Bachelor thesis

Johann Strunck

# Deep learning based grading of motionartifacts in HR-pQCT

4. September 2023

supervised by:
Prof. Dr.-Ing. Tobias Knopp
Dr. rer. nat. Martin Hofmann

Hamburg University of Technology
Institute for Biomedical Imaging
Schwarzenbergstraße 95
21073 Hamburg

University Medical Center Hamburg-Eppendorf
Section for Biomedical Imaging
Martinistraße 52
20246 Hamburg

Ich versichere an Eides statt, die vorliegende Arbeit selbstständig und nur unter Benutzung der angegebenen Quellen und Hilfsmittel angefertigt zu haben.

Hamburg, den ??.??.2010

# Inhaltsverzeichnis

# Summary

A common issue of High-resolution peripheral quantitative computed tomography (HR-pQCT) scans is the appearance of motion artifacts in Images. These artifacts can appear due to involuntary movements like twitches and spasms. Since a scan can last between 3-4 minutes its hard not to move for such a long time. Depending on the severeness of those artifacts in the resulting image, it might not be sufficient for medical use and a re scan is necessary. The decision of the severity is decided by a professional which gives the image a number from 1 to 5, where 1 equals no motion artifacts and 5 equals severe motion artifacts. The descission of severity can be biased and varies depending on the reviewing person. Studies showed that operators disagree in up to 30% of all cases. To support the descission of the operation there have been approaches by [1] and [2] to improve the confidence of the result. both methodes can be performed with the absence of a operator and results of [1] show that with crossvalidation of a operator a Convolutional Neural Network(CNN) can reach a higher accuracy than the cross validation of two operators without a CNN. The CNN still has a considerable error rate. In this paper we will propose a new CNN structure which uses state of the art methods to detect the severity of motion Scores in CT scans.

# 1 Introduction

High-resolution peripheral quantitative computed tomography (HR-pQCT) is a specialized non-invasive imaging technique that provides detailed and accurate three-dimensional images of bone and tissue microarchitecture at the peripheral skeletal sites. In our paper we focus on the radius and thibia. This advanced imaging modality offers several distinct advantages. One advantage is that HR-pQCT provides high resolution images that allow a thorough assessment of a scanned bone micro architecture. It offers precise measurement of bone mineral density(BMD) and geometric parameters such as trabecular thickness and cortical thickness. HR-pQCT has applications in both clinical and research setting and can help make more informed decissions about patient management and treatment strategies. It can provide insight into fracture or the risk of its occurrence. HR-pQCT imaging requires the patient to remain still during the scan to avoid motion artifacts. This can be challenging for certain patient populations, such as children or individuals with limited mobility. Especially with the XtremCT (Scanco Media AG),a device that was used to generated the test and training data for this paper.The XtremCT takes about 3-4 Minutes for a scan which makes it hard to hold the chosen extremity in place for such a long time.

Depending on the severity of the motion artifact the scan must be repeated. To determine the severity of the scan [3] Introduced a scale from 1 (no visible motion artefacts) to 5 (significant horizontal streaks). In clinical studies it is commonly implemented, that scans with a grading of 4 or 5 have to be repeated to mitigate the effects of the motion artifacts. However, even with a standardized scoring system, motion scoring remains subjective, and operator agreement has shown to remain only moderate, even with intensive training [], studies have shown that operators disagree in up to 30% of all cases[look at bone]. Due to this issue a objective and standardized method is desirable for the grading process. Papers like [1] and [2] have taken an approach to find a suitable methode for grading motion artifacts, with partial success.

[2] proposes a method for objective detection of subject motion. The first method measures subject motion by comparing the projections at 0° and 180° since those projections are parallelized, the should be the mirror of each other when no subject motion occurred. Therefore by comparing the difference of those two parallel projections the subject motion can be approximated. This method is called Quantitative Motion Estimate and utilizes a similarity measure like the sum of squared intensity difference(SSD) or normalized cross correlation (NCC) to compare the mirrored parallel projection image at 0° and 180°.

In the last few years a tremendous interest in machine learning emerged. Many applications have found a way in our modern society [4]. They can be found in applications like speech to text or the recommendations on e-commerce websites. The field of medical imaging is no exception to this, including computer aided diagnosis, radiomics, and medical image analysis[]. With the emerging field of deep learning in computer vision it became more attention. In 2012 the ILSVRC2012 challange was won by the CNN based Network AlexNet outperforming the runner up with 10.8 percentage points lower top-5 error score of 15.3% . Since then the application of deep learning structures like CNNs have seen rapid growth in fields like medical imaging.

...[5]

A frequently occurring issue in medical imaging is the lack of training data. In our case we had 500 labeled examples of tibia and radius.

The data that we use in this paper to compare the different approaches was provided by the Üniversitäts Klinikum Eppendorf". All 500 provided scans where generated by the Scanco XtremCT.to ensure the correctness of the labeled data three doctors of the institution labeled the data together to ensure that the data was generalized and reduce the subjective influence of the single person

If we compare the data to the amount of data used in training state of the art networks like ImageNet with many Million Examples it's a small fraction. This comes on the one hand from the fact that the labeling task in medical imaging can just be performed my professionals and therefore the labeling process is costly and just a few people can do it. Another issue is the availability of data since patient data cant be accessed and used as easy. Therefore we need to find a way to augment the data so that we don't run into problems like overfitting the network or poor generalization of the network

[1] uses the power of CNNs to train a network for grading motion artifacts. Deep learning techniques, particularly CNNs, have revolutionized medical image analysis. CNNs are a CNNs excel at tasks such as image segmentation, object detection, and classification. With their ability to automatically learn intricate patterns and features from complex visual data, enabling more accurate and efficient diagnostic processes. Even with the simplistic structure chosen in [1] the Network is reaching a accuracy of ..., this can lead to the assumtion that a more suffisticated network might be more suitable . In this paper we will build on those findings and try to create a stronger network with state of the art CNN building blocks

In this paper we will introduce a Convolutional Neural Network Structure which is designed to predict the severity of the motion artifact and compare this Structure to the findings of [1] and [2].

# 2 Literature review

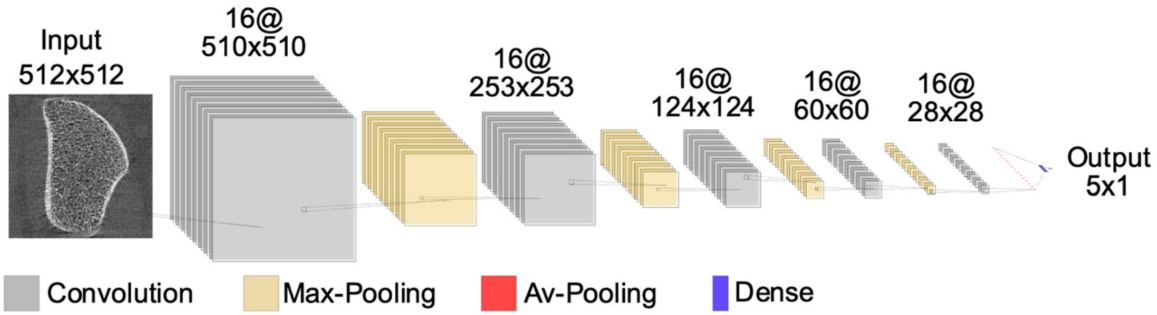


**Abbildung 2.1:** Network Structure Bone

[2] introduces a CNN in his paper which he trained to classify the severity of motion artifacts in HR-pQCT. The Network was traine with images(XtremeCT II, Scanco Media AG) from 90 patients. The size of the scans was 512 x 512 x 168. for the training he used 8 equally spaced images from every scan resulting in a database of 3312 images. The implemented Network strucuture begins with five alternating convolutional and max-pooling layers followed by a average pooling layer and is concluded by a dense layer. To extract the most important features max Pooling was used followed by a convolutional layer to aggregate them. Convolutional layer used (LeakyReLu) activation to enable faster learning while avoiding dead neurons. The classification was performed by a Fully Connected layer r integrating non-linear combinations of all high-level features using a standard Rectified Linear Unit (ReLU) activation function . On the output layer, a Softmax activation function provided an output that may be understood as a class probability, with each output larger than zero and their total always equal to one

[1] introduced a technique to measure the subject motion based on a calculating of the percent difference in each bone for the paired scans with and without visually apparent motion artifacts. Quantitative motion estimates for each motion degraded scan were calculated using two different image similarity measures: sum of squared differences(SSD) and normalized cross correlation(NCC) which were applied to the projections at 0°, 0.24° and 180° since they

are parallelized. The image at 180° was then mirrored with respect to the center of rotation on the detector to match the image at 0°.

$$SSD = \sum_{i}^{N} (f_i - g_i)^2 \text{ and } NCC = \frac{\sum_i^N (f_i \cdot g_i)}{\sqrt{\sum_i^N (f_i^2) \cdot \sum_i^N g_i^2}} \tag{2.1}$$

N is the number of voxels and $f_i$ and $g_i$ are the intensity values of the $i$th voxel. The greater the difference of the images the larger is the SSD and the larger the deviation from 1 for NCC. To account for the bone size, density positioning and other covariables the similarity measure at 0° and 180° was normalized by the similarity measure of the images at 0° and 0.24° Therefore the quantitave motion estimates(QMEs) can be calculated by:

$$QME_{SSD} = \frac{SSD_{0°vs180°}}{SSD_{0°vs0.24°}} \text{ and } QME_{NCC} = 100 \cdot \sqrt{\frac{NCC_{0°vs180°}}{NCC_{0°vs0.24°}}} \tag{2.2}$$

SSD- and NCC- based QME increases proportionally to the image quality. IT was detected that NCC is most suitable for detecting small affine translation, but is less sensitive to a large translation.

# 3
# Methods

Computer tomography(CT)uses a three dimensional radiographic imaging technique. The formation process begins with the acquisition of sequential radiographic projections captured over a range of angular positions around the object of interest. The crossectional field of view is reconstructed using established computational techniques based on the radon projection theory[6].Similar to simple radiography, the reconstructed image's intensity values represent the local radiographic attenuation: a material property related to the object's electron density (atomic number and mass density). The contrast between soft and mineralized tissue in CT is high, due to the relative electron-dense inorganic component (calcium hydroxyapatite) of the bone matrix.These principles capture high-resolution images of bone across a range of structural scales.

High resolution peripheral quantitative computer tomography(HR-pQCT) is a dedicated extremity imaging system developed to image bone microarchitecture in vivo at peripheral skeletal sites [7]. This imaging methode gives information on the bone structure and mineral density in bones like radius and tibia. This information allows the estimation of the bone strength and ability to resist fracture. The extraction of this information is possible do due to the high resolution of HR-pQCT. HR-pQCT is a low radiation dose methode,with a effective radiation dose at the distal radius and tibia of 3-5 µSV depending on the scanner generation. This is significantly less when comparing it to other common medical imaging techniques like chest X-rays with 100µSV or a hip CT scan with 2000-3000 µSV. The main field of studie using this technique is the field of osteoporosis. Osteoporosis causes bones to become weak and brittle, so brittle that a fall or even mild stresses such as bending over or coughing can cause a fracture. Six individual studies demonstrated that HR-pQCT variables could predict incident fractures in postmenopausal woman and old men, sugesting that the assesment of cortical and trabecular bone microarchitecture by HR-pQCT could improve overall fracture prediction. THere are still no widespread use of HR-pQCT since there is just a small number of devidces installed(fewer than 100 in mid 2020) therefore the main use of HR-pQCT is related to research [7].

In the recent years machine learning has become a popular in the research domain due to its versatile nature. It is a branch of artificial intelligence and computer science which focuses on the use of data and algorithms to imitate the way humans learn. A machine learning algorithm operates by processing data to identify patterns and connections. It learns from the data through a training process, adjusting its internal parameters to minimize a measure

called "loss."This loss quantifies the difference between the algorithm's predictions and the actual outcomes in the training data. The algorithm iteratively refines its parameters to reduce this loss, making its predictions more accurate. Once trained, the algorithm can apply its learning to new, unseen data, aiming to make predictions or classifications with minimized error based on its learned patterns. The process of refining the data is done by a so called optimizer. There are various different optimization algorithms like stochastic gradient descent, RMSprop, Momentum or Adam. These algorithms differ in how they calculate and apply parameter updates based on the gradients of the loss funciton with respect to the models parameters (weight and biases) during training. The optimizer seeks to find the optimal set of parameters by iteratively updating the parameters in a way that moves the model in the direciton of decreasing loss.

Gradient Descent is one of the most common ways to optimize a Neural Network [8]. It is a way to minimize a objective function by updating the parameters in the opposite direction of its gradient. There are three different variants of gradient descent that are differing in how much data they use to calculate the gradient of the cost function. There is Batch gradient descent, batch gradient descent and minibatch gradient descent. Mini batch gradient descent is a compromise between stochastic gradient descent and batch gradient descent since it just takes a small amount of data for calculating the gradient. This is more accurate then stochastic gradient descent since it calculates its gradient based on one example, therefore it isn't as fast. Compared Batch gradient descent it is faster since Batch gradient descent takes all the examples into account but therefor batch gradient descent is more accurate.Since minibatch gradient descent is a compromise of the other two variants and has a good balance between computation cost and accuracy it is the commonly used gradient descent algoritm. By calculating the gradient of the cost function and subtracting it from ... this is used to find a local minimum of the cost function, sometimes it can even find the global minimum but that is rarely the case. To ensure that the algorithm can find a minimum a learning rate $\eta$ is implemented. This makes it possible that the gradient descent method can come as close to a minimum of the cost function as possible. Even with the adjustment of the learningrate it is nearly impossible to reach the global minima since the learning algorithm usually gets caught in in a local minima or saddle point.

$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta) \tag{3.1}$$

Since the developement of gradient descent there has been a lot of progress and new, more efficient algorithms were developed.

adaptive moment estimation(Adam)[9] is a first-order gradient-based optimization technique or learning algortihm that is widely used, representing the latest trend in deep learning optimization. Adam is a deep learning strategy that was specifically designed for training deep neural networks. Its main selling points are it's memory efficiency and less computational cost compared to other optimization algorithms. It utilizes the squared gradients($v_t$) to scale

the learning rate like RMSprop and is similar to momentum by using the moving average of the gradient($m_t$).

$$m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t \tag{3.2}$$

$$v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2 \tag{3.3}$$

$m_t$ and $v_t$ are estimates of the first moment (the mean) and the second moment (the uncentered variance) these moving averages are initialized as vectors of 0's leading to moment estimates that are biased towards zero. The initialization bias can be counteracted resulting in bias-corrected estimates $\hat{m}_t$ and $\hat{v}_t$

$$\hat{m}_t = m_t/(1 - \beta_1^t) \tag{3.4}$$

$$\hat{v}_t = v_t/(1 - \beta_2^t) \tag{3.5}$$

Those moment estimates can then be used to update the parameters:

$$\omega_t = \omega_{t-1} - \alpha \cdot \hat{m}(\sqrt{\hat{v}_t} + \epsilon) \tag{3.6}$$

Good default settings are stepsize $\alpha = 0.001$, exponential decay rates for the moment estimates $\beta_1 = 0.9\beta_2 = 0.999$ and $\epsilon = 10^{-10}$.

Still using Adam is not sufficient to achieve the best results in the absence of added gradient noise [10]

Before we can train a network we first of all have to initialize the values of all the neurons weights and biases.

Convolutional Neural Networks (CNNs) are a class of deep learning models specifically designed for processing structured grid-like data, such as images, by automatically learning hierarchical patterns and features. CNNs are widely used in computer vision tasks and have revolutionized the field of image recognition, object detection, and image generation. CNNs excel at tasks like image classification, where the network assigns a label or category to an input image

The fundamental building blocks of CNNs are convolutional layers, which perform convolution operations on input data. These layers apply a set of learnable filters (also called kernels) to input images by sliding the filter over the input data and computing element-wise multiplications and summations to produce feature maps. This layers purpose is detecting different features like edges, textures, and more complex patterns. The learned features become progressively more abstract as they pass through multiple convolutional layers making it possible to classify complex structures. Two key hyperparameters that define the convolution operation are size and number of kernels. The former is typically 3 × 3, but sometimes 5 × 5 or 7 × 7. The application of a Convolutional layer on a Matrix, shrinks it in size, to mitigate this effect there are two other parameters that can be set to get the desirable output size. On the one hand we kann add Padding, this adds number rows and columns of zeros to the matrix to increase the output size. During the convolution,the filter slides over the matrix from left to

right and top to bottom, stride is defined as the step size of the filter, meaning its the definition of how many elements the filter moves to the right or bottom.
The Convolutional layer is usually followed by a Pooling layer. This layers downsample the feature maps, reducing the spatial dimensions while retaining important information.Usually The stride matches the filed size of the pooling operation, so that no feature of the previous layer is used twice. Max pooling and average pooling are the most common applied pooling operations. In the Max pooling opperation the maximum value of the current view is selected, this preserves detected features especially the most commonly used ones. The average pooling operation takes the averages of the values of the current view. During training in back propagation average pooling provides a smoother gradient compared to max pooling. It also retains information from the original image since average pooling takes the collective information into account.

The fully connected layer connects all neurons from the previous layer to all neurons in the subsequent layer, enabling the network to make high-level decisions based on the learned features. The neurons of this layer are organized in a array form. This layer is used to optimize objectives as class scores.

AlexNet is a pioneering CNN architecture that played a pivotal role in advancing the field of deep learning and computer vision [11]. Alex net won the Image Net Large Scale Visual Recognition Challange(ILSVRC) in 2012 it introduced several groundbreaking concepts, including deep architecture with multiple convolutional and fully connected layers, ReLU activation functions, dropout for regularization, and GPU acceleration for faster training. Its success highlighted the potential of deep neural networks for image classification tasks, influencing the design of subsequent CNN architectures and shaping the direction of modern deep learning research.

Due to our lack of training data a possible methode of enhancing our network is using transfer learning. This method is a common and effective strategy to train a network on a small data set. By pretraining the network on extremely large datasets like ImageNet with 1.4 million images and 1000 classes, trying to learn generic features that can be shared among networks [5].This is a unique advantage of deep learning that makes itself useful in various domain tasks with small datasets. Despite the popularity of transfer learning in medical imaging there hasn't been a lot of work studying of its effects. Usually transfer learning is performed by taking a standard IMAGENET architecture with pretrained weights and then fine tuning its parameters on the dataset which the network is supposed to detect. [12] shows on two large scale medical image networks that the gain of transfer learning on those networks is marginal. It also shows that transferusually helps large scale models, with small models showing little difference. Therefore we wont use transfer learning to enhance our network.

To ease the training process we typically normalize the initial values of our parameters by initializing them with zero mean and unit. With training on our data we would usually lose this normalization, which slows down training and amplifies changes as the network becomes

deeper. To remove those effects and therefore enhance the training stability and convergence of deep neural networks, batch normalization[13] can be employed. By standardizing the inputs within each mini-batch during training , this technique mitigates gradient related issues and accelerates convergence. Additionally it acts as a form of regularization ,curbing overfitting. With this metohde we are able to use higher learning rates and pay less attention to the initialization parameters [8].

To calculate the output from the weighted sum of the inputs from a node an activation function is needed.Those functions are used to map the input between the required values like 0 and 1 or 1 and -1. The choice of the activation function has a large impact on the capability and performance of the neural network. It can ensure a better detection of complicated patterns and even accelerate the learning process [14]. Comonly used functions are ... [15] recommends to use ELU non-linearity without batch normalization or ReLU with it. sigmoid

$$y = \frac{1}{1 + e^{-x}} \tag{3.7}$$

The ReLu function has hase significant adcantages over a sigmoid function in a neural network. The mein advantage is that ReLU function is very fast to calculate. For positive x the ReLU function has a constant gradient of 1 whereas a sigmoid function has a gradient that rapidly converges to zero. This property makes neural networks with a sigmoid activation function slower to train. The occurring phenomen is also known as vanishing gradient problem ReLu as an activation function removes this problem because the gradient of ReLU is always one for positive x so that the learningprocess wont be slowed down by a vanishing gradient.

ReLU

$$y = \max(x, 0) \tag{3.8}$$

However the 0 gradient can pose the zero gradient problem this can be compensated by adding a smaller linear term in x to give the ReLU function a nonzero slope at all points this is solved in the implementation by adding $\alpha(e^x - 1)$ for all values smaller than zero. Therefore the gradient of ELU is allways bigger than 0, tending to zero for $x \to -\infty$.

ELU

$$y = \begin{cases} x, \text{ if } x \geq 0 \\ \alpha(e^x - 1), \text{ otherwise} \end{cases} \tag{3.9}$$

Convolution based attention module(CBAM) is a simple yet effective attention module [16]

Training the model on a sizable amount of data is a good way to avoid over fitting.

A big issue in the field of medical imaging is the lack of training data. Due to the fact that just trained staff is capable of labeling the images it is expensive to label the data, also a lack of time from the staff can be an issue. But even with time and money the biggest issue in medical imaging is the lack off data. Since ...

When training a CNN with to little data we usually need to make use of the methode of early stopping. This technique is used in machine learning to prevent overfitting during model training. overfitting occurs in machine learning when a model learns to perform well on the training data, capturing noise and irrelevant patterns, but usually performs poor on new unseen data. This indicates that the model has memorized the training data instead of learning the underlying patterns, leading to reduced generalization ability and diminished perdictive accuracy on real-world examples. Early stopping involves monitoring the model's performance on a validation dataset and stopping training when the performance on a validation dataset starts to degrade. By preventing excessive training; early stopping helps the model generalize better to new data and improves its ability to make accurate prediction on new data and improves its ability to make accurate predictions.

This technique strikes a balance between training optimal performance and avoiding the point where the model starts memorizing noise in the training data.

Commonly, biological data tends to be imbalanced, often negative samples are much more numerous than positive ones [11]. When training a Network with imbalanced data, the network is prone to to bias towards the major classes, since it priorizes learning the features for detecting those classes. This also means that the network does not get enough exposure to detect minor features and therefore cant learn its distinctive features. All this leads to a higher number of false negatives. With the network trying to capture the minor features it can run into issue of overfitting the network

S... In our case we have a lack of data that is labeled with the severity level of 4 and 5. This holds a major issue. Since level 4 and 5 scans are in need of a re-scan. Thus we want to have a high accuracy score for detecting these classes to be sure that we wont unnecessarily re-scan patients since we don't want to expose them unnecessary to radiation.

If we would further train the network with the same samples the network would overfitt and lose its validity To prevent overfitting from happening we need to implement methods to augment the data, to have more of it to learn from. Since the scans have a depth of 110 layers there is a possibility of taking some of those layers to train on with this process we need to be carefull with how many slices we take since slices that lay close to each other might look to similar so that the network might tend do overfitt faster. Therefore we just take 8 slices which we do based on [2]. There are also a few other ways a very common technique of data augmentation is rotation. To augment our data we took Another common augmentation technique is to add a small amount of gaussian noise to the image, so small, that the image still looks the same way to a human. This ensures that the network does not focus on specific data points. This leads to a more robust network

adding a small amount of gaussian noise does not impact the general look of the picture which means it would still be interpreted the same way from a doctor that means that we can use this method to modify our data

| Motion Grade | Tibia rounded | Radius rounded |
|---|---|---|
| 1 | 334 | 139 |
| 2 | 97 | 169 |
| 3 | 42 | 233 |
| 4 | 24 | 62 |
| 5 | 2 | 8 |

The data was provided by the osteology department of the “Unfall Klinikum Eppendorf” and Labeled by 3 doctors of the department (need to check wether thats correct). The labeled data contained 500 Scans of the radius and 500 scans of the tibia. In 51.1 % of all scans the doctors had a consens, 57% for grading the tibia and 45.2% for grading the radius.

In 7.4% a rescan could have occured 4.6% for tibia and 10.2 percent for the radius. This is hard to compare since we have a way bigger amount of values in the domain of 3 and 4 for the radius compared to the tibia. If we therefore link the amount of values that where the rounded gathered rating was 3 or 4 with the possibility of a rescan we get : 34% for tibia and 27.9% overall this concludes to

# 4

# Experimental Setup

# 5
# Results

# 6
# Discussion

# Literaturverzeichnis


[1] Miki Sode, Andrew J. Burghardt, Jean-Baptiste Pialat, Thomas M. Link, and Sharmila Majumdar. Quantitative characterization of subject motion in HR-pQCT images of the distal radius and tibia. *Bone*, 48(6):1291–1297, jun 2011.

[2] Matthias Walle, Dominic Eggemann, Penny R. Atkins, Jack J. Kendall, Kerstin Stock, Ralph Müller, and Caitlyn J. Collins. Motion grading of high-resolution quantitative computed tomography supported by deep convolutional neural networks. *Bone*, 166:116607, jan 2023.

[3] D.E. Whittier, S.K. Boyd, A.J. Burghardt, J. Paccou, A. Ghasem-Zadeh, R. Chapurlat, K. Engelke, and M.L. Bouxsein. Guidelines for the assessment of bone density and microarchitecture in vivo using high-resolution peripheral quantitative computed tomography. *Osteoporosis International*, 31(9):1607–1627, may 2020.

[4] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015.

[5] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, 9(4):611–629, jun 2018.

[6] Lee Feldkamp, L. C. Davis, and James Kress. Practical cone-beam algorithm. *J. Opt. Soc. Am*, 1:612–619, 01 1984.

[7] J.P. van den Bergh, P. Szulc, A.M. Cheung, M. Bouxsein, K. Engelke, and R. Chapurlat. The clinical application of high-resolution peripheral computed tomography (HR-pQCT) in adults: state of the art and future directions. *Osteoporosis International*, 32(8):1465–1485, may 2021.

[8] Sebastian Ruder. An overview of gradient descent optimization algorithms, 2016.

[9] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

[10] Arvind Neelakantan, Luke Vilnis, Quoc V. Le, Ilya Sutskever, Lukasz Kaiser, Karol Kurach, and James Martens. Adding gradient noise improves learning for very deep networks, 2015.

[11] Laith Alzubaidi, Jinglan Zhang, Amjad J. Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, J. Santamaría, Mohammed A. Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8(1), mar 2021.

[12] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[13] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. February 2015.

[14] Asifullah Khan, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, 53(8):5455–5516, apr 2020.

[15] Dmytro Mishkin, Nikolay Sergievskiy, and Jiri Matas. Systematic evaluation of convolution neural network advances on the imagenet. *Computer Vision and Image Understanding*, 161:11–19, aug 2017.

[16] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module, 2018.