Anri Patron

An automatic method for assessing spiking of tibial tubercles associated with knee osteoarthritis

Master's Thesis in Information Technology

December 10, 2022

University of Jyväskylä

Faculty of Information Technology

Author: Anri Patron

Contact information: anri.a.patron@jyu.fi

Supervisors: Leevi Annala, and Sami Äyrämö

Title: An automatic method for assessing spiking of tibial tubercles associated with knee osteoarthritis

Työn nimi: Automaattinen menetelmä eminentian terävöitymisen tunnistukseen

Project: Master's Thesis

Study line: Mathematical Modeling in Science and Decision Analytics

Page count: 59+15

Abstract: Efficient and scalable early diagnostic methods are warranted due to the rising prevalence of knee osteoarthritis. Radiographic imaging is the standard procedure in osteoarthritis diagnosis. However, the circumstances for early diagnosis are problematic since the plain radiographs are insensitive to the established early signs of knee osteoarthritis. Furthermore, developing machine learning tools for radiographic knee osteoarthritis diagnosis is challenging due to noisy ground-truth. The objective of this thesis was to assess a feature called spiking of tibial tubercles, which has been hypothesized as an early sign of knee osteoarthritis. Additionally, we developed a model based on neural networks for identifying the feature in plain radiographs. Our results indicate promise in including tibial spiking as an early feature of knee osteoarthritis, and the feature is identifiable automatically. However, the work in the current thesis is limited and should be validated by future work.

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

Suomenkielinen tiivistelmä: Polvinivelrikon kasvavan esiintyvyyden vuoksi tehokkaat varhaiset diagnoosimenetelmät ovat haluttavia. Radiografia on keskeinen osa polvinivelrikon diagnostiikassa. Polvinivelrikon varhainen tunnistaminen on haastavaa, sillä tärkeimpiä polvinivelrikon merkkejä on vaikea havaita röntgenkuvista taudin varhaisessa vaiheessa. Koneoppi-

mallien kehittämistä varhaiseen polvinivelrikon tunnistamiseen vaikeuttaa lisäksi saatavilla olevan datan kohinaisuus. Tämän tutkielman tavoitteena oli tarkastella hypoteesia eminentian terävöitymisestä varhaisen polvinivelrikon piirteenä. Tutkielmassa kehitettiin myös neuroverkkopohjainen malli piirteen tunnistamiseen röntgenkuvista. Työn tulokset viittaavat eminentian terävyyden olevan yhteydessä varhaiseen polvinivelrikkoon. Tämän lisäksi piirre voidaan tunnistaa automaattisesti röntgenkuvista. Työn tuloksia voidaan pitää kuitenkin vasta alustavina.

Avainsanat: polvinivelrikko, röntgenkuvaus, eminentian terävöityminen, konvoluutioneuroverkko

Preface

I want to thank my supervisors and co-authors for their guidance and patience during the process of working on this thesis and the article contained herein. The article was submitted to and rejected by four journals before being accepted and published. The process of getting the work published was, therefore, not straightforward. While the article thesis format resulted in some strife, I am grateful for the opportunity to work on the current topic. Additionally, I want to thank everyone at the hyperspectral lab for the kind comments, advice, and countless delightful coffee breaks. Most of all, I want to thank my partner Saki for being supportive and inspiring me to do better each day.

Jyväskylä, December 10, 2022

Anri Patron

Glossary

Adam	Adaptive moments estimation (optimization method)
Anteroposterior	concerned with axis from anterior to posterior or front to back
BN	Batch normalization
CDF	Cumulative distribution function
CNN	Convolutional neural network
FNN	Feedforward neural network
Grad-CAM	Gradient-weighted Class Activation Mapping
JSN	Joint space narrowing
KL	Kellgren-Lawrence (osteoarthritis classification)
Lateral	Situated on the side
Medial	Situated near the center, opposite of lateral
OA	Osteoarthritis
Osteophyte	Bone spur or projection that form near the joints
PDF	Probability density function
ReLU	Rectified linear unit
SGD	Stochastic gradient descent
Subchondral sclerosis	Thickening of bone beneath the cartilage
Tubercle	Small prominence or protrusion

List of Figures

List of Tables

Contents

1 Introduction

Knee osteoarthritis (OA) is a chronic disease targeting the cartilage of the knee joint. OA is also found in the hands, hips, and feet (Swagerty Jr and Hellinger 2001). Symptomatically OA manifests in pain, stiffness, and impaired movement of joints (Swagerty Jr and Hellinger 2001). The prevalence of knee OA increases with age (Oliveria et al. 1995). In Finland, the age-adjusted prevalence of clinically diagnosed knee OA was found to be 6.1% for men and 8.0% for women (Arokoski et al. 2007). Risk factors of knee OA include age, female sex, obesity, and physical stress (e.g., due to occupation or sports) (Swagerty Jr and Hellinger 2001; Zhang and Jordan 2010).

Plain radiography is typically used imaging modality for diagnosis and assessing knee OA severity. Radiographic signs of knee OA include asymmetric joint space narrowing (JSN), osteophytes (bone spurs), cysts, and subchondral sclerosis (thickening of bone beneath the cartilage) (Swagerty Jr and Hellinger 2001). Kellgren-Lawrence (KL) classification (Kellgren and Lawrence 1957) is typically used for assessing the severity of radiological knee OA. KL is a semi-quantitative composite scoring emphasizing osteophyte formation and JSN (Kohn, Sassoon, and Fernando 2016). Radiographic classifications such as KL are not entirely reproducible, i.e., different physicians might disagree on a given KL-rating (Gossec et al. 2008; Culvenor et al. 2015). For example, the experienced raters produce more reproducible radiographic classifications than trainees (Günther and Sun 1999).

Automatic methods for assessing knee OA from plain radiographs were initially proposed by Oka et al. 2008; Shamir et al. 2009. The methods were suggested for improving the objectivity of assessing knee OA (Oka et al. 2008). The current state-of-the-art for classifying the knee OA severity is based on convolutional neural networks (CNNs) (Yeoh et al. 2021). The CNN-based methods are proficient in classifying the severe samples of knee OA but cannot effectively detect the early cases (Antony et al. 2016; Tiulpin et al. 2018; Chen et al. 2019). Identifying early-stage knee OA is imperative for providing effective interventions (Felson and Hodgson 2014). Radiographic classifications such as KL have limitations and, as such, should be used in conjunction with clinical examination (Kohn, Sassoon, and Fernando 2016). This is rarely considered, as the KL-grade is typically used to define the

ground-truth for knee OA severity prediction (Yeoh et al. 2021). Additionally, radiography is insensitive to the early cartilage damage and osteophyte formation (Guermazi et al. 2012; Hayashi et al. 2014). In contrast, magnetic resonance imaging (MRI) can be used to visualize the cartilage directly (Guermazi et al. 2011). However, MRI comes with drawbacks such as limited availability, high cost, and longer imaging time (Guermazi et al. 2011).

The present work is formatted as an introduction to an article by Patron et al. 2022. The article's objective was to study a radiological feature called spiking of tibial tubercles (tibial spiking in short) that has been hypothesized as a sign of early knee OA (hypothesis A) (D. Sutton 1987). Finding such early sign of knee OA would be desirable as the feature is assessable by plain radiography, which is widely available and cheap (Guermazi et al. 2011). Assuming tibial spiking is identifiable by human experts, then an automatic method can be developed, provided sufficient example data (hypothesis B). In the article, we conducted two experiments corresponding to hypotheses A and B (Patron et al. 2022).

- In experiment A, we studied the relationship between tibial spiking and early knee OA by comparing subjects with and without tibial spiking. We examined established markers of knee OA, such as osteophytes, JSN, and knee pain.
- In experiment B, we developed a CNN-based model for identifying tibial spiking. Due to the limited availability of data for the feature, we used the transfer learning approach, i.e., we utilized a pre-trained CNN as a base for our classifier.

The structure of the thesis is the following. Chapter 2 introduces the spiking of tibial tubercles and the previous research regarding the association with knee OA. The theoretical background for the methods used in the paper is provided in Chapter 3. We summarize the article's results in chapter 4. Chapter 5 contains an extended discussion of the study. We provide the conclusion in chapter 6. The article is included in appendix A.

2 Spiking of tibial tubercles

Spiking of tibial tubercles refers to abnormally tall or sharp tibial spines (Reiff, Heron, and Stoker 1991). See Fig. 1 for a comparison between spiking and non-spiking tibial spines. Tibial spiking has been hypothesized as a feature of early knee OA. Although, the origin of said hypothesis is unknown. The hypothesis has been recited by D. Sutton 1987 in a radiological textbook. Tibial spiking is also mentioned in another radiological text by Resnick and Niwayama 1988. However, neither of the authors provides evidence for the hypothesis.

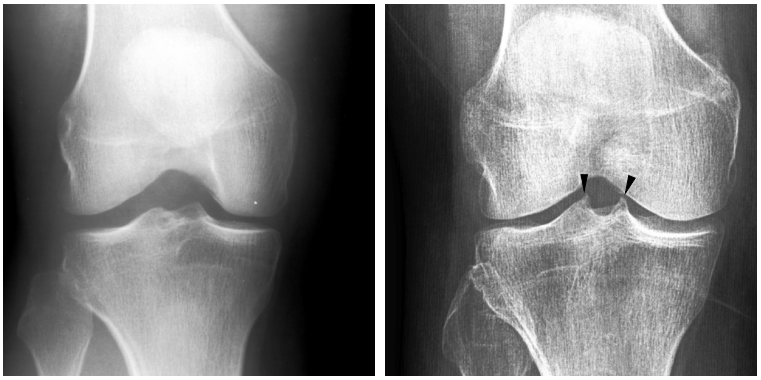


Figure 1: Comparison between knee radiographs with and without tibial spiking. Arrowheads note the spiking tibial spines (adapted from Patron et al. 2022).

The research concerning spiking of the tibial tubercles and knee OA is very sparse; to our knowledge, only five studies have been published on the subject. A summary of the published research is provided hereafter. Reiff, Heron, and Stoker 1991; Donnelly et al. 1996; Unlu et al. 2006 investigated tibial spiking by measuring the angulation of tibial spine tips and the height-width ratio of the tubercles

$$\frac{H_i}{W},\ i = \{L, M\} \tag{2.1}$$

where H_i is the height of the lateral (L) or the medial (M) spine and W is the width of the tibial plateau. The measurements were calculated using anteroposterior view radiographs. Reiff, Heron, and Stoker 1991 examined radiographs from 55 subjects and found the lengthening and sharpening of the tubercles' peaks associated with knee OA.

Donnelly et al. 1996 conducted a study with 950 subjects examining tibial spiking as a radio-

logical feature of knee OA. They found angulation of the tibial tubercle to correlate with the presence of osteophytes and knee OA status (defined by KL-grade 2 or greater). Osteophyte formation was found to correlate with the lateral tubercle height but not with the medial tubercle height (Donnelly et al. 1996). They concluded that in isolation, tibial spiking was not a reliable sign of knee OA due to a lack of strong independent association with knee pain (Donnelly et al. 1996).

Unlu et al. 2006 studied the association between tibial spiking and cartilage defects assessed via MRI. The study involved 76 knees from 47 subjects and 31 knees from a control group. The subjects with knee OA had significantly taller and sharper tibial spines compared to the controls (Unlu et al. 2006). The medial cartilage defects correlated with the medial tubercle height but not with the lateral tubercle height (Unlu et al. 2006). They also found an association between the spiking of the lateral tubercle and osteophyte formation in the tibial compartments (Unlu et al. 2006).

Hayeri et al. 2010 considered the subject from a paleopathological framework, where they examined 35 tibial bone specimens directly for signs of OA. The bone samples were assessed by evaluating the osteophytes by size and measuring the height of the tubercles. The study found spiking of the lateral tibial spine to be associated with osteophyte formation in the lateral compartment (Hayeri et al. 2010). Bastick et al. 2017 studied the characteristics associated with patients with early-onset knee or hip OA that underwent joint replacement surgery within six years. The study considered a wide variety of features. The patients who underwent total knee replacement showed more signs of radiographic OA, including tibial spiking (Bastick et al. 2017).

The results of the aforementioned studies indicate that the tibial spiking might be associated with knee OA (Reiff, Heron, and Stoker 1991; Donnelly et al. 1996; Unlu et al. 2006; Hayeri et al. 2010), but is not a reliable feature in isolation (Donnelly et al. 1996). We found no studies that examined the tibial spiking in subjects with early knee OA. Donnelly et al. 1996 stated to study tibial spiking as an early feature of knee OA but did not specifically analyze the feature in subjects with early-stage disease. Our paper aimed to fill the gap in the research regarding hypothesis A.

3 Methods

This chapter will introduce the methods used in the article (Patron et al. 2022). Section 3.1 introduces the concept of reliability and the Kappa coefficient proposed by Cohen 1960 for estimating reliability. The study used the Wilcoxon–Mann–Whitney test (Wilcoxon 1945; Mann and Whitney 1947) to compare spiking samples to non-spiking samples. The test is detailed in section 3.2. Section 3.3 introduces histogram equalization, which was used in radiograph preprocessing. Classification metrics used for measuring the performance of the developed model are summarized in section 3.4.

The model developed for tibial spiking detection is CNN-based. Consequently, we will cover the relevant details on neural networks. We provide the basics of feedforward neural networks in section 3.5, and section 3.6 details how neural networks are trained. The last section 3.7 will provide an introduction to CNNs, starting from the foundations of convolution operation and how the valid convolution differs from the convolution operation in CNNs. Subsequently, the building blocks of CNNs are introduced, i.e., convolution and pooling layers. In the article, we used ImageNet (Deng et al. 2009) pre-trained ResNeXt (Xie et al. 2017) as the foundation for the classifier. We cover the ResNeXt architecture by comparing with the related ResNet architecture (He et al. 2016) in subsection 3.7.3. Additionally, we introduce transfer learning 3.7.4, and gradient-weighted class activation mapping (Grad-CAM) 3.7.5 (Selvaraju et al. 2017). Grad-CAM was used for providing visualizations attempting to explain the model predictions.

3.1 Reliability analysis

Reliability, used in this thesis, refers to the extent to which repeated measurements or tests produce the same outcome. In classical test theory, we assume an observed random variable X is a composite of the true score T and measurement error E, formally defined as

$$X = T + E, \tag{3.1}$$

where T and E are unknown random variables (Novick 1966). The reliability of a test is the variance ratio of T and X

$$\frac{\mathrm{Var}(T)}{\mathrm{Var}(X)} \tag{3.2}$$

(Novick 1966). Since the random variable T is unknown, the reliability has to be estimated.

In medicine, the measurement is commonly produced by a human rater, e.g., a physician diagnosing a patient. Assessing the consistency of a single rater or observer is achieved by examining the intra-rater reliability. When multiple raters are employed, it is necessary to evaluate the inter-rater reliability, i.e., the consistency between different raters. In practice, reliability is evaluated by assessing the consistency of repeated measurements. For nominal variables, the most straightforward approach would be to calculate the proportion of agreed samples a to all samples N, i.e.,

$$p_a = \frac{a}{N}. \tag{3.3}$$

The naive approach is, however, inadequate as some of the agreement could be due to chance alone (Cohen 1960; Banerjee et al. 1999). To this end, Cohen 1960 introduced κ (Kappa), a chance-corrected coefficient for measuring agreement between two raters. The κ coefficient makes the following assumptions:

- The samples are independent.
- The classes are nominal scale, independent, mutually exclusive, and exhaustive.
- The raters work independently.

The formulation of κ coefficient is the following

$$\kappa = \frac{p_a - p_c}{1 - p_c}, \tag{3.4}$$

where p_a is the proportion of agreed samples and p_c is the proportion of agreement expected due to chance (Cohen 1960). The κ coefficient can be regarded as the proportion of agreement after the removal of chance agreement.

Let us consider an example. Table 1 presents an agreement matrix for two raters; note that the matrix describes a joint distribution. The proportion of agreement is simply

$$p_a = \frac{20 + 15}{50} = 0.7.$$

We assumed that the rater's responses were statistically independent. Therefore, the expected proportion of agreement due to chance is found using the marginal distributions (Cohen 1960). In the case of table 1, this means that we can calculate the p_c using the row and column proportions. To find the proportion of chance agreement p_c, we calculate the sum of the expected agreement for each category (product of corresponding p_{row} and p_{column} values), i.e.,

$$p_c = \frac{30}{50} \times \frac{25}{50} + \frac{20}{50} \times \frac{25}{50} = 0.5.$$

Now we can calculate the coefficient

$$\kappa = \frac{0.7 - 0.5}{1 - 0.5} = 0.4.$$

When observed agreement equals the expected agreement $\kappa = 0$ and for perfect agreement $\kappa = 1$ (Cohen 1960). The agreement of raters for table 1 is higher than the expected agreement due to chance, but the difference is not exceptionally high. Whether this is an acceptable level of agreement would depend on the application.

Table 1: An agreement matrix for two raters. The proportions are calculated by dividing the raw frequencies by the sample size N. The proportions for agreements on each category are found on the diagonal, and correspondingly the disagreements are found off-diagonal. Additionally, the row and column proportions are noted with p_{row} and p_{column}.

		Rater 1		
	Category	A	B	p_{row}
Rater 2	A	20/50	5/50	25/50
	B	10/50	15/50	25/50
	p_{column}	30/50	20/50	$N = 50$

The standard error of κ is approximated by

$$\text{SE} = \sqrt{\frac{p_a(1-p_a)}{N(1-p_c)^2}}, \tag{3.5}$$

where N is the sample size. Now with sufficiently large N, the sampling distribution of κ will approximate the normal distribution (Cohen 1960). Thus, the confidence interval can

be calculated equivalently for large sample sizes. For example, the 95% confidence interval (CI) is given by

$$\kappa \pm 1.96 \text{ SE}. \tag{3.6}$$

The coefficient described above is suitable for nominal variables as all disagreements are treated the same. For ordinal scale variables, Cohen 1968 introduced weighted κ, allowing the level of disagreement to be weighted differently, e.g., by distance.

3.2 Wilcoxon–Mann–Whitney two-tailed test

Let us assume samples of two independent groups, A and B. The Wilcoxon–Mann–Whitney test (also called the U-test) (Wilcoxon 1945; Mann and Whitney 1947) can be used to answer questions regarding the difference between the groups (Nachar et al. 2008). For example, if the samples are height measurements of two groups, we could test whether the samples in A are taller than samples in B or if the groups come from different populations (Nachar et al. 2008). The former describes a one-tailed test, and the latter a two-tailed test. Next, we will provide the details of the two-tailed test.

The null hypothesis H_0 of the two-tailed test stipulates that the groups are homogeneous and identically distributed, and the alternative hypothesis H_1 specifies that the groups have different distributions. Mathematically this is expressed by probabilities

$$H_0 : p(a_i > b_j) = \frac{1}{2}$$
$$H_1 : p(a_i > b_j) \neq \frac{1}{2},$$

where a_i and b_j are observations of A and B, respectively (Nachar et al. 2008). Hence, we will reject the null hypothesis if either A or B is stochastically larger than the other, i.e., the test statistic falls into either tail of the sampling distribution (Nachar et al. 2008). The U-test is a non-parametric test based on comparing the rank sums of the two samples. Non-parametric means the test does not assume any specific distribution, which is useful since the samples can be non-normal and have small sample sizes. However, the samples are assumed to be at least ordinal. The U-test, therefore, provides an alternative for the Student's t-test (Student 1908) when the assumptions are not met.

First, the observation's ranks (i.e., the ordering) must be determined. The rank sums for A and B are calculated by sorting the pooled observation values and assigning ascending ranks for each observation from 1 to N, where N is the total number of observations. The ranks of each observation are then summed for both groups. For example, considering the ranks already sorted in Table 2, the rank sums for the groups are

$$R_{\mathrm{A}} = 1+2+5+7+11+14+15+16 = 71$$
$$R_{\mathrm{B}} = 3+4+6+8+9+10+12+13 = 65,$$

where R_{A} denotes the rank sum of group A and R_{B} the rank sum of group B.

Table 2: Rank sum example: observations (measurements) from groups A and B are detailed in the score column, and the corresponding rank of the observation is in the rank column.

Group	Score	Rank	Group	Score	Rank
A	21	1	⋮	⋮	⋮
A	23	2	B	36	9
B	25	3	B	37	10
B	29	4	A	38	11
A	30	5	B	41	12
B	32	6	B	43	13
A	33	7	A	46	14
B	35	8	A	49	15
⋮	⋮	⋮	A	52	16

When ties occur between the groups (i.e., observations from different groups are tied for a rank), the average of the untied ranks is used for the tied observations (Nachar et al. 2008). For example, if k observations are tied for rank r, the tied rank is

$$\frac{1}{k}\sum_{i=0}^{k} r+i.$$

Knowing the rank sums of the variables, we can calculate the U statistic for A by

$$U_A = n_A n_B + \left(\frac{n_A(n_A+1)}{2}\right) - R_A, \tag{3.7}$$

where n_A, n_B are the sample sizes of A and B, respectively and R_A is the rank sum of A (Nachar et al. 2008). The group statistic U_A describes the number of times an observation from A proceeded an observation from B (Nachar et al. 2008). The corresponding statistic U_B can be calculated equivalently, although knowing U_A, the U_B becomes deterministic (Nachar et al. 2008). Knowing that $R_A + R_B = 1 + 2 + 3 + \cdots + N = \frac{N(N+1)}{2}$, where $N = n_A + n_B$. The statistic U_B can be calculated by

$$U_B = n_A n_B - U_A, \tag{3.8}$$

the intermediate steps of the derivation are detailed in Nachar et al. 2008. To obtain the test statistic, we choose the smaller of U_A and U_B (Nachar et al. 2008), i.e.,

$$U = \min(U_A, U_B). \tag{3.9}$$

When $n_A, n_B \geq 8$ the distribution of U is approximately normal (Mann and Whitney 1947), the standard score is therefore

$$z = \frac{U - \mu}{\sqrt{\sigma^2}}, \tag{3.10}$$

where μ is the mean and σ^2 is the variance. The mean of U is defined as

$$\mu = \frac{n_A n_B}{2} \tag{3.11}$$

and the variance of U is defined as

$$\sigma^2 = \frac{(n_A n_B)(N+1)}{12} \tag{3.12}$$

(Mann and Whitney 1947). If normal approximation cannot be justified, the distribution of U should be used instead for finding the p-value (not presented in this work). When ties are present in the ranks between the groups, the variance is adjusted by

$$\sigma^2 = \frac{n_A n_B (N+1)}{12} - \frac{n_A n_B \sum_{i=1}^{e} (d_i^3 - d_i)}{12N(N-1)}, \tag{3.13}$$

where e is the number of unique observed values, and d_i is the number of each value, indexing from lowest to highest (Lehmann and D'Abrera 1998). For example, suppose that we observed values 4, 5, 5, 5, 8, 8 then $e = 3$ and $d_1 = 1$, $d_2 = 3$, and $d_3 = 2$.

Example calculation

Using the data from Table 2, let us calculate the U statistic and find the p value for the two-tailed test. The rank sums for both groups were calculated above. The rank sums were $R_{\mathrm{A}} = 71$ and $R_{\mathrm{B}} = 65$. The statistic for A is

$$\begin{aligned} U_{\mathrm{A}} &= n_{\mathrm{A}} n_{\mathrm{B}} + \left(\frac{n_{\mathrm{A}}(n_{\mathrm{A}}+1)}{2} \right) - R_{\mathrm{A}} \\ &= 8 \times 8 + \left(\frac{8 \times (8+1)}{2} \right) - 71 \\ &= 64 + \frac{72}{2} - 71 = 29 \end{aligned}$$

and correspondingly

$$U_{\mathrm{B}} = n_{\mathrm{A}} n_{\mathrm{B}} - U_{\mathrm{A}} = 64 - 29 = 35.$$

We choose the smaller statistic $U = \min(29, 35) = 29$. The sample size is sufficiently large for normal approximation, therefore we can calculate the standard score. The mean is

$$\mu = \frac{n_{\mathrm{A}} n_{\mathrm{B}}}{2} = 32$$

and the variance is

$$\sigma^2 = \frac{(n_{\mathrm{A}} n_{\mathrm{B}})(N+1)}{12} = \frac{64 \times (16+1)}{12} = \frac{272}{3} \approx 90.67.$$

We calculate the standard score

$$z = \frac{U - \mu}{\sqrt{\sigma^2}} \approx \frac{29 - 32}{\sqrt{90.67}} \approx -0.32.$$

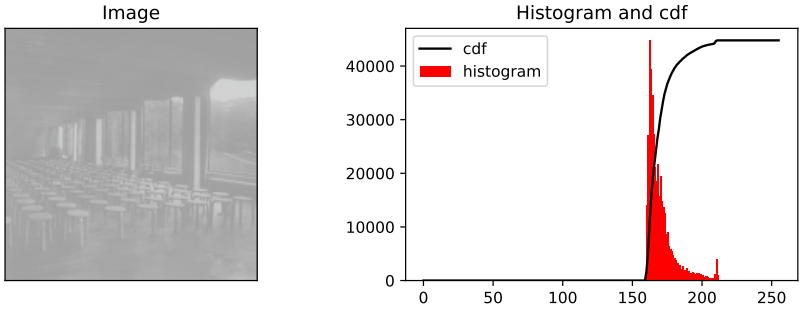
Now we can find the p-value

$$\begin{aligned} P(Z \leq |z|) &= 2P(Z \leq z) \approx 2P(Z \leq -0.32) = 2P(Z \geq 0.32) \\ &= 1 - 2P(Z \leq 0.32) = 1 - 2\Phi(0.32) \approx 1 - 2 \times 0.1255 = 0.749, \end{aligned}$$

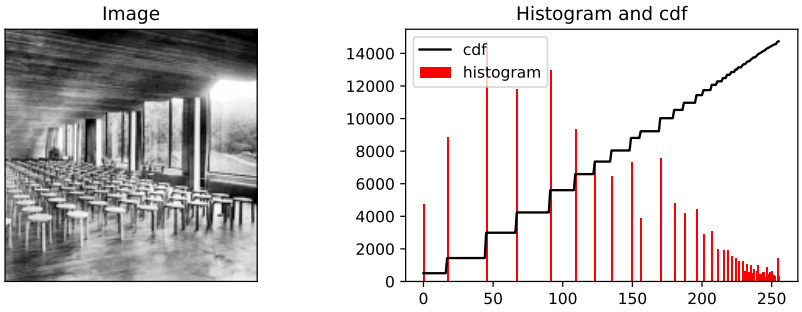
where Z is a random variable following the standard normal distribution, i.e., $Z \sim \mathcal{N}(0,1)$ and $\Phi(\cdot)$ denotes cumulative distribution function (CDF) of the standard normal distribution; the $\Phi(0.32) = 0.1255$ was found using tabled values. Suppose we use the standard level of significance 0.05 for rejecting H_0, then we would fail to reject the null hypothesis since $0.749 \not< 0.05$. Therefore, we did not find significant differences between groups A and B.

3.3 Histogram equalization

An image with low contrast has its intensity values constrained to a small range, and high contrast image has intensity values spanning a wide range (Gonzalez and Woods 2018); see Fig. 2 for an example. Histogram equalization is a method for increasing the global contrast of a grayscale image (Gonzalez and Woods 2018). The basic principle behind the method is to adjust the distribution of the pixel intensities.



(a) Original image with low contrast.



(b) Histogram equalized image.

Figure 2: Comparison between an image with a low contrast and histogram equalized image, presented with histogram and a CDF of the histogram.

The distribution of intensity levels in an image can be represented with a histogram (see Fig. 2(a)). Let I denote a grayscale image with pixel intensities ranging from 0 to $L-1$, where 0 corresponds with black and $L-1$ with white pixels. For example, in an 8-bit image, $L = 256$. We denote the number of pixels with intensity k with n_k, $k = 0, 1, \ldots, L-1$, i.e., the unnormalized histogram of I. The normalized histogram of I is defined as

$$p_k = \frac{n_k}{HW}, \tag{3.14}$$

where W is the image width and H the image height (Gonzalez and Woods 2018). The histogram equalization transforms the pixel intensities k by the function

$$T(k) = \text{round}\left((L-1)\sum_{i=0}^{k} p_i\right), \tag{3.15}$$

where the round function is used to get integers intensity values (Gonzalez and Woods 2018). See Fig. 2(b) for an example of the transformation result.

We will now briefly outline how to derive the function in Eq. 3.15. Let us consider continuous random variables X, Y in a range $[0, L-1]$. Want to find a transformation of intensity values such that the histogram of the transformed image spans the whole range $[0, L-1]$ of intensity values and the distribution of intensities is uniform (Gonzalez and Woods 2018). Such transformation is given by

$$Y = T(X) = (L-1)\int_0^X p_X(u)\mathrm{d}u, \tag{3.16}$$

where $p_X(x)$ is the probability density function (PDF) of X and the integral is the CDF of X (Gonzalez and Woods 2018). Assuming that T is differentiable, then it can be shown that

$$p_Y(y) = \frac{1}{L-1},$$

i.e., the transformation results in uniform PDF; see the proof in Gonzalez and Woods 2018. The normalized histogram p_k approximates $p_X(x)$, therefore, we can consider $\sum_{i=0}^{k} p_i$ a CDF (Gonzalez and Woods 2018). Thus Eq. 3.15 approximates the continuous transformation 3.16. Although the approximate transformation is not guaranteed to produce a uniform PDF, as can be seen in Fig. 2(b). However, the approximation will flatten the distribution of the intensity values increasing the contrast.

3.4 Classification criteria

Considering a binary classification (e.g., a test for a disease), the positive class is the detected condition, and the negative class is the absence of the condition. We assign the negative class integer label 0 and the positive class integer label 1; note that the choice of integers is arbitrary. We denote feature vectors $X = \{x_1, x_2, \ldots, x_N\}$, where $x_i \in \mathbb{R}^m$ and corresponding ground-truth labels $Y = \{y_1, y_2, \ldots, y_N\}$, where $y_i \in \{0, 1\}$, and a classifier $f : \mathbb{R}^m \to \{0, 1\}$,

that maps x_i to y_i. To assess the performance of the classifier f, we evaluate the classifier's predictions using a number of labeled samples, i.e., $\{x_i, y_i\}$ pairs. The different outcomes for the classification are the following (Runkler 2016):

1. True positive (TP): $y = 1, f(x) = 1$
2. True negative (TN): $y = 0, f(x) = 0$
3. False positive (FP): $y = 0, f(x) = 1$
4. False negative (FN): $y = 1, f(x) = 0$

FP is also referred to as a type I error, and FN as a type II error (Runkler 2016). The classification result, i.e., the frequencies of the different outcomes, are typically collected in a confusion matrix, like the one depicted in table 3.

Table 3: Confusion matrix for binary classification.

		Predicted class	
		Positive	Negative
True class	Positive	TP	FN
	Negative	FP	TN

Numerous criteria can be computed based on the counts of the different classification outcomes; some are detailed below. For a more detailed listing of classification criteria, refer to Runkler 2016. Probably the most used criteria for classification is accuracy, defined by

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.17}$$

(Runkler 2016). Accuracy is the proportion of the correct prediction counts to all predictions. If we would prefer to favor one class over the others or would like to avoid false positive errors if possible, we need to report more than simple accuracy. Otherwise, we might misrepresent the performance of our classifier.

Often it is appropriate to report sensitivity and specificity (Yerushalmy 1947). Sensitivity or true positive rate is defined as

$$\text{sensitivity} = \frac{TP}{TP + FN} \tag{3.18}$$

and specificity or true negative rate is defined as

$$\text{specificity} = \frac{TN}{TN+FP} \tag{3.19}$$

(Runkler 2016). Sensitivity informs us of the probability of classifying the positive class as positive, and specificity the probability of classifying the negative class as negative (Runkler 2016). Frequently a trade-off between sensitivity and specificity exists; by accepting lower sensitivity, a higher specificity might be achievable. If we would like to know the probability of a sample classified as positive is actually positive, e.g., a patient diagnosed as sick is sick, we can calculate the precision defined by

$$\text{precision} = \frac{TP}{TP+FP} \tag{3.20}$$

(Runkler 2016). Even a trivial classifier (e.g., a classifier that always predicts the negative class) can produce impressive-looking results on a single criterion in the right circumstances. For example, in the case of high class imbalance. Multiple performance metrics should be reported as they characterize the classifier in different ways.

3.5 Feedforward neural networks

A feedforward neural network (FNN) is a function mapping $\hat{y} = f(x, \theta)$, where x is the input tensor (a generalization of a vector in n-dimensional space), θ the parameters (usually a tensor as well), and $\hat{y}$ is the output (Goodfellow, Bengio, and Courville 2016). In the case of classification, $\hat{y}$ is a vector of class scores. In general, we want f to approximate some true function f' mapping $y = f'(x)$ (Goodfellow, Bengio, and Courville 2016). As f is an approximation, therefore, $y \approx \hat{y}$. FNNs are trained by finding suitable values for θ that produce an optimal approximation. The network is called feedforward as the information flows in one direction through the network (unlike in recurrent neural networks) (Goodfellow, Bengio, and Courville 2016). The FNN are typically composed of multiple functions $f(x) = f^{\ell}\left(f^{\ell-1}\left(\ldots\left(f^{1}(x)\right)\right)\right)$, where f^1 is the first layer of the FNN. The layers between the input and the output are often referred to as the hidden layers.

To understand better what kind of function f is, let us examine a single computational unit in these layers, called a neuron or a perceptron (McCulloch and Pitts 1943; Hebb 1949;

Rosenblatt 1958). The neuron calculates a linear combination of the elements of input vector $x \in \mathbb{R}^m$ and adds an intercept term (often referred to as bias)

$$h = \sum_{i=1}^{m} w_i x_i + b, \tag{3.21}$$

where w_i is the weight corresponding with the element x_i and b is the bias (Szeliski 2022). Matrix notation is usually preferred for brevity, i.e., the Eq. 3.21 would be denoted with

$$h = f(x; w, b) = w^T x + b,$$

where $w \in \mathbb{R}^m$ and $b \in \mathbb{R}$. Note that the parameters θ in the case of the neuron are w and b. The perceptron is a simple linear system, equivalent to a linear regression model (Goodfellow, Bengio, and Courville 2016). This comes with drawbacks, e.g., the perceptron is famously unable to learn the exclusive or (XOR) function (Minsky and Papert 1969).

FNNs are formed by stacking multiple layers of interconnected neurons in a chain (Szeliski 2022); see Fig. 3 for an example. The FNN structure comprises of an input and ℓ layers performing matrix multiplications, where $\ell \geq 1$ and the last layer is the output layer. The layers are connected so that the output of layer $l-1$ is used as the input to the following layer l, where $l = 1, \ldots, \ell$. When each neuron in a layer l is connected to each neuron in layer $l-1$ the layer is called fully connected or dense (Szeliski 2022). Usually, an activation function is added to introduce nonlinearity after the neuron output

$$a_i = \phi(h_i), \tag{3.22}$$

where $\phi(\cdot)$ is nonlinear function applied elementwise (Goodfellow, Bengio, and Courville 2016). A typical choice for an activation function is the rectified linear unit (ReLU) (Nair and Hinton 2010), defined as

$$\phi(x) = \max(0, x). \tag{3.23}$$

With the introduction of nonlinear activation, FNN can approximate nonlinear functions (Goodfellow, Bengio, and Courville 2016).

Let us consider a simple dense FNN presented in Fig 3. The network has an input size of three and two hidden layers, the first with three neurons and the second with two neurons,

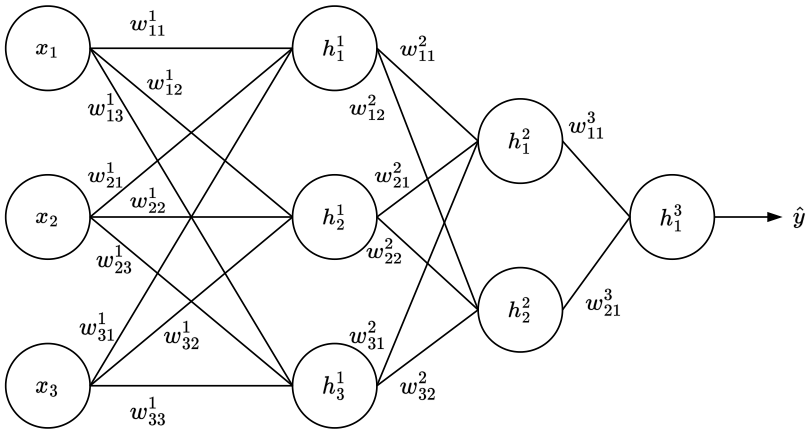


Figure 3: A simple feedforward neural network. We denote the layer number with a superscript and specify the unit with a subscript. Note that the bias is omitted in the graph.

followed by a single output. For simplicity, let us assume linear activations, the output $\hat{y}$ is given by

$$\begin{aligned} h^1 &= (W^1)^T x + b^1 \\ h^2 &= (W^2)^T h^1 + b^2 \\ \hat{y} &= (W^3)^T h^2 + b^3, \end{aligned}$$

where b^l, $l \in \{1,2,3\}$ is the bias vector, for layer l and W^l is $N_{l-1} \times N_l$ matrix, where N_l is the number of neurons in layer l. Note that the input layer size $N_0 = 3$. The bias vector $b^l \in \mathbb{R}^{N_l}$. We denote the output of hidden layers with h^1, h^2, and the output with $\hat{y}$. The weight matrix W^l is defined as

$$W^j = \begin{bmatrix} w_{11}^l & w_{12}^l & \cdots & w_{1N_l}^l \\ w_{21}^l & w_{22}^l & \cdots & w_{N_l}^l \\ \vdots & \vdots & \ddots & \vdots \\ w_{N_{l-1}1}^l & w_{N_{l-1}2}^l & \cdots & w_{N_{l-1}N_l}^l \end{bmatrix}.$$

3.6 Training neural networks

This section explains how the parameters θ (e.g., weights and biases) of neural networks are determined, usually referred to as training. Suppose we have a neural network model f and data $D = \{x_i, y_i\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^m$. For classification, y_i is usually a one-hot encoded vector with k elements, where k is the number of classes. Let gt denote the index of the ground-truth class of label i. One-hot encoding sets one element corresponding with the ground-truth $y_i^{gt} = 1$ and $y_i^{gt} = 0$ for all $j \neq gt$. We want the model to perform well on some relevant performance measurement P (e.g., classification accuracy). However, P is not optimized directly. Instead, the model is trained by minimizing a different criterion, a loss function (also called the objective or error function) $L(\theta)$, by doing so, our goal is to increase P indirectly (Goodfellow, Bengio, and Courville 2016).

In general terms, neural networks are trained by iteratively adjusting the parameters of the network until we converge to a "suitable" set of values, i.e., a local minimum. Neural networks are typically non-convex functions; therefore, we cannot guarantee convergence to a global minimum with iterative optimization methods (Goodfellow, Bengio, and Courville 2016). The parameters are optimized using example data D and a learning algorithm. We call the model training supervised learning when D includes features x and labels y (Szeliski 2022). The scope of this thesis is restricted to supervised learning. Therefore, we will not cover other machine learning tasks.

Usually, we divide D into a train, validation, and test subsets. The train samples are used, as the name implies, to train the model, i.e., to find optimal θ that minimizes the loss $L(\theta)$. The validation samples are used to tune the model hyperparameters (Goodfellow, Bengio, and Courville 2016). Hyperparameters are design decisions and are not adapted by the training algorithm (Goodfellow, Bengio, and Courville 2016), e.g., the number of layers and the number of neurons in each layer. The hyperparameter values can be determined through experimentation or optimized by another algorithm. The generalization error or the model performance, is estimated using the test samples not seen during optimizing model parameters or hyperparameters (Goodfellow, Bengio, and Courville 2016).

During training, we would like to minimize the error made by our model, i.e., we would like

to find θ such that the average of L for our data is small as possible (Szeliski 2022). Formally we define the optimization problem as

$$\min_{\theta} E(\theta) = \frac{1}{N} \sum_{i=1}^{N} L(f(x_i; \theta), y_i), \tag{3.24}$$

i.e., we minimize the expected loss (Szeliski 2022). Most of the time (for nontrivial neural networks), E is non-convex due to the nonlinear activations; thus, iterative methods are used as we are unable to solve the optimization problem 3.24 analytically (Goodfellow, Bengio, and Courville 2016). Moreover, due to the number of parameters, calculating the Hessian matrix becomes prohibitively expensive (Szeliski 2022). Therefore, we need to settle for first-order methods, i.e., gradient-based optimization algorithms.

To train our model, we apply gradient descent for updating the model's parameters θ. We start with some initial values of θ and evaluate the average loss given the data D (Szeliski 2022). To decrease the loss, we step towards the negative gradient of $E(\theta)$. We calculate the update for timestep $t+1$ by

$$\theta_{t+1} = \theta_t - \alpha \nabla_\theta E(\theta_t), \tag{3.25}$$

where $0 < \alpha \leq 1$ is the step size (Szeliski 2022). To find the gradient, we use the chain rule of calculus to find the derivative of the loss with regard to each parameter, proceeding from the outputs toward the inputs (Szeliski 2022). The backpropagation algorithm[1] (Linnainmaa 1970, 1976; Werbos 1974; Rumelhart, Hinton, and Williams 1986) describes an efficient method for calculating the gradient numerically (Szeliski 2022).

Usually, the loss is evaluated, and the parameters are updated for subsets of the training samples called mini-batches due to computational reasons (Goodfellow, Bengio, and Courville 2016). Effectively we replace the gradient with a noisy estimate. For optimizing a neural network using mini-batches instead of the whole training data, the stochastic gradient descent (SGD) is used (Goodfellow, Bengio, and Courville 2016). Loss for each mini-batch $\mathcal{B}$

1. The backpropagation algorithm has been studied by a number of people. The algorithm was first described by Linnainmaa 1970 in his master's thesis. The algorithm was later studied in the context of neural networks by Werbos 1974; Rumelhart, Hinton, and Williams 1986. The credit assignment in the field of deep learning is inconsistent; for a more detailed summary of the history of deep learning, see the survey by Schmidhuber 2015.

is calculated

$$J(\theta) = \frac{1}{n_{\mathcal{B}}} \sum_{i=1}^{n_{\mathcal{B}}} L(f(x_i;\theta), y_i), \tag{3.26}$$

where $n_{\mathcal{B}}$ is the sample size of the $\mathcal{B}$; note that to get an unbiased estimate of the gradient, the samples must be independent, i.e., drawn randomly (Goodfellow, Bengio, and Courville 2016).

In SGD, the parameters are updated iteratively until we converge to a local minimum, or another stopping condition is reached. Due to the random sampling, the stochastic gradients might have a high variance requiring a low learning rate, which can cause slow convergence (Goodfellow, Bengio, and Courville 2016). The noise added by the gradient estimator does not vanish even at the minimum (Goodfellow, Bengio, and Courville 2016). Therefore it is common to decrease the learning rate gradually over time (Goodfellow, Bengio, and Courville 2016). Additionally, saddle points,[2] prevalent in high dimensional non-convex problems present an issue for gradient descent (Dauphin et al. 2014).

3.6.1 Adam and other variants of gradient descent

This section introduces Adam and a few other variants of gradient descent used for training neural networks. For a more thorough overview of gradient descent variants, see Ruder 2016. The convergence of the regular gradient descent might sometimes be slow due to small gradients in flat areas or oscillation in ravines. (R. Sutton 1986). The momentum method (Polyak 1964) was designed to accelerate the convergence of gradient descent and average out the oscillation (Goodfellow, Bengio, and Courville 2016). We denote the gradient computed at timestep t with $g_t = \nabla_\theta J(\theta_t)$. Momentum adds a velocity term v, which is an exponentially decaying average of the negative gradient estimates, formally defined as

$$v_t = \beta v_{t-1} - \alpha g_t, \tag{3.27}$$

where $\beta \in [0,1)$ is a hyperparameter controlling the decay factor and α is the step size (Goodfellow, Bengio, and Courville 2016). The update is calculated by adding the velocity term

$$\theta_{t+1} = \theta_t + v_t. \tag{3.28}$$

2. Saddle points are critical points that are neither minima nor maxima.

A global learning rate might not be inductive to learning infrequently appearing features since they do not have a great effect on the cost (Duchi, Hazan, and Singer 2011). To this end, Duchi, Hazan, and Singer 2011 introduced the Adaptive Gradient (AdaGrad) method featuring an adaptable learning rate. The learning rate is adjusted by dividing the gradient componentwise by the square root of the running sum of squared gradients (Goodfellow, Bengio, and Courville 2016). The running sum of squared gradients is computed by

$$r_t = r_{t-1} + g_t \odot g_t, \tag{3.29}$$

where $\odot$ denotes elementwise product. The update is calculated by

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\varepsilon + \sqrt{r_t}} \odot g_t, \tag{3.30}$$

where ε is a small number added for numerical stability (Goodfellow, Bengio, and Courville 2016). Due to the scaling term, the parameters with the largest partial derivatives with respect to the loss will have the fastest reduction of learning rate (Goodfellow, Bengio, and Courville 2016). Consequently, the parameters with lower magnitude partial derivatives have a slower reduction in the learning rate. However, due to the accumulation of the whole gradient history, the method might halt convergence prematurely (Goodfellow, Bengio, and Courville 2016).

The early halting problem is addressed in root mean square propagation (RMSprop) proposed by Hinton 2012. RMSprop replaces the running sum of squared gradients of AdaGrad with a decaying sum, defined by

$$r_t = \beta r_{t-1} + (1-\beta) g_t \odot g_t, \tag{3.31}$$

where β is a hyperparameter controlling the impact of the gradient history and $\odot$ denotes elementwise product (Goodfellow, Bengio, and Courville 2016). The update is given by

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{r_t + \varepsilon}} \odot g_t. \tag{3.32}$$

(Goodfellow, Bengio, and Courville 2016).

Adam (Kingma and Ba 2014) combines the adaptive learning rate with momentum; the name derives from the phrase adaptive moment estimation. The learning rates for the parameters

are computed from exponential moving averages of the gradient (m) and the squared gradient (v), for timestep t:

$$m_t = \beta_1 m_{t-1}(1-\beta_1)g_t \tag{3.33}$$

$$v_t = \beta_2 v_{t-1}(1-\beta_2)g_t \odot g_t, \tag{3.34}$$

where β_1 and β_2 are hyperparameters controlling the decay rates of m_t and v_t, respectively (Kingma and Ba 2014). The moving averages estimate the first moment (the mean) and the second moment (centered variance), hence the algorithm's name. Especially for the initial timesteps, the estimates are biased due to initialization with zeros (Kingma and Ba 2014). The bias is corrected by

$$\hat{m}_t = \frac{m_t}{(1-\beta_1^t)} \tag{3.35}$$

$$\hat{v}_t = \frac{v_t}{(1-\beta_2^t)} \tag{3.36}$$

(Kingma and Ba 2014). The update is calculated using the bias-corrected estimates

$$\theta_{t+1} = \theta_t - \frac{\alpha \hat{m}_t}{\sqrt{\hat{v}_t + \varepsilon}}, \tag{3.37}$$

where α is the learning rate and ε is a small constant added for numerical stability (Kingma and Ba 2014).

3.6.2 Cross-entropy

The loss function has a central role in neural network training as it describes the criterion, which is optimized by gradient descent. The typical loss functions are derived from the principle of maximum likelihood; in the case of classification, the loss function is cross-entropy (Goodfellow, Bengio, and Courville 2016). Next, we briefly introduce the background of cross entropy loss.

Our set of N independent examples $X = \{x_1, x_2, \ldots, x_N\}$ are drawn from a unknown distribution $p(x)$ and our parametric model $f(X;\theta)$ describes a family of probability distributions (Goodfellow, Bengio, and Courville 2016). We want f to estimate the true distribution p. The maximum likelihood estimator for θ is the product over the individual probabilities

(Goodfellow, Bengio, and Courville 2016), defined as

$$\arg\max_{\theta} f(X;\theta) = \arg\max_{\theta} \prod_{i=1}^{N} f(x_i;\theta). \tag{3.38}$$

The product of many values is numerically inconvenient as it could either explode or vanish, which is why we take the logarithm of the likelihood to get a sum instead (Goodfellow, Bengio, and Courville 2016). This is why the cross-entropy is also called the log loss; see the derivation in Goodfellow, Bengio, and Courville 2016.

The cross-entropy is defined as

$$\mathrm{CE} = -\sum_{j=1}^{k} y_j \log(\hat{y}_j), \tag{3.39}$$

where k is the number of classes, the ground-truth y is one-hot encoded vector with k elements and $\hat{y} \in \mathbb{R}^k$ is the vector of the prediction scores for each class (Szeliski 2022); note that Eq.3.39 is defined for a single example. In classification, the softmax function is typically applied before calculating the loss. Softmax normalizes the raw neuron activations in the [0,1] range so that they sum to one. This is why softmax's output is usually thought of as likelihoods (Szeliski 2022). The Softmax function is defined as

$$\mathrm{softmax}(\hat{y})_i = \frac{\exp(\hat{y}_i)}{\sum_{j=1}^{k} \exp(\hat{y}_j)} \tag{3.40}$$

(Goodfellow, Bengio, and Courville 2016). Since the ground-truth is one-hot encoded, all but one element of y are zero. Let us denote the non-zero index with gt. After applying the softmax, the loss is simplified to

$$-\log\left(\frac{\exp(\hat{y}_{gt})}{\sum_{j=1}^{k} \exp(\hat{y}_j)}\right) = \log\left(\sum_{j=1}^{k} \exp(\hat{y}_j)\right) - \hat{y}_{gt}, \tag{3.41}$$

i.e., normalized predictions corresponding with the correct class (Szeliski 2022).

3.6.3 Batch normalization

During model training, we evaluate the gradient of the average loss and compute the updates for all parameters at once. Since the layers are functions composed together, the update can cause unexpected results, as the parameter updates assume that the other functions would be

constant (Goodfellow, Bengio, and Courville 2016). The inputs for the intermediate layers depend on the layers before. The parameter's updates change the distribution of the inputs for the proceeding layers, requiring constant adaptation to the shift (Ioffe and Szegedy 2015). Moreover, the small changes compound for the deeper layers of the network. Effectively the shifts of the input distribution force us to use a low learning rate for deep models (Ioffe and Szegedy 2015). The apparent downside is that the training will take longer. The batch normalization (BN) introduced by Ioffe and Szegedy 2015 addresses the issue by reparametrizing the inputs to the intermediate layers of a neural network.

Let $A_{\mathcal{B}} = \{a^1, \ldots, a^{n_{\mathcal{B}}}\}$ denote the activations (output of a previous layer) over a mini-batch $\mathcal{B}$, where the activations $a^j = (a_1^j, \ldots, a_d^j)$. First, each element a_i^j is normalized to have zero mean and unit variance

$$\hat{a}_i^j = \frac{a_i^j - \mu_i}{\sqrt{\sigma_i^2 + \varepsilon}}, \tag{3.42}$$

where μ_i is the mini-batch mean for element i

$$\mu_i = \frac{1}{n_{\mathcal{B}}} \sum_{j=1}^{n_{\mathcal{B}}} a_i^j,$$

and σ_i^2 is the mini-batch variance

$$\sigma_i^2 = \frac{1}{n_{\mathcal{B}}} \sum_{j=1}^{n_{\mathcal{B}}} (a_i^j - \mu_i)^2$$

and ε is a small positive number added for numerical stability (Ioffe and Szegedy 2015). Finally, the normalized inputs are reparameterized by

$$\mathrm{BN}(a_i^j) = \gamma_i \hat{a}_i^j + \beta_i, \tag{3.43}$$

where γ_i, β_i are parameters learned through backpropagation (Ioffe and Szegedy 2015).

When training, we calculate new μ_i and σ_i^2 for each mini-batch, and when testing or in production, we use running averages collected during training (Ioffe and Szegedy 2015). In essence, BN normalizes and then scales and shifts the distribution of the input. The BN procedure stabilized the distribution of the inputs by ensuring that the distribution is determined solely by the γ and β parameters (Goodfellow, Bengio, and Courville 2016). Crucially, BN was shown to allow the use of a greater learning rate and help the neural networks to converge faster (Ioffe and Szegedy 2015).

3.7 Convolutional neural networks

Convolutional neural networks (CNN) in their current form can be attributed to LeCun et al. 1989, who applied the backpropagation algorithm for learning the kernels of a shift-invariant neural network with a deep hierarchical structure named the neocognitron (Fukushima 1980; Fukushima and Miyake 1982). Goodfellow, Bengio, and Courville 2016 describe CNNs as neural networks employing convolution instead of matrix multiplication in at least one of the layers. We will start by defining convolution as a mathematical operation. In section 3.7.1, we will describe the convolution operation performed in a CNN and how it differs from the mathematical operation defined below.

Convolution of two real-valued functions f, g describes a new function

$$(f * g)(x) = \int_{-\infty}^{\infty} f(t)g(x-t)dt, \tag{3.44}$$

where t is a dummy variable for integration (Goodfellow, Bengio, and Courville 2016). We replace the integral in Eq. 3.44 with a sum to get discrete convolution. In machine learning terminology, suppose a convolution $I * K$, we denote the input I and the kernel K (Goodfellow, Bengio, and Courville 2016). In machine learning applications, the data is both discrete and finite. The discrete convolution for vectors I, K is defined as

$$G(t) = (I * K)(x) = \sum_{t} I(t)K(x-t) \tag{3.45}$$

(Goodfellow, Bengio, and Courville 2016).

If we assume a two-dimensional (2D) input I (e.g., a grayscale image described by a matrix of intensity values), then K is correspondingly a matrix. 2D convolution is defined as

$$G(i,j) = (I * K)(i,j) = \sum_{m,n} I(m,n)K(i-m, j-n) = \sum_{m,n} I(i-m, j-n)K(m,n), \tag{3.46}$$

where in the latter equality is due to the commutative property of convolution (Goodfellow, Bengio, and Courville 2016). Many deep learning libraries like PyTorch (Paszke et al. 2019) implement cross-correlation instead of convolution,[3] which corresponds with convolution without flipping the kernel (Goodfellow, Bengio, and Courville 2016). Whether the kernel is

3. See the PyTorch documentation for 2D convolution https://pytorch.org/docs/stable/generated/torch.nn.Conv2d.html.

flipped or not is irrelevant for CNN, as the kernel values are determined through backpropagation. Cross-correlation is defined as

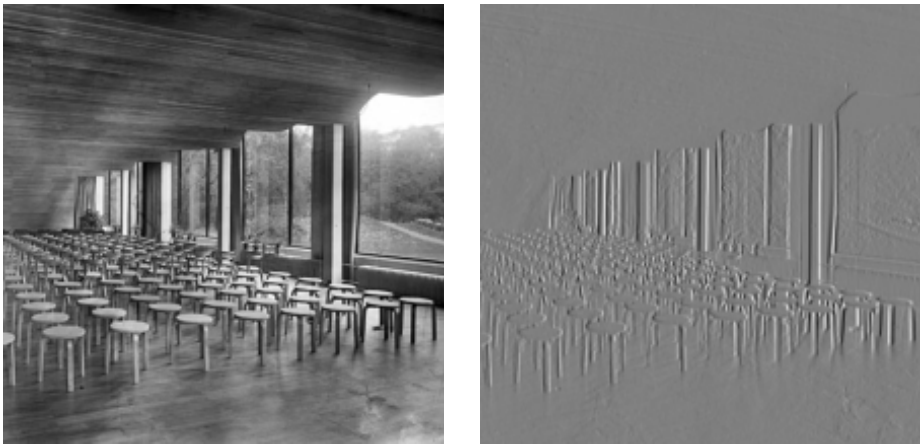
$$G(i,j) = (I \star K)(i,j) = \sum_{m,n} I(i+m, j+n)K(m,n) \tag{3.47}$$

(Goodfellow, Bengio, and Courville 2016).

Let us consider an example of valid 2D convolution (Eq. 3.46). Our input I is a grayscale image shown in Fig. 4(a) and the kernel K is a vertical Sobel filter (Szeliski 2022):

$$K = \frac{1}{8}\begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}.$$

The outcome of $I * K$ is a new image, presented in Fig. 4(b)), which describes vertical intensity changes in I, i.e., borders and lines. The convolution describes a linear transformation applied on a local region at each pixel position of I (excluding the borders).



(a) Original image. (b) Vertical Sobel filter output

Figure 4: A grayscale image convolved with Sobel filter. The input image is 374×374 and the output is 372×372. The output of a valid convolution shrinks as the pixels on the border of the input image are not processed.

3.7.1 Convolutional layer

The convolutional neural network is somewhat of a misnomer as the operation performed in CNN differs slightly from a valid convolution (Eq. 3.46). This section introduces the convolutional layer and details how it operates in contrast to the valid convolution. The convolutional layer applies the convolution operation to some input with a number of different kernels (Szeliski 2022), even though the operation is cross-correlation (Eq. 3.47), this thesis will follow the convention of calling the operation convolution. A number of different kernels are used in one layer, as one kernel specifies one kind of linear transformation extracting only one type of information (like vertical lines extracted by the vertical Sobel filter, presented in Fig. 4).

A convolutional layer consists of C_2 kernels, each kernel is convolved with an input tensor, and the layer's output is a collection of C_2 individual convolution outputs. For example, if the input is RGB-image (i.e., a three-dimensional tensor where the channel axis corresponds with the red, green, and blue intensity values), the kernels would be three-dimensional. Kernels visits each spatial position of the input image, producing a weighted sum over the local spatial and channel dimensions, all producing a 2D output (usually referred to as a feature map) (Szeliski 2022). Formally, the output of a 2D convolutional layer is given by

$$A(i,j,c_2) = \sum_{m,n,c_1} I(i+m, j+n, c_1)K(m,n,c_1,c_2) + b(c_2), \tag{3.48}$$

where c_1 iterates over the input channels and $b(\cdot)$ is a bias term (Szeliski 2022).

The number of parameters for a 2D convolutional layer can be determined by

$$(H_K \times W_K \times C_1 + 1) \times C_2, \tag{3.49}$$

where H_K is the kernel height, W_K is the kernel width, C_1 the number of input channels, and C_2 the number of kernels (Szeliski 2022); note that the addition of one is due to the bias term added by each kernel. The spatial size of the kernel is typically smaller than the input. Therefore the same kernel weights are used across the spatial dimensions of the input, i.e., the weights are shared between the inputs in a convolutional layer. The intuition is that if one kernel is useful in one part of the image, it is useful everywhere. The weight sharing allows CNNs to use fewer parameters and thus have lower memory requirements compared to dense

networks (Goodfellow, Bengio, and Courville 2016), which is why the use of feedforward neural network (FNN) for image data is impractical.

Let us consider a 64×64 grayscale image as an input. We can flatten the image to a vector with 4,096 elements. In a dense FNN, this would result in 4,097 connections with each neuron at the first hidden layer, i.e., 4,096 weight parameters and a bias. Using 100 neurons on the first hidden layer yields 409,700 parameters to optimize on the first layer alone. For example, a convolutional layer with a kernel size of 5×5 and the number of kernels $C2 = 32$ would require only 832 parameters.

Using valid convolution with a kernel of width W_K will cause the output width to shrink by $W_K - 1$ pixels (equivalent for the height) since the kernel cannot process the border pixels. Padding is used to offset the shrinking caused by convolution (Goodfellow, Bengio, and Courville 2016). Typically zero padding is used. In zero padding, as the name suggests, the input is padded with zero values (Goodfellow, Bengio, and Courville 2016). See Szeliski 2022 for other padding approaches. With padding, one can design arbitrary deep CNNs.

Suppose we want to extract more granular information from the input (i.e., downsample the output of the convolution); this can be achieved by evaluating the convolution every s:th row and column (Goodfellow, Bengio, and Courville 2016). The hyperparameter controlling the movement of the kernel is called stride. Assuming we want the stride s to be symmetric, the strided convolution can be defined as

$$A(i,j,c_2) = \sum_{m,n,c_1} I((i-1) \times s + m, (j-1) \times s + n, c_1) K(m,n,c_1,c_2) + b(c_2), \qquad (3.50)$$

where the subtraction of one from i, j is due to the indexing starting from one (Goodfellow, Bengio, and Courville 2016).

With the basic hyperparameters of the convolutional layer in place, we define how the convolutional layer alters the input volume. The output volume of the 2D convolutional layer is determined by kernel size, input height H_{in}, input width W_{in}, padding P, stride s, and the number of kernels C_2. The dimensions of the 2D convolution output is $H_{\text{out}} \times W_{\text{out}} \times C_2$

(Szeliski 2022), where

$$\left\lfloor H_{\text{out}} = \frac{H_{\text{in}} - H_K + 2P}{s} + 1 \right\rfloor, \text{ and}$$
$$\left\lfloor W_{\text{out}} = \frac{W_{\text{in}} - W_K + 2P}{s} + 1 \right\rfloor.$$

Note that we assume symmetric padding, and stride, for simplicity. Typical choices for kernel size include 3×3, 5×5, and 1×1.

1×1 convolution (Lin, Chen, and Yan 2013) is a special case of the convolution layer. When the stride $s = 1$ and the padding $P = 0$, the 1×1 convolution is in essence a matrix multiplication (Szeliski 2022), e.g., given an input tensor X with a size $H_{\text{in}} \times W_{\text{in}} \times C_1$, then 1×1 convolution is defined as

$$A_{i,j} = W_d X_{i,j}, \tag{3.51}$$

where W_d is $C_2 \times C_1$ matrix and $X_{i,j} \in \mathbb{R}^{C_1}$ denotes a vector-valued input pixel (Lin, Chen, and Yan 2013). 1×1 convolution essentially projects the input pixelwise to C_2 dimensional space. 1×1 convolution can be used to manipulate the dimensionality of the channels. Importantly, 1×1 convolution can reduce the dimensionality of the input and thus reduce the computational cost before more expensive 3×3 or 5×5 convolutions (Szegedy et al. 2015).

Typically, CNNs are used for image classification, for which multiple 2D convolutional layers are combined (Krizhevsky, Sutskever, and Hinton 2012). The general idea is to create progressively more complex feature maps by successive convolutions while shrinking the spatial resolution of the representation. CNNs usually learn kernels extracting simple shapes (e.g., lines and corners) at the initial layers; by combining the low-level features, CNNs are able to extract high-level features at the late layers (Zeiler and Fergus 2014). After a number of convolutions, we have reduced the dimensionality of the data and extracted high-level features. At this point, the transformed input is typically fed to a regular FNN for prediction (Krizhevsky, Sutskever, and Hinton 2012).

3.7.2 Pooling

We have discussed strided convolutions and 1×1 convolutions that can be used to reduce the input dimensionality. Pooling describes another way to downsample the input. In general,

pooling functions calculates summary statistics of an input in a local neighborhood (Goodfellow, Bengio, and Courville 2016). For example, assuming 2D input, max pooling (Weng, Ahuja, and Huang 1992, 1997; Riesenhuber and Poggio 1999) returns the maximum value within a square window. When pooling is used with a stride greater than one, the output is downsampled spatially. Max pooling is applied channelwise. Therefore only the spatial dimensions are affected.

The pooling functions do not have to be optimized as they do not have parameters. However, max pooling has hyperparameters, such as window size and stride. Another benefit of pooling is that it makes the CNNs invariant to small input shifts (Goodfellow, Bengio, and Courville 2016). This is useful for object recognition since we are more interested in knowing the presence of an object rather than its exact location. The results by Scherer, Müller, and Behnke 2010 indicate that the max pooling is superior to subsampling in object recognition.

Another widely used pooling function is the global average pooling (Lin, Chen, and Yan 2013), which averages the input tensor over the feature maps. Formally global average pooling is defined as

$$\mathcal{G}_c = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} A_c(i, j), \tag{3.52}$$

where, A_c is the feature map with index c and i, j indexes the height and width, respectively (Lin, Chen, and Yan 2013). This form of pooling is usually used after the last convolutional layer. Without global average pooling, one would typically flatten the tensor after the last convolutional layer and feed the resulting vector to a dense layer. Which required setting a fixed input size since too large or small input would lead to a value error in the matrix multiplication. The use of global average pooling makes the model robust toward the input size since only the number of feature maps would need to be set. The global average pooling can be used to replace the dense part of the CNN altogether. Doing so might even improve the performance (Lin, Chen, and Yan 2013).

3.7.3 ResNeXt architecture

ResNeXt is a CNN architecture introduced by Xie et al. 2017, bearing similarities to the well-known Residual Network (ResNet) (He et al. 2016); hence the name. Specifically, ResNeXt borrows the residual block architecture modifying it for increased efficiency (Xie et al. 2017). The original ResNet architecture was introduced to solve the problem of training deep neural networks. The intuition is that increasing the depth of the model should increase the model's learning capacity and, therefore, capability to overfit. However, the empirical evidence pointed otherwise; increasing the model's depth resulted in higher training error for very deep models (He et al. 2016).

The solution proposed by He et al. 2016 was the residual block architecture. Let function $\mathcal{F}(x, \{\theta_i\})$ denote a neural network's block (a number of sequential layers with possible nonlinear activations between). The residual block is formally defined as

$$a = \mathcal{F}(x, \{\theta_i\}) + x, \tag{3.53}$$

where x is the input tensor and θ_i denotes the parameters of layer i (He et al. 2016), e.g., a residual block with two dense layers with linear activations would be

$$\mathcal{F} = W_2(W_1 x + b_1) + b_2 + x.$$

Adding the input x to the output can be considered a skip connection (He et al. 2016). In Eq. 3.53 the dimensions of x and $\mathcal{F}$ must be equal. For some types of neural networks, we want to alter the dimensionality of the input (e.g., CNN typically increases the number of feature maps). When the dimensions are not equivalent, we can perform a linear projection (i.e., 1×1 convolution) to match the dimensions

$$a = \mathcal{F}(x, \{\theta_i\}) + W_d x \tag{3.54}$$

(He et al. 2016). The second design element in both architectures is the bottleneck block design (Fig. 5). Another problem with deep CNNs is the computation of late convolution layers becoming expensive as the number of feature maps expands (Szegedy et al. 2015). To this end, the ResNet utilizes bottleneck blocks (Fig. 5) in the later layers (He et al. 2016). The bottleneck block downsamples the input using 1×1 convolutions before 3×3 convolution and restores the dimensionality afterward.

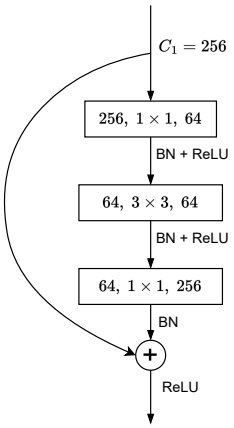


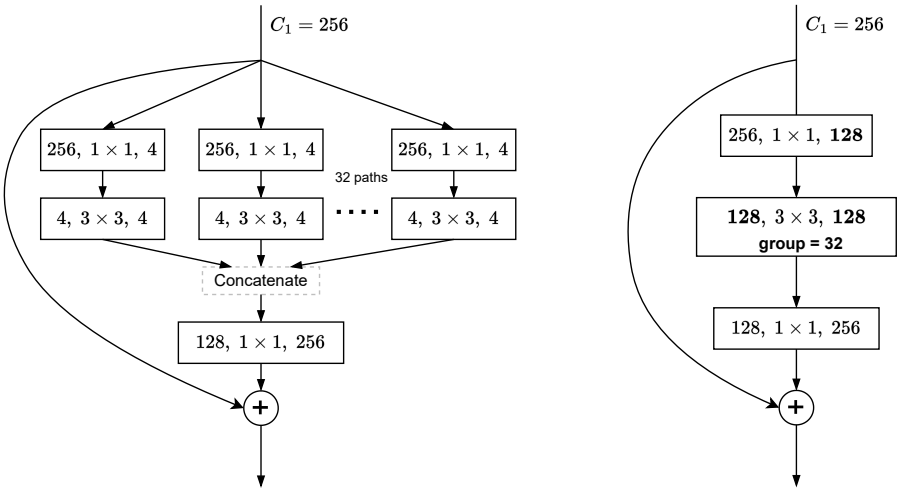
Figure 5: ResNet bottleneck block, BN denotes batch normalization. The rectangles indicate the convolutional layer, and the parameters are given in order: input channels, kernel size, output channels (adapted from He et al. 2016).

The ResNeXt block features a similar design to the ResNet bottleneck block, but the input is divided into numerous parallel paths, where the input is transformed to low-dimensional embeddings and aggregated together with the linear path (Xie et al. 2017). We denote individual transformation with $\mathcal{F}_i$. The aggregated transformation is given by

$$a = x + \sum_{i=1}^{\mathcal{C}} \mathcal{F}_i(x), \tag{3.55}$$

where $\mathcal{C}$ (cardinality) is the number of parallel paths (Xie et al. 2017). The aggregated transformation of Eq. 3.55 is realized as group convolutions (Krizhevsky, Sutskever, and Hinton 2012), where the number of groups is $\mathcal{C}$ (Xie et al. 2017). In grouped convolution, the input is split channelwise by the number of groups, the separate input groups are convolved with the corresponding group of separate kernels, and the outputs are concatenated channelwise (Krizhevsky, Sutskever, and Hinton 2012; Xie et al. 2017). The ResNeXt block with group convolution is illustrated in Fig. 6.

Now that we have introduced the design elements of ResNeXt, we can present the architecture of the model used in the article (Patron et al. 2022). The modified ResNeXt-50-32x4d architecture is presented in table 4. The ResNeXt architecture is composed by repeating a different variety of block structures after the initial convolution and pooling layers. The model uses batch normalization after each convolution operation followed by ReLU, except



(a) Block design with concatenation. (b) Block design with group convolution.

Figure 6: The ResNeXt block illustrated by group convolution, both designs are equivalent (adapted from Xie et al. 2017).

before adding the linear path ReLU is performed after the addition (similarly to as in Fig. 5) (Xie et al. 2017). The configuration ResNeXt-50-32x4d was selected as the base of the architecture simply because it performed better than the alternatives. We also experimented with ResNet (He et al. 2016), VGG (Simonyan and Zisserman 2014), and dense convolutional network (DenseNet) (Huang et al. 2017) family of models.

Table 4: The modified ResNeXt architecture used in (Patron et al. 2022), adapted from Xie et al. 2017. The convolution layer parameters are presented in the order of kernel size, and the number of kernels and $\mathcal{C}$ 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
Block1	56×56	$\left.\begin{array}{l} 1 \times 1,\ 128 \\ 3 \times 3\ 128,\ \mathcal{C} = 32 \\ 1 \times 1,\ 256 \end{array}\right\} \times 3$
Block2	28×28	$\left.\begin{array}{l} 1 \times 1,\ 256 \\ 3 \times 3\ 256,\ \mathcal{C} = 32 \\ 1 \times 1,\ 512 \end{array}\right\} \times 4$
Block3	14×14	$\left.\begin{array}{l} 1 \times 1,\ 512 \\ 3 \times 3\ 512,\ \mathcal{C} = 32 \\ 1 \times 1,\ 1024 \end{array}\right\} \times 6$
Block4	7×7	$\left.\begin{array}{l} 1 \times 1,\ 1024 \\ 3 \times 3\ 1024,\ \mathcal{C} = 32 \\ 1 \times 1,\ 2048 \end{array}\right\} \times 3$
	1×1	global average pool
Dense		2048, 2, softmax

3.7.4 Transfer learning

The growing amounts of data and computational resources have enabled neural networks to dominate on problems such as image classification (Krizhevsky, Sutskever, and Hinton 2012). However, in some domains, such as medicine, the data might be scarce or expensive to collect, hindering the applicability of data-hungry deep learning approaches. It would be desirable to transfer knowledge from one domain or task to another to improve the performance of the deep learning algorithms. This is also known as transfer learning (Bengio 2012).

The definition for transfer learning is given hereafter using the notation of Pan and Yang 2010; Weiss, Khoshgoftaar, and Wang 2016. Let $\mathcal{D}_S$ be the source domain and $\mathcal{D}_T$ the target domain, the corresponding task we denote with $\mathcal{T}_S$ and $\mathcal{T}_T$, respectively. The domain $\mathcal{D} = \{\mathcal{X}, P(X)\}$ is composed of two parts, the feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$, where the features $X = \{x_1, \ldots, x_n\} \in \mathcal{X}$. The task is formally defined as $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$, where $\mathcal{Y}$ is the label space and $f(\cdot)$ is a predictive function mapping between feature vector label pairs $\{x_i, y_i\}$, where $x_i \in X$ and $y_i \in \mathcal{Y}$. The objective of transfer learning is to improve the performance of the predictive function of the target task $f_T(\cdot)$ by using $\mathcal{D}_S$ and $\mathcal{T}_S$, where $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathcal{T}_S \neq \mathcal{T}_T$ (Pan and Yang 2010; Weiss, Khoshgoftaar, and Wang 2016).

Given the above definition, the source domain in the paper (Patron et al. 2022) is the ImageNet (Deng et al. 2009), and the source task is assigning a class label 0–999 to an image (corresponding with various objects, animals, etc.). The target domain is knee radiographs and the target task is to predict a binary spiking status. Therefore, $\mathcal{D}_S \neq \mathcal{D}_T$ and $\mathcal{T}_S \neq \mathcal{T}_T$ since $\mathcal{X}_S \neq \mathcal{X}_T$ (radiographs instead of RGB images) and $\mathcal{Y}_S \neq \mathcal{Y}_T$ (binary labels instead of [0,999]). The approach for transfer learning in the paper (Patron et al. 2022) was the network-based deep transfer learning (Pan and Yang 2010), referring to transferring a number of layers from a neural network trained on the source domain to another neural network used in the target domain. For medical image classification, the transferred layers can be used as a feature extractor or fine-tuned for the target task (Litjens et al. 2017). The latter procedure can potentially increase the performance of a deep learning model over training solely on the target task samples (Yosinski et al. 2014).

3.7.5 Gradient-weighted Class Activation Mapping

Gradient-weighted class activation mapping (Grad-CAM) (Selvaraju et al. 2017) is a method for visualizing the most influential regions of the input, given the prediction of a CNN. Grad-CAM can be considered a generalization of class activation mapping (CAM) proposed by Zhou et al. 2016. Unlike CAM, Grad-CAM does not require the model to use global average pooling and, therefore, can be applied for a broader variety of CNNs (Selvaraju et al. 2017). Grad-CAM highlights the most influential regions for predicting class u in the input. The output of Grad-CAM resembles a heatmap, usually presented over the input image.

Let us consider image classification. We denote some class with u and the score of u before softmax $\hat{y}^u$. We compute the gradient of $\hat{y}^u$ with regard to feature map activations of the last convolutional layer A^k, i.e., we compute $\frac{\partial \hat{y}^u}{\partial A^k}$. The last convolutional layer is chosen as we expect the final convolutional layer to produce the highest level feature extraction while retaining the spatial information, unlike dense layers (Selvaraju et al. 2017). The gradients are averaged over the spatial dimensions (i.e., we compute the global average pooling), producing a vector containing the importance weights ρ_k^u for each feature map:

$$\rho_k^u = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \frac{\partial y^u}{\partial A_{ij}^k}, \tag{3.56}$$

where H, W are the height and width of A^k, respectively (Selvaraju et al. 2017).

To obtain the Grad-CAM heatmap, we calculate a linear combination of the importance weights and the activation maps A followed by ReLU

$$ReLU\left(\sum_k \rho_k^u A^k\right) \tag{3.57}$$

(Selvaraju et al. 2017). Taking the ReLU will make all negative pixels zero. We are only interested in the positive pixels, i.e., the intensity of these pixels should be increased to increase $\hat{y}^u$ (Selvaraju et al. 2017). The negative pixels likely contain irrelevant information for predicting class u and thus can be zeroed (Selvaraju et al. 2017). Finally, since the size of the produced heatmap will correspond to the spatial resolution of A (which is likely very small compared to the input), therefore the heatmap needs to be interpolated to the size of the input (Selvaraju et al. 2017).

4 Results

This chapter summarizes the results of the article (Patron et al. 2022). Before the following chapters, the reader should see the article in appendix A. In the study, we used overall spiking as the model ground-truth; the inter-rater reliability for the overall spiking was 0.48 (κ). The reliability analysis results are presented in Table 5.

Table 5: Intra and inter-rater reliability (κ) for the original ratings including unsure, binary ratings with unsure omitted (denoted with o), and overall spiking (denoted with OR) with 95% confidence interval (CI). Adapted from Patron et al. 2022.

	Intra-rater κ (expert 1)	Intra-rater κ (expert 2)	Inter-rater κ
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)

In experiment A, we compared 913 knees with and without tibial spiking using the Wilcoxon–Mann–Whitney test. The sample means for spiking and control groups with U-test p-values are detailed in Table 6. The significant variables below the Bonferroni corrected $p < 0.001$ were KL-grade and BMI, and the variables below the corrected $p < 0.01$ were the medial JSN and lateral tibia compartment osteophyte scores (Patron et al. 2022).

In experiment B, we fine-tuned a pre-trained CNN for identifying tibial spiking. We evaluated the model using 191 samples. The model produced an accuracy of 0.869, sensitivity of 0.909, and specificity of 0.750 (Patron et al. 2022). More details on the model performance can be seen in Table 7.

Table 6: Mean values for spiking and control groups with U-test p-values (adapted from Patron et al. 2022).

	Spiking	Control	p-value
KL-grade	1.11	0.70	$p < 0.001$
WOMAC knee pain	2.14	1.62	0.032
BMI	29.09	27.46	$p < 0.001$
Medial JSN	0.38	0.25	$p < 0.001$
Lateral JSN	0.05	0.04	0.440
Tibia medial osteophytes	0.59	0.41	0.009
Tibia lateral osteophytes	0.40	0.21	$p < 0.001$
Femur medial osteophytes	0.48	0.27	0.007
Femur lateral osteophytes	0.41	0.22	0.009

Table 7: Classifier performance metrics computed for each data subsets (adapted from Patron et al. 2022).

	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

5 Discussion

In the study, the knees were assessed for tibial spiking by subjective visual examination (Patron et al. 2022). The subjectivity of the ratings is a concern since the raters were not blinded to the details of the hypothesis on the tibial spiking being an early sign of knee OA. The knowledge of the hypothesis could have influenced the ratings, as the raters could have looked for confirmation from the other signs of knee OA. The experts were blinded to the radiographic assessments extracted from the datasets (Patron et al. 2022). However, the radiographs themselves contain information on JSN and osteophyte formation. The bias could have been reduced by only providing an image of the tibial tubercles (e.g., by cropping) instead of showing the whole knee joint to the experts.

We found the samples from the spiking group to have significantly higher KL-grade compared to the controls (Patron et al. 2022). This finding is in accord with Donnelly et al. 1996 and corresponds with the reported association with knee OA (Reiff, Heron, and Stoker 1991; Unlu et al. 2006; Hayeri et al. 2010). We observed more severe medial JSN ratings in the spiking group (Patron et al. 2022), which supports the findings of Donnelly et al. 1996, who reported an association between the lateral tubercle angulation and medial JSN and Unlu et al. 2006 who found medial spine height to correlate with medial cartilage defects. We, however, found no difference in the lateral JSN (Patron et al. 2022), contradicting Donnelly et al. 1996, who found the height of the medial tubercle to associate with the lateral JSN.

Additionally, we found an association between tibial spiking and osteophyte formation on the lateral tibia compartment (Patron et al. 2022). Our results agree with Donnelly et al. 1996; Unlu et al. 2006; Hayeri et al. 2010. We could not confirm the association between medial osteophytes and tibial spiking reported by Donnelly et al. 1996. Alexander 1990 suggested that tibial spiking could be a form of osteophyte formation. Our findings provide support for the association between tibial spiking and osteophyte formation; therefore, the conjecture might be correct. However, the underlying causality of spiking tibial tubercles is unknown and should be investigated.

While the results for experiment B were promising, this study should be considered a proof-

of-concept rather than a deployment-ready method. Firstly, the sample size of the study can be considered limited. Despite the small sample size, the developed model was able to generalize quite well (Patron et al. 2022). However, the model should be evaluated with another dataset since the ground-truth was only based on subjective visual assessment and, therefore, might produce biased predictions. A grid-search was utilized for tuning the model hyperparameters (Patron et al. 2022), although random search provides better efficiency for optimizing neural network hyperparameters (Bergstra and Bengio 2012). Another point of concern is the reliability of the ground-truth. Although the intra-rater reliability of the ground-truth labels was similar to KL-grading (Patron et al. 2022), the label noise in the training data should be ideally zero. Deep learning algorithms seem resistant to label noise for large datasets (Rolnick et al. 2017). Our dataset was, however, relatively small. Especially the test data should be as clean as possible to produce reliable results.

The model has a lower probability of grading non-spiking knees correctly than spiking knees, indicated by lower specificity (Patron et al. 2022). The discrepancy might be explainable by the class imbalance. Additionally, due to including the unsure ratings in the non-spiking class, the non-spiking class might have included a greater quantity of borderline samples that could have been classified as doubtful spiking (Patron et al. 2022), which could have contributed to the lower specificity. The imbalance could be perhaps alleviated by using class weighting in the loss function or equalizing the sampling of the different class examples during training. Additionally, a different metric could have been chosen as the criterion in the hyperparameter optimization.

The article provided Grad-CAM visualization examples to demonstrate decision-making of the model (Patron et al. 2022). The heatmaps indicated that the model might be reliant on JSN for assessing tibial spiking, which could cause misclassifications (Patron et al. 2022). Perhaps a training scheme, where either medial or lateral condyle would be randomly covered, could force the network to not rely on features outside the tubercles themselves. Additionally, the Grad-CAM responses were high in bright areas of the radiographs, which might be indicative of sclerosis. The Grad-CAM heatmaps should be evaluated by medical experts to see if the model has learned sensible features for identifying tibial spiking.

Future studies should collect a larger quantity of samples to address the generalizability

concerns. High care should be placed on the creation of ground-truth labels, especially for clinical applications the data should consist of high-quality samples with high reliability. Ideally, the grading should be performed by a committee of experienced raters. Universally shared criteria for tibial spiking rating does not exist, to our knowledge. Therefore an atlas for grading tibial spiking would be vital for the adoption of the feature.

6 Conclusion

The article by Patron et al. 2022 examined a hypothesis on the spiking of tibial tubercles as an early sign of knee osteoarthritis. The results of the study were found to support the hypothesis (Patron et al. 2022). Therefore, adopting the feature for assessing early-onset knee osteoarthritis holds promise. The results, however, should be verified by future work due to limitations of the study by Patron et al. 2022. Currently, little is known about tibial spiking in relation to osteoarthritis. Possible future research questions include the importance of the finding, the causality, and the relation to other features of knee osteoarthritis. Additionally, a classifier based on convolutional neural networks was developed and evaluated for tibial spiking detection (Patron et al. 2022). The model was able to generalize regardless of the small number of samples available. However, the generalizability of the model should be validated using another dataset to rule out possible bias.

Bibliography

Alexander, Colin J. 1990. “Osteoarthritis: a review of old myths and current concepts.” *Skeletal radiology* 19 (5): 327–333.

Antony, Joseph, Kevin McGuinness, Noel E O’Connor, and Kieran Moran. 2016. “Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks.” In *2016 23rd International Conference on Pattern Recognition (ICPR),* 1195–1200. https://doi.org/10.1109/ICPR.2016.7899799.

Arokoski, JPA, P. Manninen, H. Kröger, M. Heliövaara, E. Nykyri, and O. Impivaara. 2007. “Hip and Knee Pain and Osteoarthritis.” *In book: Musculoskeletal disorders and diseases in Finland. Results of the Health 2000 Survey,* 37–41.

Banerjee, Mousumi, Michelle Capozzoli, Laura McSweeney, and Debajyoti Sinha. 1999. “Beyond kappa: A review of interrater agreement measures.” *Canadian Journal of Statistics* 27 (1): 3–23. https://doi.org/https://doi.org/10.2307/3315487.

Bastick, Alex N, Jurgen Damen, Rintje Agricola, Reinoud W Brouwer, Patrick JE Bindels, and Sita MA Bierma-Zeinstra. 2017. “Characteristics associated with joint replacement in early symptomatic knee or hip osteoarthritis: 6-year results from a nationwide prospective cohort study (CHECK).” 67 (663): e724–e731. https://doi.org/10.3399/bjgp17X692165.

Bengio, Yoshua. 2012. “Deep Learning of Representations for Unsupervised and Transfer Learning.” In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning,* edited by Isabelle Guyon, Gideon Dror, Vincent Lemaire, Graham Taylor, and Daniel Silver, 27:17–36. Proceedings of Machine Learning Research. Bellevue, Washington, USA: PMLR.

Bergstra, James, and Yoshua Bengio. 2012. “Random search for hyper-parameter optimization.” *Journal of machine learning research* 13 (2).

Chen, Pingjun, Linlin Gao, Xiaoshuang Shi, Kyle Allen, and Lin Yang. 2019. “Fully Automatic Knee Osteoarthritis Severity Grading Using Deep Neural Networks with a Novel Ordinal Loss.” *Computerized Medical Imaging and Graphics* 75:84–92. https://doi.org/10.1016/j.compmedimag.2019.06.002.

Cohen, Jacob. 1960. “A Coefficient of Agreement for Nominal Scales.” *Educational and Psychological Measurement* 20 (1): 37–46. https://doi.org/10.1177/001316446002000104.

———. 1968. “Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit.” *Psychological Bulletin* 70:213–220. https://doi.org/https://doi.org/10.1037/h0026256.

Culvenor, Adam G, Cathrine N Engen, Britt Elin Øiestad, Lars Engebretsen, and May Arna Risberg. 2015. “Defining the presence of radiographic knee osteoarthritis: a comparison between the Kellgren and Lawrence system and OARSI atlas criteria.” *Knee Surgery, Sports Traumatology, Arthroscopy* 23 (12): 3532–3539.

Dauphin, Yann, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. 2014. *Identifying and attacking the saddle point problem in high-dimensional non-convex optimization.* https://doi.org/10.48550/ARXIV.1406.2572.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. “ImageNet: A Large-Scale Hierarchical Image Database.” In *2009 IEEE conference on computer vision and pattern recognition,* 248–255.

Donnelly, S, D J Hart, D V Doyle, and T D Spector. 1996. “Spiking of the tibial tubercles–a radiological feature of osteoarthritis?” *Annals of the Rheumatic Diseases* 55 (2): 105–108. https://doi.org/10.1136/ard.55.2.105.

Duchi, John, Elad Hazan, and Yoram Singer. 2011. “Adaptive Subgradient Methods for Online Learning and Stochastic Optimization.” *Journal of Machine Learning Research* 12 (61): 2121–2159.

Felson, David T, and Richard Hodgson. 2014. “Identifying and treating preclinical and early osteoarthritis.” *Rheumatic Disease Clinics* 40 (4): 699–710.

Fukushima, Kunihiko. 1980. “Neocognition: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position.” *Biol. Cybern.* 36:193–202. https://doi.org/10.1007/BF00344251.

Fukushima, Kunihiko, and Sei Miyake. 1982. “Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Visual Pattern Recognition.” In *Competition and Cooperation in Neural Nets,* edited by Shun-ichi Amari and Michael A. Arbib, 267–285. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-642-46466-9.

Gonzalez, Rafael C., and Richard E. Woods. 2018. *Digital Image Processing*. 4th ed. 134–140. Pearson.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. Http://www.deeplearningbook.org. MIT Press.

Gossec, L., J.M. Jordan, S.A. Mazzuca, M.-A. Lam, M.E. Suarez-Almazor, J.B. Renner, M.A. Lopez-Olivo, G. Hawker, M. Dougados, and J.F. Maillefert. 2008. “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.” *Osteoarthritis and Cartilage* 16 (7): 742–748. https://doi.org/https://doi.org/10.1016/j.joca.2008.02.021.

Guermazi, Ali, Jingbo Niu, Daichi Hayashi, Frank W Roemer, Martin Englund, Tuhina Neogi, Piran Aliabadi, Christine E McLennan, and David T Felson. 2012. “Prevalence of abnormalities in knees detected by MRI in adults without knee osteoarthritis: population based observational study (Framingham Osteoarthritis Study).” *BMJ* 345. https://doi.org/10.1136/bmj.e5339.

Guermazi, Ali, Frank W Roemer, Deborah Burstein, and Daichi Hayashi. 2011. “Why radiography should no longer be considered a surrogate outcome measure for longitudinal assessment of cartilage in knee osteoarthritis.” *Arthritis research & therapy* 13 (6): 1–11.

Günther, Klaus P., and Yi Sun. 1999. “Reliability of radiographic assessment in hip and knee osteoarthritis.” *Osteoarthritis and Cartilage* 7 (2): 239–246. https://doi.org/https://doi.org/10.1053/joca.1998.0152.

Hayashi, D., D.T. Felson, J. Niu, D.J. Hunter, F.W. Roemer, P. Aliabadi, and A. Guermazi. 2014. “Pre-radiographic osteoarthritic changes are highly prevalent in the medial patella and medial posterior femur in older persons: Framingham OA study.” *Osteoarthritis and Cartilage* 22 (1): 76–83. https://doi.org/https://doi.org/10.1016/j.joca.2013.10.007.

Hayeri, Mohammad Reza, Masoud Shiehmorteza, Debra J Trudell, Tori Heflin, and Donald Resnick. 2010. “Proximal tibial osteophytes and their relationship with the height of the tibial spines of the intercondylar eminence: paleopathological study.” *Skeletal radiology* 39 (9): 877–881.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. “Deep Residual Learning for Image Recognition.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Hebb, D. O. 1949. *The Organization of Behavior.* New York: Wiley.

Hinton, G. 2012. “Neural Networks for Machine Learning.” Coursera Lecture, https://www.cs.toronto.edu/~hinton/coursera/lecture6/lec6.pdf.

Huang, Gao, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. “Densely Connected Convolutional Networks.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Ioffe, Sergey, and Christian Szegedy. 2015. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.” In *Proceedings of the 32nd International Conference on Machine Learning,* edited by Francis Bach and David Blei, 37:448–456. Proceedings of Machine Learning Research. Lille, France: PMLR.

Kellgren, J. H., and J. S. Lawrence. 1957. “Radiological Assessment of Osteo-Arthrosis.” *Annals of the Rheumatic Diseases* 16 (4): 494–502. https://doi.org/10.1136/ard.16.4.494.

Kingma, Diederik P., and Jimmy Ba. 2014. “Adam: A Method for Stochastic Optimization,” https://doi.org/10.48550/ARXIV.1412.6980.

Kohn, Mark D., Adam A. Sassoon, and Navin D. Fernando. 2016. “Classifications in Brief: Kellgren-Lawrence Classification of Osteoarthritis.” *Clinical Orthopaedics & Related Research* 474 (8): 1886–1893. https://doi.org/10.1007/s11999-016-4732-4.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton. 2012. “ImageNet Classification with Deep Convolutional Neural Networks.” In *Advances in Neural Information Processing Systems,* edited by F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, vol. 25. Curran Associates, Inc.

LeCun, Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. 1989. “Backpropagation Applied to Handwritten Zip Code Recognition.” *Neural Computation* 1 (4): 541–551. https://doi.org/10.1162/neco.1989.1.4.541.

Lehmann, Erich L., and Howard J. M. D’Abrera. 1998. *Nonparametrics: Statistical Methods Based on Ranks.* Revised 1st edition. 18–23. Prentice Hall.

Lin, Min, Qiang Chen, and Shuicheng Yan. 2013. *Network In Network.* https://doi.org/10.48550/ARXIV.1312.4400.

Linnainmaa, Seppo. 1970. “The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors.” Master’s Thesis (in Finnish), Univ. Helsinki.

———. 1976. “Taylor Expansion of the Accumulated Rounding Error.” *BIT Numerical Mathematics* 16 (2): 146–160. https://doi.org/10.1007/BF01931367.

Litjens, Geert, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. 2017. “A survey on deep learning in medical image analysis.” *Medical Image Analysis* 42:60–88. https://doi.org/https://doi.org/10.1016/j.media.2017.07.005.

Mann, H. B., and D. R. Whitney. 1947. “On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other.” *The Annals of Mathematical Statistics* 18 (1): 50–60.

McCulloch, Warren S., and Walter Pitts. 1943. “A Logical Calculus of the Ideas Immanent in Nervous Activity.” *The bulletin of mathematical biophysics* 5 (4): 115–133. https://doi.org/10.1007/BF02478259.

Minsky, Marvin L., and Seymour A. Papert. 1969. *Perceptrons.* Cambridge, MA, USA: MIT Press.

Nachar, Nadim, et al. 2008. “The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution.” *Tutorials in quantitative Methods for Psychology* 4 (1): 13–20.

Nair, V., and G. E. Hinton. 2010. “Rectified Linear Units Improve Restricted Boltzmann Machines.” *In Proceedings of the 27th International Conference on Machine Learning (ICML-10),* 807–814.

Novick, Melvin R. 1966. “The axioms and principal results of classical test theory.” *Journal of Mathematical Psychology* 3 (1): 1–18. https://doi.org/https://doi.org/10.1016/0022-2496(66)90002-2.

Oka, H., S. Muraki, T. Akune, A. Mabuchi, T. Suzuki, H. Yoshida, S. Yamamoto, K. Nakamura, N. Yoshimura, and H. Kawaguchi. 2008. “Fully automatic quantification of knee osteoarthritis severity on plain radiographs.” *Osteoarthritis and Cartilage* 16 (11): 1300–1306. https://doi.org/https://doi.org/10.1016/j.joca.2008.03.011.

Oliveria, Susan A., David T. Felson, John I. Reed, Priscilla A. Cirillo, and Alexander M. Walker. 1995. “Incidence of symptomatic hand, hip, and knee osteoarthritis among patients in a health maintenance organization.” *Arthritis & Rheumatism* 38 (8): 1134–1141. https://doi.org/https://doi.org/10.1002/art.1780380817.

Pan, Sinno Jialin, and Qiang Yang. 2010. “A Survey on Transfer Learning.” *IEEE Transactions on Knowledge and Data Engineering* 22 (10): 1345–1359. https://doi.org/10.1109/TKDE.2009.191.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, et al. 2019. “PyTorch: An Imperative Style, High-Performance Deep Learning Library.” In *Advances in Neural Information Processing Systems 32,* edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, 8024–8035. Curran Associates, Inc.

Patron, Anri, Leevi Annala, Olli Lainiala, Juha Paloneva, and Sami Äyrämö. 2022. “An Automatic Method for Assessing Spiking of Tibial Tubercles Associated with Knee Osteoarthritis.” *Diagnostics* 12 (11). https://doi.org/10.3390/diagnostics12112603.

Polyak, B.T. 1964. “Some methods of speeding up the convergence of iteration methods.” *USSR Computational Mathematics and Mathematical Physics* 4 (5): 1–17. https://doi.org/https://doi.org/10.1016/0041-5553(64)90137-5.

Reiff, DB, CW Heron, and DJ Stoker. 1991. “Spiking of the tubercles of the intercondylar eminence of the tibial plateau in osteoarthritis.” *The British Journal of Radiology* 64 (766): 915–917.

Resnick, Donald, and Gen Niwayama. 1988. *Diagnosis of Bone and Joint Disorders.* 2nd ed. p. 1446. W.B. Saunders.

Riesenhuber, Maximilian, and Tomaso Poggio. 1999. “Hierarchical models of object recognition in cortex.” *Nature neuroscience* 2 (11): 1019–1025.

Rolnick, David, Andreas Veit, Serge Belongie, and Nir Shavit. 2017. “Deep learning is robust to massive label noise.” *arXiv preprint arXiv:1705.10694.*

Rosenblatt, Frank. 1958. “The Perceptron: a Probabilistic Model for Information Storage and Organization in the Brain.” *Psychological Review* 65 (6): 386.

Ruder, Sebastian. 2016. “An overview of gradient descent optimization algorithms.” *arXiv preprint arXiv:1609.04747.*

Rumelhart, David E, Geoffrey E Hinton, and Ronald J Williams. 1986. “Learning representations by back-propagating errors.” *nature* 323 (6088): 533–536.

Runkler, Thomas A. 2016. *Data Analytics.* 2nd ed. 91–93. Springer.

Scherer, Dominik, Andreas Müller, and Sven Behnke. 2010. “Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition.” In *International conference on artificial neural networks,* edited by Konstantinos Diamantaras, Wlodek Duch, and Lazaros S. Iliadis, 92–101. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-642-15825-4.

Schmidhuber, Jürgen. 2015. “Deep learning in neural networks: An overview.” *Neural Networks* 61:85–117. https://doi.org/https://doi.org/10.1016/j.neunet.2014.09.003.

Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. “Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization.” In *Proceedings of the IEEE International Conference on Computer Vision (ICCV).*

Shamir, Lior, Shari M. Ling, William W. Scott, Angelo Bos, Nikita Orlov, Tomasz J. Macura, D. Mark Eckley, Luigi Ferrucci, and Ilya G. Goldberg. 2009. "Knee X-Ray Image Analysis Method for Automated Detection of Osteoarthritis." *IEEE Transactions on Biomedical Engineering* 56 (2): 407–415. https://doi.org/10.1109/TBME.2008.2006025.

Simonyan, Karen, and Andrew Zisserman. 2014. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556.*

Student. 1908. "The Probable Error of a Mean." *Biometrika* 6 (1): 1–25. Accessed October 29, 2022.

Sutton, David. 1987. *A textbook of radiology and imaging.* 4th ed. p. 113. Churchill Livingstone.

Sutton, Richard. 1986. "Two problems with back propagation and other steepest descent learning procedures for networks." In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society, 1986,* 823–832.

Swagerty Jr, Daniel L, and Deborah Hellinger. 2001. "Radiographic assessment of osteoarthritis." *American Family Physician* 64 (2): 279.

Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. "Going Deeper With Convolutions." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Szeliski, Richard. 2022. *Computer Vision: Algorithms and Applications.* 2nd ed. Springer.

Tiulpin, Aleksei, Jérôme Thevenot, Esa Rahtu, Petri Lehenkari, and Simo Saarakkala. 2018. "Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach." *Scientific reports* 8 (1): 1–10.

Unlu, Zeliha, Serdar Tarhan, Cihan Goktan, and Cigdem Tuzun. 2006. "The correlation between magnetic resonance detected cartilage defects and spiking of tibial tubercles in osteoarthritis of the knee joint." *Acta Medica Okayama* 60 (4): 207–214.

Weiss, Karl, Taghi M. Khoshgoftaar, and DingDing Wang. 2016. "A Survey of Transfer Learning." *Journal of Big Data* 3 (1): 9. https://doi.org/10.1186/s40537-016-0043-6.

Weng, J., N. Ahuja, and T.S. Huang. 1992. “Cresceptron: a self-organizing neural network which grows adaptively.” In *[Proceedings 1992] IJCNN International Joint Conference on Neural Networks,* vol. 1, 576–581 vol.1. https://doi.org/10.1109/IJCNN.1992.287150.

———. 1997. “Learning recognition and segmentation using the Cresceptron.” *International Journal of Computer Vision* 25 (2): 109–143.

Werbos, Paul. 1974. “Beyond regression:" new tools for prediction and analysis in the behavioral sciences.” PhD diss., Harvard University.

Wilcoxon, Frank. 1945. “Individual Comparisons by Ranking Methods.” *Biometrics Bulletin* 1 (6): 80–83.

Xie, Saining, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. 2017. “Aggregated Residual Transformations for Deep Neural Networks.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Yeoh, Pauline Shan Qing, Khin Wee Lai, Siew Li Goh, Khairunnisa Hasikin, Yan Chai Hum, Yee Kai Tee, and Samiappan Dhanalakshmi. 2021. “Emergence of Deep Learning in Knee Osteoarthritis Diagnosis.” Edited by Bai Yuan Ding. *Computational Intelligence and Neuroscience* 2021:1–20. https://doi.org/10.1155/2021/4931437.

Yerushalmy, Jacob. 1947. “Statistical Problems in Assessing Methods of Medical Diagnosis, with Special Reference to X-Ray Techniques.” *Public Health Reports (1896-1970)* 62 (40): 1432–1449.

Yosinski, Jason, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. “How transferable are features in deep neural networks?” In *Advances in Neural Information Processing Systems,* edited by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, vol. 27. Curran Associates, Inc.

Zeiler, Matthew D., and Rob Fergus. 2014. “Visualizing and Understanding Convolutional Networks.” In *European conference on computer vision,* edited by David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, 818–833. Cham: Springer International Publishing. ISBN: 978-3-319-10590-1.

Zhang, Yuqing, and Joanne M Jordan. 2010. “Epidemiology of osteoarthritis.” *Clinics in geriatric medicine* 26 (3): 355–369.

Zhou, Bolei, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. “Learning Deep Features for Discriminative Localization.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2921–2929. https://doi.org/10.1109/CVPR.2016.319.

Appendices

A Article: Automatic Method for Assessing Spiking of Tibial Tubercles Associated with Knee Osteoarthritis

A preprint version of the article "An Automatic Method for Assessing Spiking of Tibial Tubercles Associated with Knee Osteoarthritis", reproduced with a permission (Patron et al. 2022).

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; leevi.a.annala@jyu.fi (L.A.); sami.ayramo@jyu.fi (S.Ä.)
[2] Department of Radiology, Tampere University Hospital, 33520 Tampere, Finland; olli.lainiala@tuni.fi (O.L.)
[3] Faculty of Medicine and Health Technologies, Tampere University, 33520 Tampere, Finland
[4] Department of Surgery, Central Finland Healthcare District, 40620 Jyväskylä, Finland; juha.paloneva@ksshp.fi (J.P.)
[5] Institute of Clinical Medicine, University of Eastern Finland, 70210 Kuopio, Finland
* Correspondence: anri.a.patron@jyu.fi (A.P.)

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. *Preprints* **2022**, *12*, 0. https://doi.org/

Academic Editor: Ming-Huwi Horng

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 [8,11]:

- 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 [9]. Osteoarthritis Research Society International (OARSI) developed another radiographic atlas for OA [10]. 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.

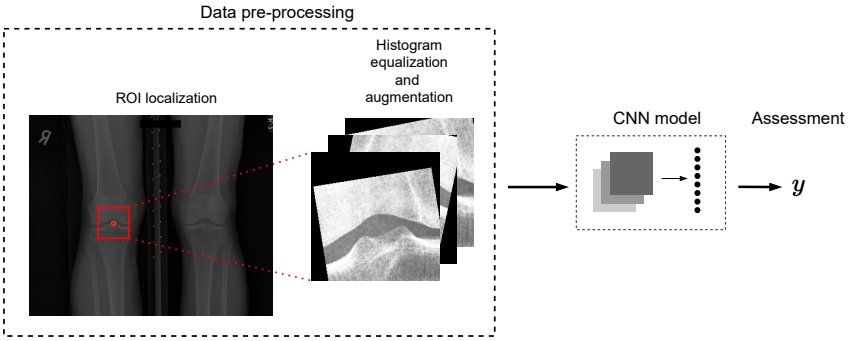


Figure 2. Tibial spiking assessment pipeline: the bilateral PA view radiographs are localized to the ROI, histogram equalization, augmentations (when applicable), and normalization are performed before feeding the data to the model.

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 [44] 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 [41].

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. [42]. 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 [42]. We used the standard cross-entropy as the loss function, optimized with Adam [43] with following parameters $\alpha = 0.0002$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$. PyTorch 1.11.0 [44] 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	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128, C = 32 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
BN2	28×28	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256, C = 32 \\ 1 \times 1, 512 \end{bmatrix} \times 4$

Table 3. *Cont.*

Block	Output Size	Architecture
BN3	14×14	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512, C = 32 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$
BN4	7×7	$\begin{bmatrix} 1 \times 1, 1024 \\ 3 \times 3\ 1024, C = 32 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$
	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.

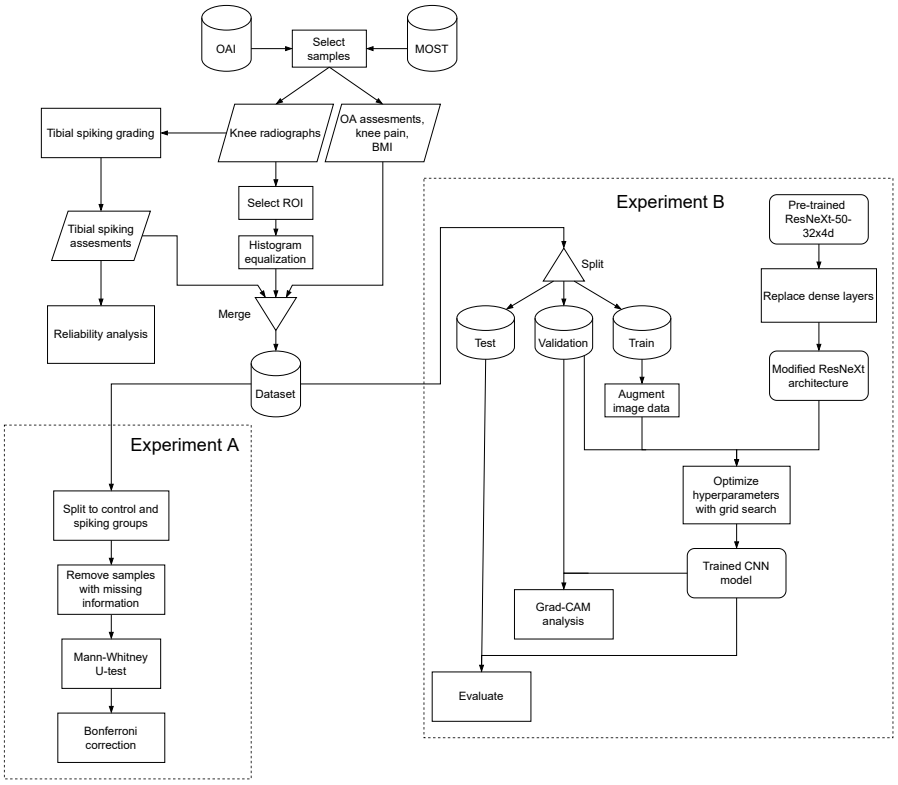


Figure 3. Flowchart of the study methodology. The processes or methods are denoted with rectangles, datasets with cylinders, data with parallelograms, and the models with rectangles with rounded corners.

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

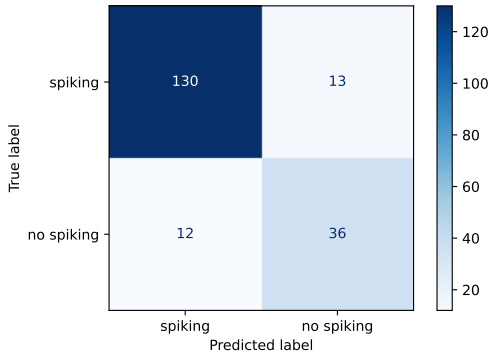


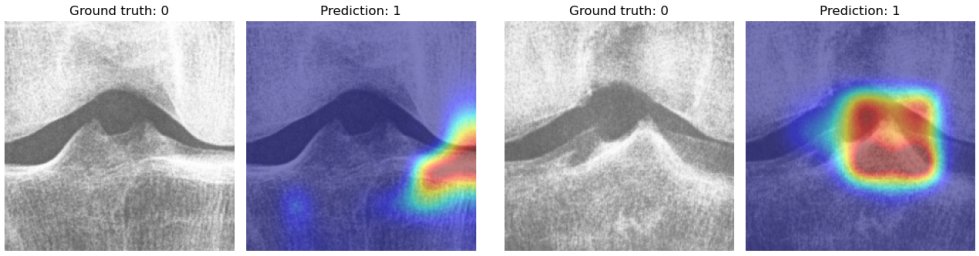
Figure 4. Confusion matrix for the test dataset.

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.

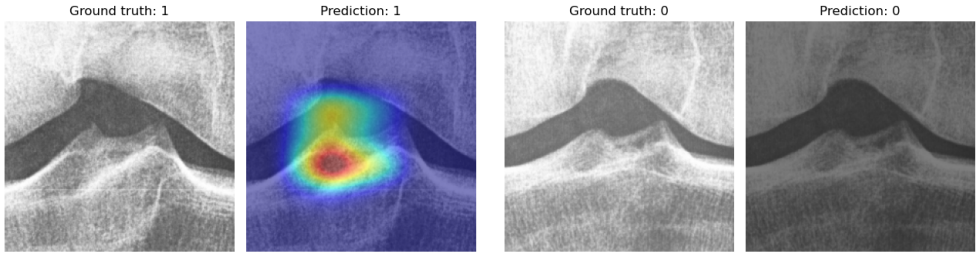
We visualized the model predictions for the validation data using Grad-CAM. By visualizing the failed predictions of the model, we can gain information on the reasons behind the failures, e.g., in Figure 5a, the network concentrates on the narrow joint space instead of on the tubercles. However, in Figure 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.; 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. https://doi.org/10.1016/S0140-6736(12)61729-2.
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. https://doi.org/10.1136/annrheumdis-2013-204763.
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. https://doi.org/10.1016/j.joca.2017.07.022.
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.
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*. https://doi.org/10.1136/bmj.e5339.
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. https://doi.org/10.1016/j.joca.2013.10.007.
8. Kellgren, J.H.; Lawrence, J.S. Radiological Assessment of Osteo-Arthrosis. *Ann. Rheum. Dis.* **1957**, *16*, 494–502. https://doi.org/10.1136/ard.16.4.494.
9. Spector, T.D.; Hochberg, M.C. Methodological problems in the epidemiological study of osteoarthritis. *Ann. Rheum. Dis.* **1994**, *53*, 143–146. https://doi.org/10.1136/ard.53.2.143.
10. Altman, R.; Gold, G. Atlas of individual radiographic features in osteoarthritis, revised. *Osteoarthr. Cartil.* **2007**, *15*, A1–A56. https://doi.org/10.1016/j.joca.2006.11.009.
11. 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. https://doi.org/10.1007/s11999-016-4732-4.
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.
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.
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. https://doi.org/10.1162/neco.1989.1.4.541.
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. https://doi.org/10.1038/nature21056.
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. https://doi.org/10.1016/j.knee.2019.11.020.
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. https://doi.org/10.1109/ICPR.2016.7899799.
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*, 1–20. https://doi.org/10.1155/2021/4931437.
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.

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. https://doi.org/10.1016/j.compmedimag.2019.06.002.
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. https://doi.org/10.3390/jcm9103341.
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.
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. https://doi.org/10.1136/ard.55.2.105.
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.
27. Eckstein, F.; Wirth, W.; Nevitt, M.C. Recent Advances in Osteoarthritis Imaging—The Osteoarthritis Initiative. *Nat. Rev. Rheumatol.* **2012**, *8*, 622–630. https://doi.org/10.1038/nrrheum.2012.113.
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 J. Inj. Funct. Rehabil.* **2013**, *5*, 647–654. https://doi.org/10.1016/j.pmrj.2013.04.014.
29. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. https://doi.org/10.1177/001316446002000104.
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.
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. https://doi.org/10.1038/s41592-019-0686-2.
34. Armstrong, R.A. When to use the Bonferroni correction. *Ophthalmic Physiol. Opt.* **2014**, *34*, 502–508. https://doi.org/10.1111/opo.12131.
35. Mason, D.; Scaramallion.; mrBean Bremen.; rhaxton.; Suever, J.; Vanessasaurus.; Orfanos, D.P.; Lemaitre, G.; Panchal, A.; Rothberg, A.; et al. *Pydicom/Pydicom*; Pydicom 2.2.2; Version 2.2.2. Available online: (accessed on 26 October 2022) https://github.com/pydicom/pydicom.
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. https://doi.org/10.1007/s11263-015-0816-y.
41. 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, Montreal, Canada, 8–13 December; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K., Eds.; Curran Associates, Inc: Red Hook, NY, USA, 2014; Volume 27.
42. 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, June 26–July 1, 2016.
43. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980v9.https://doi.org/10.48550/ARXIV.1412.6980.
44. 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 *Advances in Neural Information Processing Systems 32*; 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.
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.

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. https://doi.org/10.1016/j.joca.2008.02.021.
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.
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.