P, Nieminen HJ, Saarakkala S. Deep-learning for tidemark segmentation in human osteochondral tissues imaged with micro-computed tomography. Adv Concepts Intell Vis Syst 2020:131–8, https://doi.org/10.1007/978-3-030-40605-9_12.
39. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. Lect Notes Comput Sci 2015:9351, https://doi.org/10.1007/978-3-319-24574-4_28.
40. Deng J, Dong W, Socher R, Li L, Li Kai, ImageNet Li Fei-Fei. A large-scale hierarchical image database. IEEE Comput Soc Conf Comput Vis Pattern Recogn 2009:248–55, https://doi.org/10.1109/CVPR.2009.5206848.

41. Liu Li, Lao Songyang, Fieguth PW, Guo Yulan, Wang Xiaogang, Pietikainen M. Median robust extended local binary pattern for texture classification. TIP 2016 Mar;25(3):1368–81, https://doi.org/10.1109/TIP.2016.2522378.
42. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, *et al.* Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nature Biomed Eng 2018;2(10): 749–60, https://doi.org/10.1038/s41551-018-0304-0.
43. Bergstra J, Komer B, Eliasmith C, Yamins D, Cox DD. Hyperopt: a Python library for model selection and hyperparameter optimization. Comput Sci Discov 2015;8(1), 014008, https://doi.org/10.1088/1749-4699/8/1/014008.
44. Duan L, Xu D, Tsang IW. Learning with augmented features for heterogeneous domain adaptation. In: Proceedings of the 29th International Conference on Machine Learning (ICML); Edinburgh, Scotland 2012.
45. Li W, Duan L, Xu D, Tsang IW. Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. IEEE Trans Pattern Anal Mach Intell 2014;36(6): 1134–48, https://doi.org/10.1109/TPAMI.2013.167.
46. Ivanescu AE, Li P, George B, Brown AW, Keith SW, Raju D, *et al.* The importance of prediction model validation and assessment in obesity and nutrition research. Int J Obes 2016;40(6): 887–94, https://doi.org/10.1038/ijo.2015.214.
47. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn 1997;30(7):1145–59, https://doi.org/10.1016/S0031-3203(96)00142-2.
48. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS One 2015;10(3), e0118432, https://doi.org/10.1371/journal.pone.0118432.
49. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning (ICML). New York, NY, USA: ACM; 2006, https://doi.org/10.1145/1143844.1143874.
50. Sophia Fox AJ, Bedi A, Rodeo SA. The basic science of articular cartilage: structure, composition, and function. Sport Health 2009;1(6):461–8, https://doi.org/10.1177/1941738109350438.