JYX JYVÄSKYLÄN YLIOPISTO UNIVERSITY OF JYVÄSKYLÄ

This is a self-archived version of an original article. This version may differ from the original in pagination and typographic details.

Author(s): Patron, Anri; Annala, Leevi; Lainiala, Olli; Paloneva, Juha; Äyrämö,Sami

Title: An Automatic Method for Assessing Spiking of Tibial Tubercles Associated with Knee Osteoarthritis

Year: 2022

Version: Published version

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland

Rights: CC BY 4.0

Rights url: https://creativecommons.org/licenses/by/4.0/

Please cite the original version:

Patron, A., Annala, L., Lainiala, O., Paloneva, J., & Äyrämö, S. (2022). An Automatic Method for Assessing Spiking of Tibial Tubercles Associated with Knee Osteoarthritis. Diagnostics, 12(11), Article 2603. https://doi.org/10.3390/diagnostics12112603

diagnostics

MDPI

Article

An Automatic Method for Assessing Spiking of Tibial Tubercles Associated with Knee Osteoarthritis

Anri Patron [1,*], Leevi Annala [1], Olli Lainiala [2,3], Juha Paloneva [4,5] and Sami Äyrämö [1]

[1] Faculty of Information Technology, University of Jyväskylä, 40014 Jyväskylä, Finland
[2] Department of Radiology, Tampere University Hospital, 33520 Tampere, Finland
[3] Faculty of Medicine and Health Technologies, Tampere University, 33520 Tampere, Finland
[4] Department of Surgery, Central Finland Healthcare District, 40620 Jyväskylä, Finland
[5] Institute of Clinical Medicine, University of Eastern Finland, 70210 Kuopio, Finland
* Correspondence: anri.a.patron@jyu.fi

Abstract: Efficient and scalable early diagnostic methods for knee osteoarthritis are desired due to the disease's prevalence. The current automatic methods for detecting osteoarthritis using plain radiographs struggle to identify the subjects with early-stage disease. Tibial spiking has been hypothesized as a feature of early knee osteoarthritis. Previous research has demonstrated an association between knee osteoarthritis and tibial spiking, but the connection to the early-stage disease has not been investigated. We study tibial spiking as a feature of early knee osteoarthritis. Additionally, we develop a deep learning based model for detecting tibial spiking from plain radiographs. We collected and graded 913 knee radiographs for tibial spiking. We conducted two experiments: experiments A and B. In experiment A, we compared the subjects with and without tibial spiking using Mann-Whitney U-test. Experiment B consisted of developing and validating an interpretative deep learning based method for predicting tibial spiking. The subjects with tibial spiking had more severe Kellgren-Lawrence grade, medial joint space narrowing, and osteophyte score in the lateral tibial compartment. The developed method achieved an accuracy of 0.869. We find tibial spiking a promising feature in knee osteoarthritis diagnosis. Furthermore, the detection can be automatized.

Keywords: knee joint; osteoarthritis; radiography; tibial spiking; convolutional neural networks

check for updates

Citation: Patron, A.; Annala, L.; Lainiala, O.; Paloneva, J.; Äyrämö, S. An Automatic Method for Assessing Spiking of Tibial Tubercles Associated with Knee Osteoarthritis. *Diagnostics* **2022**, *12*, 2603. https://doi.org/10.3390/diagnostics12112603

Academic Editor: Ming-Huwi Horng

Received: 5 October 2022
Accepted: 24 October 2022
Published: 27 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Knee osteoarthritis (OA) is a highly prevalent chronic joint disease and a prominent global cause of disability. In the Global Burden of Disease 2010 study, knee and hip OA was ranked the 11th most common global cause of disability [1]. As the prevalence of OA increases with age [2], due to population aging, the burden of OA is expected to rise. Early detection of OA is imperative for maximizing the efficacy of interventions, lowering the burden of the disease and the incidence of knee joint replacement surgery [3,4].

Plain radiography is a standard imaging modality for OA diagnosis and severity assessment. Radiographic signs of knee OA include joint space narrowing (JSN), formation of osteophytes, cysts, and subchondral sclerosis [5]. Plain radiography as an imaging modality is insensitive to early signs of knee OA, such as cartilage damage and minor osteophytes [6,7], which makes the radiographic diagnosis of early knee OA challenging.

The most common classification for radiological knee OA was described by Kellgren and Lawrence (KL) [8]. KL classification consists of ordinal grades from 0 to 4, where 0 stands for no signs of OA, and each subsequent grade signifies increasing OA severity. KL is a composite grading system, defining OA by the presence of JSN and osteophytes. The definitions of KL-grades for the knee joint are the following [8,9]:

- Grade 0: No radiological signs of OA.
- Grade 1: Doubtful JSN, possible osteophytic lipping.
- Grade 2: Definite osteophytes, possible JSN.

- Grade 3: Moderate multiple osteophytes, definite JSN, some sclerosis, possible deformity of bone ends.
- Grade 4: Large osteophytes, marked JSN, severe sclerosis, definite deformity of bone ends.

The KL system has been criticized for ambiguity, e.g., in cases where JSN is present, but osteophytes are not [10]. Osteoarthritis Research Society International (OARSI) developed another radiographic atlas for OA [11]. Unlike KL, OARSI contains individual grades for JSN and osteophytes on a scale of 0–3.

To increase the reliability and objectivity of OA assessment, fully automatic methods have been developed for assessing radiographic knee OA using plain radiographs. Oka et al. [12] developed an automatic method for quantifying features of OA, such as minimum joint space width and osteophyte area using filters and differentiation. Shamir et al. [13] used various handcrafted features extracted from the plain radiographs together with a weighted nearest neighbor classifier to predict the KL-grade. More recently, convolutional neural networks (CNN) [14] have achieved success in medical image classification tasks such as malignant skin lesion classification [15] or radiographic identifying of subjects with arthroplasty [16].

Using CNNs for knee OA severity assessment from plain radiographs was initially proposed by Antony et al. [17]. The recent state-of-the-art for predicting knee OA severity using deep learning has been reviewed by Yeoh et al. [18]. CNN-based methods have been reasonably successful in assessing severe KL-grades (i.e., grades 3 and 4), but for predicting grades marking early OA, the accuracy is notably lower [17,19,20]. The limitations of automation of early OA severity assessment using plain radiographs are likely multifaceted. Firstly radiographs do not allow for direct visualization of the cartilage. Furthermore, the current CNN-based methods are constrained by the KL system. As the current methods are trained using noisy KL scores as the ground truth, the resulting models thus capture the bias inherent to the KL classification. The results by Kim et al. [21] indicate that early OA severity assessment with CNN can be improved by providing the model with additional clinical information (e.g., age, sex, and body mass index (BMI)).

The limitations of early OA assessment could perhaps be alleviated further by considering additional radiographic features of OA not incorporated in the KL system, such as the spiking of tibial tubercles that has been hypothesized as a sign of early knee OA (hypothesis A). The spiking of tibial tubercles or tibial spiking refers to the tall and angular appearance of the tibial spines (see Figure 1). The first mention of hypothesis A, to our knowledge, is from a radiological textbook by Sutton [22]. However, the author provides no evidence for the said hypothesis.

The feature was later studied by Reiff et al. [23], who examined radiographs from fifty-five subjects with established knee OA and thirty-six controls. They found the lengthening and sharpening of the peaks of the tubercles associated with knee OA. Donnelly et al. [24] conducted a study with 950 subjects examining tibial spiking as a radiological feature of OA. They found the sharpening of tibial spines to correlate with OA status (defined by KL 2 or higher) and osteophyte scores. They, however, concluded that tibial spiking is not a reliable marker for knee OA in isolation due to a lack of clear independent association with knee pain [24].

Unluet al. [25] studied the association between tibial spiking and cartilage defects assessed via magnetic resonance imaging (MRI). The study involved seventy-six knees from forty-seven subjects and thirty-one knees from sixteen controls. The subjects with knee OA had significantly higher and sharper tibial spines than the controls. They observed a correlation between cartilage defects and medial tubercle height but not with lateral tubercle height [25]. Additionally, an association between the spiking of lateral tubercle and osteophyte formation in the tibial compartments was found [25]. The latest study by Hayeri et al. [26] considered tibial spiking from a paleopathological framework, where thirty-five tibial bone specimens were directly examined for signs of OA. The study found spiking of the lateral spine associated with osteophyte formation.

Figure 1. Two knee radiographs rated for spiking of tibial tubercles are compared. The tubercles rated as spiking are indicated by arrowheads.

The previous research indicates that tibial spiking might be associated with knee OA and osteophyte formation [23–26], however, the feature might not be a reliable marker of knee OA in isolation [24]. Although osteophyte formation is one of the primary signs of radiographic OA, therefore tibial spiking might be a beneficial feature in assessing knee OA in cases where no evident osteophytes can be detected. With automatic methods, the cost of assessing radiographs for markers of OA is negligible and therefore provides a different value proposition compared to general clinical adoption. Provided that tibial spiking is identifiable by human experts, an automatic method can be developed, provided sufficient data (hypothesis B).

The aim of the present study was to evaluate the hypothesis on tibial spiking as an early sign of knee OA (hypothesis A). While the previous work on the subject indicates that tibial spiking might be a feature of knee OA [23–26], the research is still lacking. Especially whether tibial spiking is connected with the early knee OA is unclear. Furthermore, we examined the feasibility of identifying tibial spiking automatically by developing a method for assessing the feature from plain radiographs (hypothesis B), which has not been considered previously.

2. Materials and Methods

2.1. Radiographic Data

The present study utilized data from the Osteoarthritis Initiative (OAI) [27] and the Multicenter Osteoarthritis Study (MOST) [28]. OAI and MOST are longitudinal cohort studies of OA and include radiographs assessed for signs of OA at multiple time points. OAI dataset includes data from 4607 participants between ages 45–79 at baseline. MOST baseline dataset contains data from 3026 participants between ages 50–79.

We collected bilateral PA (posterior-anterior) fixed flexion knee radiographs with KL-grades 0–2 from OAI and MOST baseline datasets. The radiographs were selected randomly with an approximately equal number of samples from each KL-grade. We collected 722 radiographs from OAI and 191 radiographs from MOST, from which we used only the right knee. The collected knees with regard to the KL-grade can be seen in Table 1. Additionally, we collected assessments of radiographic knee OA, including KL-score, OARSI osteophyte and JSN scores, Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) knee pain score, and the subject BMI for each sample.

Table 1. Collected knees with regard to KL-grade.

	KL0	KL1	KL2
OAI	243	248	231
MOST	65	61	65

2.2. Data for Experiment A

As the tibial spiking was not assessed in OAI or MOST cohorts, we collected the assessments manually for all 913 knees. The assessments were performed by two physicians, a radiology resident with three years of experience in radiology (expert 1) and an experienced orthopedic surgeon (expert 2). Each knee was assessed by a single expert, blinded to the OA assessments and clinical details. Prior to grading the knees, the experts had a single session to establish a uniform view of the spiking criteria. Each spine (medial and lateral) was rated for spiking by subjective visual inspection of angulation, size, and other deformities. No angle or height measurements of the tubercles were performed in this study. The spines were graded on a binary scale of 0–1 with the possibility of giving an "unsure" rating. The unsure rating was warranted as the tibial spines might be occluded by the femur or otherwise difficult to judge, e.g., due to poor exposure. The spines were also rated unsure in borderline cases (i.e., doubtful spiking). We defined overall spiking (i.e., spiking on medial or lateral side) as

$$\text{spiking} := (\text{lateral spiking} = 1) \vee (\text{medial spiking} = 1). \tag{1}$$

2.3. Reliability

We evaluated inter-rater reliability using a subset of 205 radiographs rated by both experts in separate sessions, blinded to the assessments made by the other party. Intra-rater reliability was assessed with a subset of the knees re-rated by the same expert, blinded to the previous rating. The duplicate radiographs used for evaluating reliability were mixed among the set of regular radiographs. Additionally, the experts were blinded to the existence of duplicate radiographs. Sample sizes for evaluating intra-rater reliability were 53 and 68 for experts 1 and 2, respectively. The inter- and intra-rater reliability were measured using Cohen's κ (Kappa) [29]. For calculating κ scores we used implementation in Python (ver. 3.10.0) [30] library scikit-learn 1.0.1 [31].

The reliability analysis was performed for the original 3-way ratings (including the unsure ratings), binary (0–1) ratings where the unsure assessments were omitted, and overall spiking (defined in Equation (1)). The rating pairs were omitted if either contained an unsure rating. The sample sizes for assessing the intra-rater reliability of unsure omitted ratings were 41 and 42 for experts 1 and 2, respectively, and 124 for inter-rater reliability.

2.4. Experiment A

To evaluate the hypothesis on tibial spiking as a feature of early OA (hypothesis A), we conducted experiment A, where the differences were tested between groups with tibial spiking and a control group (i.e., subjects without tibial spiking). Given hypothesis A, we would expect the group with tibial spiking to have a higher KL-score, osteophyte-score, JSN-score, and knee pain. We combined the samples from OAI (722) and MOST (191), for 913 samples in experiment A.

We defined the inclusion criteria for the spiking group identically to overall spiking; see Equation (1). Consequently, the control group included all the samples with negative or unsure ratings. In total, 630 samples were assigned to the spiking group and 283 to the control group. For testing the group differences, we used two-tailed Mann-Whitney U-test [32]. We used the Scipy 1.7.3 [33] implementation for calculating the U-test values. To counteract the multiple comparisons problem, we applied the Bonferroni correction [34]. With significance levels $\alpha = (0.05, 0.01, 0.001)$ and the number of tests $n = 9$, the corrected significance level α_i' is

$$\alpha_i' = \frac{\alpha_i}{n}. \tag{2}$$

Some radiographs in the original OAI and MOST data were missing some OARSI assessments (JSN or osteophytes grade). We omitted the samples containing missing values for the calculations, which reduces the sample size for these variables. Sample sizes for the variables containing missing information are the following: BMI: spiking 629 and control

283, OARSI JSN variables: spiking 618 and control 281, and OARSI osteophyte variables: spiking 432 and control 145.

2.5. Data for Experiment B

We used 80% of OAI data (577 images) for model training and 20% (145 images) for model validation. All 191 images from MOST were used as the final test data. For a breakdown of each dataset split with regard to the tibial spiking rating, refer to Table 2. As the ground truth, we used the definition for overall spiking (see Equation (1)). We used the assessments from a single expert chosen randomly for the images graded by both experts.

Table 2. Tibial spiking data frequency tables for medial, lateral, and overall spiking (medial or lateral; OR). Zero indicates the absence of spiking, while one indicates the presence of spiking. The question mark indicates the unsure rating.

	Medial			**Lateral**			**OR**	
	0	**1**	**?**	**0**	**1**	**?**	**0**	**1**
Train	216	281	80	253	279	45	188	389
Validation	57	71	17	53	77	15	47	98
Test	66	108	17	59	116	16	48	143
Total	339	460	114	365	472	76	283	630

The image data in the OAI and MOST datasets were stored as Digital Imaging and Communications in Medicine (DICOM) image format. We used pydicom 2.2.2 [35] for reading the DICOM pixel data. The original bilateral PA images were first localized to the region of interest (ROI), i.e., the right knee joint area. The images were localized by manually annotating a center point in the valley between medial and lateral tibial tubercles and calculating square ROI of size 300 × 300 from the center point. For annotating the ROI centers, we used labelme 5.0.1 [36]. Finally, we downsampled the ROI images to an input size of 224 × 224 by bilinear interpolation.

Following the localization, we inverted the pixel values of images with MONOCHROME1 photometric interpretation (i.e., we converted black-on-white images to white-on-black). We performed histogram equalization using OpenCV Python library [37] to improve the image contrast. All image sample pixel intensities were normalized by subtracting the mean and dividing by the standard deviation, which were calculated from the set of training samples.

We replaced the original training samples with augmented samples in a one-to-one fashion. The augmentations included flipping the images horizontally with a probability of 0.5 and performing random affine transformation (i.e., scaling, spatial translation, and rotation) to introduce variance among the training samples. The degree of rotation was sampled from a range of (–12, 12). The degree of spatial translation was sampled from (–11.2, 11.2) and (–2.24, 2.24) for the horizontal and vertical axis, respectively. The scale factor was sampled from (0.8, 1.2). The augmentations were resampled for each training iteration. For an illustration of the data processing pipeline, see Figure 2.

2.6. Experiment B

We developed and evaluated a model for identifying tibial spiking from plain radiographs to determine whether tibial spiking is detectable with automatic methods (hypothesis B). For the classification model, we fine-tuned (i.e., domain adapted) a CNN model ResNeXt-50-32x4d introduced by Xie et al. [38] pre-trained on around 1.2 million color images from ImageNet [39] challenge [40] for assessing tibial spiking. The CNN model implementation from Torchvision [41] version 0.12.0 was used. We modified the model by replacing the dense layer with two dense units, followed by softmax. We, therefore, initialized the weights for all other layers pre-trained on ImageNet. The motivation for the procedure is to improve the performance of CNN by utilizing the features learned from another dataset [42].

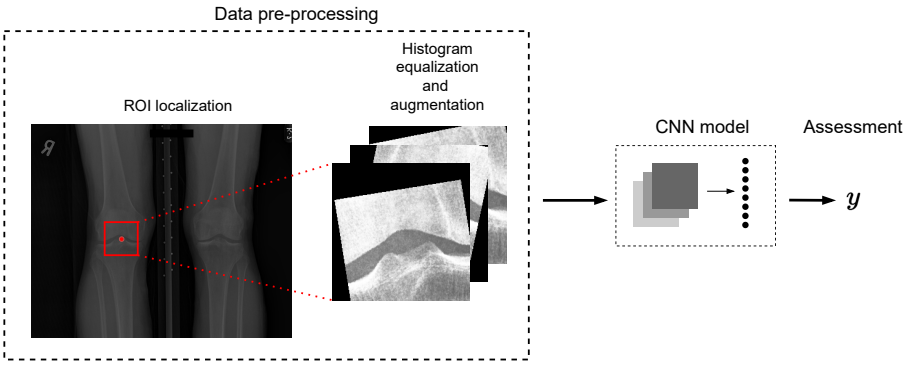


Figure 2. Tibial spiking assessment pipeline: the bilateral PA view radiographs are localized to the ROI, histogram equalization, augmentations (when applicable), and normalization are performed before feeding the data to the model.

The CNN model architecture is detailed in Table 3. The bottleneck blocks, i.e., BN1–4 in Table 3 featured a shortcut connections [38] similar to He et al. [43]. The shortcut connection for a block $\mathcal{B}$ is defined as

$$y = \mathcal{B}(x) + x, \tag{3}$$

where y is the output and x is the input. Note that the dimensions of the block $\mathcal{B}$ and the identity x must be equal in Equation (3). When this is not the case, a linear projection $W_d x$ is used to match the dimensions before adding the identity [43]. We used the standard cross-entropy as the loss function, optimized with Adam [44] with following parameters $\alpha = 0.0002$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$. PyTorch 1.11.0 [41] was used as the model training and testing framework.

Table 3. The architecture of the used model consists of sequential bottleneck (BN) blocks after the initial convolution and max pooling. The convolution layer parameters are presented in the order of kernel size, and the number of kernels and C denotes the number of grouped convolutions. A bracket followed by $\times$ k indicates the block is repeated k times. The dense layer has an input size of 2048 and an output size of two. The spatial output size of the block is presented in the middle column.

Block	Output Size	Architecture
Conv1	112×112	7×7, 64, stride 2
Pool	56×56	3×3 max pool, stride 2
BN1	56×56	$\left.\begin{matrix} 1 \times 1, 128 \\ 3 \times 3, 128, C = 32 \\ 1 \times 1, 256 \end{matrix}\right\} \times 3$
BN2	28×28	$\left.\begin{matrix} 1 \times 1, 256 \\ 3 \times 3, 256, C = 32 \\ 1 \times 1, 512 \end{matrix}\right\} \times 4$
BN3	14×14	$\left.\begin{matrix} 1 \times 1, 512 \\ 3 \times 3, 512, C = 32 \\ 1 \times 1, 1024 \end{matrix}\right\} \times 6$
BN4	7×7	$\left.\begin{matrix} 1 \times 1, 1024 \\ 3 \times 3\ 1024, C = 32 \\ 1 \times 1, 2048 \end{matrix}\right\} \times 3$
	1×1	global average pool
Dense		2048, 2, softmax

We performed a grid search to determine suitable hyperparameters for fine-tuning the classification model. As the model selection criteria, we used validation accuracy. The parameter space used in the grid search can be seen in Table 4. The best-performing model

was trained for ten epochs with a batch-size four and a learning rate (α) of 0.0002. The learning rate was decayed every four epochs by a factor of 0.115.

Table 4. Grid search parameter space. Step-size defines the interval for decaying the learning rate specified by gamma.

	Values
Epochs	1, 2, 3, . . . , 23, 24, 25
Batch-size	4, 5, 6
Learning-rate	7×10^{-3}, 8×10^{-3}, 9×10^{-3}, 10^{-4}, 2×10^{-4}, 3×10^{-4}, 4×10^{-4}
Step-size	4, 5, 6
Gamma	0.115, 0.12, 0.125

We used Gradient-weighted Class Activation Mapping (Grad-CAM) [45] implementation TorchCAM (ver. 0.3.1a0) [46] for visualizing the model predictions. Grad-CAM provides visual explainability by highlighting the regions from the input image strongly influencing the output [45]. The heatmaps generated by Grad-CAM provide transparency into the model prediction-making and enable the developers to determine how the model is able to discern the classes or fails to do so. For the users, the visual explanations enable the building of trust in the classifier system. The study workflow and the methodology are summarized in Figure 3. The analysis code and the trained model for detecting tibial spiking have been made available on GitHub: https://github.com/AI-hub-keskisuomi/AI_hub_keskisuomi/tree/main/WP3_knee_osteoarthritis/tibial_spiking_grading (accessed on 4 October 2022).

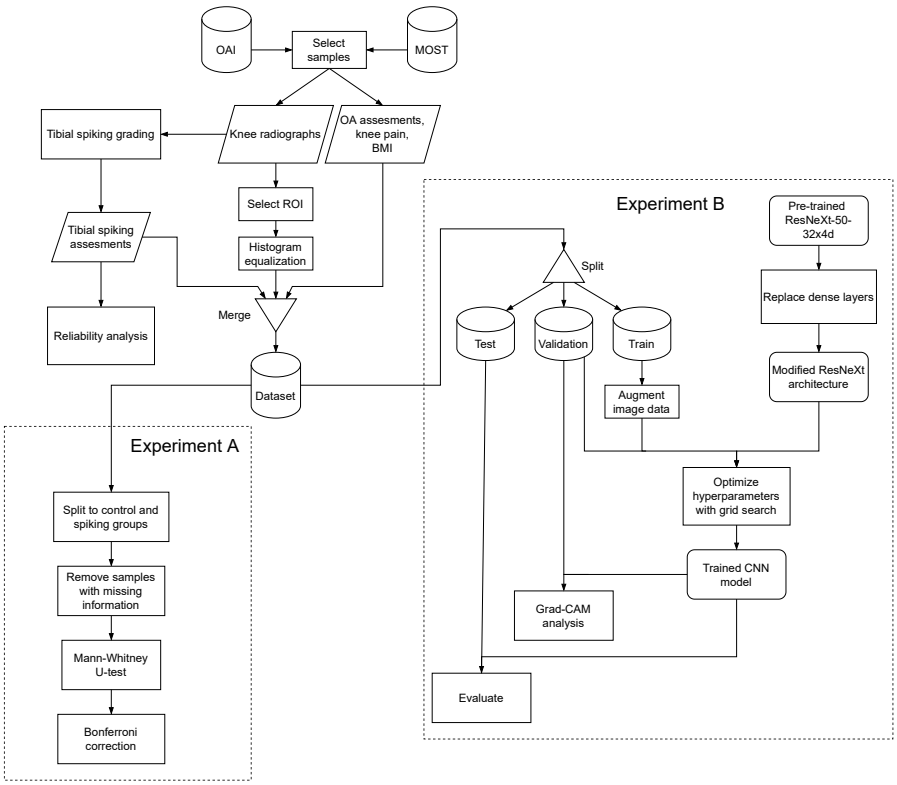


Figure 3. Flowchart of the study methodology. The processes or methods are denoted with rectangles, datasets with cylinders, data with parallelograms, and the models with rectangles with rounded corners.

3. Results and Discussion

3.1. Reliability

The reproducibility of tibial spiking grading is detailed in Table 5. According to a frequently used scale reported by Landis and Koch [47] for interpreting κ values, the range 0.21–0.40 represents fair agreement, 0.41–0.60 moderate agreement, 0.61–80 substantial agreement, and 0.81–1.00 almost perfect agreement. Albeit, the ranges are arbitrary according to Landis and Koch [47] but are often used to discuss reliability analysis results.

Table 5. Intra and inter-rater reliability (κ) for 3-way ratings, binary ratings with unsure omitted (denoted with o), and overall spiking (denoted with OR) with 95% confidence interval (CI).

	Intra-Rater Reliability (Expert 1)	Intra-Rater Reliability (Expert 2)	Inter-Rater Reliability
Medial	0.61 (0.58–0.64)	0.52 (0.50–0.54)	0.34 (0.33–0.35)
Medial (o)	0.78 (0.75–0.82)	0.94 (0.92–0.96)	0.59 (0.58–0.61)
Lateral	0.59 (0.56–0.62)	0.75 (0.73–0.76)	0.55 (0.55–0.56)
Lateral (o)	0.71 (0.67–0.74)	1.00 (1.00–1.00)	0.75 (0.74–0.76)
OR	0.53 (0.50–0.57)	0.69 (0.67–0.72)	0.48 (0.47–0.49)

Inter-rater reliability for the lateral spiking was in the moderate range, similar to the KL-grade reliability evaluated in previous studies [48,49]. However, the inter-rater reliability of the medial side assessments was only fair. The medial spiking ratings contained more unsure ratings, indicating that the medial spines were more challenging to assess. The proximate cause for the discrepancy can only be conjectured. Nevertheless, the medial tibial tubercles are more prominent and thus are more likely to be occluded by the femur. After omitting the unsure ratings, the inter-rater reliability was comparable to the KL grading reliability.

The intra-rater reliability of both experts was moderate to substantial for the 3-way ratings. The intra-rater reliability of expert 2 was lower for the medial side. However, expert 1 was equally consistent in assessing the lateral and the medial sides. The disparity in the ratings of expert 2 seems to be explainable by the unsure ratings. The ratings for the knees graded twice of expert 2 contained 38% unsure ratings, while assessments of expert 1 contained 23% unsure ratings. The intra-rater reliability of expert 2 was near-perfect after omitting the unsure ratings.

Overall, the intra-rater reliability for assessing tibial spiking was comparable to KL grading (0.50 weighted κ with 95% confidence interval (CI) of (0.25–0.75)) [48]. After removing the unsure ratings, the intra-rater reliability of expert 2 exceeded the κ reported by Gossec et al. [48]. Although, intra-rater reliability of assessing tibial spiking seems less reliable than KL grading when compared against the KL intra-rater reliability of 0.97 weighted κ with 95% CI of (0.92–1.0) reported by Culvenor et al. [49]. After omitting the unsure ratings, the intra-rater reliability of expert 2 was comparable to the weighted κ obtained by Culvenor et al. [49].

3.2. Experiment A

The sample means for spiking and control groups jointly with U-test significance levels are detailed in Table 6. After applying the Bonferroni correction, the significant variables below $p < 0.001$ were KL-grade and BMI, and the variables below $p < 0.01$ were OARSI Medial JSN and tibia lateral osteophytes score.

Table 6. Mean values for spiking and control groups with U-test significance levels (without Bonferroni correction * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$).

	Spiking	Control
KL-grade ***	1.11	0.70
WOMAC knee pain *	2.14	1.62
BMI ***	29.09	27.46
Medial JSN ***	0.38	0.25
Lateral JSN	0.05	0.04
Tibia medial osteophytes **	0.59	0.41
Tibia lateral osteophytes ***	0.40	0.21
Femur medial osteophytes **	0.48	0.27
Femur lateral osteophytes **	0.41	0.22

We found significant group differences supporting hypothesis A. Higher KL, OARSI medial JSN and lateral tibia osteophyte grades in the spiking group support the previously reported evidence for the association between tibial spiking and knee OA [23,26]. However, there was no difference in the less prominent [5] lateral JSN. Considering the spiking group's higher mean KL-grade, the higher BMI in the spiking group is consistent with the literature on the risk factors of knee OA [50]. Our results partly confirm the association between tibial spiking and osteophytes reported previously [24–26]. However, we could not confirm an association between tibial spiking and knee pain.

3.3. Experiment B

We evaluated the top model from the grid search using 191 samples from the MOST dataset. The model produced an accuracy of 0.869 with sensitivity of 0.909 and specificity of 0.750. More details on the model performance can be seen in Table 7. Confusion matrix for the test data is presented in Figure 4.

Table 7. Classifier performance metrics.

	Accuracy	Loss	Sensitivity	Specificity	Precision
Train	0.872	0.300	0.882	0.851	0.925
Validation	0.869	0.399	0.929	0.745	0.883
Test	0.869	0.314	0.909	0.750	0.915

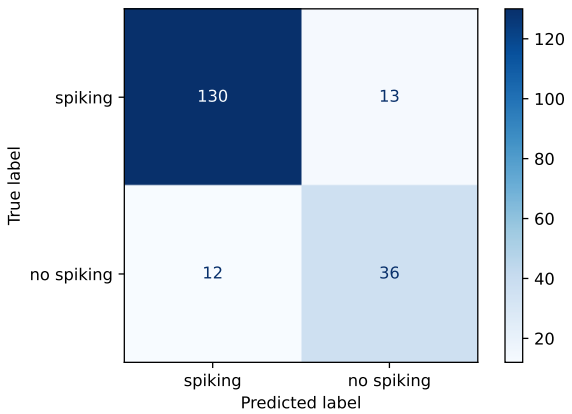


Figure 4. Confusion matrix for the test dataset.

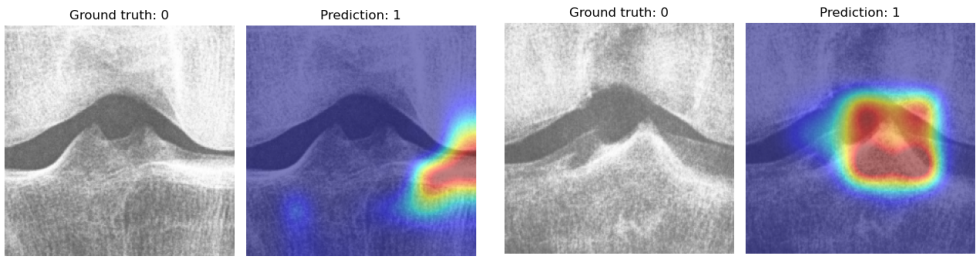
The developed model obtained lower specificity than sensitivity, meaning the model suffers to a greater extent from type I error (i.e., the model predicted spiking when none was present), which might reflect the class imbalance in the training samples. Additionally, due to how the ground truth was constructed, the no-spiking class might have contained

more borderline cases (a subset of the unsure ratings could have been regarded as doubtful spiking). In future studies, more data should be accumulated to address the asymmetry.

Like knee OA severity grading, automating tibial spiking detection lacks the "ideal" ground truth (i.e., 100% reliable labels). Consequently, the models derived will be constrained by the quality of data available. The ground truth's inter-rater reliability (κ) was 0.48, i.e., moderate. Currently, a universally agreed-upon atlas for assessing tibial spiking does not exist. The lack of shared criteria for grading tibial spiking casts doubt on the generalizability of the model developed in the present work, as different experts might have divergent views on how the spiking of tibial tubercles manifests in the radiographs. Nevertheless, the results of experiment A indicate that the spiking the model was trained to detect is associated with the other signs of knee OA.

It should be emphasized that calling the feature tibial "spiking" is somewhat imprecise in the present work, as the feature was graded based not only on the angulation of the spines but also on the length and bony growth on the spine peaks. Therefore, the feature might be more explicitly considered as "osteophytic" abnormalities in the tibial tubercles. Alexander [51] speculated that tibial spiking might be a type of osteophyte formation. Our results give some support for the theory. However, more research on the topic is required as the exact mechanism behind tibial spiking is unknown.

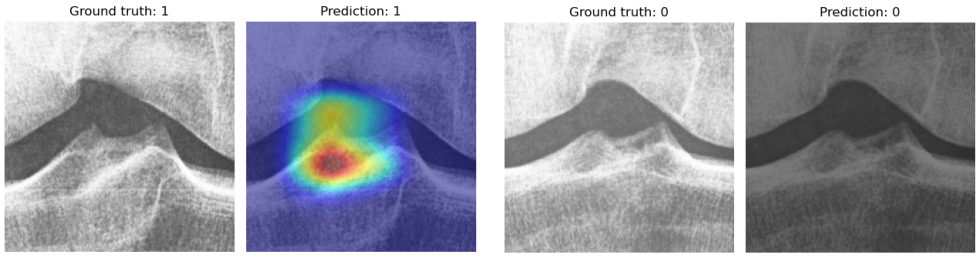
We visualized the model predictions for the validation data using Grad-CAM. By visualizing the failed predictions of the model, we can gain information on the reasons behind the failures, e.g., in Figure 5a, the network concentrates on the narrow joint space instead of on the tubercles. However, in Figure 5b, the model concentrates on the medial tibial tubercle predicting spiking while the ground truth was non-spiking. The heatmap indicates a strong influence of the medial tubercle; the assessments could be re-evaluated in cases where the model has an apparent disagreement with the expert. The model used in this manner can provide a second opinion for a physician.



(**a**) (**b**)

Figure 5. Grad-CAM visualizations for incorrect predictions. In subfigure (**a**), the model rates the knee spiking based on the narrow appearance of the medial joint space, which indicates that the model has learned the association between tibial spiking and medial JSN. In subfigure (**b**), the model makes a spiking assessment based on the medial tubercle.

The successful predictions can be visualized to identify how the model is able to discern the spiking samples from the non-spiking samples. In Figure 6a, the model has concentrated on the vicinity of the lateral tubercle. In Figure 6b, the model could not identify any spiking features and consequently predicted non-spiking.



(**a**) (**b**)

Figure 6. Grad-CAM visualizations for correct predictions. In subfigure (**a**), the lateral tubercle had the highest contribution to the prediction. In subfigure (**b**), no influential region was found from a non-spiking sample, as can be observed from the blank heatmap.

4. Conclusions

Our results indicate an association between the spiking of tibial tubercles and early knee OA. Adopting spiking of tibial tubercles as additional information in the diagnosis of knee OA seems promising. Although, additional research on the characteristics of tibial spiking and guidelines for assessing the feature is needed. The model developed for automatically identifying tibial spiking was able to generalize despite the modest number of training samples. The analysis using Grad-CAM revealed that the developed method is somewhat reliant on the JSN, which could lead to misclassifications. In the future, more data on tibial spiking should be acquired to develop better tools for healthcare.

Author Contributions: Conceptualization, J.P. and S.Ä.; Methodology, A.P. and L.A.; Software, A.P.; Validation, A.P.; Formal Analysis, A.P.; Investigation, A.P.; Resources, S.Ä.; Data Curation, O.L., J.P. and A.P.; Writing—Original Draft Preparation, A.P.; Writing—Review & Editing, L.A., O.L., J.P. and S.Ä.; Visualization, A.P.; Supervision, L.A. and S.Ä.; Project Administration, L.A., J.P. and S.Ä.; Funding Acquisition, S.Ä. All authors have read and agreed to the published version of the manuscript.

Funding: The work is related to the AI Hub Central Finland project that has received funding from the Council of Tampere Region and European Regional Development Fund and Leverage from the EU 2014-2020. This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available in a publicly accessible repository that does not issue DOIs Publicly available datasets were analyzed in this study. OAI data can be found here: https://nda.nih.gov/oai/ (accessed on 4 October 2022). MOST data was available at: https://most.ucsf.edu (currently unavailable as of 4 October 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Vos, T.; Flaxman, A.D.; Naghavi, M.; Lozano, R.; Michaud, C.; Ezzati, M.; Shibuya, K.; Salomon, J.A.; Abdalla, S.; Aboyans, V.; et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: A systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **2012**, *380*, 2163–2196. [CrossRef]
2. Cross, M.; Smith, E.; Hoy, D.; Nolte, S.; Ackerman, I.; Fransen, M.; Bridgett, L.; Williams, S.; Guillemin, F.; Hill, C.L.; et al. The global burden of hip and knee osteoarthritis: Estimates from the Global Burden of Disease 2010 study. *Ann. Rheum. Dis.* **2014**, *73*, 1323–1330. [CrossRef]
3. Inacio, M.; Paxton, E.; Graves, S.; Namba, R.; Nemes, S. Projected increase in total knee arthroplasty in the United States—An alternative projection model. *Osteoarthr. Cartil.* **2017**, *25*, 1797–1803. [CrossRef] [PubMed]
4. Pamilo, K.J.; Haapakoski, J.; Sokka-Isler, T.; Remes, V.; Paloneva, J. Rapid rise in prevalence of knee replacements and decrease in revision burden over past 3 decades in Finland: A register-based analysis. *Acta Orthop.* **2022**, *93*, 382. [CrossRef] [PubMed]

5. Swagerty, D.L., Jr.; Hellinger, D. Radiographic assessment of osteoarthritis. *Am. Fam. Physician* **2001**, *64*, 279.
6. Guermazi, A.; Niu, J.; Hayashi, D.; Roemer, F.W.; Englund, M.; Neogi, T.; Aliabadi, P.; McLennan, C.E.; Felson, D.T. Prevalence of abnormalities in knees detected by MRI in adults without knee osteoarthritis: Population based observational study (Framingham Osteoarthritis Study). *BMJ* **2012**, *345*, e5339. [CrossRef]
7. Hayashi, D.; Felson, D.; Niu, J.; Hunter, D.; Roemer, F.; Aliabadi, P.; Guermazi, A. Pre-radiographic osteoarthritic changes are highly prevalent in the medial patella and medial posterior femur in older persons: Framingham OA study. *Osteoarthr. Cartil.* **2014**, *22*, 76–83. [CrossRef] [PubMed]
8. Kellgren, J.H.; Lawrence, J.S. Radiological Assessment of Osteo-Arthrosis. *Ann. Rheum. Dis.* **1957**, *16*, 494–502. [CrossRef]
9. Kohn, M.D.; Sassoon, A.A.; Fernando, N.D. Classifications in Brief: Kellgren-Lawrence Classification of Osteoarthritis. *Clin. Orthop. Relat. Res.* **2016**, *474*, 1886–1893. [CrossRef] [PubMed]
10. Spector, T.D.; Hochberg, M.C. Methodological problems in the epidemiological study of osteoarthritis. *Ann. Rheum. Dis.* **1994**, *53*, 143–146. [CrossRef] [PubMed]
11. Altman, R.; Gold, G. Atlas of individual radiographic features in osteoarthritis, revised. *Osteoarthr. Cartil.* **2007**, *15*, A1–A56. [CrossRef]
12. Oka, H.; Muraki, S.; Akune, T.; Mabuchi, A.; Suzuki, T.; Yoshida, H.; Yamamoto, S.; Nakamura, K.; Yoshimura, N.; Kawaguchi, H. Fully automatic quantification of knee osteoarthritis severity on plain radiographs. *Osteoarthr. Cartil.* **2008**, *16*, 1300–1306. [CrossRef] [PubMed]
13. Shamir, L.; Ling, S.M.; Scott, W.W.; Bos, A.; Orlov, N.; Macura, T.J.; Eckley, D.M.; Ferrucci, L.; Goldberg, I.G. Knee X-ray image analysis method for automated detection of osteoarthritis. *IEEE Trans. Biomed. Eng.* **2008**, *56*, 407–415. [CrossRef]
14. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput.* **1989**, *1*, 541–551. [CrossRef]
15. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. *Nature* **2017**, *542*, 115–118. [CrossRef]
16. Yi, P.H.; Wei, J.; Kim, T.K.; Sair, H.I.; Hui, F.K.; Hager, G.D.; Fritz, J.; Oni, J.K. Automated detection & classification of knee arthroplasty using deep learning. *Knee* **2020**, *27*, 535–542. [CrossRef] [PubMed]
17. Antony, J.; McGuinness, K.; O'Connor, N.E.; Moran, K. Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 1195–1200. [CrossRef]
18. Yeoh, P.S.Q.; Lai, K.W.; Goh, S.L.; Hasikin, K.; Hum, Y.C.; Tee, Y.K.; Dhanalakshmi, S. Emergence of Deep Learning in Knee Osteoarthritis Diagnosis. *Comput. Intell. Neurosci.* **2021**, *2021*, 4931437. [CrossRef]
19. 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*, 1–10. [CrossRef]
20. Chen, P.; Gao, L.; Shi, X.; Allen, K.; Yang, L. Fully Automatic Knee Osteoarthritis Severity Grading Using Deep Neural Networks with a Novel Ordinal Loss. *Comput. Med. Imaging Graph.* **2019**, *75*, 84–92. [CrossRef] [PubMed]
21. Kim, D.H.; Lee, K.J.; Choi, D.; Lee, J.I.; Choi, H.G.; Lee, Y.S. Can Additional Patient Information Improve the Diagnostic Performance of Deep Learning for the Interpretation of Knee Osteoarthritis Severity. *J. Clin. Med.* **2020**, *9*, 3341. [CrossRef] [PubMed]
22. Sutton, D. *A Textbook of Radiology and Imaging*, 4th ed.; Churchill Livingstone: London, UK, 1987; p. 113.
23. Reiff, D.; Heron, C.; Stoker, D. Spiking of the tubercles of the intercondylar eminence of the tibial plateau in osteoarthritis. *Br. J. Radiol.* **1991**, *64*, 915–917. [CrossRef] [PubMed]
24. Donnelly, S.; Hart, D.J.; Doyle, D.V.; Spector, T.D. Spiking of the tibial tubercles–a radiological feature of osteoarthritis? *Ann. Rheum. Dis.* **1996**, *55*, 105–108. [CrossRef] [PubMed]
25. Unlu, Z.; Tarhan, S.; Goktan, C.; Tuzun, C. The correlation between magnetic resonance detected cartilage defects and spiking of tibial tubercles in osteoarthritis of the knee joint. *Acta Medica Okayama* **2006**, *60*, 207–214.
26. Hayeri, M.R.; Shiehmorteza, M.; Trudell, D.J.; Heflin, T.; Resnick, D. Proximal tibial osteophytes and their relationship with the height of the tibial spines of the intercondylar eminence: Paleopathological study. *Skelet. Radiol.* **2010**, *39*, 877–881. [CrossRef] [PubMed]
27. Eckstein, F.; Wirth, W.; Nevitt, M.C. Recent Advances in Osteoarthritis Imaging—The Osteoarthritis Initiative. *Nat. Rev. Rheumatol.* **2012**, *8*, 622–630. [CrossRef] [PubMed]
28. Segal, N.A.; Nevitt, M.C.; Gross, K.D.; Gross, K.D.; Hietpas, J.; Glass, N.A.; Lewis, C.E.; Torner, J.C. The Multicenter Osteoarthritis Study: Opportunities for Rehabilitation Research. *PM&R* **2013**, *5*, 647–654. [CrossRef]
29. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
30. Python Core Team. *Python: A Dynamic, Open Source Programming Language*; Python Core Team: 2015.
31. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
32. Mann, H.B.; Whitney, D.R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann. Math. Stat.* **1947**, *18*, 50–60. [CrossRef]

33. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261–272. [CrossRef] [PubMed]
34. Armstrong, R.A. When to use the Bonferroni correction. *Ophthalmic Physiol. Opt.* **2014**, *34*, 502–508. [CrossRef] [PubMed]
35. *Pydicom/Pydicom*; Pydicom 2.2.2; Version 2.2.2. Available online: https://github.com/pydicom/pydicom (accessed on 26 October 2022).
36. Wada, K. Labelme: Image Polygonal Annotation with Python, 2022. Version 5.0.1. Available online: https://github.com/wkentaro/labelme (accessed on 26 October 2022).
37. Bradski, G. The OpenCV Library, 2022. Version 4.5.5.64. Available online: https://github.com/opencv/opencv-python (accessed on 26 October 2022).
38. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. *arXiv* **2016**, arXiv:1611.05431.
39. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Li, F.-F. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR09, Miami, FL, USA, 20–25 June 2009.
40. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
41. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Proceedings of the Advances in Neural Information Processing Systems 32, Montreal, QC, Canada, 8–14 December 2019; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2019; pp. 8024–8035.
42. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How transferable are features in deep neural networks? In Proceedings of the Advances in Neural Information Processing Systems 27, Montreal, QC, Canada, 8–13 December 2014; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2014; Volume 27.
43. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June 26–1 July 2016.
44. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980v9.
45. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
46. Fernandez, F.G. TorchCAM: Class Activation Explorer. 2020. Available online: https://github.com/frgfm/torch-cam (accessed on 26 October 2022).
47. Landis, J.R.; Koch, G.G. The Measurement of Observer Agreement for Categorical Data. *Biometrics* **1977**, *33*, 159–174. [CrossRef] [PubMed]
48. Gossec, L.; Jordan, J.; Mazzuca, S.; Lam, M.A.; Suarez-Almazor, M.; Renner, J.; Lopez-Olivo, M.; Hawker, G.; Dougados, M.; Maillefert, J. Comparative evaluation of three semi-quantitative radiographic grading techniques for knee osteoarthritis in terms of validity and reproducibility in 1759 X-rays: Report of the OARSI–OMERACT task force: Extended report. *Osteoarthr. Cartil.* **2008**, *16*, 742–748. [CrossRef] [PubMed]
49. Culvenor, A.G.; Engen, C.N.; Øiestad, B.E.; Engebretsen, L.; Risberg, M.A. Defining the presence of radiographic knee osteoarthritis: A comparison between the Kellgren and Lawrence system and OARSI atlas criteria. *Knee Surgery Sport. Traumatol. Arthrosc.* **2015**, *23*, 3532–3539. [CrossRef]
50. Bijlsma, J.W.; Berenbaum, F.; Lafeber, F.P. Osteoarthritis: An update with relevance for clinical practice. *Lancet* **2011**, *377*, 2115–2126.
51. Alexander, C.J. Osteoarthritis: A review of old myths and current concepts. *Skelet. Radiol.* **1990**, *19*, 327–333. [CrossRef]