UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR
NEURO- UND BIOINFORMATIK

From the Institute for Neuro- and Bioinformatics
of the University of Lübeck
Director: Prof. Dr. rer. nat. Thomas Martinetz

An Intelligent X-ray Assistant: Optimizing Diagnostic Quality and Image Acquisition in Digital Radiography

Dissertation
for
Fulfillment of Requirements for the Doctoral Degree
of the University of Lübeck

from the Department of Computer Sciences and Technical Engineering

Submitted by
Dominik Mairhöfer
from Aschaffenburg

Lübeck, 2026

First referee: Prof. Dr. rer. nat. Thomas Martinetz
Second referee: Prof. Dr.-Ing. habil. Marcin Grzegorzek

Date of oral examination: 28.04.2026

Approved for printing. Lübeck, 04.05.2026

Abstract

Radiographs are the most commonly used modality in diagnostic imaging. Although they are of fundamental importance, there are constantly images acquired that cannot be used for diagnosis. Since the development of digital radiography, such unusable images have mainly been caused by incorrect patient positioning or improper collimation. Importantly, these errors can directly impact patients' health. Repeating the procedure increases radiation exposure, delays treatment, and extends hospital stays. Poor-quality images can even lead to misdiagnosis and incorrect treatment if unnoticed. Ultimately, images that cannot be used for diagnostic purposes also harm the hospital by increasing costs, staff workload, and room occupancy.

To improve radiograph image quality, this thesis develops deep learning based methods for assistance systems in the radiography process. First, a developed framework is used to learn how to automatically assess the quality of radiographs. While radiologists can usually assess quality immediately, radiographers often come to different judgments. The automatic assessment system can provide support here and spare patients unnecessary repeat examinations. The results show that deep learning models are capable of assessing quality at the level of a radiologist across different parts of the body.

Based on these results, instead of assessing radiographs, the quality of radiographs is predicted using depth images. While automatic assessment allows low-quality images to be detected immediately, it cannot directly prevent them from being produced. To achieve this, depth cameras were used to record the patient's pose, and neural networks were then used to predict the quality of the resulting radiograph. This type of positioning assessment makes it possible to acquire a radiograph only when a poor-quality image is unlikely to result. Although quality can only be determined on the radiographs themselves, the results show that a similarly accurate prediction is possible on depth images.

In addition to positioning, the irradiation area is another key factor affecting radiograph quality. Therefore, a system is developed to automatically determine the required irradiation area. Depth cameras are used to capture images of the patients, while simultaneously acquired radiographs are used to label the minimally necessary irradiation area. The system learns to predict the required area automatically and demonstrates performance comparable to that of trained radiographers.

Overall, this thesis presents approaches to solving the most common causes of low-quality radiographs today. It presents the first studies to systematically evaluate radiographs or patient poses in terms of their suitability for diagnosis.

Zusammenfassung

Röntgenbilder sind die meistgenutzte Modalität in der diagnostischen Bildgebung. Obwohl sie von fundamentaler Bedeutung sind, kommt es kontinuierlich zu Bildern, die nicht für eine Diagnose verwendet werden können. Seit der Entwicklung der digitalen Radiografie sind solche unbrauchbaren Bilder hauptsächlich auf eine falsche Positionierung des Patienten oder eine unsachgemäße Kollimation zurückzuführen. Entscheidend ist, dass diese Fehler jedoch direkte Auswirkungen auf die Gesundheit der Patienten haben können. Die Wiederholung der Aufnahme erhöht die Strahlenbelastung, verzögert die Behandlung und verlängert den Krankenhausaufenthalt. Aufnahmen von schlechter Qualität können sogar zu Fehldiagnosen und falschen Behandlungen führen, wenn sie unbemerkt bleiben. Letztendlich schaden Aufnahmen, die nicht diagnostisch verwendet werden können, auch dem Krankenhaus, da sie die Kosten, die Arbeitsbelastung des Personals und die Raumbelegung erhöhen.

Um die Röntgenbildqualität zu verbessern, werden in dieser Arbeit Deep Learning basierte Methoden für Assistenzsysteme im Röntgenprozess entwickelt. Zuerst wird ein Framework entwickelt und verwendet, um zu lernen, die Qualität von Röntgenaufnahmen automatisch zu bewerten. Während Radiologen die Qualität in der Regel sofort einschätzen können, kommen die röntgenden MTRAs oft zu anderen Einschätzungen. Das automatische Bewertungssystem kann hier unterstützen und Patienten vor unnötigen Neuaufnahmen bewahren. Die Ergebnisse zeigen, dass Deep Learning Modelle in der Lage sind, die Qualität auf dem Niveau eines Radiologen für verschiedene Körperteile zu beurteilen.

Auf Grundlage dieser Ergebnisse erfolgt statt einer Bewertung von Röntgenbildern eine Vorhersage der Röntgenbildqualität anhand von Tiefenbildern. Während durch eine automatische Bewertung schlechte Bilder sofort erkannt werden können, kann deren Entstehen nicht direkt verhindert werden. Um dies zu erreichen, wurde mittels Tiefenkameras die Pose des Patienten aufgenommen und anschließend durch neuronale Netze vorhergesagt, welche Qualität das entstehende Röntgenbild haben wird. Eine solche Bewertung der Positionierung ermöglicht es, erst dann ein Röntgenbild aufzunehmen, sobald ein schlechtes resultierendes Röntgenbild unwahrscheinlich ist. Obwohl die Qualität nur anhand der Röntgenbilder selbst bestimmt werden kann, zeigen die Ergebnisse, dass eine ähnlich genaue Vorhersage auf Tiefenbildern möglich ist.

Neben der Positionierung ist der Bestrahlungsbereich ein weiterer bestimmender Faktor der Röntgenbildqualität. Daher wurde ein System entwickelt, das den erforder-

lichen Bestrahlungsbereich automatisch bestimmt. Mit Tiefenkameras werden Bilder der Patienten aufgenommen, während gleichzeitig angefertigte Röntgenaufnahmen dazu dienen, den minimal erforderlichen Bestrahlungsbereich zu kennzeichnen. Das System lernt, den erforderlichen Bereich automatisch vorherzusagen, und zeigt eine Leistung, die mit der von ausgebildeten MTRAs vergleichbar ist.

Insgesamt präsentiert diese Arbeit Ansätze zur Lösung der häufigsten Ursachen für die schlechte Qualität von Röntgenbildern. Sie stellt die ersten Studien vor, die Röntgenbilder oder Patientenpositionen systematisch hinsichtlich ihrer Eignung für die Diagnose bewerten.

Contents

Chapter 1

Introduction

1.1 Motivation

In clinical radiography, the quality of radiographs has direct and severe consequences for patient health and safety. When a radiograph does not meet diagnostic quality standards due to incorrect patient positioning, inadequate collimation, or technical errors, it must be rejected or repeated. These retakes expose patients to additional ionizing radiation without diagnostic benefit, violating the fundamental principle of radiation protection to keep exposure as low as reasonably achievable. Apart from radiation exposure, poor-quality radiographs that are not immediately rejected lead to interpretation errors with potentially fatal consequences. A study from 2013 shows that radiographic interpretation errors contribute to 40,000–80,000 deaths annually in the United States [Lee et al., 2013], with poor image quality being a documented contributing factor to systematic errors in radiology [Waite et al., 2017]. These diagnostic errors manifest as missed diagnoses, incorrect diagnoses, delayed treatment initiation, prolonged hospital stays, and unnecessary follow-up appointments, which all compromise treatment outcomes and patient safety.

The quality problem extends beyond patient harm to create cascading operational difficulties throughout the healthcare system. A high-quality radiograph can be evaluated in 22–80 seconds [Abozeed et al., 2024; Jeong et al., 2025], as standardized images allow radiologists to leverage fast, automatic cognitive processes that enable efficient workflow [Lee et al., 2013]. Poor-quality or non-standardized images disrupt this automaticity, forcing radiologists to engage slower, deliberative processes that require significantly greater mental effort and time, thereby reducing diagnostic throughput and increasing cognitive workload. Moreover, radiographs unsuitable for diagnosis still generate administrative overhead. The radiologist must document the quality deficiency, potentially consult the attending physician or the responsible radiographer, and await a repeat examination before completing the diagnostic report. For radiographers, retakes create immediate operational pressure in tightly scheduled hospital workflows, leading to equipment backlogs, appointment delays, and increased workplace stress. Economically, each repeat appointment costs approximately 40 € [Breitwieser et al., 2025; Winberg et al., 2025], and when multiplied by the frequency of quality-related retakes,

represents a significant financial burden for healthcare institutions. Beyond internal institutional concerns, quality assurance is also externally mandated; in Germany, for example, radiograph quality is verified by regulatory authorities through randomized audits [Deutschland, 1988; Gemeinsamer Bundesausschuss, 2023].

These individual and institutional challenges reflect a broader systemic crisis in radiology that continues to worsen. Recent data from England [The Royal College of Radiologists, 2025a] reveals that in 2024, the number of fully trained clinical radiologists is 29% lower than required to meet safe and sustainable service demand, with projections indicating this shortage will reach 39% by 2029. The consequences of this workforce shortage are already severe. In 2024, 976,000 medical images remained uninterpreted for more than a month, of which over 500,000 were radiographs [NHS England, 2024; The Royal College of Radiologists, 2025b]. This backlog represents a 28% increase compared to the previous year, demonstrating an accelerating crisis. While workforce shortages represent the primary driver, poor radiograph quality compounds the problem by reducing the effective capacity of the already limited radiologist workforce. Every minute lost due to poor-quality images is time that cannot be spent evaluating diagnostic examinations, effectively increasing staff shortages. The issue becomes even more pressing when considering future demographic trends, as aging populations require increasingly frequent medical imaging, while the capacity to interpret these images stagnates or declines. This is further complicated by the fact that even automated systems based on artificial intelligence cannot compensate for staff shortages if the quality of radiographs remains poor, as AI models that have been trained predominantly with high-quality standardized images are unreliable when dealing with non-standardized or poor-quality images.

Reported rejection rates in clinical practice vary from 9% to 17% [Serra et al., 2024], with additional unreported deletions occurring in digital radiography systems at rates of approximately 11%, reaching up to 21% for certain examinations such as knee radiographs [Hofmann et al., 2015]. Notably, the nature of quality deficiencies has shifted with the transition to digital radiography. While many technical parameters that influence image quality can now be corrected post-acquisition and have therefore become less critical, incorrect patient positioning has emerged as the dominant cause, responsible for 76% of rejections [Serra et al., 2024]. This change requires a focus on diagnostic quality rather than solely evaluating technical imaging parameters. This means assessing the images in terms of the extent to which the image content, in particular the anatomical positioning and anatomical features depicted, allows for a reliable clinical diagnosis. Given these high rejection rates and their profound impact on patient safety, healthcare workflows, and the already strained radiology workforce, this thesis focuses on developing automated assistance systems specifically targeting the assessment of patient positioning in radiography. Machine learning, and particularly deep learning, has demonstrated transformative potential in medical imaging through its ability to recognize complex patterns in visual data, making it well-suited to evaluate

the semantic content of radiographs rather than merely technical parameters. By deploying these technologies in the clinical workflow, during or even before image acquisition, quality issues could be detected and even prevented. Although there are a few proprietary commercial systems available to assist with collimation, there remains a significant gap in accessible, scientifically validated solutions that can provide positioning assistance.

The methods developed in this work aim to support radiographers during image acquisition through real-time feedback on diagnostic quality, enable systematic quality assurance through automated assessment, and provide educational support for training programs. By ensuring that radiographs meet diagnostic quality standards before reaching the radiologist, these approaches can reduce patient radiation exposure from retakes, decrease interpretation errors, alleviate workflow pressure on both radiographers and radiologists, and ultimately contribute to better patient outcomes in an increasingly strained healthcare system.

1.2 Objectives

The main objective of this work is to improve the diagnostic quality of radiographs. To achieve this, one objective is to shift the focus from the technical evaluation of radiographs to an evaluation of their diagnostic quality based on the visible anatomical structures. Furthermore, depth cameras are to be used to improve the workflow for both patients and radiographers. The objective here is to investigate how data from depth cameras on X-ray machines can be used to improve radiograph quality. Since both annotated medical data and real patient data are often scarce, the final objective is to investigate what added value can be achieved even with limited data. Below, each objective is briefly described in more detail.

Diagnostic Quality as a Metric While substantial research exists on assessing radiograph quality through technical parameters and their automated evaluation [Takaki et al., 2020; Wang et al., 2020], minimal work addresses the automated assessment of diagnostic quality. This refers to the degree to which a radiograph can actually support clinical diagnosis. Rejection rates reaching 17% [Serra et al., 2024], along with additional uncounted deleted images, demonstrate that numerous non-diagnostic radiographs are produced, indicating a clear need for quality metrics beyond technical parameters alone. Although technical parameters influence overall image quality, they have become less critical with digital X-ray technology. In fact, patient positioning has emerged as the primary determinant of diagnostic quality and the most frequent cause of rejections and retakes.

Therefore, this work aims to directly quantify how well radiographs can be used for diagnosis. Since no prior work in this area exists, this requires first collecting and

labeling data, which is then used to train automated quality assessment models. Such automated assessment can provide real-time feedback to radiographers after each image acquisition, reducing delayed treatments and misdiagnoses. Additionally, it can serve as a retrospective quality control tool, systematically reviewing acquired images to identify systematic patterns of poor quality. Finally, it can provide guidance for trainees and less experienced radiographers to improve patient positioning, ultimately enhancing treatment outcomes.

Beyond evaluating acquired radiographs, a further goal is to predict diagnostic quality before exposure. By assessing quality prior to image acquisition, such a system can prevent non-diagnostic radiographs from being taken, thereby minimizing patient radiation exposure.

Depth Imaging in Clinical Routine While modern X-ray devices often have RGB cameras built in, depth cameras are less common, and many systems contain no cameras at all. Moreover, existing depth camera systems are proprietary, with minimal public research on their application in radiography. In various clinical settings beyond radiology, depth cameras have already demonstrated their value for pose determination [Hansen et al., 2019], weight estimation [Bigalke et al., 2021], and respiratory monitoring [Takamoto et al., 2020]. This thesis investigates use cases for depth cameras in the radiograph acquisition process. This requires placing cameras in the X-ray room and evaluating the influence of their positioning. Furthermore, steps in the acquisition process that benefit from depth image feedback must be identified and assessed. Based on these findings, data must be collected and methods developed to extract the necessary feedback. Where possible, the collected data should originate from clinical routine or closely replicate it.

Limited Data Most deep learning approaches require large amounts of annotated training data. Since both annotated medical data and clinical routine data are often sparse, this work focuses on limited data settings. Several factors contribute to data scarcity in medicine. First, labeling requires expert knowledge, which is extremely costly given the high value of medical experts' time. Second, privacy concerns limit data collection, as medical data is typically linked to personal information and is considered particularly sensitive. Third, high regulatory hurdles in the medical field impose ethical and legal constraints, particularly when data collection interferes with clinical workflow. These factors combined typically restrict medical data to retrospective use under strict anonymization. Given these constraints, advancing deep learning-based research in this domain requires working with limited data. Therefore, one objective is to explore what can be achieved with minimal data points and whether the required data amount can be estimated.

1.3 Organization and Contributions

The remainder of this thesis is organized into five chapters. Chapter 2 provides fundamental knowledge about radiography, computed tomography, and the clinical workflow for radiograph acquisition. It further describes the basics of depth imaging, introduces deep learning concepts, and presents network architectures used in this work. Chapters 3 to 5 present the main research and findings addressing the objectives described above. Each of these chapters begins with a brief introduction to the chapter's main topic and structure. The sections within each chapter investigate specific objectives and represent self-contained works. Finally, Chapter 6 discusses the results in a global context, highlights limitations, and provides an outlook on potential further research. Below, a brief summary of the contributions and results of each main chapter is provided.

Chapter 3 This chapter introduces the first method to automatically assess the diagnostic quality of radiographs. First, the factors determining diagnostic quality are described in detail. Since no public dataset exists, radiographs of the ankle are collected and labeled by four radiologists. Based on this data, a novel framework for assessing radiographs is developed, achieving accuracy comparable to that of radiologists. The framework consists of multiple stages, refining the task into subproblems before the actual quality assessment. Subsequently, the framework is extended to knee and wrist radiographs to test generalizability. Furthermore, the amount of labeled data required and the labeling effort are examined. In the chapter's final section, clinical use cases are explored. First, a retrospective quality analysis of ankle, knee, and wrist radiographs spanning over ten years is conducted. Based on this data, impact factors affecting quality are analyzed to uncover systematic causes of poor-quality images. Second, a live implementation is deployed to a clinical emergency X-ray room to provide radiographers with feedback on acquired radiographs and is subsequently evaluated. The presented methods were published in

[Mairhöfer et al., 2021] Mairhöfer, D., Laufer, M., Simon, P. M., Sieren, M., Bischof, A., Käster, T., Barth, E., Barkhausen, J., and Martinetz, T. "An AI-based Framework for Diagnostic Quality Assessment of Ankle Radiographs". In: *Proceedings of the Fourth Conference on Medical Imaging with Deep Learning (MIDL 2021)*. PMLR, 2021, pp. 484–496

and are currently under review as

[Gerdes et al., 2026] Gerdes, H.*, Mairhöfer, D.*, Laufer, M., Reis, F. L., Bischof, A., Wegner, F., Käster, T., Barth, E., Barkhausen, J., Martinetz, T., and Sieren, M. *Generalizable Deep Learning Framework for Diagnostic Quality Assessment of Musculoskeletal Radiographs.* *Authors contributed equally. 2026. **Under Review at Radiology: Artificial Intelligence** .

Chapter 4 While Chapter 3 demonstrates that accurate quality assessment is possible, this assessment cannot directly prevent poor images. Chapter 4 extends the approach of the previous chapter by predicting image quality even before acquisition. To prevent patients from being irradiated in a pose that would result in a non-diagnostic radiograph, quality assessment must occur before exposure. To achieve this, depth cameras are attached to an X-ray machine, and the depth images are used to predict the quality of the resulting radiograph, thereby assessing the patient's pose. The first dataset linking depth images with radiograph quality is created by X-raying human feet preparations in different poses while simultaneously capturing depth images. Since data from clinical routine could not be used, anatomical preparations were employed instead.

The issue of limited data availability is discussed subsequently in the chapter. Since using anatomical preparations does not scale well, a method to generate synthetic data from CT scans is developed. Combining the extraction of a 3D model of the foot from CT scans with digitally reconstructed radiographs enables the generation of corresponding depth images and radiographs in various poses and diagnostic qualities. The generated data can be used for pretraining and reduces the required amount of real training data. The presented methods were published in:

[Laufer et al., 2024b] Laufer, M.*, Mairhöfer, D.*, Sieren, M., Gerdes, H., Reis, F. L., Bischof, A., Käster, T., Barth, E., Barkhausen, J., and Martinetz, T. "Patient Pose Assessment in Radiography Using Time-of-Flight Cameras". In: *Medical Imaging 2024: Image Processing*. Vol. 12926. *Authors contributed equally. SPIE, 2024, pp. 385–393

[Laufer et al., 2025] Laufer, M., Mairhöfer, D., Sieren, M., Gerdes, H., Leal dos Reis, F., Bischof, A., Käster, T., Barth, E., Barkhausen, J., and Martinetz, T. "Synthetic Data Generated from CT Scans for Patient Pose Assessment". In: *Proceedings of the Eighth Conference on Medical Imaging with Deep Learning (MIDL 2025)*. Proceedings not yet published. PMLR, 2025

Further, the presented methods are part of the following granted patent:

[Laufer et al., 2024a] Laufer, M., Mairhöfer, D., Bischof, A., Käster, T., Sieren, M., Reis, F. L., Gerdes, H. W., and Simon, P. "Verfahren zur Erzeugung von Trainingsdaten für ein KI-basiertes Assistenzsystem und Vorrichtung zur Unterstützung der Röntgendiagnostik". DE102022133272A1. 2024

Chapter 5 Besides patient pose, another crucial factor affecting quality is collimation. Since collimation defines the area to be irradiated, it determines which anatomical regions are visible, and radiographs with missing or truncated relevant anatomical regions are a common reason for repeats. This chapter therefore investigates how

optimal collimation can be predicted, again using depth cameras to capture the patient. As with diagnostic quality, the optimal area to be irradiated, and thus the collimation, can only be labeled directly on the radiograph. Following promising results from a preceding experiment where clinical routine was recreated without actual X-raying, a clinical study was performed to obtain real clinical data. The resulting pairs of depth images and radiographs, labeled regarding their optimal collimation, were used to learn collimation prediction. Even with very limited data, results on par with radiographers were achieved. The presented methods were published in:

[Mairhöfer et al., 2024] Mairhöfer, D.*, Laufer, M.*, Berkel, L., Bischof, A., Barth, E., Barkhausen, J., and Martinetz, T. "AI-based Collimation Optimization for X-Ray Imaging Using Time-of-Flight Cameras". In: *ESANN 2024 Proceedings*. *Authors contributed equally. Ciaco - i6doc.com, 2024, pp. 703–708

[Mairhöfer et al., 2026] Mairhöfer, D.*, Laufer, M.*, Berkel, L., Sieren, M., Bischof, A., Barth, E., Barkhausen, J., and Martinetz, T. "AI-based Collimation Optimization for X-ray Imaging Using Depth Cameras". *Neurocomputing* 661, 2026. *Authors contributed equally. P. 131881

Chapter 2

Background

2.1 X-ray-based Imaging

X-ray-based imaging systems are the standard method for imaging procedures in medicine worldwide. While other methods such as ultrasound or magnetic resonance imaging exist, conventional radiography and fluoroscopy account for more than 80% of all images. Another 8% of all images are computed tomography (CT) [Kjelle et al., 2024]. The methods used in this study are limited to conventional radiography and CT scans.

2.1.1 Radiography

In conventional radiography, electrons accelerated in a vacuum tube collide with an anode, generating X-rays. These rays generated in the X-ray source are directed at the body part to be examined and penetrate the body, where they are attenuated to varying degrees depending on tissue density. The emerging radiation is captured by a detector, which displays the differences in intensity as a two-dimensional shadow image. This imaging technique is used in a wide range of medical situations, from the emergency room for critical issues to follow-up examinations after surgery. While the issues that can be clarified by radiography are diverse, the process of acquiring radiographs is quite uniform. A schematic flow for this process is shown in Figure 2.1 and described in the following paragraphs.

Step 1: Patient Arrival Since conventional radiography is usually performed using stationary X-ray devices, the process begins with the patient arriving at the X-ray room.

Step 2: Positioning and Collimation The second step consists of positioning the X-ray machine and the patient and adjusting the collimation. While there are fixed positions for positioning the X-ray machine for many issues, the patient must be aligned manually on an individual basis [Becht et al., 2019]. For this purpose, the body part to be examined is positioned between the detector and the X-ray source and then

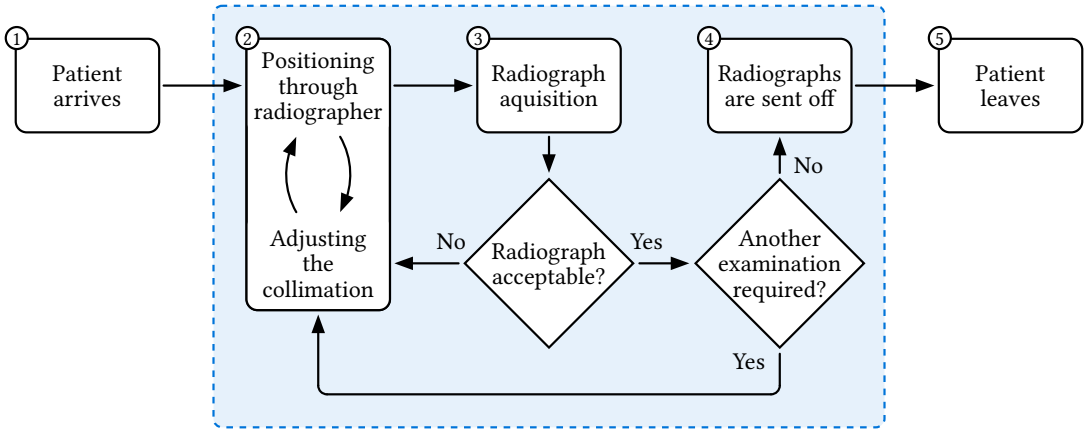


Figure 2.1: The clinical workflow of radiograph acquisition. First, the patient arrives at the X-ray room. The radiographer then positions the patient in front of the X-ray device, moves the patient into the correct pose, and defines the area to be irradiated through adjustments in the collimation. If the radiographer is satisfied with the positioning and collimation, the radiograph is acquired. The radiographer then checks if the radiograph is acceptable for a diagnosis and, if not, repeats the acquisition process. Otherwise, the acquisition process is repeated for any further required examinations. When no more examinations are required, all radiographs are sent off into the hospital's PACS. Finally, the patient leaves.

aligned so that the resulting radiograph complies with standards as closely as possible. In order to ensure that only the area relevant to the diagnosis is irradiated, collimation is used to adjust the irradiation area. The collimator, typically consisting of movable metal plates, is located between the X-ray source and the area of the patient to be irradiated. By moving the plates into the X-ray beam, the effective exposure area can be limited to avoid unnecessary radiation exposure. This is shown schematically in Figure 2.2. The positioning of the patient and the adjustment of the collimation are interdependent and can therefore be adjusted several times until both are satisfactory.

Step 3: Radiograph Acquisition Once positioning and collimation have been adjusted, the radiographer leaves the patient and moves to an area protected from scattered radiation, where the radiograph acquisition is triggered. With digital radiographs, the image is displayed a few seconds later, and the radiographer decides whether it is of sufficient quality for a diagnosis. The radiograph is checked by the radiographer to ensure that all anatomical areas are visible and that they are displayed correctly. If the radiograph is insufficient, the previous step of positioning and collimation is repeated, and another radiograph is taken.

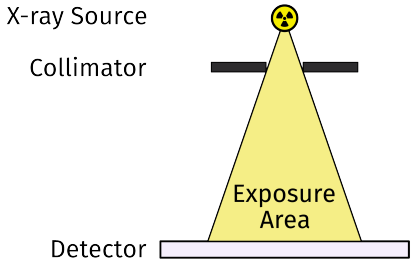


Figure 2.2: The use of a collimator to define the exposure area is shown. Movable metal plates, placed between the X-ray source and the detector, are used to limit the size of the exposure area.

Step 4: Next Examination and Dispatch If a radiograph of sufficient quality is obtained, the next examination is performed. As a rule, two perpendicular radiographs are taken for a single body part. If there are several medical issues to be addressed, multiple body parts are examined. Once all necessary examinations have been performed, the radiographer sends the radiographs. Only at this point are the radiographs transferred from the X-ray system and stored in the archiving system (PACS).

Step 5: Patient Departure Finally the patient leaves the X-ray room, and the process repeats for the next patient.

The radiographs acquired in step 3 are saved according to the DICOM standard [NEMA, 2025]. These files not only contain the image itself but are also enriched with numerous metadata. These contain information about the patient, such as age, sex, or weight; the question to be examined; the specific examination performed; the treating doctor; and technical information about the equipment used. While identifying metadata must be removed for privacy reasons before the data can be used for research, the remaining metadata can be used to obtain valuable additional information about the image itself. However, it should be noted that some of this metadata is created manually and is not standardized, which means that this information may be unreliable.

2.1.2 Computed Tomography

Computed tomography scans can be considered an extension of conventional 2D radiography. X-rays are generated in the same manner, passing through the patient, being attenuated, and subsequently captured by a detector. The key difference is that it provides 3D imaging.

To achieve this, in principle, the X-ray source and the detector rotate around the patient and take many one-dimensional intensity profiles. These 1D profiles are used

to reconstruct a two-dimensional intensity image, known as a CT slice, through back projection. This process can be repeated for several adjacent positions by incrementally moving the patient to obtain a 3D reconstruction consisting of multiple 2D slices. Modern CT scanners use more complex imaging techniques, such as multislice CT, where the detector captures narrow 2D projections, and helical CT, where the patient is moving continuously.

In contrast to radiographs, where the absolute intensity values have no interpretable meaning, the values in a CT scan are standardized. The values in the CT scan are given in Hounsfield units (HU), which are a measure of tissue density. They are calibrated so that air has a value of -1,000 HU and water has a value of 0 HU. These values allow the images to be interpreted not only visually but also to draw direct conclusions about tissue types.

Another interesting feature of CT is that it can be used to create radiographs from any chosen perspective. This creation of radiographs from CT scans, known as digitally reconstructed radiographs (DRR), can be performed using forward projection or physical simulation.

2.2 Depth Imaging

For a wide range of applications, it is beneficial to have not only color data but also 3D spatial information. There are various methods for capturing depth information for a scene, such as stereo vision, structured light, time-of-flight (ToF), or LiDAR. In this work, ToF and stereo vision sensors are used to capture depth images of patients.

The captured depth information can be encoded in different formats like depth images or point clouds. Depth maps are analogous to natural images, but each pixel stores a scalar value representing the distance, typically in millimeters. An important concept that appears in depth images is that of invalid points. When creating depth images, the information required to calculate the depth for certain pixels may be missing, incorrect, or ambiguous. If this is recognized during depth calculation, the corresponding pixel can be marked as invalid, often represented as zero in the depth images. Unlike color images, where even over- or underexposed pixels provide a tendency of light intensity, depth images can contain pixels with a total lack of information.

Point clouds, on the other hand, consist of a set of points in a 3D coordinate system. This offers the benefit of representing the real 3D structure of the scene and not being only a 2D projection. Using the point clouds, it is easy to combine multiple simultaneously made measurements into one large point cloud. Further, from the point cloud, new depth images can be created through 2D projections using different viewpoints or other camera parameters. However, it should be noted that point clouds from single depth images do not represent true 3D imaging, as the depth information is only available from a single perspective. This is often referred to as 2.5D. As the

number of points in point clouds is not required to be the same for every capture, pixels for which no depth could be calculated can simply be left out. Both formats, depth images and point clouds, can be converted into each other when the intrinsic camera parameters are known.

2.2.1 Stereo Vision

Stereo vision sensors consist of two camera sensors. To estimate the depth, both sensors capture an image, which is mostly a monochrome image. These images are then aligned, and for each pixel in one image, the corresponding pixel in the other image is matched. Then, based on the disparity between both pixels and the camera's parameters, the depth can be calculated. As some recognizable features are required for a reliable matching between both cameras' pixels, this technique is error-prone for textureless parts of the images, for example, uniform-colored smooth walls. To prevent this, many stereo vision cameras additionally use an infrared light emitter, which projects a random light pattern on the scene. Such a pattern ensures that some texture is everywhere, which improves the reliability of the matching. Another important reason for missing data in the depth image is occlusion. If a certain part of the scene is only visible from one camera but not from the other, there can be no matching for these pixels.

2.2.2 Time-of-Flight

ToF cameras function by measuring the time it takes for the light to travel from the source to an object and back to the sensor. Although there are ToF cameras that measure this time directly (dToF), they require very accurate time measurement and sensitive sensors, which is not suitable for high-resolution applications. Many ToF cameras therefore use an indirect method, such as the continuous wave (CW-ToF) method, which measures a phase shift rather than time directly. To achieve this, the brightness of the emitted light is modulated, and the phase shift between the emitted and received light is measured, from which the distance can be calculated. In contrast to stereo vision, this eliminates the need for error-prone matching. However, even when measuring with ToF cameras, errors can occur that lead to invalid values. If multipath interference occurs, the sensor captures not only direct reflections but also light that has traveled along multiple paths. This can lead to ambiguous phase shifts and incorrect distance values. Similarly, very dark or reflective surfaces can result in a signal that is too low or saturated to calculate the phase shift correctly.

2.3 Deep Learning

Deep learning is a machine learning method that uses artificial multilayer neural networks. In general, supervised machine learning tries to find a function $f \in \mathcal{F}$

that approximates an, in most cases, unknown function $\mathbf{y}_i = f^*(\mathbf{x}_i)$ based on training samples $(\mathbf{x}_i, \mathbf{y}_i), i = 1, \ldots, p$ and a loss function $L(\mathbf{y}_i, f(\mathbf{x}_i))$ that measures the error between the learned function's output and target value.

For deep learning, the set of possible functions $\mathcal{F}$ is a neural network architecture, whereas $f(\cdot;\theta)$ is an instance of the architecture with the concrete weights θ, which can be adapted. The basic building block of neural networks is the artificial neuron, consisting of learnable weights $\mathbf{w} \in \mathbb{R}^n$ and a bias $b \in \mathbb{R}$. The neuron receives an input vector $\mathbf{x} \in \mathbb{R}^n$, which is multiplied by the neuron's weights and then summed up. The bias is subtracted, and the result is propagated through an activation function $\sigma(\cdot)$. This can be described in a formula as follows:

$$\mathbf{y} = \sigma(\mathbf{w}^\top \mathbf{x} - b)$$

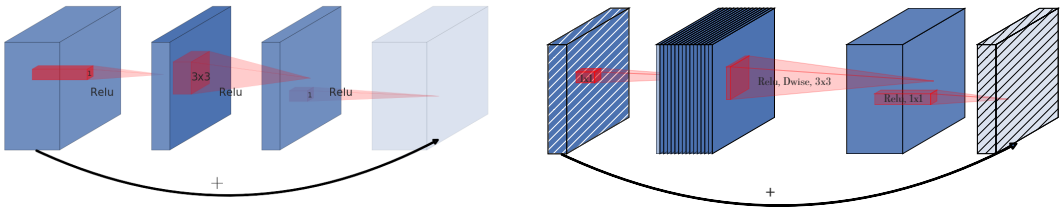
Combining several such neurons results in a fully connected layer, and combining such layers in series results in a multilayer perceptron (MLP). In such MLPs, each neuron is connected to every neuron in the previous layer, which quickly leads to a very high number of parameters in layers with many neurons. Convolutional neural networks (CNNs) were developed to reduce the number of parameters and achieve desired properties for spatial data, such as translation invariance. CNNs mainly consist of convolution and pooling layers and generally only use fully connected layers at the very end of the network. Unlike fully connected layers, a convolutional layer uses a sparse set of weights called kernels. Instead of processing the entire input at once, these kernels act on a small local area of the previous layer called a receptive field or patch. It is crucial that these weights are not unique to a single location, but rather the same kernel slides across the entire input and is applied at each point to generate an output, which is referred to as weight sharing. While convolutional layers allow new features to be learned from the previous layer, pooling layers apply a fixed operation, like max pooling, on each patch to reduce the spatial resolution.

2.3.1 Deep Learning Architectures

These building blocks form the basis for modern network architectures. Some architectures used in this work are briefly presented below.

EfficientNet

The EfficientNet presented in Tan et al. [2019b] is a family of convolutional network architectures found through a neural architecture search. The architecture search optimized accuracy and FLOPS to find a model that is as accurate as possible while computational costs are still low. Their found base architecture is similar to Tan et al. [2019a], which itself is heavily based on the MobileNetV2 architecture presented in Sandler et al. [2018]. A key building block for all these architectures is the inverted



(a) **(b)**

Figure 2.3: Panel (a) shows a normal residual bottleneck block. A high-dimensional feature block is narrowed down by a 1×1 convolution and the result is then transformed by a 3×3 convolution. Finally, the low-dimensional features are transformed to high-dimensional features by a 1×1 convolution on which the block's input is added. The inverted residual block shown in panel (b) starts with low-dimensional features, which are transformed to a higher dimension by a 1×1 convolution. On the high-dimensional features, a depthwise convolution is used, which means that for each feature map independently, a single kernel is used, producing a new feature map. The still high-dimensional features are then transformed in a lower feature space with a 1×1 convolution. Note that after the last 1×1 convolution, the block's input is added, but no non-linearity is applied. Figure taken from Sandler et al. [2018].

residual block. This block follows the idea that the creation of new features should happen in a high-dimensional feature space instead of a low-dimensional one, like in the normal residual bottleneck block. Instead of reducing the feature dimension, applying a spatial convolution, and then increasing the feature dimension, the inverted residual block first increases the feature dimension, applies a spatial convolution, and then again decreases the feature dimension. Since experiments showed that using a non-linearity like ReLU in a low-dimensional feature space leads to worse results, no non-linearity is used after reducing the feature dimension. Furthermore, the use of depthwise convolution reduces the parameters compared to a standard residual block while only slightly affecting the results. The design of both blocks is shown in Figure 2.3.

The family of EfficientNets contains architectures of different sizes, from small networks (EfficientNet-B0) to large (EfficientNet-B7). The EfficientNet-B0, as the smallest network, has 5.7 million parameters and achieves a better accuracy on ImageNet [Deng et al., 2009] than a ResNet-50 or a DenseNet-169. In terms of accuracy per parameter, it is 4.9 times more efficient than the ResNet-50 and 2.6 times more efficient than the DenseNet-169. As the small number of parameters reduces the risks of overfitting on small datasets, the EfficientNet-B0 is used for most regression or classification tasks in this work.

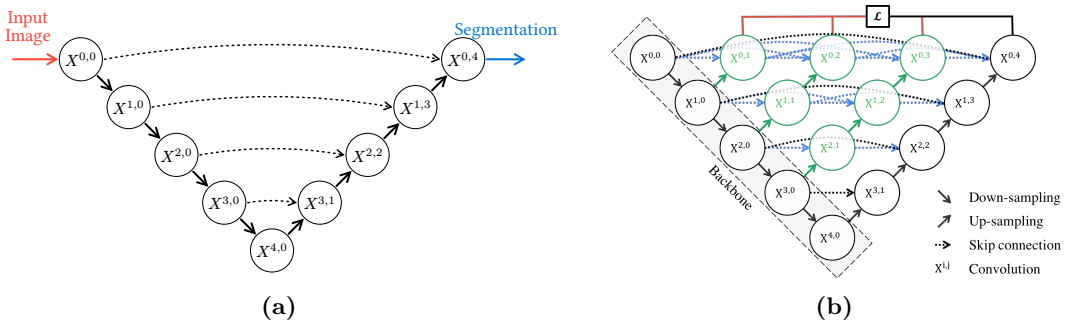


Figure 2.4: The U-Net architecture [Ronneberger et al., 2015] shown in panel (a) receives an image as input, which is initially transformed by convolutions and then four times spatially downsampled through max pooling and transformed with convolutions. The resulting representation is then four times upsampled and again transformed after each step. Each pair of representations on the same spatial level is connected through a skip connection. The convolution blocks always consist of two consecutive convolutions. The inputs are marked in red and the outputs in blue. Panel (b) shows the UNet++ structure from Zhou et al. [2018]. In addition to the U-Net, there are multiple intermediate upsampling steps, marked in green. This results in multiple nested U-Nets of different depths.

UNet++

The UNet++ is a segmentation network presented by Zhou et al. [2018] and is still one of the state-of-the-art segmentation models. The base of its architecture is the U-Net [Ronneberger et al., 2015], which is outlined in Figure 2.4a. The U-Net consists of an encoder and a decoder part. The encoder follows a typical CNN structure, built of blocks of convolution layers followed by pooling layers, removing spatial resolution while increasing the feature dimension. The decoder takes the encoded features and, at each step, restores spatial information through a transposed convolution. Importantly, the input to each decoder is concatenated with the output of the corresponding encoder step, skipping the deeper encoder and decoder layers.

A downside of the U-Net's architecture is that there are long-range skip connections. For example, the output layer $X^{0,4}$ in Figure 2.4a receives highly processed features from the previous decoder and, at the same time, features through the skip connection that are only processed in a single convolution block. To counter this, the UNet++ extends the architecture by adding convolutions and skip connections in-between both sides of the U-Net's U-structure, which results in multiple nested U-Nets with different depths. As shown in Figure 2.4b, every decoder stage now additionally receives the output of all intermediate decoders on the same spatial level.

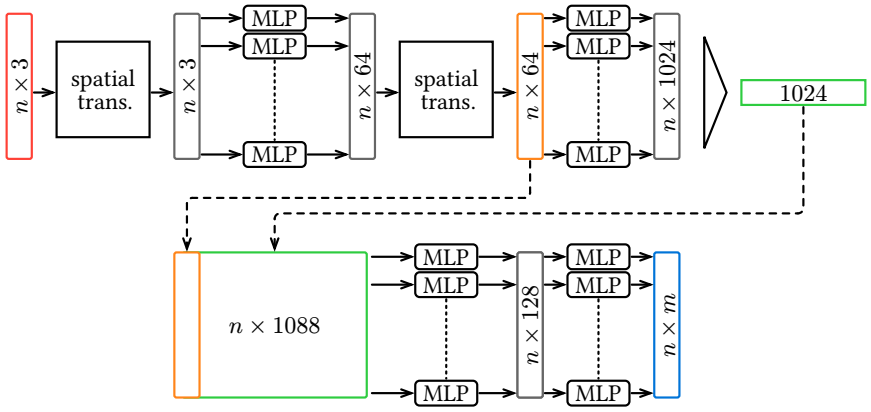


Figure 2.5: The PointNet architecture [Qi et al., 2017a] receives an $n \times 3$ input, where n is the number of points and the three dimensions are their (x, y, z) coordinates. The points are transformed spatially, and features are extracted by feeding each point into the same MLP. This is repeated twice. On the output, a global max pooling is applied for each feature dimension to get a global feature representation (green). For segmentation, this representation is concatenated with the local feature of each point from a previous feature representation (orange). These combinations of local and global features are transformed into the final output by two additional MLPs. The inputs are marked in red and the outputs in blue. The figure is adapted from Qi et al. [2017a].

PointNet

CNNs require a spatial arrangement of data, which is not always available. For example, point clouds consist of a set of points that cannot be arranged in an orderly sequence. Every possible sequence represents the same point cloud. To process such data, architectures such as PointNet [Qi et al., 2017a], which is shown in Figure 2.5, have been developed that do not require assumptions about spatial arrangement. The architecture receives a set of points as input. These points are transformed spatially by a small network that learns an affine transformation. Each transformed point is fed into a shared MLP so that for each point, new features are extracted independently. Both steps, the spatial transformation and feature extraction, are then repeated. To get a global feature vector, max pooling is applied over the features of all points. If not used for classification but for pointwise segmentation, the global feature vector is concatenated to each pointwise feature vector from the second spatial transformation. The concatenated features for each point are transformed by a shared MLP independently twice, resulting in the final output.

The PointNet++ architecture extends the PointNet by concatenating multiple PointNets and grouping the resulting representations. The PVConv layer [Liu et al., 2019] simultaneously applies a convolution on the voxelized point representations and an MLP on each individual point representation. The PVCNN++ [Liu et al., 2019] combines both

by replacing all MLP layers in the PointNet++ with PVConv layers. The PVCNN++ model is used in this work to minimize inference time and memory consumption to enable real-time clinical use. Since predicting the optimal collimation can be modeled as a part segmentation task, we extended the PointNet++ part segmentation architecture with the PVConv layer of the PVCNN.

2.3.2 Regularization

When working with neural networks with millions of adjustable parameters, there is a risk of overfitting. Overfitting in the context of machine learning means that a model is too closely adapted to the training data. The model then predicts the training data accurately and makes only a small error on it but is unable to predict unknown data well. The model therefore does not generalize but rather learns by memorization. To counter this, there are multiple different regularization techniques. Since this work often only had to deal with particularly small datasets, all of these techniques were used to varying degrees during training.

Dropout

The dropout technique, presented by Srivastava et al. [2014], randomly sets activations of single neurons to zero. It can be considered as a separate layer in the model architecture that is parameterized with a dropout probability p. While training, the dropout layer receives the output from the previous layer and, for its output, zeros out every value with the probability p. During evaluation, no activations are zeroed out, but the whole output is scaled down with the multiplier $1 - p$ to ensure the total activations are consistent with training-time values. Such dropout layers are typically added after fully connected layers at the end of the network. Using dropout during training prevents the co-adaptation of neurons, forcing the network to learn more robust and redundant representations of the data. As the dropped neurons are chosen randomly for each batch, the training can be interpreted as training an ensemble of models, each with a different number of neurons in the affected fully connected layer.

Stochastic Depth

Huang et al. [2016] presented the stochastic depth regularization, which follows a similar idea as dropout. Instead of zeroing out the activation of single neurons, stochastic depth removes entire layers during training. This requires the layers on which stochastic depth is applied to have a skip connection, as in the residual blocks shown in Figure 2.3. These skip connections, adding the input to the block's output, allow the output to be set to zero while still having an uninterrupted signal through the entire network. Stochastic depth, like dropout, is parameterized with a drop probability p. However, this is not applied uniformly across all layers but scaled linearly, with the first layer having a drop

probability of zero and the last layer having a probability of p. As with dropout, during evaluation, no layer outputs are dropped, but the layer's output is scaled with $1 - p_l$, where p_l is the layer's scaled drop probability. Again, such training can be considered as training an ensemble of models, but with stochastic depth, the ensemble consists of models with different numbers of layers. In addition to reducing overfitting, training with stochastic depth reduces the computational costs and stabilizes the gradients as the networks are shallower.

Weight Decay

Weight decay was first introduced by Hinton [1987]. Instead of altering the network structure, weight decay adds a penalty to weight updates depending on the magnitude of the weight. Using weight decay, the loss becomes

$$L(\mathbf{y}_i, f(\mathbf{x}_i), \theta) = L_{\text{orig}}(\mathbf{y}_i, f(\mathbf{x}_i)) + \frac{\lambda}{2} \sum_{w_i \in \theta} w_i^2$$

where L_{orig} is the original used loss function and θ are the model's weights. When gradient descent is used to update the weights, the weight decay leads to the added term $-\lambda w_i$ in the weight update:

$$\Delta w_i \propto -\frac{\partial L_{\text{orig}}}{\partial w_i} - \lambda w_i$$

The intuition behind this regularization is that weights can only contribute to the output if they are relevant for the prediction, as all weights are pushed to zero unless there is a strong signal from the original loss term. Further, as larger weights lead to a higher sensitivity to small changes in the inputs, the weights are kept as small as possible, which may lead to a better generalization.

Chapter 3

Diagnostic Quality of Radiographs

As one of the most frequently used imaging modalities worldwide [Kjelle et al., 2024], radiographs are of significant importance for diagnosis and treatment planning. For these tasks, a high diagnostic image quality is mandatory. For a radiograph, diagnostic quality refers to its suitability for clinical interpretation by a radiologist. This implies that high diagnostic quality requires more than a technically flawless image; the visible anatomical structures must also fulfill specific diagnostic requirements. These requirements are crucial, as standardized protocols for diagnostic interpretation are followed for a diagnosis. Strash et al. [2004] showed that the diagnostic image quality of the radiographs is a key factor for standardized image content, thereby allowing compliance with these standardized protocols for their interpretation. A radiograph may therefore be of perfect technical quality but can nevertheless be worthless for diagnostic purposes if relevant anatomical structures are not visible due to misalignment.

As of today, incorrect patient positioning is the primary cause of inadequate diagnostic quality, which can lead to radiograph rejection or misdiagnosis [Atkinson et al., 2020; Foos et al., 2009; Little et al., 2017; Serra et al., 2024]. The incorrect positioning of the patient can affect the radiograph in various ways. Firstly, areas relevant to the diagnosis may not be imaged at all. In this case, the collimation is incorrectly adjusted or the patient is mispositioned. These issues are addressed in Chapter 5. Secondly, the patient's pose may be incorrect. In this case, the relevant areas are imaged, but the bones are misaligned, leading to overlapping anatomical structures in the radiograph. Such radiographs deviate from standardized views and are therefore non-diagnostic [Barile et al., 2017; Nguyen et al., 2020]. While there is extensive research assessing radiograph quality based on technical factors such as contrast and noise [Takaki et al., 2020], these parameters are less important with digital radiography. Variables like exposure time are largely automated, and issues such as overexposure can be digitally corrected [Lin et al., 2012; Samei et al., 2014; Willis et al., 2018].

Currently, radiographers have to decide if the quality of the radiograph suffices for the diagnosis or if the imaging process must be repeated, which happens at Step 3 in the workflow shown in Figure 2.1. Being unable to immediately judge the diagnostic quality correctly can result in various disadvantages. Misjudging image quality as sufficient can lead to misdiagnoses and incorrect treatments if the treating radiologist does not

recognize the poor quality. For instance, treatment decisions for fractures often rely on radiographic evidence of displacement, which demands precise and reproducible imaging [Lalone et al., 2015; Lichtman et al., 2011]. Alternatively, it leads to additional time and financial costs if the radiologist has to schedule a new examination and prolong patient hospitalization [Pinto et al., 2018]. In the worst case, if the image quality is misjudged as inadequate, the radiographer would take a second radiograph even though the first one was sufficient, thereby exposing the patient to radiation again. Reasons that radiographers misjudge the image quality may be time pressure, inexperience, or overtiredness. While expert radiologists can typically detect inadequate positioning rapidly with high confidence, this assessment is typically delayed, as images are not assessed by the radiologist immediately after the acquisition.

To prevent these errors and to establish a quality control mechanism, an automated quality assessment can help. However, this remains difficult due to the absence of reliable quantitative metrics. Although deep learning has been increasingly applied across various radiological tasks [Choy et al., 2018; Saba et al., 2019], research on automated assessment of radiograph quality as of today is limited to technical features while overlooking anatomical alignment.

This chapter proposes using anatomical structures as more suitable features. In Section 3.1, the first AI-based framework for diagnostic quality assessment based on anatomical features is presented. For evaluation of the framework, a new dataset consisting of ankle radiographs is collected and annotated by multiple radiologists regarding the diagnostic quality. The results show that, based on such anatomical features, a modular deep-learning framework can serve as a quality control mechanism for the diagnostic quality of radiographs. Subsequently, in Section 3.2, the framework is expanded and refined to work with different body parts. Next to the generalizability, the amount of data and the labeling effort needed to achieve a reliable prediction are investigated. The findings show that the framework not only works reliably across different body parts but also that good predictions can be achieved with minimal time and data effort. Finally, Section 3.3 presents different clinical use cases. Based on real clinical data, factors that influence diagnostic quality are identified. Furthermore, the implementation of a system that provides radiographers with immediate feedback on the diagnostic quality of radiographs is described, and its effects are examined.

3.1 An AI-Based Framework for Diagnostic Quality Assessment of Ankle Radiographs

In this section, we propose the first AI-based framework consisting of regression, classification, and segmentation neural networks, which assesses the diagnostic quality of ankle radiographs based on anatomical features. To the best of our knowledge,

there were no previous approaches assessing the quality of a radiograph based on this criterion.

The framework was tested on a new dataset containing radiographs of ankles, with 950 radiographs in two different radiographic views (*anterior-posterior* and *lateral*), all labeled by four radiologists. Using this framework, radiographers will be able to immediately get an initial quality assessment of the acquired radiographs without relying on a radiologist. Besides reducing judgment errors made by radiographers, the framework can be used as a quality control mechanism to detect causes for low-quality radiographs. Upon evaluation of the framework, an average accuracy of 94.1%, which surpasses the average performance of expert radiologists, was achieved.

In the following, Section 3.1.1 reviews prior and related work. The framework itself is then introduced in Section 3.1.2. Section 3.1.3 describes the used datasets and the data annotation process. The experiments and their results, based on the framework and the datasets, are presented in Section 3.1.4 and are finally discussed in Section 3.1.5.

This section has been published as:

[Mairhöfer et al., 2021] Mairhöfer, D., Laufer, M., Simon, P. M., Sieren, M., Bischof, A., Käster, T., Barth, E., Barkhausen, J., and Martinetz, T. "An AI-based Framework for Diagnostic Quality Assessment of Ankle Radiographs". In: *Proceedings of the Fourth Conference on Medical Imaging with Deep Learning (MIDL 2021)*. PMLR, 2021, pp. 484–496.

According to the Contributor Roles Taxonomy (CRediT), the contributions of the author of this thesis to the publication are: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – original draft (together with M.L.), Writing – review & editing (together with M.L., E.B., T.M.).

3.1.1 Related Work

In recent years, deep learning has become more common in radiology [Choy et al., 2018; Saba et al., 2019]. Researchers working with radiographs successfully applied neural networks to detect fractures [Lindsey et al., 2018; Thian et al., 2019], classify body parts [Agunwa et al., 2019], and radiographic views [Fang et al., 2020], to facilitate the work process in radiology. Although our proposed framework also includes radiographic view recognition, these steps are only part of the preprocessing for assessing diagnostic quality. Distinct from Fang et al. [2020], where only a single step is used for recognition, we use multiple steps containing different networks and do not include laterality recognition.

Esses et al. [2018] and Wang et al. [2020] focus on automated diagnostic quality evaluation of MRI images using neural networks. Due to the different modalities of the imaging systems, one cannot easily transfer the results to radiographs.

Approaches that automatically assess the perceptual quality of radiographs only take technical parameters such as noise and contrast into account and rely on conventional

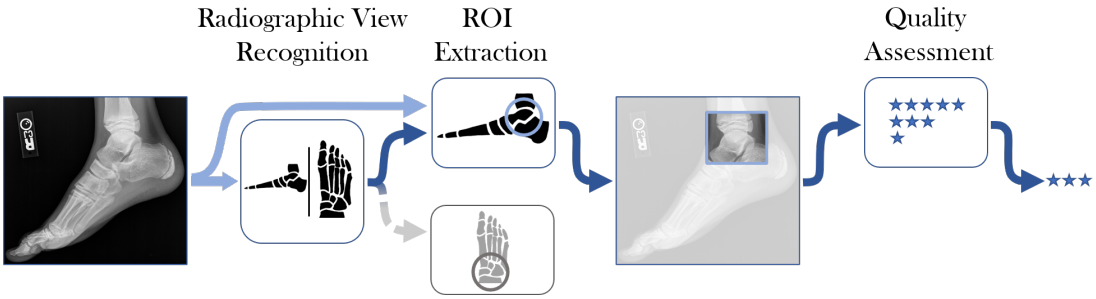


Figure 3.1: Schematic flow of a radiograph through the framework. For each radiograph, the framework decides first which radiographic view was used. Depending on that decision, the radiograph is passed to the corresponding region of interest (ROI) segmentation network. After segmentation, the resulting ROI is fed into the final quality prediction network, which outputs the quality assessment.

computer vision methods [Samei et al., 2014; Willis et al., 2018]. Takaki et al. [2020] present a deep learning approach to calculate the target exposure index for chest radiographs based on the perceptual quality of small patches.

To the best of our knowledge, there are no studies considering anatomical features for the diagnostic quality of radiographs in a deep learning framework.

3.1.2 Proposed Framework

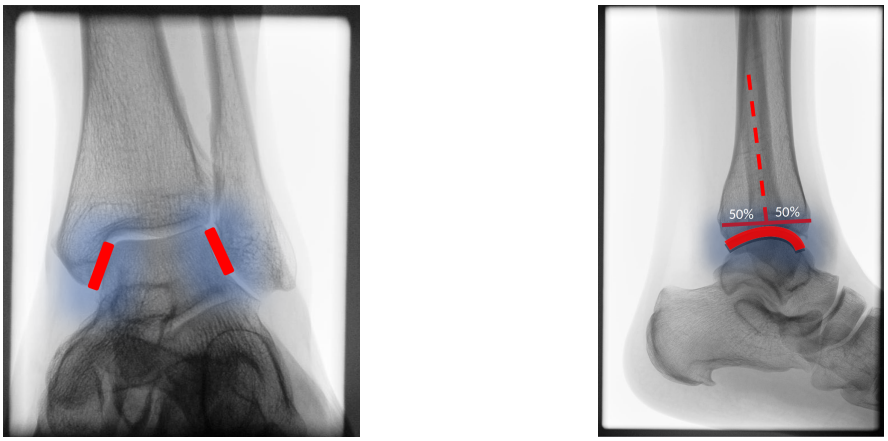
To solve the challenge of diagnostic quality assessment and to standardize the required steps, we propose a framework of several neural networks that can process radiographs of ankles and output their diagnostic quality. It consists of the following steps: *Recognition of Radiographic View*, *Extraction of the Region of Interest (ROI)*, and *Quality Assessment*.

The first step relies on the fact that radiographs can be ordered hierarchically by radiographic view. Quality prediction strongly depends on the view, as the corresponding criteria may differ between radiographic views. The second step prepares the input for quality assessment by removing unnecessary information.

Each individual step can be used independently. But only within the whole framework do they provide the possibility to decide whether an ankle radiograph is of high or low diagnostic quality, thereby directly supporting the radiographers in their decision-making process. A complete overview of the framework is shown in Figure 3.1. In the following, we describe each step in more detail.

Recognition of the Radiographic View

The first step of the proposed framework consists of recognizing the radiographic view of the radiograph. This classification task is essential since radiographs of different



(a) **(b)**

Figure 3.2: In panel (a), the most relevant anatomical structures in the *AP* radiographic view are highlighted. These include the joint gap between the medial malleolus and talus as well as the lateral malleolus and talus. In panel (b), the joint space between the distal tibia and the talus is highlighted as the most relevant structure for the *LAT* radiographic view.

radiographic views differ in quality assessment characteristics, as shown in Figure 3.2. By dividing the quality assessment of an ankle radiograph into view-specific tasks, we facilitate the networks' learning process, as the radiographs then belong to the same domain for each network.

Extraction of the ROI

While the entire radiograph is relevant for diagnosis, only a fraction is required for assessing the quality of the standard view (red marks in Figure 3.2). Based on this fact, the next step in the framework is to segment this ROI, which contains the most information relevant for the diagnostic quality. Example ROIs are shown in Figure 3.3a for the *AP* and in Figure 3.3b for the *LAT* view. Since there are different quality characteristics in the radiographic views, we trained neural networks individually for each view. Besides removing irrelevant information, the benefit of extracting ROIs is that the subsequent quality assessment can operate on a standardized size and resolution of the relevant image part.

Quality Assessment

Obtaining standardized ROIs of a particular radiographic view is the basis for assessing the diagnostic quality accurately. We use two different neural networks, one for each of the two radiographic views. These are trained individually on the *anterior-posterior* (*AP*) and *lateral* (*LAT*) ROI and output the quality on a continuous scale from 1 to 3.

3.1.3 Datasets

To test the framework presented in Section 3.1.2, two datasets were created. The first one is a collection of ankle radiographs as DICOM images and associated metadata. The second one, which, to the best of our knowledge, did not exist previously in this or similar form, contains radiographs labeled by radiologists according to diagnostic quality based on anatomical features. Both datasets contain radiographs from five different imaging systems.

Weakly Labeled Dataset for Recognition of the Radiographic View

We used a dataset of 26,542 ankle radiographs provided by the University Hospital Schleswig-Holstein, Campus Lübeck. From those radiographs we extracted labels for the radiographic view (LAT or AP) using a keyword matching on the metadata. The resulting dataset contains approximately 12,000 radiographs for each view. Since creating the metadata is mostly done manually and the content is not standardized, we assume that not all labels are accurate.

Diagnostic Quality Dataset

To learn the relationship between the radiographs and the quality, an annotated dataset is needed. To create such a dataset, four radiologists labeled 950 ankle radiographs, containing 475 for LAT and AP each.

The radiologists determined which objective criteria a radiograph of an ankle has to fulfill to be of high diagnostic quality. One important criterion, for instance, is the complete visibility of the joint gap between the medial malleolus and talus. According to those criteria, each radiograph was labeled by each radiologist as *1* if the radiograph fulfilled the criteria perfectly, *2* if partly, and *3* if the criteria were not met and a new radiograph would have to be taken. To determine whether a radiograph can be used for a diagnosis, classes *1* and *2* were grouped under the label *diagnostic*, and class *3* was labeled as *non-diagnostic*. If the labels differed greatly, the radiologists had a consensus meeting. Of the 475×4 labels assigned for the AP radiographs, 37% are *1*s, 53% are *2*s, and 10% are *3*s. For the LAT view, 17% of the assigned labels are *1*s, 55% are *2*s, and 28% are *3*s. Examples for the three classes can be seen in Figure 3.4 (a-c) for the AP view and Figure 3.5 (a-c) for the LAT view.

Additionally, each of the 950 radiographs was labeled with an ROI. As previously described, only a fraction of the radiograph is relevant for the diagnostic quality. Therefore, the ROI was labeled as a square containing only the most relevant information. In Figure 3.3, which shows examples of ground truth ROI labels, it can be seen that the size of each ROI is highly dependent on the image content. Examples of the extracted ROIs can be seen in Figure 3.4 and Figure 3.5.

3.1.4 Experiments and Results

To evaluate the framework described in Section 3.1.2, each step was implemented using PyTorch, and evaluated on the datasets of Section 3.1.3. To improve quality control measurements, we tested each step individually. Because of the relatively small datasets, we used the EfficientNet-B0 [Tan et al., 2019b] for classification. For segmentation, a DeepLabV3 [Chen et al., 2017] with a ResNet-50 [He et al., 2016] backbone was used. Both networks were not pretrained. For all experiments, we padded the input radiograph with zeros to obtain the desired size while maintaining the aspect ratio. Furthermore, the training radiographs were augmented with random cropping, histogram normalization, Gaussian noise, blurring, horizontal flipping, and rotation. Training and test datasets were split with an 80/20 ratio.

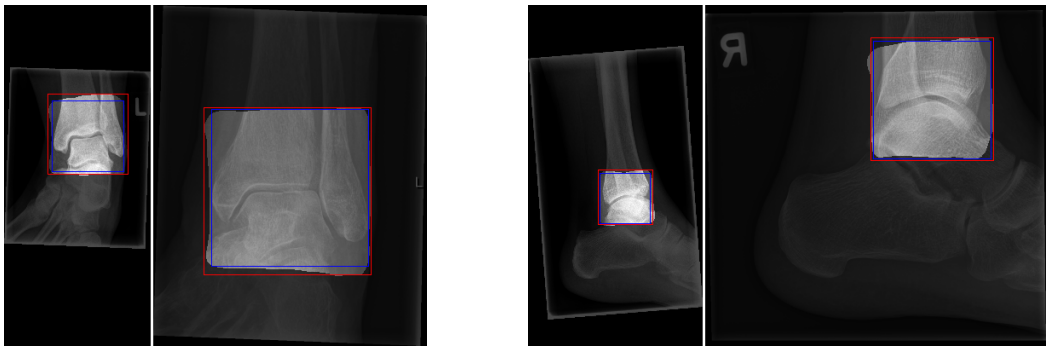
Recognition of the Radiographic View

For the recognition of different radiographic views, the weakly labeled dataset was used. Therefore, the last layer of the EfficientNet-B0 was modified to output two classes, either *LAT* or *AP*, which was followed by a softmax layer to obtain class probabilities. The model was trained using the cross-entropy as the loss function and stochastic gradient descent (SGD) as the optimizer using a learning rate of $1 \cdot 10^{-3}$, a momentum of 0.9, a weight decay of $1 \cdot 10^{-5}$, and a batch size of 8 over 500,000 iterations. To reduce possible overfitting, the stochastic depth drop connect rate [Huang et al., 2016] was set to 0.4. The resulting input size of the radiographs, after augmentation, was 224×224 pixels.

Training with these parameters resulted in an accuracy of 98.4% for the test set and 98.5% for the training set. The results must be interpreted with caution due to the potentially incorrectly assigned labels in the weakly labeled dataset. It may be that (i) the model predicts the correct class but the label is assigned incorrectly (e.g. prediction *AP*, weak label *AP*, true label *LAT*) or that (ii) the model predicts the incorrect class and the label is also assigned incorrectly (e.g. prediction *AP*, weak label *LAT*, true label *AP*). Reviewing the resulting radiographs for case (ii) revealed 54 incorrect labels for the test set and 244 for the training set. Taking this into account, the accuracy increased to 99.5% for the test set and 99.7% for the training set. Although the actual accuracy may be slightly lower due to errors of case (i), these results clearly demonstrate that a recognition of the radiographic view can be achieved with high precision.

Extraction of the ROI

To segment the ROI, a DeepLabV3 was trained with the labels from the diagnostic quality dataset. The target feature map is binary, with 0 for *non-ROI* and 1 for *ROI*. As segmentation output, we used a single feature map, followed by a sigmoid function,



(a) (b)

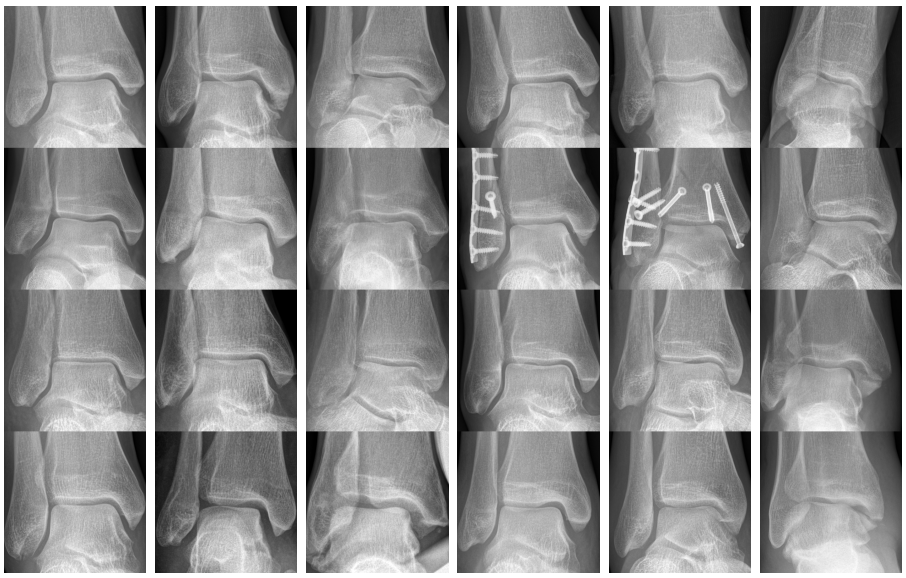
Figure 3.3: Panel (a) shows two radiographs in the *AP* view. Their labeled ROI is marked with a blue box, and the predicted ROI with a red box. The predicted segmentation mask used to construct the red box is highlighted. The same is shown in panel (b) for the *LAT* view. Both examples also show that the proportion of ROI in the radiograph can vary greatly.

to get pixel-wise outputs from 0 to 1. For the training, we used the mean over the pixel-wise squared error, optimized with the Adam optimizer, a learning rate of $1 \cdot 10^{-4}$, a weight decay of $1 \cdot 10^{-4}$, and a batch size of 4 over 50,000 iterations. For this task, the input size after augmentation was 400×400 pixels. This training was done separately for *LAT* and *AP* views. Given the small dataset, we used a random sub-sampling validation over 12 different dataset splits.

To measure the accuracy of the predicted ROIs, the Dice score was calculated. If a pixel value of the output feature map was above a threshold of 0.7, the pixel was classified as part of the ROI. Over all 12 dataset splits, the mean Dice score was 94.17% on the *AP* views and 85.91% on the *LAT* views. One reason for the lower Dice score on the *LAT* views might be that the ROIs on the *LAT* views are significantly smaller than on the *AP* view and thus harder to predict. Regardless of this difference in the Dice score, the resulting segmentations are sufficient to get bounding boxes of the ROIs. To extract bounding boxes based on the segmentation, first the smallest fitting rectangle of the segmentation is calculated and then rotated to be horizontal. Examples with the labeled and the predicted ROIs are shown in Figure 3.3.

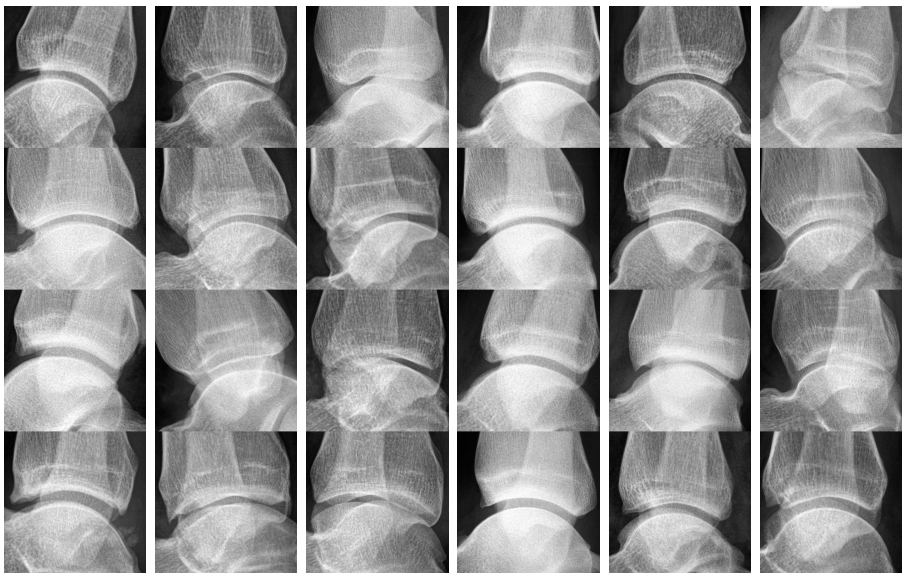
Quality Assessment

For the quality assessment task, an EfficientNet-B0 was used. To preserve the intrinsic order of the classes, we modeled the task as a regression. One benefit of using regression is that we obtain intermediate scores. We also trained classification networks using the earth mover's distance, but this led to slightly worse results. The model was trained using the mean squared error (MSE) as loss and the mean label of the four radiologists as target. The loss was minimized by SGD using a learning rate of $1 \cdot 10^{-3}$, a momentum



(a) **(b)** **(c)** **(d)** **(e)** **(f)**

Figure 3.4: Example ROIs of the labeled dataset in the *anterior-posterior* view. Columns (a), (b), and (c) show examples with expert labels *1*, *2*, and *3*. Columns (d), (e), and (f) show unlabeled ROIs for which the framework predicts the quality classes *1*, *2*, and *3*.



(a) **(b)** **(c)** **(d)** **(e)** **(f)**

Figure 3.5: Example ROIs of the labeled dataset in the *lateral* view. Columns (a), (b), and (c) show examples with expert labels *1*, *2*, and *3*. Columns (d), (e), and (f) show unlabeled ROIs for which the framework predicts the quality classes *1*, *2*, and *3*.

of 0.9, a weight decay of $1 \cdot 10^{-3}$, and a batch size of 16 over 500,000 iterations. As with the recognition of the radiographic view, the input size was 224×224 pixels. The same random sub-sampling validation as for the extraction of the ROI was used for testing.

To evaluate the accuracy of the model, an output was counted as correct if the nearest class to the continuous output was the class of the label. Evaluation on the test set resulted in a mean accuracy of 93.0% for the *AP* view and 95.1% for the *LAT* view, with a mean absolute error of 0.19 for *AP* and 0.20 for *LAT*. Over the 12 runs, the standard deviation is 0.025 and 0.02 and the median accuracy is 93.4% and 95.4% for *AP* and *LAT*, respectively. The classification into *diagnostic* and *non-diagnostic* resulted in an accuracy of 97.8% for the *AP* view and 93.2% for the *LAT* view. This accuracy shift is because there are different distributions of *1*s and *3*s in the *AP* and *LAT* parts of the dataset.

To evaluate whether the accuracy of the quality assessment benefits from the steps described in Section 3.1.2, we repeated the training with and without these steps. The results, which are given in Table 3.1, show that each step of the pipeline improves the accuracy. Overall, the mean accuracy improves from 82.4% to 94.1% when all steps are included. While the benefit of training separately for the different views is small, the extraction of ROIs appears to be necessary to obtain high accuracy. When trained without the previous view recognition, a single model is trained on the combined *AP* and *LAT* data to predict the quality of both views. For this, each view is sampled equally often.

To get an estimate on how accurate the labels are, we tested each labeling radiologist against the others, taking one label as the prediction and the mean of the remaining three as ground truth. If the difference between prediction and ground truth was at least 1, the prediction was counted as incorrect. This resulted in a mean accuracy of 92.6% for *AP* and 90.1% for *LAT*. Across the four radiologists, the standard deviation is 0.026 and 0.037 for *AP* and *LAT*, respectively. The mean accuracy over both views is 94.1% for the networks and 91.4% for the radiologists. Although our method for evaluating the performance of the radiologists is based on only four experts, it should suffice as a first estimate.

A visual comparison of the expert labels and framework predictions on the unlabeled dataset is shown in Figure 3.4 for the *AP* view and Figure 3.5 for the *LAT* view. For further illustration, the ROIs with the highest error between expert label and predicted quality are shown in Figure 3.6. It should be noted that there is no clear pattern that explains the deviation.

3.1.5 Discussion

The aim was to develop a framework for automatic quality assessment and to evaluate how well it performs. Our results demonstrate that the model's predictive accuracy (93.0% *anterior-posterior*, 95.1% *lateral*) exceeds the inter-observer agreement of the

Table 3.1: Accuracy of quality assessment depending on the steps *View Recognition* and *ROI Extraction* described in Section 3.1.2. Not training separately for *AP* and *LAT* and not extracting the ROI leads to the lowest accuracy. Both steps on their own increased the accuracy, while using both provided the best result.

View Recog.	ROI Ext.	Accuracy		
		mean	*AP*	*LAT*
✗	✗	82.4%	80.3%	84.5%
✓	✗	85.1%	82.9%	87.2%
✗	✓	92.4%	92.2%	92.5%
✓	✓	94.1%	93.0%	95.1%

Table 3.2: Overview of all steps in the framework and their results. The results for the *View Recognition* and the *Quality Assessment* are the achieved accuracy. For the *ROI Extraction* the result is the achieved Dice score. The *AP* and *LAT* results are not from the same model because we trained individually for each view. Since this is not the case for the *View Recognition*, there is only a single accuracy.

Step	Accuracy or Dice		
	mean	*AP*	*LAT*
View Recognition	99.5%	–	–
ROI Extraction	90.1%	94.2%	85.9%
Quality Assessment	94.1%	93.0%	95.1%

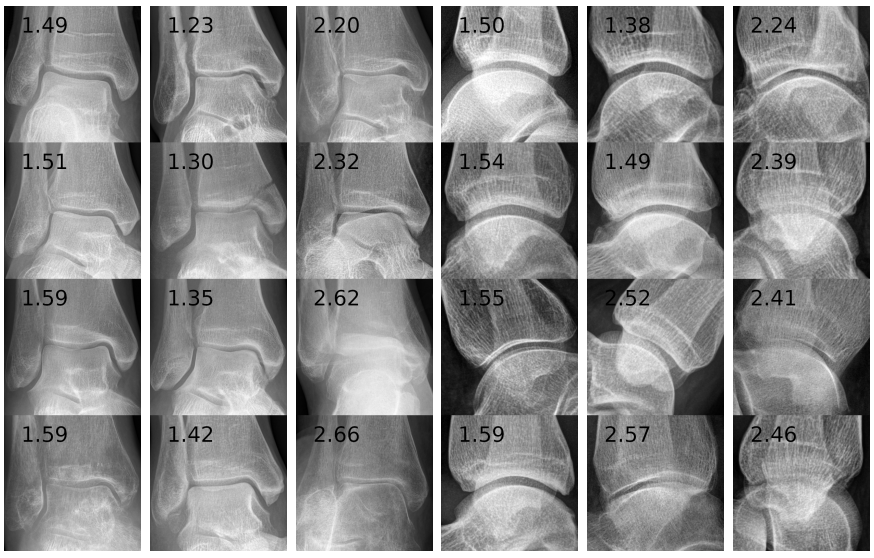


Figure 3.6: ROIs of the labeled dataset with the highest prediction errors. Columns (a), (b), and (c) show examples in the *anterior-posterior* view with expert labels *1*, *2*, and *3*. Columns (d), (e), and (f) show examples in the *lateral* view with expert labels *1*, *2*, and *3*. The quality predicted by the framework is printed on each ROI. For each class and view, the four ROIs with the highest error between expert label and predicted quality are shown.

radiologists (92.6% *anterior-posterior*, 90.1% *lateral*). The results of the individual steps included in the framework are summarized in Table 3.2. With this framework, it is now possible for radiographers to get initial feedback immediately on the same level of expertise as they would get from a radiologist. These results support our view that anatomical features can be learned and are therefore suitable for the automatic assessment of diagnostic quality. To achieve these results, an initial separation of the radiographs into *lateral* and *anterior-posterior* was necessary. This task could be achieved with an accuracy of 99.5%.

If our framework had already been in place when acquiring the 950 radiographs of our dataset, 80.0% of the non-diagnostic radiographs would have been immediately and correctly recognized as such. Since 12.9% of the dataset are non-diagnostic radiographs, for every 100 radiographs, the number of additional needed appointments for examinations could have been decreased from 13 to only 3.

Regarding scalability, our experiments show that approximately 500 labeled radiographs per radiographic view are sufficient to train a network to the accuracy level of an expert. We assume that the framework can be transferred to radiographs of other body parts. In addition to its use in day-to-day operations, the framework can potentially help to comply with quality standards and optimize clinical routine.

3.2 Generalizable Deep Learning Framework for Diagnostic Quality Assessment of Musculoskeletal Radiographs

In Section 3.1, the first approach for assessing the diagnostic quality regarding patient positioning was presented, and an AI-based framework for assessing ankle radiographs was proposed. By using this framework, a prediction of the diagnostic quality at the accuracy level of expert radiologists was achieved. While this shows prototypically that the diagnostic quality regarding patient positioning can be predicted for ankle radiographs, these results may not necessarily generalize to other anatomical regions. Further, even if the results transfer to other body parts, the question remains whether the approach scales well regarding the required number of annotated radiographs and the required labeling time.

This section closes this gap by extending the quality prediction to ankle, knee, and wrist radiographs. Further, the labeling process is refined, allowing for a more precise quality annotation while at the same time measuring the required cost for labeling. We analyze the necessary amount of training data to examine the scalability and thus the practicality of the developed framework. Publishing the trained models simplifies and accelerates further research regarding the diagnostic quality of radiographs. Furthermore, it enables a broader community to automatically assess diagnostic quality for the first time. The trained models and the code for inference can be found GitHub[1].

This section is currently under review as:

[Gerdes et al., 2026] Gerdes, H.*, Mairhöfer, D.*, Laufer, M., Reis, F. L., Bischof, A., Wegner, F., Käster, T., Barth, E., Barkhausen, J., Martinetz, T., and Sieren, M. *Generalizable Deep Learning Framework for Diagnostic Quality Assessment of Musculoskeletal Radiographs.* *Authors contributed equally. 2026. **Under Review at Radiology: Artificial Intelligence**.

According to the Contributor Roles Taxonomy (CRediT), the contributions of the author of this thesis to the publication are: Conceptualization (together with H.G.), Data curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – original draft (together with H.G.), Writing – review & editing (together with H.G., M.L., M.S., T.M.)

3.2.1 Related Work

While there is some preliminary earlier work regarding the diagnostic quality of radiographs, the topic remained largely unexplored, despite its importance for reliable diagnoses and exposure minimization. Krönke et al. [2022] attempt to assess the

[1] `https://github.com/DominikMa/xray-quality-assessment`

diagnostic quality of ankle radiographs by predicting different rotation angles of the ankle. While they predict the ankle rotation, it remains unclear how the diagnostic quality can be derived from the ankle rotation. Köpnick et al. [2023] follow a similar pipeline as presented in Mairhöfer et al. [2021] but use only lateral ankle radiographs. They confirm the results using a real ankle phantom.

Chen et al. [2022] proposed an enhanced U-Net framework to evaluate lumbar spine radiographs by segmenting anatomical landmarks and verifying whether each view (*AP*, *LAT*, oblique) met standardized positioning criteria. Their system demonstrated strong performance in identifying incomplete or misaligned lumbar spine images and was capable of operating in near-real time, highlighting the feasibility of anatomically grounded quality control in a clinical setting.

At a conceptual level, the work by Poggenborg et al. [2021] emphasized the importance of radiograph quality control for both clinical safety and AI development. Their findings stressed that training models on low-quality data can lead to unreliable outcomes, advocating for automated filtering of suboptimal studies. Similarly, Kashyap et al. [2019] discussed the potential of AI for point-of-care radiograph quality assessment, proposing a general framework for integrating image feedback into acquisition workflows. However, both studies remained conceptual or relied on proof-of-concept prototypes, without large-scale evaluation across multiple body parts.

3.2.2 Dataset and Annotation

Study Population and Data Acquisition

Two datasets were constructed for this study. All radiographs were acquired using modern digital imaging systems (DigitalDiagnost C90 & PCR Eleva, Philips, Netherlands; Fluorospot Compact FD, Siemens, Germany).

The first dataset consists of 24,431 ankle, 7,810 knee, and 12,291 wrist radiographs exported from the local PACS archive. Selection was randomized and independent of clinical indication or pathology to ensure a representative clinical sample. Keyword matching on the DICOM metadata was used to generate weak labels for the body part and view (anterior-posterior (*AP*) or lateral (*LAT*)).

The second dataset was annotated manually and comprised 2,000 radiographs of the ankle, 800 of the knee, and 800 of the wrist, which were selected at random but with a balanced distribution between *AP* and *LAT* radiographs. For this dataset, radiographs of patients under the age of 16, patients with severe joint deformities, or patients with implants obscuring the joint space were excluded.

Data Annotation

Four radiologists, each with 6–25 years of experience in musculoskeletal imaging, independently rated the radiographs. A semi-quantitative five-point scale from 1.0

(optimal) to 3.0 (non-diagnostic), in 0.5 increments, was used. Compared to Section 3.1, labeling was done in finer granularity, as many radiographs were not clearly assignable to one of the three categories. Standardized quality criteria guided the assessment, including joint-specific benchmarks. The assessment time per radiograph was recorded for each radiologist. Figure 3.7 shows examples of radiographs with different ratings.

For the final rating, the mean of all four radiologists' ratings was used. If the maximum inter-rater deviation exceeded 1.0, the radiograph was reviewed in a consensus meeting. Radiographs were additionally classified as *diagnostic* or *non-diagnostic*; a radiograph was considered *diagnostic* if its quality rating was in the range [1, 2.5).

Each radiograph was further annotated with a square region of interest (ROI) capturing the anatomical structures relevant for quality assessment. Examples of these ROIs are shown in Figure 3.8.

Detailed Criteria for the Quality Assessment The following criteria are the standards established by radiologists to determine the diagnostic quality of a radiograph.

Ankle *AP* The distal tibia and fibula should be seen in profile without superimposition and with clear visualization of the joint space.

Ankle *LAT* The distal fibula should be superimposed with the posterior aspect of the distal tibia. The tibial joint surface should form a smooth arc with the talar dome, showing an open joint space without evidence of overlay or irregularity.

Knee *AP* The femoral condyles and tibial plateau should be symmetrical with a clearly defined joint space and without overlap of other structures. The head of the fibula is slightly superimposed by the lateral tibial plateau. The patella is in a neutral position, superimposing the distal femur on the superior portion of the radiograph.

Knee *LAT* The femoral condyles should be superimposed and aligned parallel to each other, whereby the femoral condyles should appear as smooth, continuous arcs. The patellofemoral and femorotibial joint spaces should be clearly visualized with no superimposition. The head of the fibula is slightly superimposed by the tibia.

Wrist *AP* Each carpal bone should be clearly visible without overlap. The distal radioulnar joint should demonstrate smooth articulation between the ulnar head and the radius with minimal or no superimposition. Nearly equal distances exist among the proximal metacarpals.

Wrist *LAT* The carpal bones overlap, including the pisiform and the distal part of the scaphoid. There is a superimposition of the ulnar head and the distal radius. The ulnar styloid is shown posteriorly.

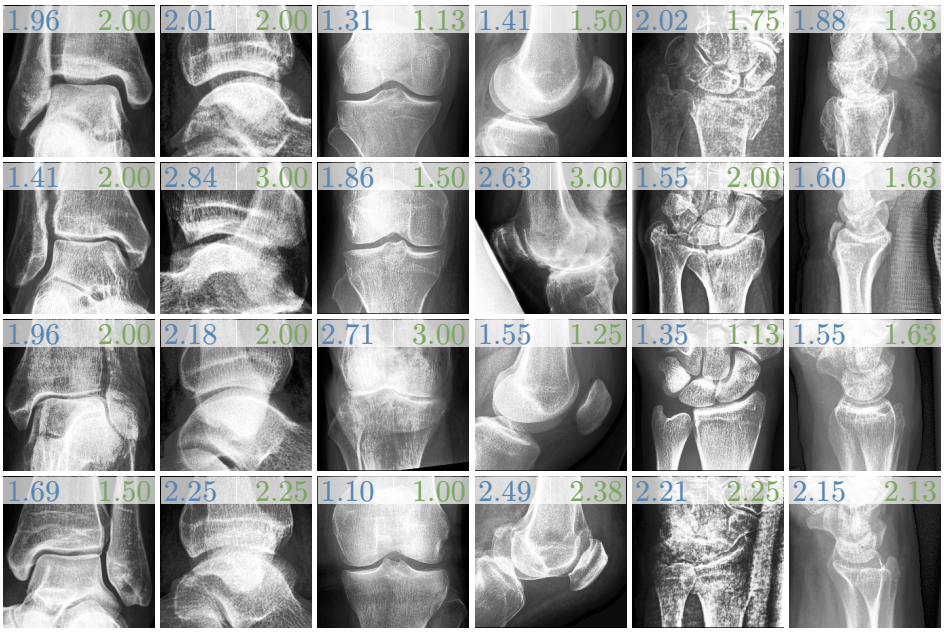


Figure 3.7: This figure shows randomly selected ankle, knee, and wrist ROIs, each in *AP* and *LAT* position, annotated with their labeled quality (green, top right) and their predicted quality (blue, top left).

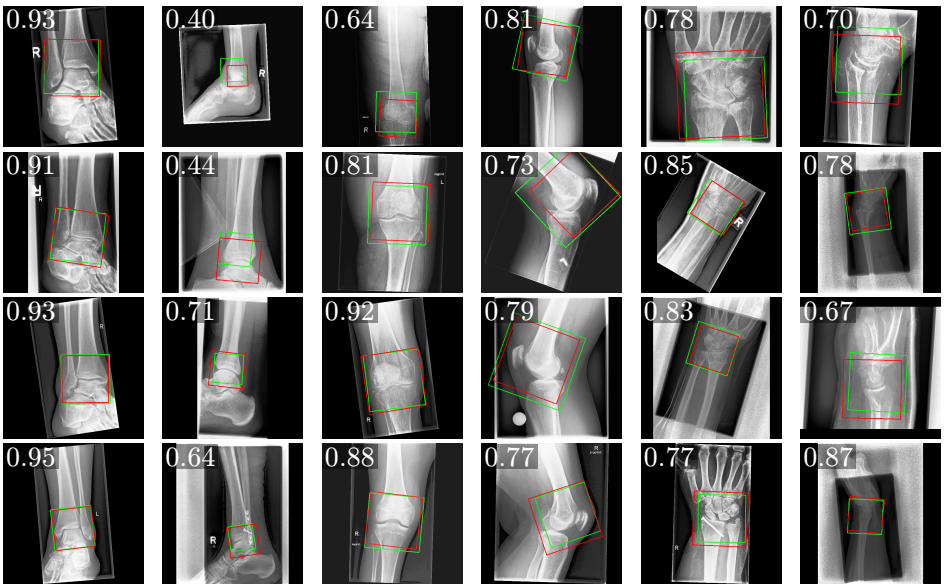


Figure 3.8: This figure shows randomly selected ankle, knee, and wrist radiographs annotated with their labeled ROI in red and their predicted ROI in green. The intersection over union between both is shown in the top left of each radiograph.

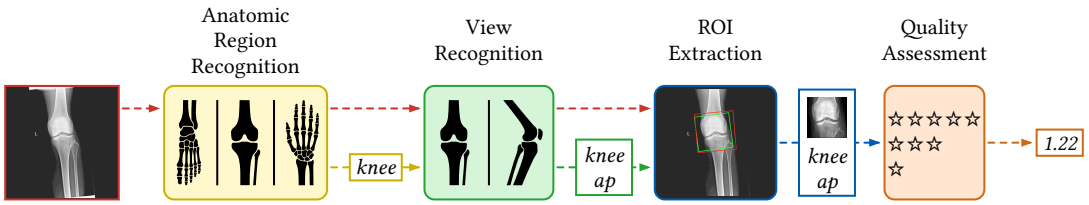


Figure 3.9: The figure shows a schematic overview of the quality assessment framework. First, the body part of the input radiograph is detected. In the second step, based on the body part, the view is detected. Using the body part and view information in the third step, the relevant region of interest (ROI) on the input radiograph is chosen and sampled into a new image. In the fourth step, the ROI image is used to assess the quality.

3.2.3 Quality Assessment Framework

The automatic quality assessment pipeline is based on the framework described in Section 3.1.2. As the framework assesses multiple body parts, not only ankles, body part recognition is added as a preceding step. In addition, the ROI prediction is changed from a segmentation to a prediction of transformation parameters to allow for rotated ROIs. The updated framework consists of four sequential stages: (1) *Body Part Recognition*, (2) *View Recognition*, (3) *ROI Extraction*, and (4) *Quality Assessment*. A schematic overview is shown in Figure 3.9.

Body Part Recognition The weakly annotated dataset was divided into training and test sets, with 80% set aside for training and 20% for testing. An EfficientNet-B0 [Tan et al., 2019b] was trained to classify radiographs into the classes *ankle*, *knee*, and *wrist*, minimizing the cross-entropy loss.

View Recognition After the radiographs are grouped according to their examined body part, the view is classified into *AP* and *LAT*. For each body part, a separate EfficientNet-B0 was trained, again using the weakly annotated dataset, split at an 80:20 ratio. This modular approach breaks the task down into simpler subtasks, as each network only needs to learn features of a single body part.

The training procedure for the body part and the view recognition is identical. An EfficientNet-B0 [Tan et al., 2019b] was optimized for 200,000 iterations using stochastic gradient descent with an initial learning rate of $1 \cdot 10^{-3}$, a weight decay of $1 \cdot 10^{-4}$, and a momentum of 0.9 to minimize the cross-entropy loss. The initial learning rate was warmed up linearly from $1 \cdot 10^{-5}$ for 1,000 iterations and was decreased during training by a factor of 0.3 when the loss did not improve for over 90 minutes. The batch size of each iteration was 8. The model weights were initialized randomly. The dataset was randomly split, with each radiograph having an 80% probability of being in the training

set and a 20% probability of being in the test set. During training, radiographs of all classes were sampled equally often, avoiding a bias towards a body part or a view.

For augmentation, the following transformations were applied to the original radiograph:

- histogram equalization with a probability of 5%
- zero padding the shorter side
- random-sized crop with a fixed aspect ratio
- adding Gaussian noise with a probability of 70%
- augmenting brightness and contrast with a probability of 50%
- horizontal flipping with a probability of 50%
- resizing to 224×224

ROI Extraction While the whole radiograph may be relevant for diagnosis, only specific anatomical subregions are relevant for assessing patient positioning and diagnostic quality. For each body part and view, a CNN was trained to output the four parameters needed for the construction of a similarity transformation matrix transforming the original radiograph to the annotated ROI: scaling, x- and y-translation, and rotation. Examples of annotated and predicted ROIs are shown in Figure 3.8.

In Section 3.1.2, this step was modeled as segmentation, which had the disadvantage that rotations in the image could not be uniquely converted into bounding boxes. In particular, if the rotation exceeds 90 degrees, the resulting ROI would also be rotated. Learning a transformation ensures that all ROIs not only show the same area but are also rotated in the same way.

To extract an ROI from the original radiograph, a possibly rotated square ROI was first annotated on the radiograph. The radiograph was then zero-padded to have equal width and height, and the similarity transformation from the padded radiograph's image corners to the coordinates of the annotated ROI was calculated. Since the padded radiograph and the ROI are both squares, the transformation only consists of a scaling factor, a rotation, and a 2D translation. The calculated transformation matrix was decomposed into these factors, which were then used as targets for the training of the model.

The CNN used has eight convolution layers, whose parameters are shown in Table 3.3. Each convolution layer is followed by a batch normalization and a ReLU activation. The output of the last convolution layer is max pooled to a spatial resolution of 128×128, flattened, and then fed into a three-layer MLP projecting the flattened representation to 32, 8, and finally 4 features. In the MLP, all but the last projections are followed by ReLU activation functions.

For optimization, the absolute error between the model's output and the four transformation factors is minimized using the Adam optimizer [Kingma et al., 2017] for 1,000,000 iterations with a batch size of 64 using learning rate initialization and

Table 3.3: ROI Extraction CNN Layers. The table shows the parameters of the convolution layers of the ROI extraction network. For all layers, zero padding, keeping the output at the same size as the input, is used.

Layer Number	Kernel Size	Output Channels
1	9	32
2	9	32
3	3	32
4	9	32
5	3	32
6	7	64
7	3	64
8	7	2

scheduling as for the body part and view recognition. The same dataset split and image augmentation as for the body part and view recognition were used except for the input size, which is 512×512 instead of 224×224.

Quality Assessment The quality assessment, as the final step of the pipeline, used the extracted ROI as input. An EfficientNet-B0 [Tan et al., 2019b] was optimized for each body part and each view in a regression task to predict the mean annotated quality. Therefore, the mean squared error (MSE) was minimized over 500,000 iterations using a stochastic gradient optimization. The hyperparameters and learning rate scheduling are the same as for the body part and view recognition, except for the batch size, which is 16. Likewise, the weights were initialized randomly, and the same data augmentation is used. Unlike than the previous steps, the input image is the ROI extracted out of the original radiograph. The data was again split at an 80:20 ratio, and the training was repeated 10 times with different random splits implementing a Monte Carlo cross-validation.

To assess scalability, we systematically reduced the training samples, simulating restricted availability of labeled data. We trained with 10%, 20%, 40%, 60%, and 80% of the available data, keeping the test set constant across experiments. Each subset experiment was repeated 30 times with different random splits. These repetitions were trained for 200,000 iterations instead of the original 500,000 for computational cost savings.

Statistical Analysis and Result Evaluation

The accuracy and F_1 scores were calculated for the body part and view classifications. As body part prediction is a multi-class classification, micro- and macro-F_1 scores are reported. As view prediction is binary, recall and precision are also reported.

ROI performance was quantified by comparing the predicted and annotated transformation parameters. To compare these, the rotation error in degrees, the size relative to the target size, and the translation error of the ROI center in pixels are calculated. In addition, the intersection over union (IoU) between the predicted and labeled ROI is measured.

As the quality prediction is modeled as a regression, the prediction error is calculated as the mean absolute error (MAE). Furthermore, we define the quality of a radiograph as correctly predicted if the absolute error between prediction and label is less than 0.5. We refer to this metric as *Correct*. We choose this threshold because it also corresponds to the granularity of the labeling by the radiologists. For binary classification into *diagnostic* and *non-diagnostic* radiographs, we calculate accuracy, called *Diagnostic Accuracy*, precision, recall, and F1-score.

To quantify interrater reliability and the alignment of the model with human ratings, intraclass correlation coefficients (ICC) and their 95% confidence intervals were calculated based on a two-way random effects model for absolute agreement with mean-rating (k = 2), following Koo et al. [2016].

All reported results are for the predictions on the test set. All results regarding the quality assessment are reported with the standard deviation over the experiment repetitions.

3.2.4 Results

Labeling

Quality labels were assigned to 3,474 of the 3,600 selected radiographs; the remainder were excluded due to labeling ambiguity or quality issues. A rating of 2.0 was most frequently assigned across all body parts and views. Fewer radiographs were rated at the extremes of the scale, with 1.0 being more common than 3.0. This distribution was expected, since radiographs with low diagnostic quality are avoided in clinical practice. The distribution of ratings is shown in Table 3.4.

The median time required to label a single radiograph was 12.5 seconds. While individual labeling times varied between radiologists (24.0 s, 13.9 s, 10.3 s, and 8.9 s), no systematic differences were observed regarding body part, view, or assigned rating.

Inter-rater agreement among radiologists was good to excellent across all joints, as shown by the intraclass correlation coefficients in Table 3.5. This shows that the defined quality criteria can be applied consistently across different raters and radiographs.

Body Part and View Prediction

The model predicting the body part of the radiograph achieved an accuracy of 99.88% and micro- and macro-F_1 scores of 99.88% and 99.86%, respectively. The confusion matrix is provided in Table 3.6.

Table 3.4: Distribution of diagnostic quality ratings by body part and view. The table presents the number of radiographs assigned to each quality rating, stratified by body part (ankle, knee, wrist) and view (anterior-posterior (AP), lateral (LAT)). Ratings range from 1.0 (optimal diagnostic quality) to 3.0 (non-diagnostic), in increments of 0.5. Ratings of 1.5 and 2.5 represent intermediate quality levels. These distributions highlight the predominance of moderate- to high-quality images in the clinical dataset.

Rating	Ankle		Knee		Wrist	
	AP	LAT	AP	LAT	AP	LAT
1.0	1,087	737	318	160	329	211
1.5	0	0	408	412	485	311
2.0	1,337	1,409	475	559	434	500
2.5	0	0	220	335	121	292
3.0	504	840	154	138	44	102

Table 3.5: Inter-rater agreement by body part and view. Intraclass correlation coefficients (ICC) are reported for each body part and view (AP and LAT), based on a two-way random-effects model with mean-rating (k = 2) and absolute agreement. The table includes the ICC values along with their corresponding 95% confidence intervals. Results indicate good to excellent agreement for all body parts and views.

Metric	Ankle		Knee		Wrist	
	AP	LAT	AP	LAT	AP	LAT
ICC (2k)	0.90	0.88	0.92	0.89	0.85	0.73
95% CI_{min}	0.88	0.85	0.90	0.85	0.78	0.60
95% CI_{max}	0.91	0.89	0.94	0.92	0.89	0.80

Table 3.6: The table shows the confusion matrix for the body part recognition with the body parts ankle, knee, and wrist. The results show that a near-perfect distinction can be made. Mismatches between label and prediction were manually verified and are the result of wrong labels that have been produced by keyword matching.

Label \ Prediction	Ankle	Knee	Wrist
Ankle	4,806	2	5
Knee	2	1,587	0
Wrist	0	2	2,430

Table 3.7: The table shows the results for the binary classification of the view into *AP* and *LAT*. The results for the body parts ankle, knee, and wrist are all similarly high.

Metric	Ankle	Knee	Wrist
Accuracy	0.9983	0.9975	0.9967
Precision	0.9979	1.0000	0.9960
Recall	0.9988	0.9950	0.9976
F_1-score	0.9983	0.9975	0.9968

The view prediction resulted in a mean accuracy of 99.75% over the three body parts, an F_1-score of 0.9975, recall of 0.9971, and precision of 0.998. The results for the individual body parts are shown in Table 3.7. These results show that a high-quality prediction model can be learned for body part prediction and view prediction even when using a weakly labeled dataset.

ROI Prediction

Predicting the ROI can be split into the prediction of rotation, translation, and cropping. For the rotation, the trained model has an average absolute error of 1.55 degrees. Measuring the translation error between the centers of the predicted and labeled ROI results in an average error of 35.04 pixels. The average size of the predicted ROI relative to the labeled ROI is 102.44%, while the average IoU is 0.81. These results show that the deviations between the predicted ROI and the labeled ROI are small. Even when the predicted ROI does not align perfectly with the labeled one, it still contains the relevant anatomical structures, as it has a similar position, size, and rotation. Visual examples of predicted and labeled ROIs are shown in Figure 3.8. Furthermore, Figure 3.10 shows the ten worst predictions for each body part and view. It can be seen that the *LAT* view of the knee has a high variability, leading to worse results. With the exception of a few outliers, even for the worst predictions, the labeled ROI is always included in the predicted ROI. The results for all individual body parts and views can be found in Table 3.8.

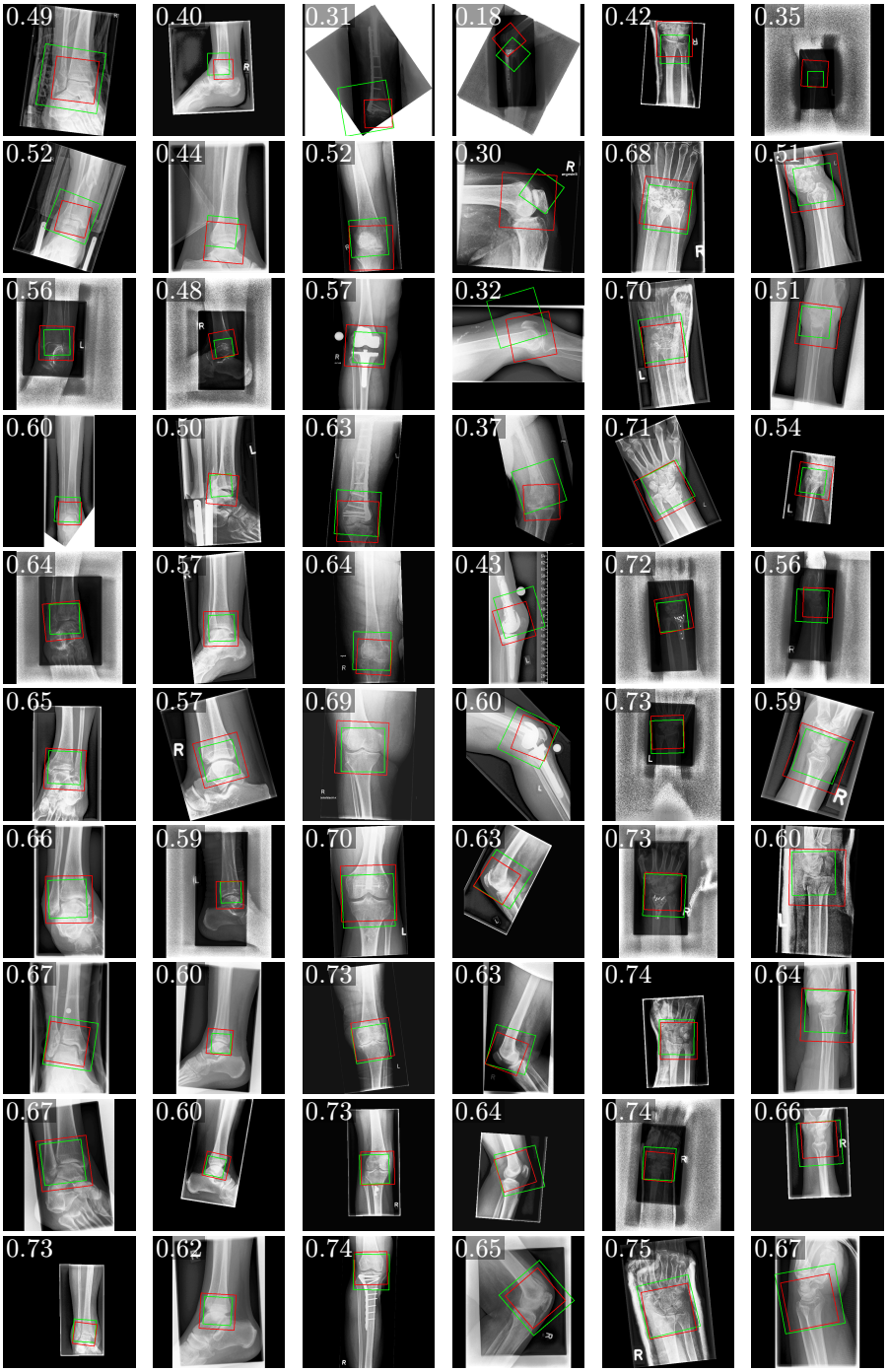


Figure 3.10: This figure shows failure cases for ankle, knee, and wrist radiographs annotated with their labeled ROI in red and their predicted ROI in green The shown radiographs are the ones with the lowest IoU between label and prediction.

Table 3.8: The table shows the results for the extraction of the region of interest (ROI) out of the whole radiograph. The *Rotation Error* (degree), *Center Error* (distance in pixels), *Intersection over Union*, and *Relative Size* (predicted area divided by annotated area) are measured.

Metric	Ankle		Knee		Wrist	
	AP	*LAT*	*AP*	*LAT*	*AP*	*LAT*
Rotation Error	1.65	0.99	1.73	2.51	1.25	1.15
Center Error	18.31	20.91	43.11	66.22	29.87	31.84
IoU	0.87	0.78	0.83	0.77	0.82	0.77
Relative Size	1.00	1.02	1.00	1.15	1.00	0.98

Quality Assessment

The final step of the quality assessment framework, the quality prediction itself, achieved an MAE of less than 0.23. The accuracy, using the previously defined *Correct* metric, ranges from 89.77% for the ankle *AP* view to 94.35% for the wrist *AP* view. Both metrics demonstrate the prediction of diagnostic quality with a low error regardless of the body part and view. Furthermore, the classification into *diagnostic* and *non-diagnostic* radiographs achieves accuracies ranging from 87.76% for the wrist *LAT* view to 96.17% for the wrist *AP* view. For the classification task, the recall is of particular interest, since a false negative would mean that a *diagnostic* radiograph would be classified as *non-diagnostic*, possibly leading to an unnecessary imaging repetition and radiation exposure. Recall values range from 96.75% for the ankle *AP* view to 100% for the wrist *AP* view. The results for all metrics, body parts, and views are shown in Table 3.9. In Figure 3.7 randomly selected radiographs with their labeled and predicted quality are shown. Figure 3.11 shows the ten worst predictions for each body part and view. It can be seen that even among the 60 worst predictions, only five have an error greater than 1.0. Furthermore, there are only five instances where a *diagnostic* image is incorrectly considered *non-diagnostic*.

Intraclass correlation The ICC between the radiologists' annotation and the model's prediction shows excellent reliability for the ankle and knee and good reliability for the wrist. All 95% confidence interval ranges are above 0.75, supporting the strong reliability of the model's quality predictions. Table 3.10 shows all results for the individual body parts and views.

Table 3.9: The table summarizes model performance for ankle, knee, and wrist radiographs in anterior-posterior (*AP*) and lateral (*LAT*) views. Reported metrics include mean absolute error (MAE), percentage of correctly predicted quality scores (Correct), and classification metrics for binary diagnostic vs. non-diagnostic categorization: accuracy, F_1-score, recall, and precision. All values are shown as mean ± standard deviation across cross-validation runs. All values are percentages except MAE.

Metric	Ankle		Knee		Wrist	
	AP	*LAT*	*AP*	*LAT*	*AP*	*LAT*
MAE	0.22±0.01	0.21±0.01	0.22±0.01	0.20±0.01	0.20±0.01	0.23±0.01
Correct	89.77±2.14	91.50±2.31	92.04±2.44	93.90±2.29	94.35±2.72	91.87±2.81
Accuracy	91.05±1.12	91.97±1.41	91.36±2.61	89.53±2.57	96.17±1.25	87.76±3.37
F_1-score	94.17±0.81	94.19±0.86	94.98±1.51	93.97±1.65	98.04±0.65	93.41±1.94
Recall	96.75±1.26	97.09±1.20	97.25±2.59	97.07±3.50	100.00±0.00	99.17±1.42
Precision	91.78±2.13	91.49±1.91	92.94±3.20	91.21±3.09	96.16±1.25	88.36±3.38

Table 3.10: Agreement between model predictions and expert ratings across body parts and views. Intraclass correlation coefficients (ICC, two-way random-effects model, mean-rating, absolute agreement) are reported for each body part and view (*AP* and *LAT*). The table includes ICC values along with their corresponding 95% confidence intervals, reflecting the consistency between predicted and annotated diagnostic quality scores.

Metric	Ankle		Knee		Wrist	
	AP	*LAT*	*AP*	*LAT*	*AP*	*LAT*
ICC (2k)	0.95	0.95	0.93	0.92	0.87	0.81
95% CI_{min}	0.94	0.95	0.92	0.90	0.84	0.76
95% CI_{max}	0.96	0.96	0.95	0.93	0.90	0.85

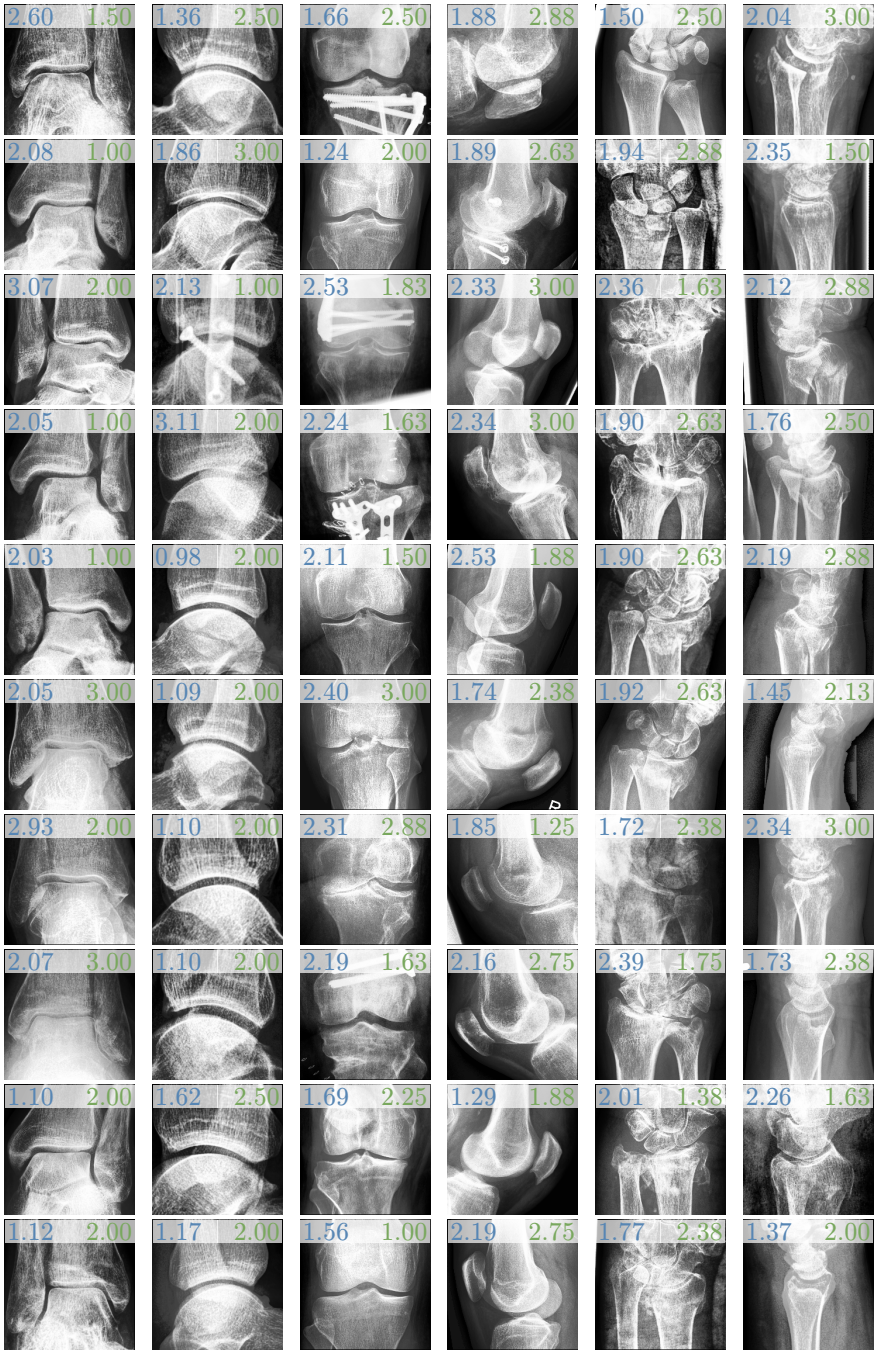


Figure 3.11: This figure shows failure cases for ankle, knee, and wrist ROIs annotated with their labeled quality (green, top right) and their predicted quality (blue, top left). The shown radiographs are the ones with the largest deviation between label and prediction. There are only five radiographs with a deviation larger than 1.0 for the ankle for both *AP* and *LAT*. There are none for the knee and the wrist.

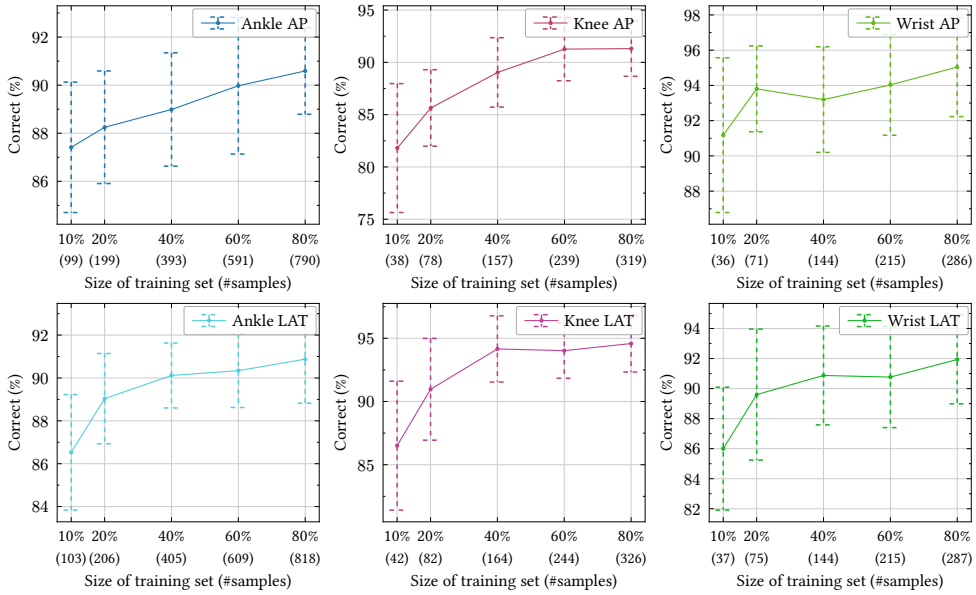


Figure 3.12: The plot shows the percentage of correctly predicted quality scores (*Correct*) for ankle, knee, and wrist radiographs in *AP* and *LAT* views at varying training data proportions. The vertical lines indicate standard deviations.

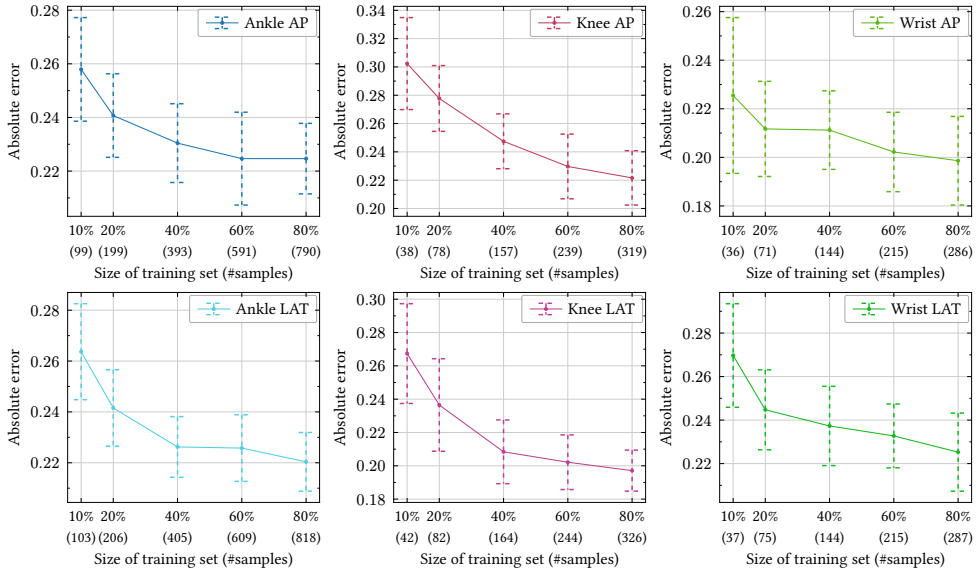


Figure 3.13: The plots show the absolute error for ankle, knee, and wrist radiographs in *AP* and *LAT* views at varying training data proportions. The vertical lines indicate standard deviation.

Impact of Dataset Size To evaluate the scalability of the approach, model performance was assessed using subsets of the training data (10%, 20%, 40%, 60%, 80%), while keeping the test set fixed. The results, visually presented in Figure 3.12 for the *Correct* metric and in Figure 3.13 for the MAE, show that, on average, the results improve with more training data. For the *Correct* metric, there are a few cases in which the average accuracy is slightly worse with more data than with less data. Given the high standard deviation, this appears to be a random fluctuation. This is underlined by the fact that the MAE is constantly improving with more data.

However, even with only 10% of the data, the model achieved robust performance. Using 40% of the data for training, the model predicts on average over 90% of the radiographs correctly. Considering the time required for labeling a radiograph, this results in a workload of about two hours for a single radiologist to annotate data for both views of a single body part with an expected correctness of above 90% for the predictions. The results for all metrics and sizes can be found in Table 3.11.

3.2.5 Discussion

This work presents a scalable deep learning framework for assessing the diagnostic quality of radiographs based on anatomical positioning. We demonstrate that automated quality assessment based on radiographic alignment is not only feasible but also generalizable. The models maintained high accuracy across body parts and views, with minimal annotation effort. Furthermore, we analyzed the impact of dataset size and showed that reliable predictions can be achieved with modest training volumes, enabling practical implementation across institutions and applications.

This study provides the first comprehensive evidence that positioning-based quality assessment is not only feasible but generalizable across a broader set of body parts and standard views while at the same time being scalable with modest annotation effort. Additionally, through the publication of the trained models, we enable a wider community to accelerate research on the diagnostic quality of radiographs.

Clinical application A key clinical application of this approach lies in providing immediate quality feedback directly after image acquisition. In current workflows, radiographs are often reviewed by radiologists with a delay, which can lead to postponed repeat imaging, misdiagnosis, or treatment delays, particularly critical in emergency settings where rapid, accurate diagnosis is essential [Ritchie et al., 2025]. By integrating real-time AI-based quality assessment into acquisition systems, suboptimal positioning could be identified and corrected immediately while the patient is still present. Additionally, such feedback systems might offer significant educational value for radiographers, who can develop a more intuitive understanding of positioning criteria through consistent, objective evaluation. Looking forward, our work may also serve as a foundational step toward future solutions that utilize external camera systems or real-time motion

Table 3.11: The table shows the results for the quality assessment for all body parts and views dependent on different sizes of the training set. The size is given as a portion of the total available labeled data. Overall, metrics improve when the size is increased, while the improvements decrease with larger sizes.

Size	Ankle		Knee		Wrist	
	AP	*LAT*	*AP*	*LAT*	*AP*	*LAT*
MAE						
0.1	0.26±0.02	0.26±0.02	0.30±0.03	0.27±0.03	0.23±0.03	0.27±0.02
0.2	0.24±0.02	0.24±0.02	0.28±0.02	0.24±0.03	0.21±0.02	0.24±0.02
0.4	0.23±0.01	0.23±0.01	0.25±0.02	0.21±0.02	0.21±0.02	0.24±0.02
0.6	0.22±0.02	0.23±0.01	0.23±0.02	0.20±0.02	0.20±0.02	0.23±0.01
0.8	0.22±0.01	0.22±0.01	0.22±0.02	0.20±0.01	0.20±0.02	0.23±0.02
Correct in %						
0.1	0.87±0.03	0.87±0.03	0.82±0.06	0.87±0.05	0.91±0.04	0.86±0.04
0.2	0.88±0.02	0.89±0.02	0.86±0.04	0.91±0.04	0.94±0.02	0.90±0.04
0.4	0.89±0.02	0.90±0.02	0.89±0.03	0.94±0.03	0.93±0.03	0.91±0.03
0.6	0.90±0.03	0.90±0.02	0.91±0.03	0.94±0.02	0.94±0.03	0.91±0.03
0.8	0.91±0.02	0.91±0.02	0.91±0.03	0.95±0.02	0.95±0.03	0.92±0.03
Diagnostic Accuracy in %						
0.1	0.90±0.02	0.87±0.02	0.88±0.04	0.87±0.03	0.96±0.02	0.86±0.03
0.2	0.90±0.02	0.89±0.02	0.90±0.03	0.88±0.03	0.97±0.02	0.87±0.04
0.4	0.91±0.02	0.91±0.02	0.91±0.03	0.90±0.03	0.97±0.02	0.87±0.03
0.6	0.91±0.02	0.91±0.02	0.92±0.03	0.89±0.03	0.97±0.01	0.88±0.03
0.8	0.92±0.02	0.92±0.02	0.92±0.03	0.90±0.03	0.97±0.02	0.88±0.03
Diagnostic F_1-Score in %						
0.1	0.94±0.01	0.91±0.02	0.93±0.03	0.93±0.02	0.98±0.01	0.93±0.02
0.2	0.94±0.01	0.92±0.02	0.94±0.02	0.93±0.02	0.98±0.01	0.93±0.02
0.4	0.94±0.01	0.93±0.02	0.95±0.02	0.94±0.02	0.98±0.01	0.93±0.02
0.6	0.94±0.01	0.93±0.01	0.95±0.02	0.94±0.02	0.98±0.01	0.93±0.02
0.8	0.95±0.01	0.94±0.01	0.95±0.02	0.94±0.02	0.98±0.01	0.93±0.02
Diagnostic Recall in %						
0.1	0.97±0.01	0.95±0.02	0.98±0.02	0.97±0.03	1.00±0.01	0.98±0.02
0.2	0.97±0.02	0.96±0.02	0.98±0.02	0.96±0.03	1.00±0.00	0.99±0.02
0.4	0.97±0.01	0.96±0.01	0.98±0.02	0.96±0.03	1.00±0.00	0.99±0.01
0.6	0.97±0.01	0.96±0.02	0.97±0.03	0.96±0.03	1.00±0.00	0.99±0.01
0.8	0.97±0.01	0.96±0.02	0.98±0.02	0.96±0.02	1.00±0.00	0.99±0.02
Diagnostic Precision in %						
0.1	0.90±0.03	0.87±0.03	0.90±0.05	0.89±0.04	0.96±0.01	0.88±0.04
0.2	0.91±0.03	0.89±0.03	0.91±0.04	0.90±0.03	0.97±0.01	0.88±0.04
0.4	0.92±0.03	0.90±0.03	0.92±0.03	0.92±0.03	0.97±0.02	0.88±0.03
0.6	0.92±0.02	0.91±0.03	0.93±0.04	0.92±0.03	0.97±0.01	0.88±0.03
0.8	0.92±0.02	0.92±0.02	0.93±0.03	0.92±0.03	0.97±0.02	0.89±0.03

tracking to assess or even predict radiographic quality before acquisition. This would further streamline the imaging process and enhances consistency in diagnostic imaging.

Beyond real-time feedback, another promising application is dataset curation. AI-driven quality filters can automatically exclude low-quality radiographs from training sets for diagnostic or segmentation models, thereby improving the robustness and clinical relevance of downstream AI tools. Moreover, the low annotation burden required to extend our model to additional body parts supports its scalability for large-scale retrospective analyses or institution-wide deployment. This opens the door to standardized, objective quality control across imaging departments.

Limitations While showing promising results, there are still several limitations. First, the data was sourced from a single institution; however, radiographs were acquired using various vendor systems and from different imaging sites, which enhances generalizability. Second, although this study evaluated three specific anatomical regions, radiographic imaging is utilized across the entire body. Future research could incorporate additional body parts to further assess the model's generalizability. Third, ground truth labels were based on expert consensus, which, despite high inter-rater reliability, may reflect institution-specific practices or subjective criteria.

Conclusion This study demonstrates that deep learning can be used to assess the diagnostic quality of musculoskeletal radiographs based on anatomical alignment with high accuracy and efficiency. The framework generalizes across multiple joints and views, requires limited annotation effort, and supports both real-time clinical feedback and automated dataset curation. These findings contribute to ongoing efforts to improve radiographic quality, reduce repeat imaging, and enable standardized image assessment in routine care.

3.3 Quality Impact Factors and Instant Feedback

Being able to automatically assess the quality of a radiograph opens the possibility for various use cases. Two of these use cases have been explored and evaluated in greater detail within the scope of this work.

First, in Section 3.3.1, a retrospective analysis of the diagnostic quality of radiographs over a longer period of time is carried out. This makes it possible to assess the overall quality of the radiographs taken as part of quality assurance and to ensure a minimum level of quality. Influencing factors can be identified and used to optimize quality. Secondly, in Section 3.3.2, a clinical instant feedback system is implemented and evaluated. Such a system can be used to provide feedback to the radiographers immediately after the radiograph is taken, allowing them to react based on the quality. It may further help to train inexperienced radiographers or provide motivation through positive feedback.

3.3.1 Impact Factors on Diagnostic Quality

For a medical clinic, it is important to ensure a high diagnostic quality of radiographs. On the one hand, good diagnostic quality is necessary for a correct diagnosis and is therefore crucial for the treatment and health of patients. On the other hand, poor diagnostic quality can lead to higher costs and additional work. Patients have to undergo repeated radiation examinations, requiring new appointments and the reallocation of equipment, rooms, and staff. Additionally, poor treatment can also damage reputation. Ultimately, it is not only in the hospital's own interest to ensure the quality of radiographs. In Germany, regulatory authorities also verify quality through randomized audits [Deutschland, 1988; Gemeinsamer Bundesausschuss, 2023].

To ensure high quality, continuous quality assessment must be carried out to identify poor-quality radiographs, determine the reasons for this, and take corrective action. While it is possible to do this manually, an automated quality assessment relieves this burden and opens the possibility to review every single radiograph. Such an automated assessment can further be used to identify potential impact factors by correlating the radiographs' metadata with their quality. In this evaluation, we analyzed the quality of 72,299 radiographs spanning a decade to evaluate various impact factors.

Dataset

The dataset exported from the local university hospital contains 72,299 radiographs of the ankle, knee, and wrist of the *AP* and *LAT* views. The first exported radiograph was taken on January 1, 2013, while the last was from December 31, 2022, spanning a range of ten years. The radiographs were acquired from different stations using four different imaging systems (27,744 DigitalDiagnost C90, 11,782 DigitalDiagnost, 26,846 Fluorospot Compact FD, and 5,886 PCR Eleva) produced by the two manufacturers,

Table 3.12: Number of radiographs for each body part and view contained in the large unlabeled dataset exported from the local university hospital.

View	Ankle	Knee	Wrist
AP	11,847	13,010	9,260
LAT	11,719	14,215	12,248

Philips and SIEMENS. Besides filtering the imaged body part and view and excluding patients under the age of 16, no further filtering for joint deformities or presence of implants was done. The number of radiographs for each body part and view is given in Table 3.12. The dataset is about evenly distributed regarding laterality (36,297 left; 36,002 right) and patient's sex (32,873 male; 39,378 female). 44,555 radiographs are computed radiographs, while 27,744 are digital radiographs.

The following DICOM tags were extracted from the radiographs and used for further analyses: *Study Date*, *Study Time*, *Modality*, *Performed Station AE Title*, *Manufacturer*, *Manufacturer's Model Name*, *Patient's Sex*, *Patient's Age*, and *Patient's Weight*. As not all of these tags are available for every radiograph, the number of radiographs used for each evaluation varies. This applies in particular to the patient's weight.

Results

Patient's Characteristics A factor impacting the diagnostic quality might be the patients themselves. Age, sex, and weight can be extracted from the DICOM metadata for patient characteristics. Regarding age, a higher age correlates with a worse diagnostic quality, as can be seen in Figure 3.14. There is a clear monotonic decline in quality from patients in their 20s to patients in their 90s. A similar pattern is shown in Figure 3.15 for the patients' weight. The quality of radiographs from patients with lower weight is slightly better than those from patients with higher weight. The only exception to this is patients weighing less than 50 kg. Regarding sex, there is only a slight difference in the quality of radiographs for male and female patients, with the quality for male patients being better, as shown in Figure 3.16.

Time and Location In addition to the patient's own characteristics, the time of day or the specific station within the hospital can also influence quality. The quality as a function of the hour of the day is shown in Figure 3.17. It can be seen that the quality between 00:00 and 08:00 is better than the median, worse than the median between 08:00 and 16:00, and then better again between 16:00 and 00:00. Another influencing factor is the station where the radiograph is acquired. In larger hospitals, different stations treat different cases; some, for example, only treat patients from the emergency room, while others only perform scheduled examinations of hospitalized patients. That the resulting quality varies between the stations can be seen in Figure 3.18.

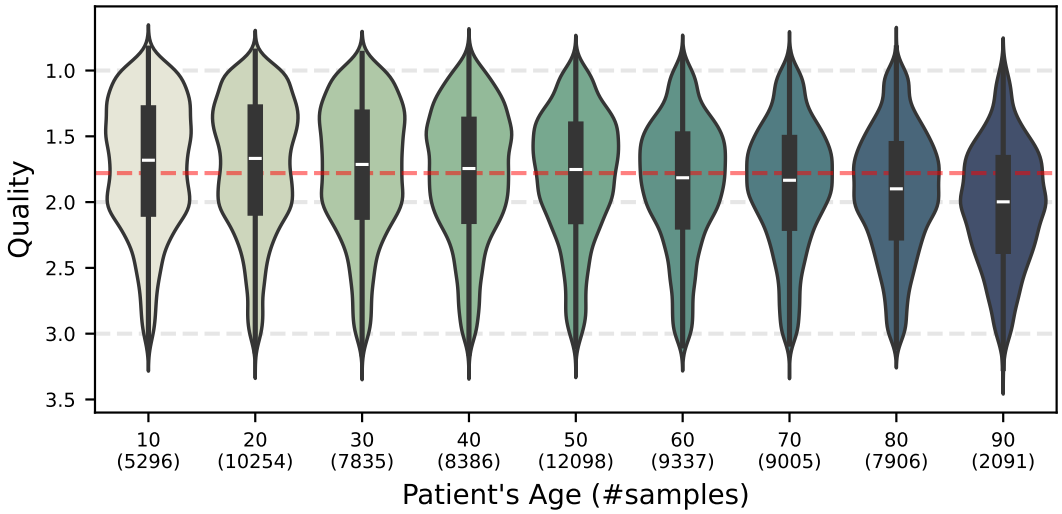


Figure 3.14: The predicted quality of radiographs, stratified by 10-year age groups. For each group, a box plot and the estimated distribution are shown. The red line marks the median quality over all radiographs. It can be seen that the quality declines with the patient's age.

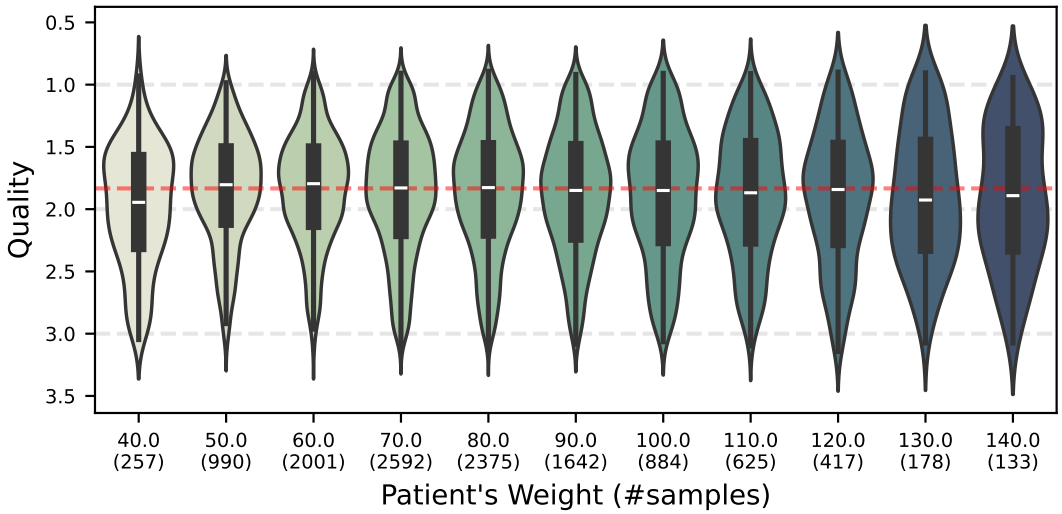


Figure 3.15: The predicted quality of radiographs, stratified by 10-kg weight groups. Patients' weights were rounded down to the nearest multiple of ten. For each group, a box plot and the estimated distribution are shown. The red line marks the median quality over all radiographs. Except for patients weighing between 40 and 50 kg, the quality decreases with the patient's weight. Radiographs of patients weighing between 40 and 50 kg are of poorer diagnostic quality.

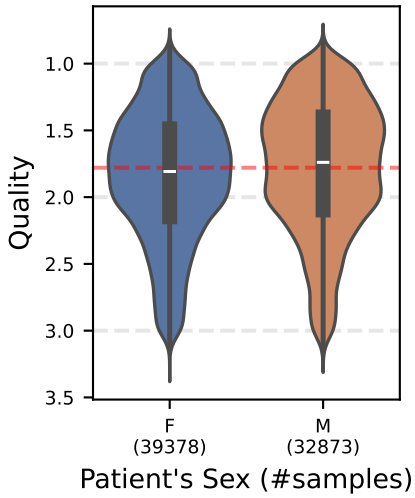


Figure 3.16: The predicted quality of radiographs, stratified by sex. For both groups, a box plot and the estimated distribution are shown. The red line marks the median quality over all radiographs. For male patients the median quality is slightly better than for female patients.

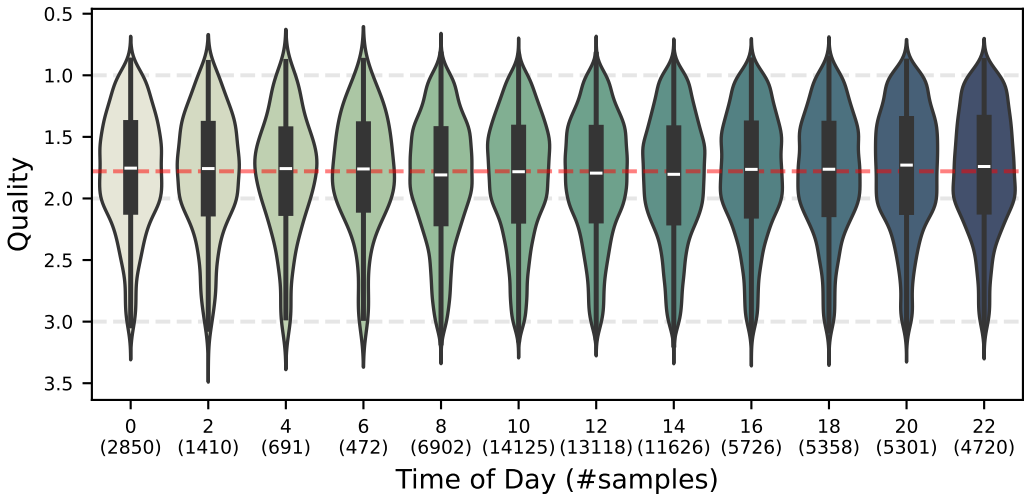


Figure 3.17: The predicted quality of radiographs, stratified by time of day in two-hour intervals. For each group, a box plot and the estimated distribution are shown. The red line marks the median quality over all radiographs. Radiographs acquired between 08:00 and 16:00 have a slightly worse quality.

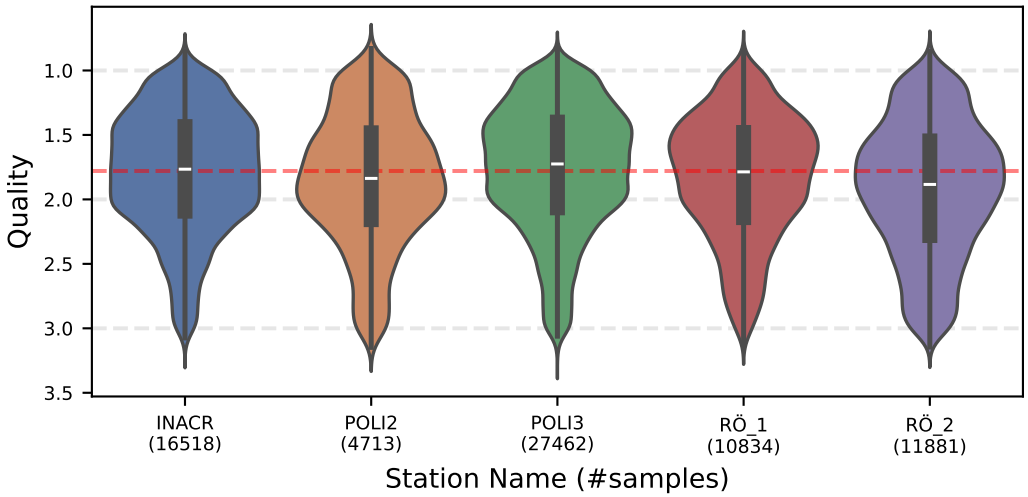


Figure 3.18: The predicted quality of radiographs, stratified by hospital station. For each group, a box plot and the estimated distribution are shown. The red line marks the median quality over all radiographs. It can be seen that the quality differs between station. Differences may reflect how patients are distributed across hospital stations (e.g., emergencies or inpatients).

Discussion

Although the data show that there are several measurable factors that correlate with the quality of radiographs, these do not necessarily have a causal relationship.

The two most relevant factors presented are the age and weight of the patient. It seems likely that positioning younger patients is easier, as they are more mobile and have fewer pre-existing medical conditions. Since radiographers have to physically touch patients and feel for bone structures and joints when positioning them, it seems equally obvious that positioning heavier patients is more difficult. The difference regarding sex might result from different pain tolerance levels. Maintaining a certain position during an injury can cause pain, and since males and females have different pain tolerances [Fillingim et al., 2002], it may be easier for males to maintain the position. Another explanation is that the generally smaller size of female bones and joints may increase the difficulty of precise positioning compared to male anatomy. Although these factors influence the quality of the radiograph, they cannot be influenced by the radiographers. However, one possible intervention would be to train staff to take more time for positioning patients with risk factors.

The fact that quality depends on the time of day could be because the number of patients is significantly higher during the day. There is more time pressure on the radiographers, which can lead to poorer positioning. Outside core working hours, radiographs are mainly taken for emergencies, so there is more time available for

individual patients. Another explanation is that scheduled radiographs for inpatients are only taken during working hours. Since patients who are treated as inpatients may be more seriously injured, it is more difficult to position them. In this case, the influence of time would merely be a confounding factor and would have no causal connection. Ultimately, the quality also depends on the station where the radiograph was taken. While there are noticeable differences, the cause can only be clarified with internal hospital knowledge. Patients might be assigned to different stations according to specific criteria, which leads to the observable differences. The names POLI2 and POLI3 indicate stations for outpatient treatment, while INACR is likely the emergency room.

All in all, the evaluation shows that there are various factors that could influence diagnostic quality. Whether these factors are causally related cannot be readily assessed. However, automatic quality assessment makes it possible to identify these factors, find potential problems, and decide whether quality assurance measures can be used to improve quality. For further and more detailed investigation, it would be possible to evaluate the quality depending on the radiographer who performed the examination. Such an evaluation would allow staff to be trained specifically.

3.3.2 Instant Diagnostic Quality Feedback

In addition to retrospective diagnostic quality analysis, another application of automated quality assessment is an instant feedback system. In current clinical workflows, there is usually a significant time gap between the acquisition of the radiograph and its use for diagnosis. Although radiographers can assess whether a radiograph has good quality, this assessment does not necessarily correspond to that of radiologists. The datasets from the previous sections show that radiographs with poor diagnostic quality are regularly accepted by radiographers. It could therefore be helpful for radiographers to receive instant feedback on how a radiologist would assess the quality. The immediate feedback would allow radiographers to reevaluate their own assessment if it deviates from the prediction. Over time, radiographers could learn to better assess the quality required by radiologists. Ultimately, immediate feedback could also improve patient positioning. Inexperienced radiographers in particular would be able to check their positioning objectively.

The following describes a live feedback system that has been implemented and is used at the local university hospital. The system provides feedback on diagnostic quality after the examination. After one and a half years of operation, the system's impact on quality was evaluated. Furthermore, the radiographers were surveyed to measure acceptance and determine any possible changes in working methods regarding patient positioning.

Figure 3.19: This figure shows the setup in the X-ray room. Next to the screens for controlling the X-ray device and the list of patients for upcoming examinations, an additional screen for the automatic quality assessment, marked in red, is added. On this screen, radiographs from ankle or knee examinations, together with their quality prediction, are shown.

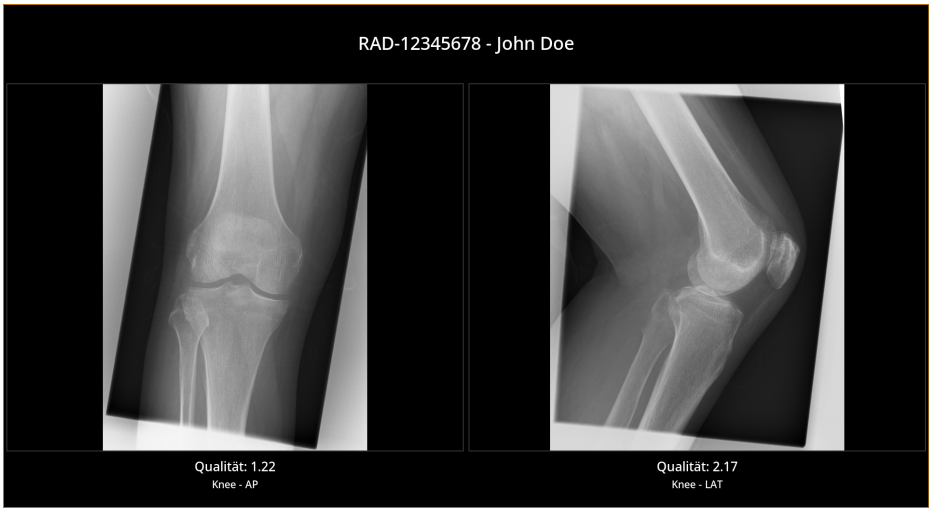


Figure 3.20: Screenshot of the quality prediction tool's GUI as seen by the radiographers. The tool shows an identification number together with the patient's name in the header. Below, all ankle and knee radiographs of the current examination appear. For each radiograph, the predicted quality is shown as a numerical value as well as the predicted body part and view.

Table 3.13: Number of radiographs for each body part and view contained in the three-year dataset for evaluation of the feedback system.

View	Ankle	Knee	Wrist
AP	2,562	3,303	1,735
LAT	2,538	3,368	1,755

Setup

To add the automatic quality assessment into the radiographers' workflow, a new computer with a screen was added to the station. A DICOM server ran on the computer, accepting incoming radiographs. Upon the arrival of new radiographs, any radiographs of body parts other than the ankle, knee, or wrist were filtered out. The radiographs were then fed into the quality assessment framework, predicting the body part, view, and quality. After prediction, the GUI shows all ankle and knee radiographs from the current examination together with their predicted quality. No feedback is displayed for the radiographs of the wrist in order to have an unbiased comparison group. Starting from the arrival of new radiographs, this process takes approximately two seconds. The computer controlling the X-ray device was configured to send the final examination radiographs to the quality assessment DICOM node, in addition to the local PACS node. It is important to note that the examination's radiographs were only sent after all required examinations were done, which corresponds to Step 4 in the workflow visualization shown in Figure 2.1. A more effective integration would be to display the radiographs immediately after acquisition (Step 3 in Figure 2.1), but this was not technically feasible.

The physical setup is shown in Figure 3.19 and a screenshot of the GUI is shown in Figure 3.20. In order to assess whether displaying the quality has an impact on behavior, only the quality prediction for the ankle and knee joints was integrated, allowing the wrists to serve as a control group.

Changes in Diagnostic Quality

Dataset For evaluation, all ankle, knee, and wrist radiographs acquired in this particular station over a span of three years were exported. This includes the 18 months before the introduction of quality assessment and the 18 months following its introduction. In total, 15,261 radiographs were exported. Table 3.13 shows how the radiographs are distributed across the different body parts and views. The radiographs were acquired between January 1, 2023, and December 1, 2025.

Results As shown in Figure 3.21 and Figure 3.22, the quality varies between the periods with and without quality assessment. In the six-month periods without quality

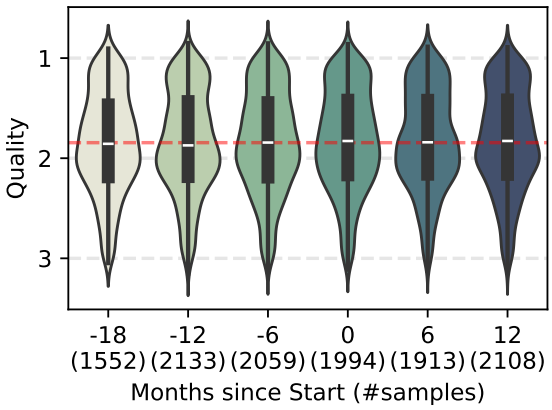


Figure 3.21: Predicted quality stratified in six-month intervals from the start of the assessment (Month 0). For each group, a box plot and the estimated distribution are shown. The red line marks the median quality over all radiographs. It is visible that the median quality for all six-month groups after the start of the quality assessment is better than the overall median quality.

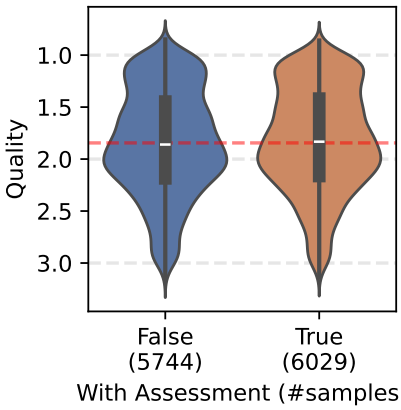


Figure 3.22: Predicted quality stratified by the presence or absence of quality assessment. For both groups, a box plot and the estimated distribution are shown. The red line marks the median quality over all radiographs. Visibly, the median quality is better in the group with quality assessment.

assessment, the median quality is mostly worse than the overall median. All six-month periods with quality assessment have a median quality better than the overall median. The median quality of the group with quality assessment was 0.027 better. The Mann-Whitney U test [Mann et al., 1947] was used to compare the groups with and without quality assessment. The null hypothesis that both distributions are equal was rejected with a p-value of 0.0276 at a significance level of $\alpha = 5\%$. It can therefore be assumed that there is a difference in distribution between the groups. However, the same analysis of the wrist data shows that these have also improved over time. Here, the null hypothesis is rejected with a p-value of $5.02 \cdot 10^{-6}$, and the average improvement is 0.044. This indicates that there is likely another influencing factor and that the feedback display is not the primary cause for the change in quality.

Radiographers' Feedback

Questionnaire Beyond evaluating the system's objective impact, understanding its perception and acceptance among practicing radiographers is essential. To assess this, the radiographers were asked to complete a questionnaire. The questionnaire consists of the following twelve statements as Likert-type items, each rated on a 5-point scale.

The questionnaire statements are provided below in English, followed by the original German text in parentheses.

1. Automatic quality assessment provides helpful feedback for inexperienced radiographers. (Die automatische Bewertung der Qualität ist hilfreiches Feedback für unerfahrene MTRAs.)

2. Automatic quality assessment disrupts the workflow. (Die automatische Bewertung der Qualität ist störend für den Arbeitsablauf.)

3. The quality shown matches my own perception. (Die angezeigte Qualität stimmt mit meiner eigenen Wahrnehmung überein.)

4. Automatic assessment makes me pay closer attention to positioning. (Die automatische Bewertung lässt mich genauer auf die Positionierung achten.)

5. When an ankle or knee joint is X-rayed, I look at the evaluation afterward. (Wird ein Sprung- oder Kniegelenk geröntgt, schaue ich mir danach die Bewertung an.)

6. I am pleased when the radiograph is rated as being of good quality. (Ich freue mich, wenn das Röntgenbild mit einer guten Qualität bewertet wurde.)

7. If the radiograph is assessed as poor quality, I consider whether the positioning was correct. (Wird das Röntgenbild mit einer schlechten Qualität beurteilt, überlege ich, ob die Positionierung korrekt war.)

8. I would like to see assessments displayed for more body parts. (Ich würde mir wünschen, dass für mehr Körperteile eine Bewertung angezeigt wird.)

9. The assessed quality should be displayed immediately after recording, rather than after all recordings have been submitted. (Die bewertete Qualität sollte direkt nach der Aufnahme und nicht erst nach dem Abschicken aller Aufnahmen angezeigt werden.)

10. Predicting quality while positioning the patient would make positioning easier. (Eine Vorhersage der Qualität schon während des Ausrichtens des Patienten würde die Positionierung erleichtern.)

11. I feel controlled in my work by the automatic assessment. (Ich fühle mich durch die automatische Bewertung in meiner Arbeit kontrolliert.)

12. Through the assessment, I learn how to position patients better. (Durch die Bewertung lerne ich, Patienten besser zu positionieren.)

Each printed questionnaire contained the statements in random order.

Each item could be rated with the choices:

1. I completely disagree (Stimme überhaupt nicht zu)
2. I do not agree (Stimme nicht zu)
3. Neutral (Neutral)
4. I agree (Stimme zu)
5. I completely agree (Stimme voll und ganz zu)

Results A total of eight questionnaires were completed. For each statement, the mean and standard deviation were calculated and are shown in Table 3.14. An important point that can be concluded from the results is that the radiographers do not reject the use of the system or perceive it as negative. It is neither perceived as disruptive (statement 2) nor controlling (statement 11). Even though the system is not perceived negatively, the radiographers do not consider it to be a helpful support for themselves. In their opinion, they do not improve themselves (statement 12), do not pay closer attention to positioning (statement 4), and do not reevaluate their positioning in the event of a poorly rated radiograph (statement 7). The reason for this could be that the displayed quality usually does not correspond to the perception of the radiographers (statement 3). However, radiographers are undecided whether the system provides helpful feedback for inexperienced radiographers (statement 1), and even if they do not agree with the predicted quality, they still actively use the system (statement 5) and are pleased when the assessed quality is good (statement 6). Finally, the radiographers would welcome an expansion and improvement of the system. The system should be expanded to cover more body parts (statement 8), and the assessment should be done right after the image is taken, not just at the end of the treatment (statement 9). The radiographers also considered the idea of displaying the expected quality of the radiographs during positioning to be a valuable feature (statement 10).

Discussion

Both the quantitative evaluation of the change in quality before and after the launch of the system and the accompanying questionnaire provided valuable insights. The most important result of the evaluation is that the improvement in quality might not be thanks to instant feedback.

This is consistent with the statements made by the radiographers that the displayed quality does not make them pay more attention to their positioning and that they do not reconsider it even if the rating is poor. Since the radiographers find that the quality displayed does not match their own perception, they may simply ignore the feedback. Somewhat contradictory is their statement that they will be pleased if the assessment

Table 3.14: The mean and standard deviation for each of the twelve Likert-type items across eight respondents. A negative value means that most radiographers do not agree with the statement. A positive value shows agreement. The statement that the radiographers are pleased when a radiograph is rated as of good quality (number 6) received the highest agreement. The statement that the radiographers feel controlled through the system (number 11) received the lowest agreement.

Number	Mean	Std
1	0.00	0.93
2	−1.25	0.71
3	−0.71	0.76
4	−0.38	0.92
5	0.75	0.89
6	1.71	0.49
7	−0.43	0.98
8	0.57	1.27
9	1.29	0.49
10	0.88	0.83
11	−0.88	1.25
12	−0.86	0.90

results in a good quality, which suggests that they want good feedback. Since the assessed quality reflects relatively well the assessment by radiologists, this shows that the perception of diagnostic quality by radiographers differs from that of radiologists. This is consistent with previous studies showing that assessments of the diagnostic quality of radiographs by radiologists and radiology assistants do not agree [Decoster et al., 2023; Decoster et al., 2025]. Automated quality assessment may therefore not only help improve quality but also help radiographers better understand the needs of radiologists.

Finally, the evaluation of the questionnaire shows that expanding and improving the system would increase acceptance. Currently, the system can only assess the radiographs after they have been sent. This is a major limitation, as feedback on an individual radiograph is only provided at the end of the examination. This delay can make it difficult to learn from the feedback and may be one reason why radiographers do not respond to poor quality ratings. Assessing the quality immediately after the radiograph is taken would therefore be a significant improvement, which is only prevented by the proprietary X-ray system, which does not allow access to radiographs that have not been sent. Another important improvement would be to predict diagnostic quality based on patient positioning before a radiograph is acquired. This could further assist radiographers through direct feedback and potentially prevent non-diagnostic radiographs from being taken in the first place.

Chapter 4

Assessment of Patient Positions

The previous chapter showed that the correct pose of the patient during radiography is of critical importance to ensure an adequate diagnostic quality of radiographs. However, correct patient positioning is not a standardized process, often resulting in inadequate radiographs and repeated radiation exposure. Nevertheless, a radiograph of high diagnostic quality is mandatory for a reliable diagnosis and correct treatment planning.

The framework presented earlier is capable of assessing the quality of radiographs with high accuracy and can also be used to indirectly improve patient positioning. While this can prevent the radiographer from noticing the poor quality or the patient from undergoing unnecessary multiple X-ray examinations, it cannot proactively prevent poor positioning. In the worst case, a poor-quality radiograph has already been acquired, and the patient must be X-rayed again.

To avoid such cases, it would be necessary to be able to estimate the quality of the radiograph before it was even taken. To protect the patient and support the radiographers further, an automated assessment of the patient's pose can help improve the diagnostic quality of radiographs.

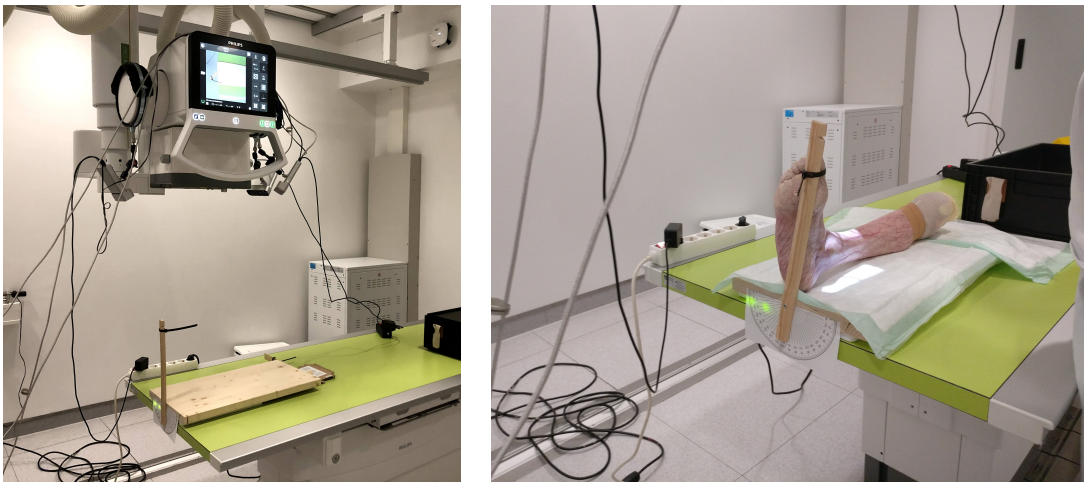
In this chapter, a method to predict the diagnostic quality of radiographs based on depth images is presented in Section 4.1. It shows for the upper ankle that depth images of the patient's pose contain relevant features to assess whether the resulting radiograph will have a high diagnostic quality or not. As data acquisition for this research is particularly hard, subsequently Section 4.2 introduces a way to generate corresponding synthetic radiographs and depth images from CT scans. These synthetic images can then be annotated and used for pretraining to lower the needed amount of real training data.

Preliminary work

Since there are existing guidelines on how patients should be positioned by radiographers, the first step was to check whether these could be transferred to an automated system. In the simplest case, an automated system could ensure that the existing guidelines are adhered to as closely as possible. The guideline for radiographers in German hospitals for *AP* imaging of the ankle joint is that the patient should flex their foot so that the

sole and the table are at a 90-degree angle to each other. The foot is also rotated inward by approximately 15 degrees so that the malleoli are parallel [Becht et al., 2019]. Given such clear rules for positioning, it seems logical to derive the parameters of the patient's pose from depth images and then use them to determine quality. In Krönke et al. [2022] an approach to derive the parameters of the patient's pose from a radiograph is presented; however, no method is suggested for transferring the parameters into a diagnostic quality.

However, an initial experiment revealed that a simple rotation of approximately 15 degrees is not always optimal for positioning. For this experiment, a device was designed in which an anatomical preparation of a foot with a lower leg was fixed in place. The device allows the foot to be rotated precisely around the axis of the leg without changing the positioning in any other way. Figure 4.1 shows the device on the X-ray table beneath the X-ray machine and the device with the foot fixed in place. To determine how rotation and the diagnostic quality of the resulting images are related, the foot was rotated in one-degree increments, and a radiograph was taken for each rotation. This was done for three different flexions of the foot. The flexion describes how strongly the foot is pulled toward the shinbone. The resulting radiographs were then evaluated by radiologists according to their quality. The results, presented in Figure 4.2, show that the range of rotation in which optimal quality is achieved not only differs depending on flexion but is largely disjoint. These results raise the question of whether it is possible to determine quality directly from the parameters of the pose. For this reason, an end-to-end approach is used in the following, which directly predicts diagnostic quality from depth images.



(a) (b)

Figure 4.1: In panel (a) the device on the X-ray table below the X-ray machine is shown. An anatomical preparation can be fixed to the wooden plate, which can then be rotated by turning the rod at the end of the plate. In panel (b) an anatomical preparation is fixed on the plate. Markings below the rod allow the rotation to be read accurately in degrees. The detector is placed beneath the plate.

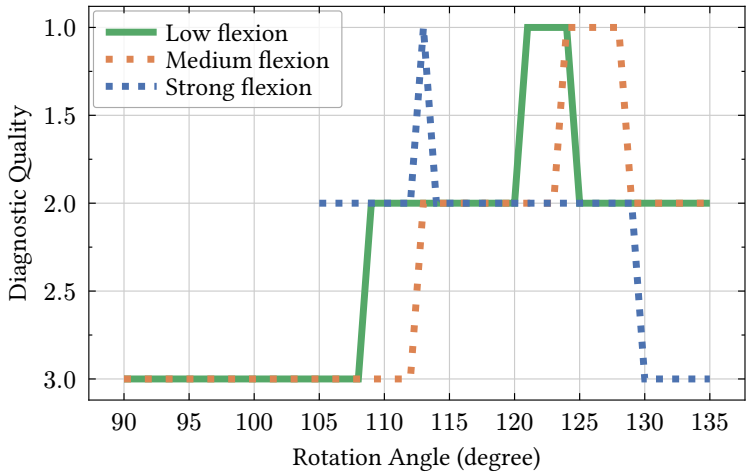


Figure 4.2: The diagnostic quality of the radiographs as a function of rotation of the foot preparation. It can be seen that the rotation ranges resulting in an optimal quality are mostly disjunct, depending on the flexion of the foot. With strong flexion, the range for optimal quality is significantly smaller than with the other two flexions.

4.1 Patient Pose Assessment in Radiography Using Time-of-Flight Cameras

In this work, we propose a novel approach using Time-of-Flight (ToF) cameras to assess the patient's pose and thereby predict the expected diagnostic quality of the radiograph before it is even captured.

In order to make such a prediction, it is necessary to have training data from radiographs and associated depth images. The radiographs are mandatory for obtaining a ground truth label, as only they contain the information about diagnostic quality. Since no such dataset exists, a new dataset was created linking the two together. However, due to regulatory and legal challenges, the acquisition of depth images of the patient's pose and their corresponding radiographs is not readily possible in clinical practice. In particular, taking radiographs of subjects without an indication is problematic, and intentionally taking radiographs of subjects in non-diagnostic poses, which are necessary for training, is ethically difficult to justify.

Therefore, as a first step towards this goal, in a study with two anatomical preparations of the lower leg, we acquired 10,440 depth images of 87 different poses of the upper ankle joint in the *anterior-posterior* (*AP*) view, using two ToF cameras. The radiographs were labeled by radiologists for their diagnostic quality related to the patient's pose. These labels serve as quality labels for the corresponding depth image. Using this dataset, we trained deep neural networks that correctly assessed the diagnostic quality of a pose with a mean accuracy of 90.2%. This demonstrates that shared features for pose assessment across patients exist and can be learned.

Using this approach, radiographers would get immediate feedback for quality control during positioning. This would reduce the number of radiographs with inadequate diagnostic quality and ultimately reduce radiation dose, misdiagnosis, and the risk of mistreatment, as well as time and cost.

This section has been published as:

[Laufer et al., 2024b] Laufer, M.*, Mairhöfer, D.*, Sieren, M., Gerdes, H., Reis, F. L., Bischof, A., Käster, T., Barth, E., Barkhausen, J., and Martinetz, T. "Patient Pose Assessment in Radiography Using Time-of-Flight Cameras". In: *Medical Imaging 2024: Image Processing*. Vol. 12926. *Authors contributed equally. SPIE, 2024, pp. 385–393.

According to the Contributor Roles Taxonomy (CRediT), the contributions of the author of this thesis to the publication are: Conceptualization (together with M.L.), Data curation (together with M.L.), Investigation (together with M.L.), Methodology (together with M.L.), Software, Validation, Visualization (together with M.L.), Writing – original draft (together with M.L.), Writing – review & editing (together with M.L., E.B., T.M.)

4.1.1 Related Work

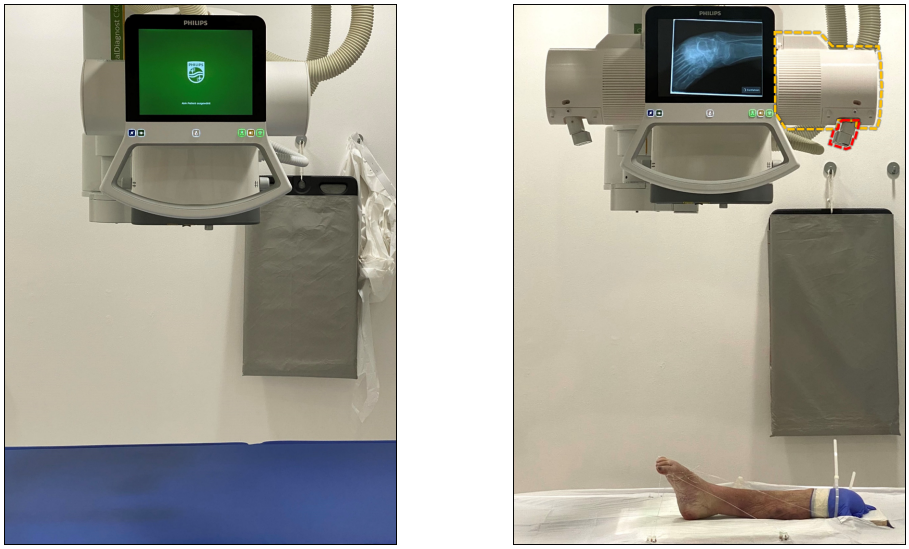
In Little et al. [2017], it is shown that without quality control, radiologists have to reject 25% of the radiographs due to their insufficient diagnostic quality. After the introduction of special interventions, such as training of the radiographers, this rate could be reduced to 13% [Little et al., 2017], which is still above the target level. Furthermore, Little et al. [2017] and Atkinson et al. [2020] identified incorrect positioning as by far the most frequent reason for rejection, ranging from 49% to 67.5%. While there is already research on automated quality control of radiographs [Köpnick et al., 2023; Krönke et al., 2022; Mairhöfer et al., 2021; Meng et al., 2022], these approaches can detect but not prevent radiographs of insufficient diagnostic quality.

Many applications in the clinical context show that the use of ToF cameras can be beneficial for both patients and hospitals. Among other applications, they are used to determine poses [Hansen et al., 2019; Kadkhodamohammadi et al., 2017; Srivastav et al., 2021], monitor patient breathing [Takamoto et al., 2020], and improve the planning and monitoring of radiotherapy treatments [Placht et al., 2012].

However, to our knowledge, there is no system, either with or without the use of ToF cameras, that is capable of automatically assessing a patient's pose with respect to the expected quality of the radiograph.

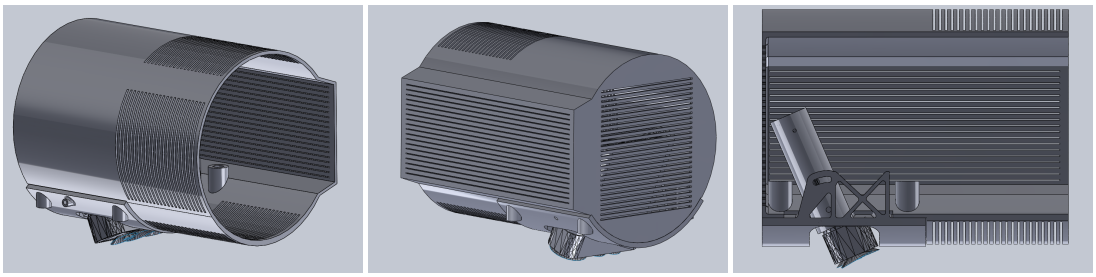
4.1.2 Dataset

To evaluate if it is possible to predict the quality of radiographs based on depth images of the patient's pose, depth images and radiographs must be captured simultaneously. The radiographs will then be labeled regarding their diagnostic quality, which is related only to the patient's pose and therefore directly defines the quality of the pose. This is necessary because there is no generally valid correlation between positioning parameters and diagnostic radiograph quality. As it is not allowed to capture radiographs of humans without an indication, studies with healthy subjects are not possible. Due to legal, data protection, and regulatory hurdles, it is not readily possible to acquire data in clinical practice either. In particular, intentionally taking radiographs of patients in poses with insufficient diagnostic quality is ethically unacceptable but necessary for training. Therefore, we used two anatomical preparations of the lower leg in a pilot study and captured them in different poses using two Azure Kinect DK ToF cameras (see Figure 4.3). These cameras provide depth images containing the measured distance in millimeters for each pixel. To be able to mount the cameras and keep them in a fixed position, a 3D-printed camera mount was developed. Using this mount, whose design is shown in Figure 4.4, the cameras were attached to both ends of the X-ray tube. This position was chosen because the focus-detector distance is standardized, and therefore the cameras also have a defined distance to the target anatomy. As the X-ray device is aligned with the patient at the time of the exposure, the target anatomy is



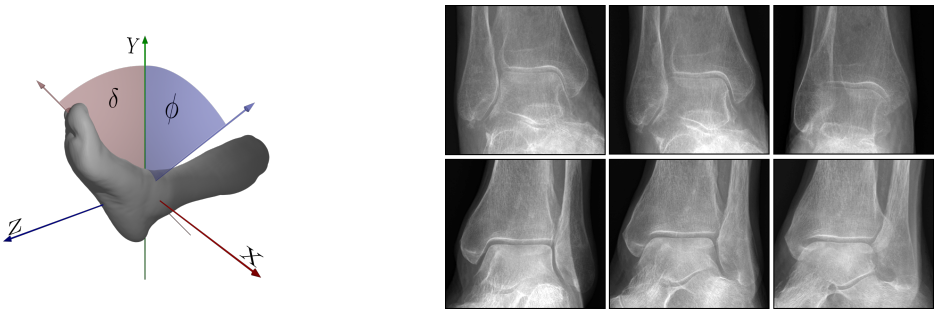
(a) (b)

Figure 4.3: Panel (a) shows the original X-ray device without the 3D-printed mount and the two ToF cameras. Panel (b) shows the experimental setup of the pilot study. The orange-outlined area is one side of the 3D-printed mount. The red-outlined area marks one ToF camera. The same structure is mounted mirrored on the other side. The anatomical preparation is fixed in its pose and is located under the Philips DigitalDiagnost C90 X-ray device. For each pose, both a radiograph and depth images were taken of the anatomical preparations.



(a) (b) (c)

Figure 4.4: Design of the 3D-printed camera mount. Panels (a) and (b) show the mount from the front and rear, while panel (c) shows a cross-section. The camera is mounted in a removable holder, which allows the camera to be removed without having to dismantle the entire mount. The angle of the camera can be adjusted and fixed within the camera holder.



(a) (b)

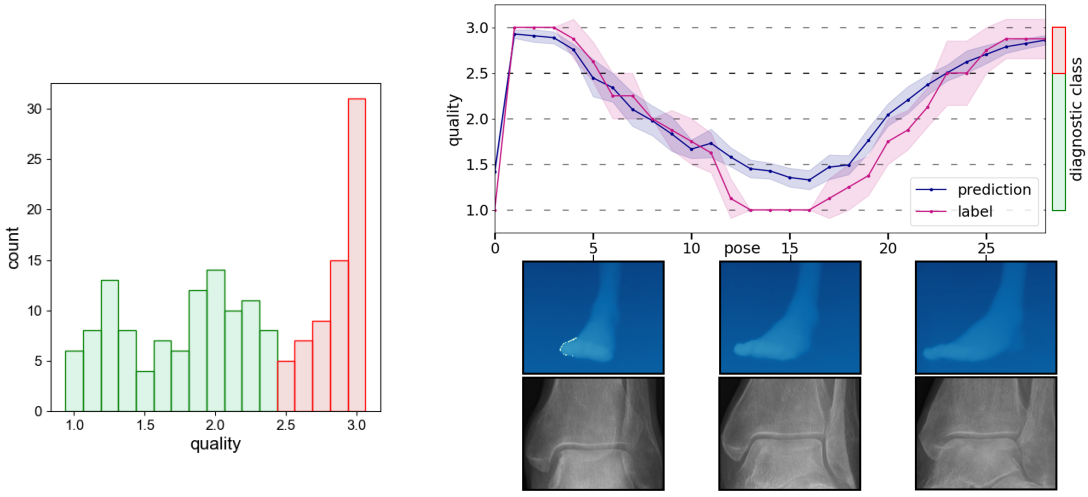
Figure 4.5: The image in panel (a) shows the axes of rotation of the preparations used in the pilot study to generate the different poses. Panel (b) shows examples of radiographs of different diagnostic qualities for both anatomical preparations (rows): The first column shows images with a diagnostic quality of 1, the second with a quality of 2, and the third with a diagnostic quality of 3.

always in the camera's field of view, regardless of where the patient is located in the room. The orientation of the cameras was chosen so that the patient's target anatomy is approximately in the center of the image if the focus-detector distance is in the range of 1 m to 2 m, as is the case for most anatomies. For this study, two cameras were used in order to be independent of the rotation of the X-ray device but also to investigate whether one perspective is more important than the other.

The pilot study was carried out in such a way that it replicates the clinical X-ray process. This includes the pose of the anatomical preparations, the distance of the X-ray machine from the detector, and the height and arrangement of the X-ray table. The anatomical preparations were provided by the local Institute of Anatomy for scientific research and were subsequently made accessible to us for this pilot study. This process has received approval from the local ethics committee and complies with HIPAA guidelines. This study was carried out in a university clinic using a Philips DigitalDiagnost C90.

The feet, i.e., a left and a right foot of two different Caucasian women of normal build, aged 69 and 86, were positioned in three different flexions, ranging from maximal flexion to maximal extension. To achieve this, we rotated the foot around X by the angle δ while keeping the lower leg fixed. The axis and angles are visualized in Figure 4.5a. For each flexion, the feet were rotated medially around the longitudinal axis Z in 29 steps by ϕ, for practical reasons starting from the ideal pose, resulting in 87 typical poses for each foot. Since rotation around the Y axis would have no effect on the visibility of the joint gap, we kept it fixed.

Since we captured 60 depth images per pose using two cameras, we ultimately acquired a total of 10,440 depth images with a resolution of 640×576 and 174 corresponding radiographs. These radiographs were then labeled by four radiologists regarding their



(a) **(b)**

Figure 4.6: The histogram in panel (a) shows the distribution of the averaged diagnostic quality labels of the radiographs. The green-marked bars represent the qualities belonging to the class *diagnostic* and the red ones accordingly to the class *non-diagnostic*. It can be seen that the distribution of the labels is not uniform, with a peak at 3. Panel (b) shows the averaged quality predictions from the *score fusion* architecture, the averaged labels from radiologists, and their corresponding standard deviation for all poses of one foot and one flexion. The poses were achieved by medial rotation of the foot around the longitudinal axis (Z in Figure 4.5a), resulting in a rise and fall in quality. For specific poses, the depth images and radiographs are provided below the graph. It is visible that small changes in the pose have large effects on the quality.

diagnostic quality. The evaluation scale was between 1 and 3, with 0.5 increments, whereby a quality of 1 is ideal and one of 3 is insufficient. In this context, one essential criterion for the diagnostic quality assessment of radiographs of the upper ankle joint is the visibility of the joint gap. Examples of labeled radiographs are shown in Figure 4.5b. Furthermore, all radiographs with labels in the interval [1.0, 2.5) are grouped as class *diagnostic* and from [2.5, 3.0] as class *non-diagnostic*, i.e., of unusable quality. For the distribution of the labels, see Figure 4.6a.

Due to the anatomy of the upper ankle joint, the range of poses that lead to the visibility of the joint gap, and thus to an assessment of 1, is significantly smaller than for the other classes. To counteract this, poses in this range were captured more frequently. The labels of the different poses of one flexion, as well as selected corresponding depth images and radiographs, are shown in Figure 4.6b. The dataset as well as the code can be found on GitHub[2].

[2] `https://github.com/INB-KI-SIGS/patient-pose-assessment`

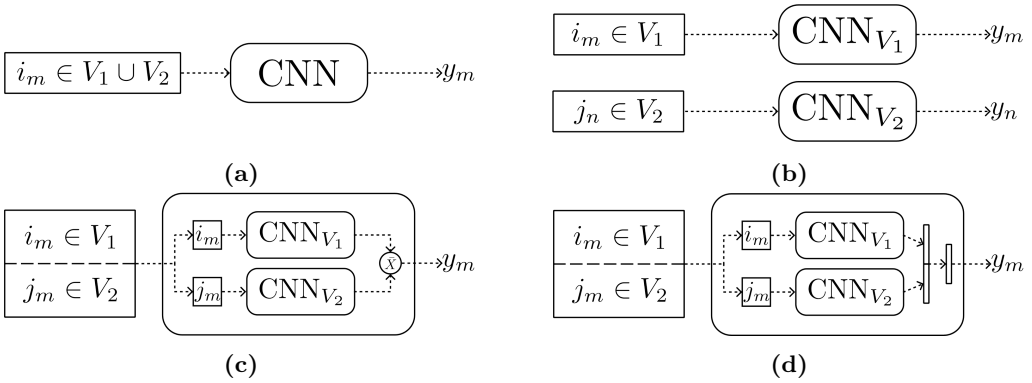


Figure 4.7: This figure shows the used architectures. *Single network* in panel (a): Input images are drawn from both viewpoints and fed into one CNN. *Separate network* in panel (b): From each viewpoint, images are drawn individually and fed into separate CNNs. *Score fusion* in panel (c): Corresponding images from both viewpoints are drawn and fed into one CNN containing separate branches; the output scores are averaged. *Late fusion* in panel (d): Like *score fusion*, but feature fusion is done in a linear layer that reduces 512 features to 256. V_1 and V_2 are sets of all images of view 1 and view 2.

4.1.3 Methods

To evaluate whether it is possible to assess the patient's pose based on depth images, using the dataset described in Section 4.1.2, we implemented and tested different architectures suitable for this problem. The following sections first explain the architectures in depth and then describe the details of the training process.

Neural-Network Architectures

Since our dataset was acquired using two ToF cameras facing opposite directions, the scene was captured from two different views simultaneously, leading to a multi-view task. To address this, we have implemented and tested different architectures. They can be divided into *fusion* architectures, which use both views as input simultaneously and fuse their features, and *no-fusion* architectures, which use only one of the two views at a time. They are visualized in Figure 4.7.

The *no-fusion* architectures consist of the *single network* and the *separate network* architecture. The *single network* architecture served as the baseline: The whole dataset containing images from both views was used to train a single network. This architecture cannot use the information shared between the two views. In the *separate network* architecture, we trained both views separately using two independent networks — one for each view. Due to the separation while training, it is possible to assess the importance of each view.

To use the information shared between the two views, we implemented two *fusion* architectures, similar to Seeland et al. [2021], called *late fusion* and *score fusion* architecture. Both use two separate branches of the same network architecture — one for each view. For *late fusion*, the outputs of the branches are feature vectors, which are then concatenated for feature fusion and propagated through a linear layer, which yields a quality score. For *score fusion*, each branch outputs a score individually, and the two scores are then aggregated into a quality score by calculating their mean. Both architectures were trained and evaluated with and without shared weights between both branches.

Training

We used the PyTorch framework version 1.11.0 [Paszke et al., 2019] to implement all architectures. Due to the small size of our dataset, the compact yet powerful EfficientNet-B0 [Tan et al., 2019b] was used as the base network. Since our dataset consists of depth images and models pretrained on ImageNet [Deng et al., 2009] did not perform better, we trained all models from scratch. Because the labels have an intrinsic order, we modeled the task as regression. To increase the comparability between architectures, all EfficientNet-B0 used the same initialization. All models were trained using the Mean Squared Error (MSE) as the loss function, comparing the predicted quality with the averaged labels of the radiologists. The Adam [Kingma et al., 2017] optimizer with a learning rate of $1 \cdot 10^{-3}$ and a batch size of 16 was used for training for 200,000 iterations. In the last 10,000 iterations, the learning rate was decreased four times by a factor of 10. For each branch of the architectures, the dropout rate [Srivastava et al., 2014] was set to 0.5 and the stochastic depth probability [Huang et al., 2016] to 0.2. To ensure that the *fusion* architectures always use information from both camera views, we randomly set all features of one branch, before fusion, with a chosen probability, to zero. We call this *branch dropout*. For the *late fusion* architecture, this affects the branch feature vectors, and for the *score fusion*, the branch scores. When evaluating several *branch dropout* probabilities, a value of 0 led to the worst results, and values of 0.2 and 0.4 led to the best results for *late fusion* and *score fusion*, respectively.

Furthermore, in order to narrow down the image to the area relevant for the task, a Region of Interest (ROI) was cut out and scaled to a fixed input size of 336×336 pixels. Of the 60 images that were captured per pose, 5 were randomly selected for each camera. In addition, the images were randomly rotated, flipped horizontally, blurred, and cropped during training. To further reduce the influence of the background, the areas in the image beyond a distance of 1.5 meters were set to 0. Furthermore, in addition to Gaussian noise, we randomly set pixels with high range gradients to zero to introduce so-called invalid points, which are characteristic of depth images captured with ToF cameras. The values were scaled to the range 0 to 1, and then their mean

was subtracted. For architectures with multiple branches, each of the two images in a pair was augmented in the same way.

Since the classes of the dataset are unequally distributed, the label interval was divided into 9 equal sections, which were sampled uniformly during training to avoid a class-specific bias. To evaluate how well we generalize to other feet, we followed a leave-one-out cross-validation strategy by leaving out one of the feet while training. In our case this leads to a training set containing images of one foot and a test set containing the other.

4.1.4 Results

In order to reduce the influence of statistical outliers, each cross-validation was repeated with random initializations ten times for each architecture. The results of the ten cross-validation runs were averaged. Since the *single network* and *separate network* architectures output one score per view for each pose, we used three different methods to evaluate the scores:

Divided The corresponding scores for the individual views are evaluated independently.

Paired The corresponding scores for the individual views are averaged and then evaluated. By doing this, both scores can be combined to obtain a potentially more accurate prediction.

View The evaluation is done for both views independently. This applies only to the *separate network* architecture.

Each experiment was evaluated with respect to the following metrics:

Mean Absolute Error We used the MAE to measure the difference between label and predicted quality.

Correlation r_s To additionally evaluate how well the order of predictions is preserved with respect to the labels, we calculated Spearman's rank correlation coefficient.

Accuracy An output is classified as correct if the absolute difference between the label and the predicted quality is smaller than 0.5. This threshold is chosen because the labeling was done in increments of 0.5.

Diagnostic Accuracy By additionally classifying images into *diagnostic* and *non-diagnostic*, the problem becomes a two-class classification problem. For this, an output was classified as correct if both the prediction and the label are greater/smaller than the diagnostic threshold of 2.5. Since errors in predictions are of varying severity, we also calculated the sensitivity and specificity.

Table 4.1: The results for the *single network* and *separate network* architectures evaluated with respect to the metrics and methods from Section 4.1.4. It shows that *separate network* using the *paired* evaluation method achieves the best results. All values are percentages except MAE and Correlation.

Metric	Single Network		Separate Network			
	divided	paired	view 1	view 2	divided	paired
MAE	0.29±0.04	0.24±0.04	0.24±0.03	0.32±0.05	0.28±0.03	**0.23±0.03**
Correlation r_s	0.81±0.06	0.89±0.03	0.88±0.03	0.80±0.05	0.84±0.03	**0.90±0.03**
Accuracy	79.08±4.14	86.94±6.17	86.63±4.44	75.06±5.96	80.84±4.26	**89.53±4.73**
Diag. Acc.	87.34±3.29	92.60±2.61	93.20±2.28	86.28±3.08	89.74±1.99	**93.51±1.50**
Sensitivity	77.62±6.33	85.33±5.95	86.26±5.44	79.93±5.05	83.10±3.81	**89.91±4.78**
Specificity	95.90±2.73	98.13±1.69	**98.58±2.14**	92.02±3.88	95.30±2.53	95.88±2.43

The results of all metrics for the *no-fusion* architectures are shown in Table 4.1 and for the *fusion* architectures in Table 4.2. The tables show that the assessment of a patient's pose is possible with an accuracy of 90.2%, whereas the baseline only achieves an accuracy of 79.1%. This ranks the baseline among the worst when using *divided* evaluation. *Paired* evaluation leads to an increase of 7.9 percentage points (pp). The same can be seen in the *separate network* architecture, where the increase is even higher with 8.7 pp. When comparing the two architectures in terms of both evaluation methods, the *separate network* improves accuracy by up to 2.6 pp. Looking at the two individual views in the *separate network* architecture, *view 1* yields significantly better results for all metrics than *view 2*. Furthermore, it becomes apparent that images from *view 1* are sufficient to achieve the results of the baseline with the *paired* evaluation method.

Table 4.2 shows that for both *fusion* architectures, the results with shared weights are significantly worse than without. This is comprehensible because our setup provides fixed viewpoints, and it may be beneficial for the model to learn features from these independently. The results of the architectures with shared weights are on par with those of the baseline. The *score fusion* architecture without shared weights outperforms the *late fusion* architecture by 5.8 pp. Overall, better results can be achieved with the *fusion* architectures than with the *no-fusion* architectures when they are evaluated with the *divided* method. In general, *score fusion* without shared weights and *separate network* using *paired* evaluation achieve the best results, outperforming the baseline by about 11.0 pp. These comparisons largely apply to the other metrics.

While accuracy is a good measure of how accurately quality can be predicted over the full label interval, it is more critical to distinguish *diagnostic* poses from *non-diagnostic* ones, which we measured with the diagnostic accuracy. Our results show that diagnostic accuracy always outperforms accuracy, reaching values of up to 93.5%.

Table 4.2: The results for the *fusion* architectures evaluated with respect to the metrics and methods from Section 4.1.4. It shows that *score fusion* using non-shared weights achieves the best results. All values are percentages except MAE.

Metric	Late Fusion		Score Fusion	
	shared weights	no shared weights	shared weights	no shared weights
MAE	0.30±0.05	0.26±0.04	0.28±0.04	**0.23±0.03**
Correlation r_s	0.82±0.05	0.87±0.04	0.83±0.04	**0.88±0.02**
Accuracy	78.80±7.91	84.34±5.44	79.05±6.25	**90.17±5.67**
Diag. Acc.	89.70±3.33	92.68±3.12	91.59±4.52	**93.33±2.09**
Sensitivity	83.78±8.55	88.53±6.85	85.15±9.68	**90.25±5.62**
Specificity	94.16±3.58	95.77±2.94	**96.71±2.00**	95.47±2.74

This is reasonable since the diagnostic quality does not penalize errors above and below the 2.5 threshold. Since it is worse if a *non-diagnostic* pose is predicted as *diagnostic* than the other way around, the sensitivity of the diagnostic accuracy is more critical than the specificity. Even though in our case the specificity is always slightly higher than the sensitivity, the latter still achieves values up to 90.3%. Finally, the high correlation values confirm that high/low predicted quality is associated with high/low labeled quality.

4.1.5 Conclusions

We have shown how to help radiographers find an adequate pose before taking the radiograph by capturing depth images of the patient's pose and predicting the diagnostic quality. Since no datasets exist that link patient poses to the quality of their radiographs, we created one using two anatomical preparations of the lower leg. Although a larger dataset would be desirable, this is not feasible due to limited availability of anatomical preparations and regulatory hurdles. Nonetheless, our results show that what is learned from one of the preparations can be transferred to the other. This shows that pose-specific features exist in depth images that allow conclusions to be drawn about the diagnostic quality of the corresponding radiographs. When examining different architectures, differences in the achieved accuracy have further shown that not every viewpoint contains these features equally. It became clear that it is still useful to aggregate the views in order to achieve the best possible results. Another critical point is that each view has an independent feature extractor (no shared weights). Taking these points into consideration, we achieve a maximum accuracy of 90.2%.

Looking at the metric of diagnostic accuracy, which is the most relevant to the patient, we have shown that it is possible to detect 90.3% of *non-diagnostic* poses with this approach. This would translate into a significant reduction in radiation exposure

in the clinical setting. By continuing to identify 95.5% of *diagnostic* poses as such, the X-ray workflow is not unnecessarily disturbed if the radiographers position the patient well themselves. Thus, this system could potentially add significant value to the X-ray imaging procedure. Although further data is needed to verify that the system works reliably regardless of age, gender, and race, our proof of concept demonstrates feasibility.

4.2 Synthetic Data Generated from CT Scans for Patient Pose Assessment

By attaching two Time-of-Flight (ToF) cameras to the X-ray device, we were able to show in the previous section that depth images of anatomical preparations of upper ankle joints contain information that can lead to high-accuracy pose assessment. In order to learn the mapping between the depth image of the pose and the diagnostic quality of the radiograph, pairs of X-ray images and depth images must be acquired simultaneously. Furthermore, the radiographs must be annotated with their diagnostic quality. The depth image and the annotation can then be used to train neural networks to predict the diagnostic quality of the radiograph before the radiograph is even taken.

Due to regulatory hurdles, however, it is difficult in practice to acquire the required depth images and corresponding radiographs. Using cameras in live clinical practice is not possible for data protection and regulatory reasons, and working with anatomical preparations is not a scalable solution.

To address this challenge, in this work we present a framework that synthetically generates the required image pairs of depth images and radiographs from computed tomography (CT) scans. CT scans that have already been taken can thus be used retrospectively to create a large dataset, which makes the approach scalable. It is furthermore possible to intentionally generate non-diagnostic poses by selectively adjusting the CT scans. We show that by pretraining on our generated synthetic dataset of upper ankle joints, the pose assessment of real upper ankle joints can be improved by up to 11 percentage points (pp).

The methods described are part of the following granted patent:

[Laufer et al., 2024a] Laufer, M., Mairhöfer, D., Bischof, A., Käster, T., Sieren, M., Reis, F. L., Gerdes, H. W., and Simon, P. “Verfahren zur Erzeugung von Trainingsdaten für ein KI-basiertes Assistenzsystem und Vorrichtung zur Unterstützung der Röntgendiagnostik”. DE102022133272A1. 2024.

This section has been published as:

[Laufer et al., 2025] Laufer, M., Mairhöfer, D., Sieren, M., Gerdes, H., Leal dos Reis, F., Bischof, A., Käster, T., Barth, E., Barkhausen, J., and Martinetz, T. “Synthetic Data Generated from CT Scans for Patient Pose Assessment”. In: *Proceedings of the Eighth Conference on Medical Imaging with Deep Learning (MIDL 2025)*. Proceedings not yet published. PMLR, 2025.

According to the Contributor Roles Taxonomy (CRediT), the contributions of the author of this thesis to the publication are: Conceptualization (together with M.L.), Data curation (together with M.L.), Investigation (together with M.L.), Writing – original draft (together with M.L.), Writing – review & editing (together with M.L., E.B., T.M.)

4.2.1 Related Work

For the generation of synthetic radiographs from CT scans, known as digitally reconstructed radiographs (DRRs), two main approaches are used: The ray tracing approach, using forward projection, casts a ray through the CT volume for each pixel on the detector and accumulates the intensity of the values along the path. Although ray tracing is computationally efficient, it is incapable of modeling scattering or beam hardening [Russakoff et al., 2005]. In the approach presented in Unberath et al. [2018], a radiograph generated by forward projection is combined with deep learning-based scatter and noise estimation. In contrast, the Monte Carlo (MC) approach simulates the transport of photons across the CT scan to model the photon-matter interaction and therefore requires material properties for each CT voxel [Badal et al., 2009]. Such simulations result in realistic DRRs; however, they are computationally more expensive than forward projection. Although there is little research on generating synthetic depth images from CT scans, more works are investigating the generation of point clouds from CT scans. A voxel-based approach to generate surfaces from medical 3D data is the Marching Cubes Algorithm (MCA) [Lorensen et al., 1987]. Saiti et al. [2022] use the MCA to create synthetic point clouds from CT scans to learn multimodal registration with point clouds and CT scans. To the best of our knowledge, there is no framework that generates pairs of synthetic depth images *and* corresponding radiographs for different views from CT scans and uses them for training patient pose assessment models.

4.2.2 Framework

The generation of synthetic radiographs and depth images from CT scans involves several steps. First, the surface of the target anatomy is extracted from the CT scan as a point cloud. The point cloud is then augmented to simulate different body types. These point clouds are placed on a pre-recorded point cloud of an imaging table in an X-ray room and rotated to create poses of different diagnostic quality. A synthetic depth image is generated for each pose by 2D projections of the point clouds. The corresponding synthetic radiograph is generated for each pose from the CT scan using an MC simulation. Our framework, which is implemented via Open3D's graphical visualization, is shown in Figure 4.8 and examples of synthetically generated depth images and radiographs are shown in Figure 4.10.

Preprocessing

In order to generate a synthetic depth image of the target anatomy from a CT scan, the target anatomy must first be extracted from the CT scan. For this, the scan is converted to a point cloud using MCA. The threshold value of the MCA is set to -500 Hounsfield units (HU) so that the air around the patient is removed. The point

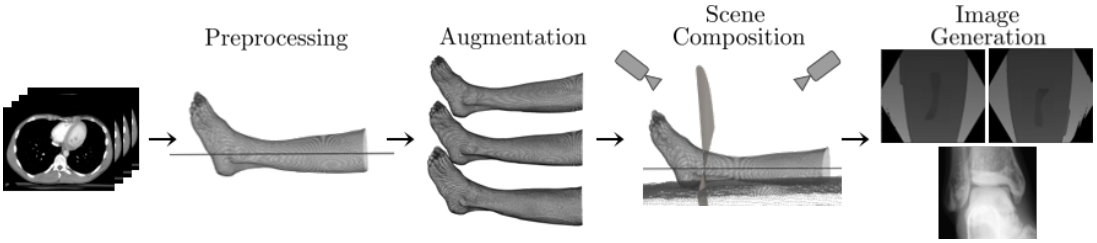


Figure 4.8: Schematic overview of the framework. The CT scan as input is passed through the steps described in Section 4.2.2 to generate synthetic radiographs and depth images.

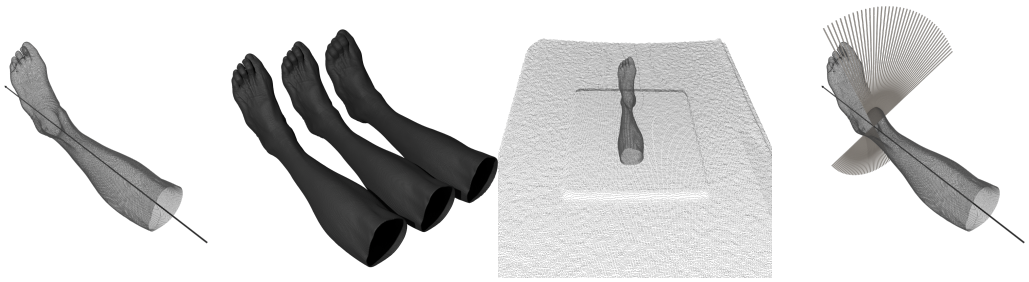
cloud is then cropped to the target anatomy to simplify the subsequent steps and calculations. Since only the surface of the target anatomy is relevant for the synthetic depth image, additional MCA runs, the clustering algorithm DBSCAN [Ester et al., 1996], and cropping are applied to remove the imaging table and points that do not belong to the surface of the target anatomy. To generate poses with different diagnostic quality, including inadequate quality, the target anatomy can be brought into other poses by specific rotations of the point cloud. The axis of rotation is strongly dependent on the target anatomy. For the upper ankle joint, it is the longitudinal axis, which is positioned in the point cloud to mimic human leg rotations. This axis passes through the center of the upper ankle joint. See Figure 4.9a for illustrations.

Augmentation

By determining the normal vectors of the point cloud, it is possible to shift the point cloud both outwards and inwards along the direction of the normal vector, resulting in two additional point clouds. This allows us to simulate patients with different shapes and to increase the amount of data; see Figure 4.9b. The diagnostic-quality label applies to all augmentations of a particular pose, as it can be assumed that the anatomy that determines the quality, in particular the position of the bones in relation to each other, does not change with minor displacements along the normal direction.

Scene Composition and Synthetic Depth Image Generation

To generate realistic synthetic depth images, it is beneficial to embed the target anatomy in a realistic scene. This can be achieved by recording the X-ray room in advance, including the imaging table, so that the target anatomy can then be placed on the imaging table under the X-ray device; see Figure 4.9c. The exact position of the target anatomy is selected so that the X-ray beam of the X-ray device passes through the axis of rotation of the target anatomy. Since a realistic environment and positioning of the target anatomy have been established, it is possible for the user to easily determine the range of rotation of the target anatomy to include non-diagnostic poses. From the



(a) (b) (c) (d)

Figure 4.9: This figure illustrates the different steps described in Section 4.2.2. Panel (a) shows the target anatomy cut out of the CT scan as a point cloud and the rotation axis, which passes through the center of the upper ankle joint. Panel (b) shows the augmentation of the point clouds by moving the points along the direction of the normal vectors. Panel (c) shows the target anatomy combined with the previously acquired X-ray room including the imaging table and detector. Panel (d) sketches the positions of the X-ray device, which change due to the medial rotation around the longitudinal axis of rotation.

point clouds of the target anatomy and the X-ray room, a 2D projection yields the synthetic depth image by taking into account intrinsic and extrinsic camera parameters and distortion coefficients. This way, synthetic depth images can be generated for any desired angles of rotation, augmentations, and camera views.

Synthetic Radiograph Generation

When generating synthetic radiographs, it is important that all depicted anatomical features are identical to those of real radiographs in the same position to avoid false labels of the corresponding depth images. Generation based on a physical model was therefore preferred to methods based on deep learning. Since only a region smaller than the one depicted in the point clouds must be visible in the radiograph, the target anatomy is cropped out of the CT in the first step. The cropped CT voxels are then converted into material and mass-density voxels. The material voxels each contain one of the materials air, soft tissue, bone, or titanium, based on the HU value of the respective voxel. Similarly, the mass-density voxels contain the density of the material adjusted by the HU value.

For each pair of material and mass-density voxels, multiple radiographs corresponding to different positions are generated. Instead of changing the position of the target anatomy, which would require rotation and interpolation of the voxels, the position of the X-ray device relative to the target anatomy is changed; see Figure 4.9d. To generate the radiograph, the MCGPU tool by Badal et al. [2011] was used for an MC simulation together with the material properties from the PENELOPE 2006 [Salvat et al., 2006] material files to simulate $2 \cdot 10^{10}$ X-ray beam paths. While more simulated paths would

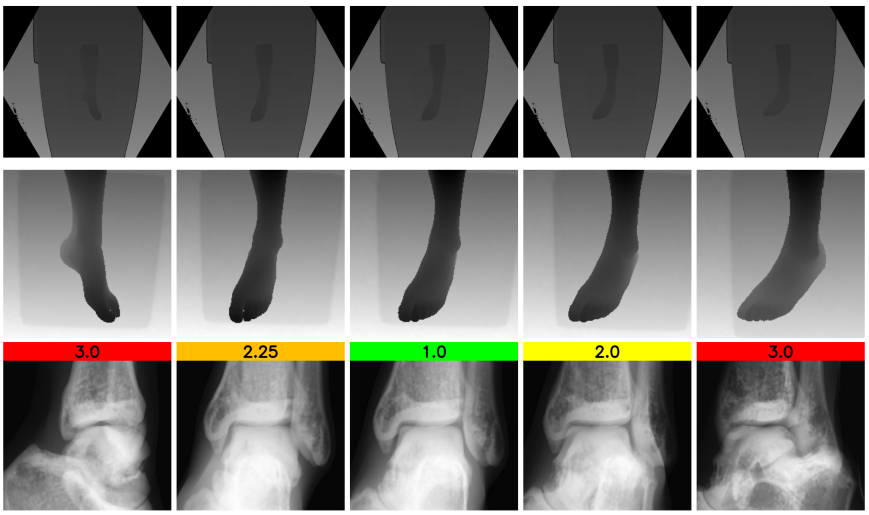


Figure 4.10: This figure shows synthetic images generated by using the framework. The first row shows the synthetic depth images that were generated with different rotations of the target anatomy. The second row shows the ROIs for the synthetic depth images from the first row, which are used for training. The third row shows the synthetic radiographs corresponding to the synthetic depth images, including their diagnostic quality. Note that only small rotations are necessary to change a diagnostic quality of 1 to a diagnostic quality of 2.

reduce the noise in the generated radiographs, the simulation time would increase. The use of $2 \cdot 10^{10}$ simulated paths allowed us to create realistic radiographs in a reasonable amount of time. The resulting raw image is then converted to a synthetic radiograph using a non-linear value mapping to obtain a look similar to a real radiograph. The synthetic radiograph is not as detailed as a real radiograph, but expert radiologists have validated that the visual quality is suitable for assessing the diagnostic quality for a given pose.

4.2.3 Datasets

The two datasets used in this paper consist of depth images from two camera views and corresponding radiographs. Each of these radiographs was assessed by four radiologists to determine its diagnostic quality on a scale of 1 to 3 in steps of 0.5. A diagnostic quality of 1 is ideal and a diagnostic quality of 3 is inadequate. The deciding factor in the assessment of the upper ankle joint is the visibility of the joint space. Radiographs with a label in the interval of $[1, 2.5)$ can be further classified as *diagnostic* and anything with a label greater than that as *non-diagnostic*.

Synthetic Dataset

Using the framework proposed in Section 4.2.2, we were able to generate pairs of synthetic radiographs and depth images of upper ankle joints from ten CT scans of different patients in 3,077 different poses. The anonymized CT scans were selected to contain flexed upper ankle positions and exclude clutter such as tubes or screws. From the ten CT scans, a total of 17 upper ankle joints were extracted and rotated medially around the longitudinal axis in a range of 90°. A synthetic depth image was generated from two camera views V_1 and V_2 for each half degree, i.e., a total of 181 poses per foot. This was done for each of the three augmentations, resulting in a total of 18,462 depth images. Since the synthetic radiograph is the same for all camera views and augmentations, one synthetic radiograph was created for each of the 181 poses, resulting in a total of 3,077 synthetic radiographs. To the best of our knowledge, this is by far the largest dataset linking depth images with diagnostic quality labels.

Anatomical Preparations Dataset

As presented in Section 4.1, we captured two anatomical preparations, a left and a right lower leg of two women, in 174 different poses using two ToF cameras. Parallel to the depth images from two different views, a radiograph of the upper ankle joint was also taken. The preparations were rotated medially around the longitudinal axis and flexed in three different positions of the ankle joint. In contrast to the generated synthetic data, this dataset provides real data.

4.2.4 Experiments and Results

The two experiments carried out are designed to answer the following questions:

Experiment 1: Can a neural network be trained on the generated synthetic depth images to assess the pose with high accuracy?

Experiment 2: If so, can the neural network be finetuned on real data in order to improve the results of patient pose assessment?

Both experiments were modeled as regressions to better reflect the nature of the labels. While the depth images served as input for the models, the models output a single continuous value between 1 and 3 as the diagnostic quality. All experiments were implemented using PyTorch and repeated ten times with different seeds. The results were then averaged.

Experiment 1

The first experiment tests whether relevant features can be learned with only the synthetic dataset. In addition, to clarify whether the augmentation of the depth images has a benefit, the training was carried out with and without augmentation.

Training Two EfficientNet-B0 models [Tan et al., 2019b], called CNN_{V_1} and CNN_{V_2}, were used. Following the separate network architecture (see Figure 4.11a) described in Section 4.1.3, the CNN_{V_1} is only trained on the depth images from camera view V_1 while the CNN_{V_2} is trained on images from camera view V_2. The models each receive a single-channel depth image i, resized to 336×336 pixels, as input and output a single continuous value y. The models are trained to minimize the mean squared error between the model's output and the averaged diagnostic quality that had been assigned to the input depth image. The Adam optimizer [Kingma et al., 2017] was used with an initial learning rate of $1 \cdot 10^{-3}$, which was decreased four times by a factor of 10 in the last 10,000 steps. We used a batch size of 16 and trained each model for 200,000 steps. For this experiment, a three-fold cross-validation was performed on the synthetic dataset. For each cross-validation run, the test set consisted of three randomly selected upper ankle joints so that no subject from the test set would appear in the training set. Note that possible differences in the results in comparison to Laufer et al. [2024b] are due to different implementations and updated libraries. The results are shown in Table 4.3.

Experiment 2

To answer the second question, we evaluated whether pretraining with the synthetic data improves performance when finetuning on real data (see Section 4.2.3). In addition, we compare these results with those of a model that was trained from scratch on the anatomical preparations dataset without any pretraining and with those of a model that was pretrained on ImageNet [Deng et al., 2009] and then finetuned with the anatomical preparations dataset. Furthermore, as in Experiment 1, we evaluated whether the augmentation of the depth data during pretraining influences the results.

Pretraining There are four alternative ways to initialize weights before finetuning. For training *from scratch*, without any pretraining, the models are randomly initialized. For the *pretraining on ImageNet*, both models CNN_{V_1} and CNN_{V_2} are initialized with the ImageNet weights. When pretraining using the synthetic dataset, there are two possible ways, as this dataset also contains depth images of two views: In the *unified camera view approach*, both models are initialized with identical weights $W_{V_1 \cup V_2}$ obtained by pretraining with images from both camera views V_1 and V_2 of the synthetic dataset; see Figure 4.11c. With the *camera view specific approach*, each model CNN_{V_1}

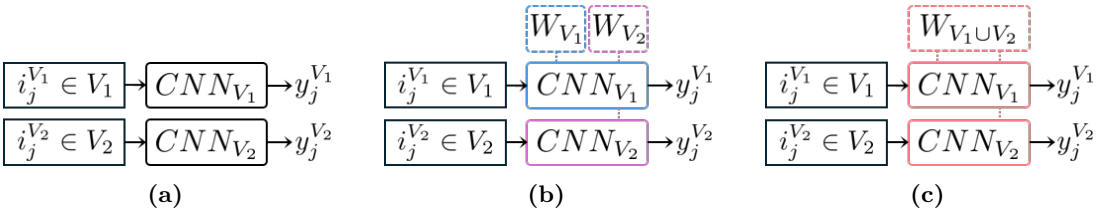


Figure 4.11: Panel (a) shows the separate network architecture: individual images $i_j^{V_1}$ and $i_j^{V_2}$ from each camera view V_1 and V_2 are used to train two separate CNNs. $y_j^{V_1}$ and $y_j^{V_2}$ are the continuous outputs of the networks, ranging from 1 to 3. Panel (b) shows the *camera view specific approach*, where the CNNs are initialized with the weights W_{V_1} and W_{V_2} obtained by pretraining with the corresponding view. CNN_{V_1} would therefore only be initialized with weights W_{V_1} that were obtained by pretraining with images only from V_1. Panel (c) shows the *unified camera view approach*, where both CNNs are initialized with the same weights $W_{V_1 \cup V_2}$ obtained by pretraining with images from both views $V_1 \cup V_2$.

and CNN_{V_2} is initialized with weights obtained by pretraining on synthetic depth images from only camera view V_1 or V_2, respectively; see Figure 4.11b. This approach effectively halves the amount of pretraining data, but the finetuning is more specific. The pretraining on the synthetic dataset was performed as described in Section 4.2.4, except that the whole dataset was used as training set.

Finetuning The training from scratch and finetuning on the anatomical preparations dataset were performed with the same hyperparameters as in Experiment 1 in order to be comparable, using the separate network architecture shown in Figure 4.11a. As the anatomical preparations dataset consists of only two preparations, we trained on one preparation and tested on the other and vice versa. The results are shown in Table 4.4.

Results

The high accuracy of 87.6% in Experiment 1 based on the synthetic dataset (see Table 4.3) shows that synthetic depth images can be used to learn to assess poses. It further shows that training with augmented synthetic depth images results in an improvement in all metrics compared to training without augmentation. Compared to the diagnostic accuracy of just 59.3% in the case of a simple baseline that only predicts the mean of all labels, the proposed approach leads to a significant improvement. The results in Table 4.4 show that pretraining with the synthetic data improves training on real data. Although the sensitivity of the diagnostic quality is highest (93.79%) for the model pretrained on ImageNet, the other metrics show that there is no benefit from pretraining with ImageNet compared to training from scratch. This suggests that the features learned on ImageNet are not useful for the task. However, pretraining on the synthetic depth images is useful since, except for sensitivity, all metrics are improved

Table 4.3: Results of the experiments conducted solely on the synthetic dataset and evaluated by the described metrics regarding the impact of augmentation of the synthetic depth images. Note that training with augmented data can improve the results for all metrics. All values are percentages except MAE and Correlation.

Metric	without augmentation	with augmentation
MAE	0.23±0.02	**0.21±0.02**
Correlation r_s	0.85±0.03	**0.87±0.02**
Accuracy	85.29±2.34	**87.60±2.60**
Diag. Acc.	85.45±2.77	**86.96±2.47**
Sens.	91.50±1.50	**92.82±1.80**
Spec.	76.90±6.01	**79.00±5.88**

Table 4.4: Results of the experiments without pretraining, with pretraining on ImageNet, and with pretraining with the synthetic dataset and subsequent finetuning on the anatomical preparations dataset, using the metrics and methods described in Section 4.2.4. Note that pretraining with synthetic data outperforms models without pretraining and with pretraining on ImageNet. All values are percentages except MAE and Correlation.

Metric	from scratch	pretrained onImageNet	pretrained on synthetic dataset			
			unified camera view		camera view specific	
			wo/ aug.	w/ aug.	wo/ aug.	w/ aug.
MAE	0.28±0.04	0.31±0.05	**0.22±0.04**	0.24±0.04	0.23±0.03	**0.22±0.06**
Correlation r_s	0.88±0.04	0.90±0.03	**0.92±0.03**	0.90±0.04	0.90±0.03	0.91±0.04
Accuracy	79.25±6.98	77.37±8.22	89.03±5.51	86.91±6.14	**90.45±4.73**	90.07±7.64
Diag. Acc.	89.08±3.75	84.20±2.43	90.93±3.00	91.91±4.64	**92.80±2.14**	92.43±4.05
Sens.	91.21±4.23	**93.79±3.19**	92.65±7.45	90.90±10.29	87.64±4.51	91.16±6.72
Spec.	88.66±4.61	80.05±4.60	89.84±5.49	92.24±6.59	**95.34±3.91**	93.10±6.27

over those obtained without pretraining. Most importantly, the accuracy increases by 11 pp to 90.45% for the *camera view specific approach* without augmentation. With the same model, the diagnostic accuracy can also be improved by 3 pp.

Moreover, the *camera view specific approach* is performing slightly better compared to the *unified camera view approach* according to the more important accuracy metrics, which is presumably due to the higher similarity of the pretraining and finetuning dataset. In contrast to the results from Table 4.3, the augmentation of the synthetic depth images does not yield any benefits in these experiments. This is possibly due to an insufficient variance in the shape of the rather few anatomical preparations.

Note that in this work we obtain an overall accuracy of 90.45% when predicting diagnostic quality based on depth images. When comparing with the accuracy (93.0%) obtained for predicting the same diagnostic quality by using real radiographs as shown in Section 3.1, we can conclude that a similar quality assessment is possible with only depth images, although it is significantly more difficult.

4.2.5 Conclusion and Outlook

In this work, we presented a framework for generating both synthetic depth images and radiographs from CT scans. We have shown that by pretraining on such a synthetic dataset relevant features can be learned, which are useful for the assessment of patients' poses on real data. This makes it easier to assess whether the patient's pose would lead to a radiograph with inadequate diagnostic quality before it is taken and thus protect the patient from unnecessary radiation due to a retake.

The main advantage of this framework is that the data acquisition problem, which is often critical in the medical context, can be solved by using already available CT scans. It is also possible to adapt the synthetic training data to different X-ray rooms and different ToF camera setups in order to generate realistic and case-specific training data. Furthermore, it is possible to investigate which camera positions and how many cameras are best suited for the pose assessment task.

Although here we have only shown this for upper ankle joints, we believe that the framework can also be applied to other anatomies. Ideally, our novel way of generating radiographs and depth images can be used beyond this pose-assessment application.

Chapter 5

Collimation Optimization

5.1 AI-based Collimation Optimization for X-ray Imaging Using Depth Cameras

The previous chapters explored how the diagnostic quality of radiographs can be assessed automatically and how the quality can be estimated even before the radiograph is taken. Both aim to optimize a critical factor in diagnostic quality, the patient's pose, in order to support radiographers and improve treatment. However, in addition to the patient's pose, collimation is another important factor for diagnostic quality and one of the main causes for a retake and the associated costs [Little et al., 2017]. Collimation during radiography is the process of defining the area to be radiated and influences treatment in several ways. On the one hand, it is crucial for diagnostic quality that all anatomical areas necessary for diagnosis are imaged. If this is not the case, an incorrect or no diagnosis may be made, treatment may be delayed, or the patient may have to undergo additional X-ray examinations. In addition, though, the collimation setting also directly affects the radiation dose to which the patient is exposed. Therefore, to ensure patient safety, the collimated field should be kept as small as possible while remaining large enough to cover all diagnostically essential anatomy. This goal is known as the *ALARA* (As Low As Reasonably Achievable) principle regarding the effective radiation dose. Consequently, precise collimation is an essential tool for meeting the requirements of this standard.

In clinical practice, collimation is done by radiographers. However, time pressure, inexperience, patients with irregular body shapes, e.g., because of obesity, or patient movement after collimation are reasons why collimation is often not optimal. Various studies show that frequently the radiation dose is unnecessarily high due to incorrect collimation [Farzanegan et al., 2020; Karami et al., 2016; Karami et al., 2017; Little et al., 2017]. In Karami et al. [2017] it is shown that, on the other hand, in more than 830 radiographs of the lumbar spine, the average irradiated area was 1.26 times larger than the diagnostically relevant area. This implies unnecessary irradiation of organs that are particularly sensitive to radiation, such as the ovaries. To protect the patient and to support the radiographer, AI-based optimization of the collimation during the

X-ray process would help. Rasche et al. [2024] showed that automatic collimation assistants can save radiographers time in their workflow and reduce machine interaction.

In this chapter, an approach that uses depth cameras to capture depth images of the patient and train deep neural networks to directly predict the optimal collimation is proposed. The optimal collimation can only be labeled on the radiograph, as only on the radiograph is it possible to define the areas relevant for the diagnosis. Therefore, corresponding pairs of depth images and radiographs taken at the same time are required for training. As it is not permitted to radiograph people without an indication, it is necessary to mount a depth camera on the X-ray device during ongoing clinical practice in the form of a clinical study to capture these image pairs of people who have an indication. Due to the high regulatory and legal challenges of such an experimental setup, we preceded such a clinical study with a proof of concept based only on the acquisition of depth images. The resulting dataset consists of a total of 1,020 depth images of the anterior-posterior (AP) view of the abdomen and the upper ankle joint of four subjects during a recreation of the clinical process of radiographing these anatomical sites. For this dataset, the correct collimation was directly labeled based on the depth information in close exchange with experienced radiographers. The next step was to conduct the clinical study by mounting a depth camera in clinical practice to take corresponding depth and radiographs of patients after the ethics application had been approved. With the help of this setup, we were able to capture 295 depth images and 59 radiographs of the thorax in posterior anterior (PA) and laterolateral (LL) views of patients in real clinical practice. For this dataset, the optimal collimation could be labeled on the radiograph, which is the gold standard, and then transferred to the depth image. The labeling was performed by an experienced radiologist. Using these two datasets, we were able to train neural networks to predict the optimal collimation on depth images and point clouds as input modalities and thoroughly evaluate our approach for different anatomical sites and architectures. An application of this approach in clinical practice could support the collimation decisions of radiographers, reducing the overall radiation dose for the patient and improving the overall workflow and diagnostic quality.

This section has been published as:

[Mairhöfer et al., 2024] Mairhöfer, D.*, Laufer, M.*, Berkel, L., Bischof, A., Barth, E., Barkhausen, J., and Martinetz, T. “AI-based Collimation Optimization for X-Ray Imaging Using Time-of-Flight Cameras”. In: *ESANN 2024 Proceedings*. *Authors contributed equally. Ciaco - i6doc.com, 2024, pp. 703–708.

[Mairhöfer et al., 2026] Mairhöfer, D.*, Laufer, M.*, Berkel, L., Sieren, M., Bischof, A., Barth, E., Barkhausen, J., and Martinetz, T. “AI-based Collimation Optimization for X-ray Imaging Using Depth Cameras”. *Neurocomputing* 661, 2026. *Authors contributed equally. P. 131881

According to the Contributor Roles Taxonomy (CRediT), the contributions of the author of this thesis to the publications are: Conceptualization (together with M.L.), Data curation, Investigation, Methodology (together with M.L.), Software, Visualization, Writing – original draft (together with M.L.), Writing – review & editing (together with M.L., E.B., T.M.)

5.1.1 Related Work

While there is already research regarding the determination of the optimal collimation on radiographs [Berg et al., 2020; Elgaard et al., 2022], at this point the possibly bad radiograph has already been made. In Sénégas et al. [2018] an approach is presented to learn the optimal collimation before taking the radiograph based on depth images. Since the authors' dataset consists of 177 depth images and radiographs of the chest in *AP* and lateral view, they were able to label on the radiograph. However, for this, they used landmarks as handpicked features instead of directly labeling the collimation. This leads to dividing the problem into a landmark detection, implemented as boosted tree classifiers, and a multivariate regression problem for the actual collimation area prediction. With such an approach, an end-to-end training becomes impossible. In contrast, we present an end-to-end trainable deep learning approach that uses depth information from multiple cameras for predicting optimal collimation and evaluate its performance for three anatomical sites. Furthermore, by labeling only the actual collimation area and not additional landmarks, our approach requires less labeling effort. To our knowledge, there are no other studies that automatically optimize the collimation before exposure to radiation.

5.1.2 Datasets

Since there was no public dataset in which subjects were captured under an X-ray device using depth cameras, two novel datasets were acquired for this work.

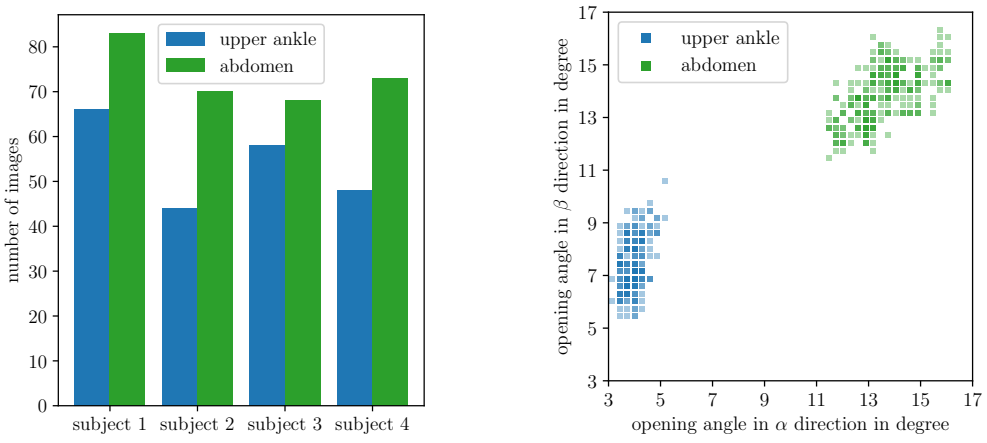
Clinical Setting Depth Dataset

This dataset was designed to answer the question of how well collimation areas can generally be predicted using depth images in a clinical setting. For this purpose, two Microsoft Azure Kinect Time-of-Flight (ToF) cameras were attached directly to the X-ray tube in an X-ray room used in everyday clinical practice. The two cameras were mounted on both ends of the X-ray tube to avoid possible occlusions of the patient. The experimental setup can be seen in Figure 5.1a. This study was carried out in a university clinic using a Philips DigitalDiagnost C90. Using this setup, four subjects (three male, one female) were captured in various typical poses of the upper ankle or abdomen in each case in the *AP* view on the X-ray table. A notable feature of our dataset is the inclusion of positions in which subjects had not yet fully moved to the

(a) **(b)**

Figure 5.1: In panels (a) and (b) the X-ray source is located in the horizontal tube, while the collimator is contained in the cube below the tube. The experimental setup used to acquire the *Clinical Setting Depth Dataset*, including the mounted Azure Kinect ToF cameras for acquisition, is shown in panel (a). One camera was mounted at each end of the X-ray tube so that the center of the collimation area is always in the field of view. In panel (b) the setup used to acquire the *Clinical Practice Depth Radiograph Dataset*, including the Intel DS435i mounted to the back of the collimator, is shown.

correct position for radiography. After sorting out images with motion artifacts or occlusions, a total of 510 depth image pairs (216 upper ankle, 294 abdomen) were selected for the dataset. In Figure 5.2a, the distribution over the individual subjects is shown. To aggregate information from both cameras, these pairs of depth images were converted into one registered point cloud. As this dataset does not contain radiographs, collimation was labeled on these merged point clouds in close exchange with experienced radiographers. To perform the labeling, the position of a virtual X-ray device and thus the position of the central beam as well as the two aperture angles relative to the point cloud were adjusted. Using the position, rotation, and opening angle, the label can be determined for each point of the point cloud. In Figure 5.2b, the distribution over both opening angles and their frequency is shown. For the upper ankle AP view, it can be seen that the angular range for β, i.e., the opening in the longitudinal axis, is large, while it is small for α, i.e., the opening in the transversal axis. This fits the anatomical conditions, since α depends on the width of the lower leg, while β is influenced by the length and posture of the lower leg. In contrast, the abdominal AP view shows a more linear relationship. A larger upper body requires a larger collimation in both angles. Based on these data, 10 augmented point clouds and 10 augmented depth images, together with the augmented label, were created from each merged point cloud for training. Including the augmented data, the dataset contains 2,376 labeled point clouds and depth images of the upper ankle AP view and 3,289 of the abdomen AP view. The following sections describe the details for the generation of the augmented data of the point clouds and the depth images, respectively. Figure 5.3 visualizes the described dataset creation process.



(a) **(b)**

Figure 5.2: This figure shows the distribution of the *Clinical Setting Depth Dataset* over the subjects as well as over the collimation opening angles. As shown in panel (a) there are more image pairs of the abdomen view than of the upper ankle view, while the distribution over the subjects is quite even. An even distribution helps to avoid a bias towards a single subject. The frequencies of opening angles in the α and β directions are shown in panel (b). For the upper ankle view, the range for the opening angles is highly variable. Note that the range for the angle β is larger than for angle α. The opposite is the case with the abdomen view. Here, α and β are more similar. Color coding indicates the frequency.

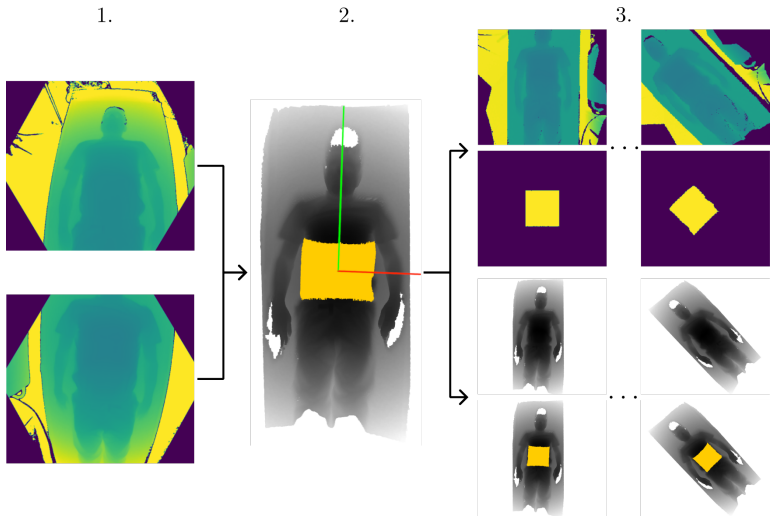


Figure 5.3: The creation of the *Clinical Setting Depth Dataset* can be divided into three steps. In step 1, pairs of depth images were acquired. In step 2, the depth images were merged into a registered point cloud. Then, optimal collimation was labeled based on this point cloud. In step 3, 10 different augmented depth images and point clouds were created from the original point cloud.

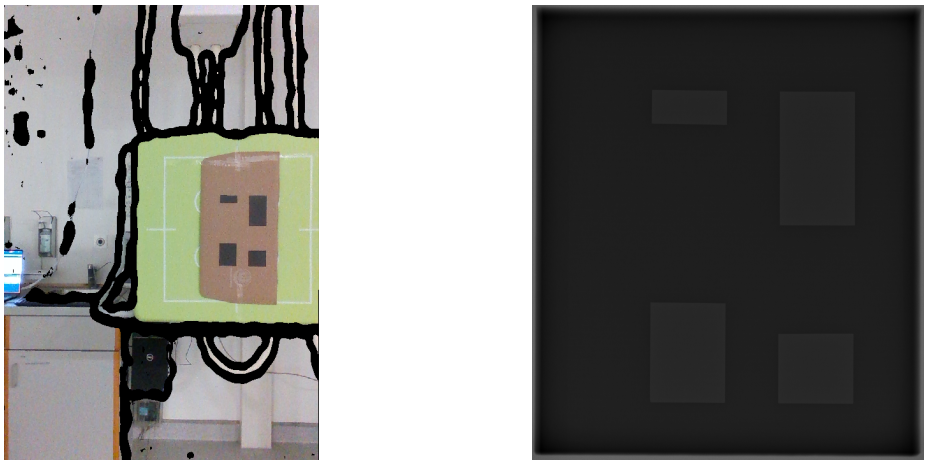
Point Cloud Generation Since the original point clouds contain a lot of unnecessary information, all points further than 1.5 m from both ToF cameras were removed. Before removing the points, the positions of the camera are translated randomly by up to 25 cm, which corresponds to a translation of the X-ray device before the exposure. From the remaining points, 75,000 points were then sampled using standard farthest-point-sampling to create a uniformly distributed point cloud.

Depth Image Generation To obtain a single depth image containing information from both cameras and avoiding a multiview problem, a synthetic depth image was rendered from the merged point clouds. For this, the origin of the central beam was chosen as the camera position, since the optimal collimation can be determined from this perspective. To create the synthetic depth image, the labeled optimal position of the X-ray device is randomly translated by up to 25 cm, and then the depth image is rendered from the new position. While the Azure Kinect captures depth images at a resolution of 640×576, we rendered the images at half that resolution, since the image size used during training is significantly smaller than the native resolution.

Clinical Practice Depth Radiograph Dataset

Because optimal collimation can only be correctly labeled on a radiograph, the corresponding radiograph must be acquired together with the depth image. An ethics approval was granted by the local ethics committee for the conduct of this study in a university hospital. Due to the size and weight of the Azure Kinect, we attached an Intel DS435i directly to the back of the collimator instead. This position allows easier subsequent matching of the radiograph and the depth image, even when the collimator is rotated. The setup is shown in Figure 5.1b. All patients under the age of 18, pregnant patients, and emergency patients were excluded from the study. In addition, only examinations with a wall or table detector were used. With this setup, a dataset of 29 radiographs of the thorax in *LL* view and 30 radiographs of the thorax in *PA* view were acquired. For each radiograph, the five depth images closest in time to the radiograph were selected as corresponding depth images. Using multiple depth images taken over a period of about one second allows it to be more robust against noise. The depth images have a resolution of 1280×720. In total, we were able to create a dataset of 59 radiographs and 295 corresponding depth images of examinations of the thorax. For this dataset, the optimal collimation was labeled by a radiologist directly on each of the radiographs by drawing a polygon. It is possible that parts of the polygon of optimal collimation lie outside the actual radiograph, for example, if the radiographer has selected a collimation that is too small.

In order to be able to transfer the radiographs into the coordinate system of the depth camera, a previous camera alignment was necessary. First, metallic markers were attached onto a cardboard board to obtain features that are visible in the camera images

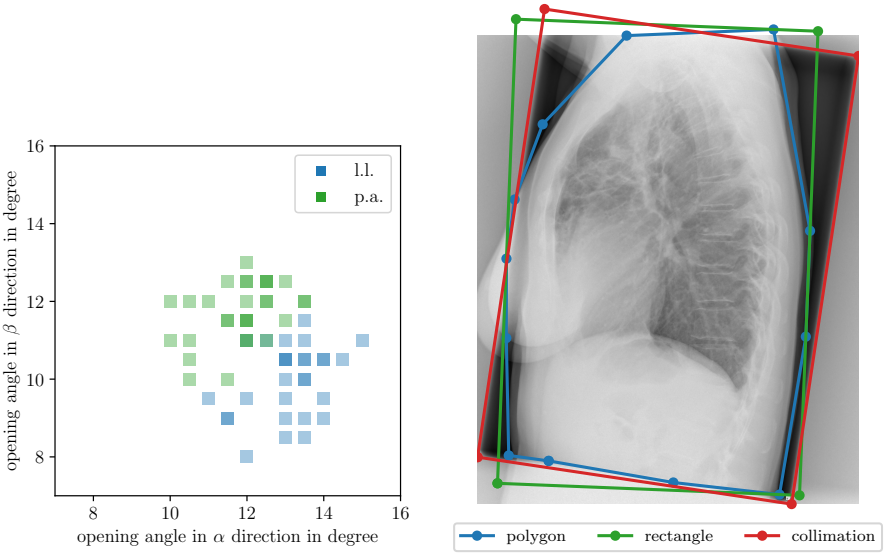


(a) **(b)**

Figure 5.4: Example images taken for the camera calibration. Panel (a) shows the RGB image from the depth sensor's viewpoint from the attached camera. One can see the cardboard board with metallic markers in front of the detector. Panel (b) The markers are clearly visible in the color image and in the radiograph, which allows an image alignment.

and also on the radiograph. This marked board was X-rayed while being recorded with the attached camera from different positions and distances, resulting in image pairs like the one shown in Figure 5.4. Using the visible features of the color image and the radiograph, both can be aligned. Since the vector orthogonal to the radiograph points from the image center toward the X-ray source, and the distance between the detector and the X-ray source is known, the position of the X-ray source can be derived. As the relative position of the X-ray source and the camera does not change, this position can be used to position new radiographs correctly in the camera's 3D coordinate system.

By converting each depth image into a point cloud, projecting the corresponding radiograph and the optimal collimation into the same 3D space, and knowing the position of the X-ray source, we define a cone that starts from the X-ray source and ends at the points of the optimal collimation. All points inside this cone are therefore points that should be irradiated, and all points outside of it are points that should not be irradiated. From the point cloud and the corresponding labeled points, the depth image can then be projected back again, with the corresponding label, to be used for training. In Figure 5.5a, the distribution over both labeled collimation opening angles and their frequency is shown.



(a) **(b)**

Figure 5.5: The distribution of the *Clinical Practice Depth Radiograph Dataset* over the opening angles of the rectangle annotation is shown in panel (a). While the rectangle annotation of the *PA* view is relatively square, i.e. α and β values are similar in size, the α value predominates in the *LL* view, which leads to a more elongated collimation; see 5.5b. Color coding indicates the frequency. The three different label types of a radiograph from the *Clinical Practice Depth Radiograph Dataset*, are visualized in panel (b). The polygon annotation defines the optimal collimation of the radiologist, the rectangle annotation marks the minimal rectangle spanned by the polygon annotation, and the applied collimation is set by the radiographers. Note that both the polygon and the rectangle annotation may be located outside the radiograph if important areas are missing on the radiograph.

5.1.3 Experiments and Training

To evaluate if it is possible to predict the optimal collimation using depth information, we implemented and tested different architectures suitable for our task using the two datasets described in Section 5.1.2. In all cases, the prediction was modeled as a segmentation task using the pixel-wise/point-wise labels. An alternative approach would be to train a regression to directly predict the position, rotation, and opening angles of the collimation. We decided against regression and in favor of segmentation for several reasons:

1. As the labels show, for example, in Figure 5.5b, optimal collimation is usually not rectangular. The restriction to rectangular collimation is based on the need for simple and quick adjustment by the radiographer. However, apart from

conventional X-ray diagnostics, for example in radiotherapy, multileaf collimators are also used, which can limit the radiation to irregular shapes. Predicting the area to be irradiated as segmentation makes it possible to irradiate even such irregular shapes, provided this is technically possible. If this is not the case, the smallest rectangle to be irradiated can simply be derived from the segmentation.

2. Even a rectangular optimal collimation will only automatically result in a rectangular label on the depth image if the depth camera is in the same position as the X-ray source. However, this is not possible. While it still would be possible to predict the collimation parameters from the depth image, this would require the network to learn a translation from the position of the depth camera to the X-ray source.

3. The task itself is highly spatially dependent. To predict the correct collimation, the network needs to learn where each anatomical structure is located. While this information is directly provided by segmentation masks, it is not by collimation parameters. Predicting spatially dependent collimation parameters from a one-dimensional feature vector, which is the output size in most classification or regression CNN models, may even degrade convergence and performance.

For the point clouds as input data, we used the PVCNN++ model [Liu et al., 2019] with a PointNet++ [Qi et al., 2017b] as backbone. For the depth images as input data, we used a UNet++ [Zhou et al., 2018] as a state-of-the-art segmentation model. Both architectures are described in Section 2.3.

Training using the Clinical Setting Depth Dataset

To answer the question of how well collimation areas can generally be predicted, we used the depth images and point clouds of the *Clinical Setting Depth Dataset*. Since there are no widely used pre-trained models for depth data, the models were trained from scratch. The training process was identical for all models of one input modality but differed slightly between image- and point cloud-based networks. Both networks were trained independently for both anatomical sites using the Cross Entropy Loss as the loss function for a point-wise/pixel-wise classification in the classes *collimation* and *no collimation*. For the point cloud models, the Adam optimizer [Kingma et al., 2017] with an initial learning rate of $1 \cdot 10^{-3}$ and a batch size of 16 was used for training for 400,000 iterations, which translates into approximately 2,700 epochs. This number of steps was chosen because validation accuracy has consistently increased until it plateaued at the end. The learning rate was decreased 3 times by a factor of 10, at 250,000, 300,000, and 350,000 iterations. From the merged point cloud, which was downsampled to 75,000 points, 4,096 points were randomly resampled during training and testing. The point clouds were furthermore augmented by adding jitter and applying random

translation and rotation transformations, simulating noise and different positions of the patient relative to the X-ray device. For normalization, the mean of the point cloud was subtracted. For the depth images, we used the stochastic gradient descent with $1 \cdot 10^{-3}$ as the learning rate, a momentum of 0.9, and a weight decay of $1 \cdot 10^{-6}$. The network was trained for 350,000 iterations using a batch size of 32 with the learning rate decreased 3 times by a factor of 10 at 200,000, 275,000, and 325,000 iterations. While training, the depth images were randomly augmented using horizontal flipping, rotation, noise, the addition of invalid points, and cropping. After augmentation, the depth images were normalized to the interval $[-0.5, 0.5]$ and resized to 224×224. From the four subjects of the dataset, one was randomly chosen before the training process and specified as the test set and one as the validation set. For the final training process, we trained on all three subjects and tested on the initially selected test subject. To reduce the influence of statistical outliers due to randomly initialized networks, each experiment was repeated 5 times for each architecture using the fixed hyperparameters but different initialization. The results of the repetitions were averaged. The results of the experiments using this dataset are presented in Section 5.1.4.

Training using the Clinical Practice Depth Radiograph Dataset

Since the initial results indicated that depth images are preferable to point clouds for rather square collimation areas, such as the abdomen, we have confined ourselves to depth images as input using the *Clinical Practice Depth Radiograph Dataset* for the thorax examinations.

To ensure optimal comparability with the results for the *Clinical Setting Depth Dataset*, the network architecture and all hyperparameters were kept identical. In this case too, individual networks were trained five times with different seeds for the different views, and the results were then averaged. The depth images were randomly sampled for each of the five repetitions into training and test data with a probability of 80% for the training set and 20% for the test set. Since the hyperparameters were identical and there was no optimization of the training process, we did not use a validation set.

Since the labels for the dataset were created directly on the radiograph as polygons, there are three possible ways to train with them:

Polygon The minimal area spanned by the polygons defines the minimal optimal collimation. This area is the minimum area that must be irradiated to obtain a radiograph of good diagnostic quality.

Rectangle The minimal rectangle spanned by the polygon, which would be the realistically achievable collimation. In conventional X-ray devices, collimation can only be defined as rectangular. The smallest rectangle that encloses the annotated polygon defines the smallest area that must be irradiated and that is feasible in practice.

Table 5.1: Results for the upper ankle and abdomen for both input modalities on the *Clinical Setting Depth Dataset*, respectively. It is visible that high IoU scores can be achieved for both anatomical sites. All values are percentages except FN/FP Ratio.

Metric	upper ankle		abdomen	
	point cloud	depth image	point cloud	depth image
IoU	81.63±0.70	78.76±0.31	83.81±0.61	88.47±0.48
Accuracy	99.39±0.03	99.68±0.01	98.12±0.08	98.85±0.05
Sensitivity	92.37±0.54	91.21±0.41	92.50±0.26	96.11±0.38
Specificity	99.61±0.03	99.79±0.01	98.80±0.08	99.15±0.05
FN/FP Ratio	0.60±0.07	0.57±0.05	0.75±0.05	0.48±0.06

Applied Collimation The collimation applied by the radiographer at the time of the examination and used for the radiograph. This collimation can be extracted from the metadata of the radiograph and is used to compare how much clinical practice differs from optimal and predicted collimation.

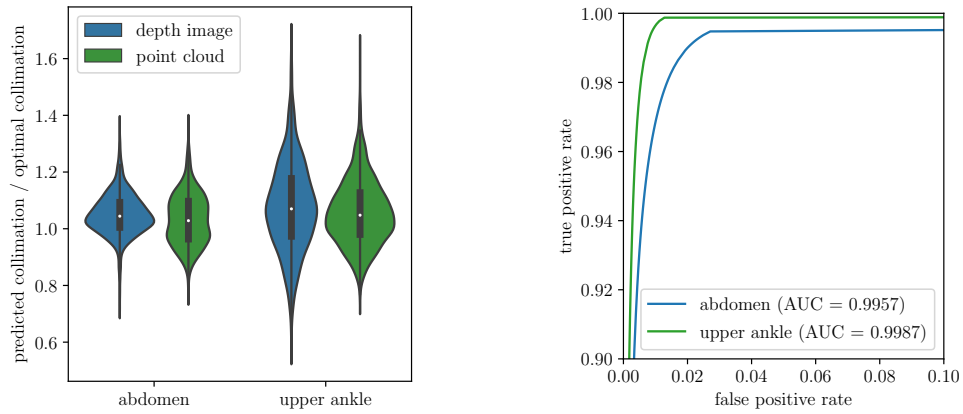
A visual representation of the labels on a radiograph can be seen in Figure 5.5b. All experiments with this dataset were trained using all three labels, and the results are presented in Section 5.1.4.

5.1.4 Results

We modeled the problem as a segmentation task and measured the accuracy of the prediction using the intersection over union (IoU) metric. To enable a more precise evaluation about a too small/large collimation, the specificity and sensitivity are also measured in addition to the pixel-wise accuracy. A sensitivity close to 100% indicates that the area of labeled collimation is almost completely correctly covered. A high specificity, on the other hand, indicates that not much area was exposed unnecessarily. Since it is preferable to choose a collimation slightly larger than absolutely necessary than to have to repeat the acquisition due to a collimation being too small, a higher sensitivity is more important than a higher specificity.

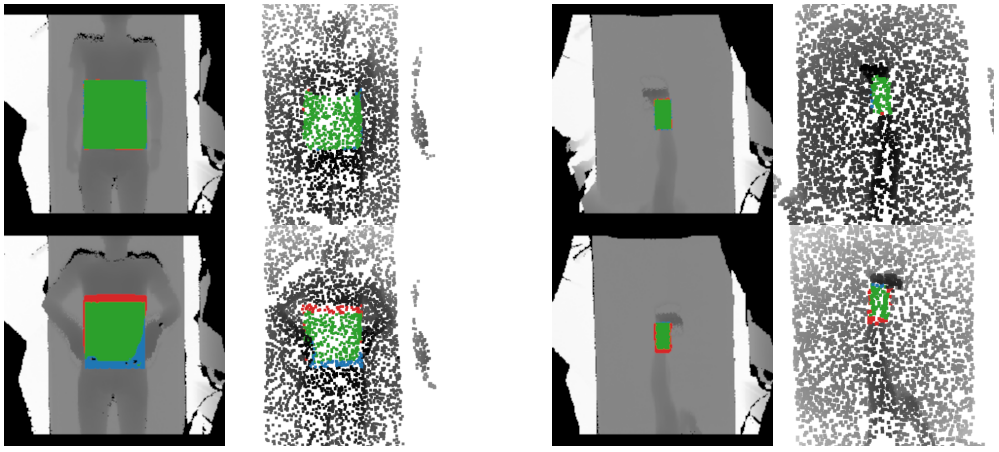
Results for the Clinical Setting Depth Dataset

Table 5.1 shows that for both point clouds and depth images, high IoU scores are achieved for both anatomical sites. As can be seen in Figure 5.6a, the collimation prediction is on average only 4.09% too large for the abdomen and 6.88% for the upper ankle regarding the optimal collimation. While not directly comparable due to the different anatomical sites, this is significantly less than the 26% reported in Karami et al. [2017]. Furthermore, the results show that the sensitivity is always slightly worse than



(a) **(b)**

Figure 5.6: The distribution of the ratios of the predicted collimation to the optimal collimation based on the *Clinical Setting Depth Dataset* is shown in panel (a). The majority of the results are slightly above the ideal value of 1 with the mean value of 104.97% and 107.69% for the abdomen and upper ankle on depth images, as well as 103.22% and 106.07% on point clouds. This slightly larger collimation prediction is more suitable for practical use than a collimation that is too small. The cropped ROC curve for depth images is shown in panel (b).



(a) **(b)**

Figure 5.7: These figures show examples of collimation predictions on point clouds and depth images from the *Clinical Setting Depth Dataset* for abdomen in panel (a) and the upper ankle in panel (b). The top row shows examples of predictions with high IoU scores, and the bottom row shows examples with low IoU scores. True positive pixels and points are colored in green, false positives in red, and false negatives in blue. On both input modalities the prediction errors are similar. In panel (a), it can be seen that the hands, which are held in front of the abdomen, interfere with the prediction.

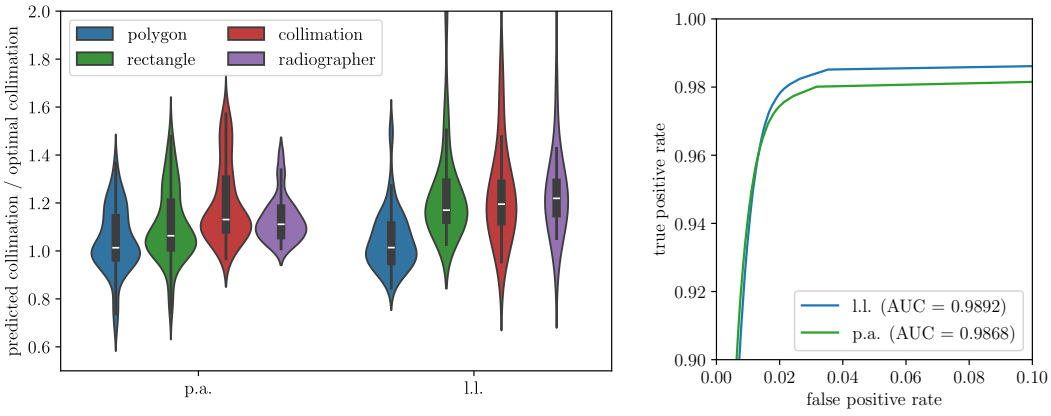
Table 5.2: The results compare the applied collimation, which is done by the radiographers with the polygon and rectangle annotation for the two views of the thorax examinations of the *Clinical Practice Depth Radiograph Dataset*. It can be seen that the polygon annotation differs from the applied collimation and that the rectangle annotation is closer to the applied collimation than the polygon annotation. All values are percentages except FN/FP Ratio.

Metric	thorax *PA*		thorax *LL*	
	polygon	rectangle	polygon	rectangle
IoU	84.43	87.55	77.88	85.96
Accuracy	98.61	98.89	98.09	98.76
Sensitivity	97.31	96.94	98.07	94.06
Specificity	98.69	99.04	98.01	99.12
FN/FP Ratio	0.22	0.43	0.12	1.76

the specificity. This can be explained by the fact that we have significantly more true negatives than true positives, as our *collimation* area is smaller than our *no collimation* area. The results also show that the absolute number of false negatives is already significantly below the false positives (see FN/FP Ratio in Table 5.1). Furthermore, it is possible to adjust the two values using threshold adjustment. In Figure 5.6b, the relationship between specificity and sensitivity regarding different thresholds can be seen for depth images. The mean AUC value of 99.72% shows that a correct classification can be achieved with a high probability. The difference in the accuracy of the collimation prediction between the two anatomical sites is explained by the difference in collimation size. In Figure 5.7, prediction examples on point clouds and depth images from the *Clinical Setting Depth Dataset* of the same pose are shown.

Results for the Clinical Practice Depth Radiograph Dataset

When training with the data from the *Clinical Practice Depth Radiograph Dataset*, it first helps to compare the three types of used labels. Table 5.2 compares the collimation applied by the radiographers with the optimal collimation labeled by the radiologists as polygon annotation and the rectangle of this annotation. The deviation of up to 22 percentage points of the IoU (thorax *LL* with the polygon annotation) illustrates the gap between the applied collimation and the optimal collimation in clinical practice. However, it is practically impossible to always achieve optimal collimation since, in practice, only rectangular collimations can be applied. A comparison with the rectangle annotation shows significantly smaller deviations in the collimations. If Figure 5.8a is considered, it becomes clear that the applied collimation is often larger than the optimal collimation, which leads to unnecessary radiation dose for the patient. This confirms the results of the study by Karami et al. [2017]. Nevertheless, the probability



(a) **(b)**

Figure 5.8: The distribution of the ratios of the predicted collimations to the optimal collimations based on the training with the different labels of the *Clinical Practice Depth Radiograph Dataset* is shown in panel (a). The purple distribution of the radiographers represents the ratio of the actually applied collimation of the radiographers to the optimal collimation, without training at all. The majority of the results are slightly above the ideal value of 1. Note that the radiographers usually apply a collimation that is too large and that the ratio is best when training with the optimal annotation. The cropped ROC curve for depth images and different discrimination thresholds is shown in panel (b).

of a radiograph retake is reduced if a safety buffer is included in the collimation, which reduces the severity of a too large collimation.

The results after training with the three different label types and testing on the same label types are shown in Table 5.3. The high accuracy of over 98% and an IoU of over 81% for all experiments clearly show that it is possible to learn differently sized relevant areas. This supports the results from the training with the *Clinical Setting Depth Dataset*. There is only a minor difference in IoU between the two views, with *PA* being slightly better. One reason for this is that the labels in the *PA* view are mostly square, while the *LL* labels often contain irregular shapes. Furthermore, there is a substantial difference in the FN/FP ratios between the two views. The ROC curve in Figure 5.8b also demonstrates that a high AUC value with an average of 0.988 can also be achieved for the thorax by using the polygon annotation as label.

Since the applied collimation of the radiographers is already close to the rectangular annotation, i.e. the practically optimal possible collimation, the question arises whether the polygon annotation of a radiologist is really necessary and whether one could not just learn with the applied collimation. This would reduce the effort required for labeling. To determine whether this approach would be sufficient, we evaluated the models trained on the three different label types only on the rectangle annotation to

Table 5.3: The results for the two views of the thorax examinations of the *Clinical Practice Depth Radiograph Dataset*. Each of the three annotation types (polygon, rectangle, and applied collimation (app. coll.)) was used for training. The same annotation type was used for training and testing. Note that the various labels can be predicted with high performance.

Metric	thorax *PA*			thorax *LL*		
	polygon	rectangle	app. coll.	polygon	rectangle	app. coll.
IoU	83.48±2.84	84.38±2.13	85.07±3.28	82.79±2.46	81.45±2.66	81.95±2.58
Acc.	98.80±0.07	98.82±0.08	98.73±0.15	98.52±0.26	98.15±0.20	98.08±0.18
Sens.	92.46±4.00	93.38±3.40	93.83±3.51	91.49±1.69	90.28±1.55	90.53±1.70
Spec.	99.19±0.24	99.18±0.26	99.11±0.23	99.17±0.17	98.94±0.15	98.90±0.23
FN/FP	0.72±0.50	0.67±0.45	0.66±0.45	1.00±0.25	0.94±0.16	0.99±0.31

Table 5.4: The results for the two views of the thorax examinations of the *Clinical Practice Depth Radiograph Dataset*. Each type of labeling (polygon, rectangle, and applied collimation (app. coll.)) used for the training was tested on the rectangle annotation. Note that using the rectangle annotations for training yields the best results. All values are percentages except FN/FP Ratio.

Metric	thorax *PA*			thorax *LL*		
	polygon	rectangle	app. coll.	polygon	rectangle	app. coll.
IoU	82.63±2.70	84.38±2.13	81.45±3.70	76.42±2.60	81.45±2.66	79.64±1.32
Acc.	98.70±0.04	98.82±0.08	98.46±0.23	97.66±0.38	98.15±0.20	97.83±0.49
Sens.	89.94±3.98	93.38±3.40	95.18±3.30	80.38±2.52	90.28±1.55	90.83±1.44
Spec.	99.31±0.23	99.18±0.26	98.65±0.26	99.50±0.09	98.94±0.15	98.57±0.32
FN/FP	1.26±0.77	0.67±0.45	0.26±0.19	4.21±0.44	0.94±0.16	0.69±0.09

Table 5.5: The results for the two views of the thorax examinations of the *Clinical Practice Depth Radiograph Dataset*. Each type of labeling (polygon, rectangle, and applied collimation (app. coll.)) used for the training was tested on the polygon annotation. It is visible that using the polygon annotations for training yields the best results. All values are percentages except FN/FP Ratio.

Metric	thorax *PA*			thorax *LL*		
	polygon	rectangle	app. coll.	polygon	rectangle	app. coll.
IoU	83.48±2.84	83.53±2.05	78.98±3.91	82.79±2.46	77.27±3.74	75.20±3.57
Acc.	98.80±0.07	98.76±0.12	98.25±0.25	98.52±0.26	97.85±0.25	97.50±0.38
Sens.	92.46±4.00	94.99±3.20	95.84±3.38	91.49±1.69	96.02±0.73	96.13±1.00
Spec.	99.19±0.24	98.96±0.28	98.36±0.29	99.17±0.17	97.98±0.29	97.60±0.35
FN/FP	0.72±0.50	0.36±0.28	0.17±0.15	1.00±0.25	0.18±0.04	0.15±0.03

(a) (b)

Figure 5.9: These figures show examples of collimation predictions and depth images from the *Clinical Practice Depth Radiograph Dataset* for thorax in *PA* view in panel (a) and thorax in *LL* view in panel (b). The top row shows examples of predictions with high IoU scores and the bottom row shows examples with low IoU scores. True positive pixels and points are colored in green, false positives in red, and false negatives in blue. On both input views the prediction errors are similar. It is noteworthy that the images with a low IoU in the bottom row all have a high number of invalid points in the area of the collimator. These occur when the collimator light, which indicates the area to be irradiated, is turned on during the examination and hinders the stereo depth sensor in determining a depth due to the brightly illuminated homogeneous surface. Since these invalid pixels mainly occur in the center of the collimation and are not taken into account in the evaluation, the class ratio of the prediction of the pixels changes. Nevertheless, it is apparent that in most cases the edges of the collimation also match to a large extent.

assess the learned collimation regarding the realistically achievable optimal collimation. Table 5.4 shows that results worsen when learning directly on the collimation used. This underlines the importance of labeling the optimal collimation on the radiograph rather than on the depth image. We also performed the same experiment by evaluating on the polygon annotation instead of the rectangle annotation. These results are shown in Table 5.5, and they confirm the necessity of labeling an optimal collimation.

In Figure 5.9, prediction examples on depth images from the *Clinical Practice Depth Radiograph Dataset* are shown. The results of Table 5.3 and Table 5.4 together show that it is possible to train networks that can predict the respective label type accurately, but not each label type represents the optimal collimation equally well.

5.1.5 Discussion and Outlook

In this work, we present an AI-based approach to predict the optimal collimation during the X-ray process based on depth information. We have acquired two novel datasets in a clinical setting and in clinical practice, which contain depth images of the abdomen, upper ankle, and thorax examinations. For the second dataset, the corresponding radiographs were also available so that these could be used to create the optimal collimation labels. We have evaluated state-of-the-art neural networks for depth images and point clouds to predict the optimal collimation as segmentation.

Our results show that we are able to accurately predict collimation in various shapes for different anatomical sites and views. We found that it is also beneficial to annotate the optimal label on the radiograph. In our clinical study, we were able to show that radiographers do not achieve optimal collimation but often come close to the optimal collimation that is practically possible.

In terms of optimal collimation, our approach achieves performance on par with radiographers. Given the limited data available, which is a consequence of the extensive clinical study process, these results seem remarkable. While the generalizability of the results is limited by the small amount of data, there are several indications that the approach can be applied to different anatomical sites and also works with limited data. Firstly, the model shows similarly good results across the four different anatomical sites and view combinations, even when the optimal collimation changes in position, size, and shape, when different camera parameters are used, and the surrounding background changes. This shows that the approach is robust to changes in these parameters. Secondly, the results of the *Clinical Practice Depth Radiograph Dataset* show that the approach can be applied to real-world data. This dataset includes individuals of different ages, heights, and weights with different clothing, impairments, and medical conditions. The fact that the results are not worse than those of the *Clinical Setting Depth Dataset*, which only includes four people with similar body shapes, suggests that generalization across these variables is possible. Finally, the restriction that the same individual may not appear in both the training and test sets ensures that the model has

not learned any person-specific features, but rather features that are also meaningful for unseen individuals.

Despite all these indications of generalizability, there are some limitations. Even though various individuals were included under real-life conditions, the small size of the dataset does not cover all the variability, e.g., underweight or overweight people, whose body shape may differ significantly from that of the people in the dataset. Another limiting factor is that all images were taken in a single hospital, using a single X-ray machine in one room. Even though the background for images of different anatomical sites varies, it is usually very similar for the same anatomical site. Standardized nationwide rules for positioning patients and X-ray equipment and adjusting collimation reduce differences between hospitals, but differences will remain in clinical practice. To include these in the dataset and be able to generalize across different hospitals, multi-center data collection is mandatory. It may not be possible to generalize across multiple countries using a single trained model, as there are different standards and regulations for collimation settings.

Even though the radiographers in our study came close to the practically possible optimal collimation, the patient is often still exposed to an unnecessarily high radiation dose due to oversized collimation. This is because conventional X-ray imaging currently only allows rectangular collimation application. But even the use of a multileaf collimator, as used in radiotherapy, would only provide limited assistance, as manual adjustment by the radiographer would be extremely time-consuming. An AI assistant trained with the optimal collimation that can apply this polygonal collimation automatically would remove this restriction, reduce radiation exposure to a minimum, and further improve the workflow in radiography.

Chapter 6

Summary and Discussion

This chapter provides a comprehensive summary and discussion of the work presented in this thesis. Section 6.1 summarizes the main findings and key results of each chapter. Subsequently, Section 6.2 addresses the research objectives outlined in the introduction and discusses how the findings relate to these objectives. Finally, Section 6.3 discusses the limitations of this work and identifies directions for future research.

6.1 Summary

Chapter 3 introduces diagnostic quality as a quantifiable metric for radiographs. The first dataset linking ankle radiographs and diagnostic quality is created. To automatically assess the quality, a new framework is developed that evaluates radiograph quality on par with radiologists. The framework first classifies radiographs according to their view and then extracts a region of interest from each radiograph. This ensures highly consistent image content, thereby facilitating subsequent quality assessment. The framework correctly predicts the quality in 94.1% of cases, while radiologists achieve an estimated correctness of only 91.4%, demonstrating the high accuracy achievable for this task. Had the framework been in place when the radiographs of this dataset were acquired, 80% of non-diagnostic images would have been immediately identified as such. This means that for every 100 radiographs, the number of additional appointments required for re-examination could have been reduced from 13 to only 3.

Following the success with ankle radiographs, the framework is extended and refined to work with knee and wrist radiographs, and the dataset is expanded to include a total of 3,500 labeled radiographs across three body parts and two views, comprising 12,000 individual labels. The results demonstrate that the framework generalizes well across different body parts. A mean intraclass correlation coefficient (ICC) of 90.5% between the prediction models and the radiologists, with no single value lower than 76%, indicates excellent agreement. The agreement between the models and the radiologists is even better than the agreement among radiologists themselves, who achieve only a mean ICC of 85.4%. A further analysis of the required amount of annotated data reveals that a labeling effort of two hours from a single radiologist is sufficient to achieve 90% accuracy for two views of a single body part.

Furthermore, two clinical use cases are evaluated. A retrospective analysis of 10 years of radiograph data reveals that diagnostic quality is influenced by multiple factors. First, age, weight, and sex all correlate with quality. While these parameters cannot be adjusted, they can be taken into account during the examination. Second, radiographs acquired during normal working hours from 08:00 to 16:00 are slightly worse than those taken outside of working hours. Finally, the station where the radiograph is acquired also influences the quality. Whether these last two factors are confounding variables or reflect systematic underlying causes remains an open question.

The final part of this chapter presents a feedback system evaluated live in clinical routine. After running for 1.5 years, no significant change could be detected that could be attributed to the feedback system. While the quality improves slightly, this is also observed for wrist radiographs, which were not included in the feedback system. A questionnaire revealed that the radiographers do not agree with the displayed feedback and did not allow it to influence their positioning in any way. As Decoster et al. [2025] demonstrated that quality perception between radiographers and radiologists not only differs but does not correlate at all, this underscores the importance of such feedback and might explain why the radiographers do not show any reaction to the feedback. Despite not agreeing with the feedback, the feedback system was received positively overall.

Chapter 4 extends the quality assessment from radiographs after acquisition to depth images before acquisition. Building on the knowledge gained that diagnostic quality can be precisely assessed on radiographs, the next step involves predicting quality from depth images of the patient's pose. For this, two anatomical preparations were X-rayed in different positions while being captured by depth cameras to collect a new dataset linking depth images to diagnostic quality. Using this dataset, it is demonstrated that even with limited data for the ankle, the quality of the resulting radiographs can be predicted with high accuracy. As the training data never contained images of both feet, it could be shown to some extent that the learned features generalize to different individuals.

As data acquisition using anatomical preparations does not scale, synthetic training data was generated. For this purpose, a framework to generate depth images and corresponding synthetic radiographs from CT scans was developed. From the CT scans containing the lower extremities, a 3D representation of the lower leg was extracted, which was then placed in a prepared scene resembling the X-ray room. From this setup, new depth images can be generated from any chosen position. Additionally, digitally reconstructed radiographs (DRRs) can be created from the CT scans. If the virtual X-ray position is chosen to be identically aligned with the camera's position, as in a real setting, corresponding radiographs and depth images can be generated. Using 19 CT scans, 3,077 DRRs with corresponding depth images were generated, labeled, and used to train the model to predict quality based on the depth image. Pretraining on

these synthetic data showed substantial improvement when applied to the real data captured from the preparations. Correct prediction of the quality was achieved in 90.5% of all cases. Considering that this performance is similarly high as a prediction on the radiographs themselves, this result is remarkable.

Chapter 5 applies depth cameras to address incorrectly set collimation, which is another factor that degrades the quality of radiographs. Two new datasets are collected: one resembling clinical practice without X-raying and one clinical study gathering real clinical data. For the first dataset, the collimation was labeled directly on depth images taken in abdomen and ankle X-ray examination positions from four subjects. The depth images contain approximately the same information that radiographers have available during positioning. For this dataset, a segmentation of the area to be irradiated was already possible with an intersection over union (IoU) of 82%. The second dataset comprises 59 radiographs from thorax examinations split into two views, together with depth images captured simultaneously. As the radiographs corresponding to the depth images were available for the second dataset, the collimation could be labeled based on what radiologists actually need to see for diagnosis. Through a known alignment between the X-ray source and depth camera, the collimation labeled on the radiograph could be transferred to the depth images and used for training. The results demonstrate that even with this limited data, a prediction with an IoU of 83% can be achieved. Again, as no subjects were shared between training and test sets, this demonstrates generalization across different patients.

6.2 Research Findings

This section revisits the objectives presented in Section 1.2. For each objective, it evaluates how the research contributes to its achievement and summarizes the corresponding findings.

Diagnostic Quality as a Metric This work introduced diagnostic quality as a quantifiable metric for radiograph assessment based on visible anatomical structures rather than technical parameters. As demonstrated in Chapter 3, diagnostic quality can serve as an automatically assessed metric to quantify the usability of radiographs. This automated quality assessment achieved a high mean accuracy of 92.24% across ankle, knee, and wrist radiographs. A mean ICC of 90.5% between the predictions and the radiologists indicates excellent agreement, further demonstrating that automated quality assessment can serve as a reliable metric. Notably, the agreement between the predictions and the radiologists exceeds the ICC among radiologists themselves, which is only 86.17%. These findings demonstrate that automated prediction can be reliably used to provide feedback or analyze quality retrospectively. The proposed framework generalizes well across body parts, requiring only approximately two hours of labeling

effort to integrate an additional body part. This work provided the first research on automated diagnostic quality assessment and inspired subsequent research groups to pursue similar approaches. For example, Krönke et al. [2022] estimated the pose of an ankle from a radiograph, measuring rotation and flexion. Lysdahlgaard et al. [2023] pursued a similar goal of predicting diagnostic quality, though they labeled radiographs using radiographers rather than radiologists. Köpnick et al. [2023] employed a similar pipeline, confirming the finding that rotation alone is insufficient for quality prediction. Additionally, Gabryś et al. [2025] employed a similar metric, and Ameli et al. [2025] adapted the framework for dental panoramic radiographs. These works demonstrate that quality prediction based on diagnostic utility is gaining momentum in the field.

Beyond demonstrating high prediction accuracy, this work showed that quality assessment can be practically deployed in a feedback system. The feedback system was accepted by radiographers, who expressed interest in extending it to additional body parts. This indicates that automated diagnostic quality assessment can be used beyond academic research and could have a real impact in clinical practice. A retrospective analysis revealed further insights into factors correlating with diagnostic quality. Patient characteristics such as age, weight, and sex were found to correlate with the quality of acquired radiographs, as did the time of day and the station at which the radiograph was acquired. While these factors are difficult to interpret without the context of a specific hospital, the findings demonstrate that influencing factors exist that could be analyzed for quality assurance or improvement.

The research presented in Chapter 4 further showed that diagnostic quality can be measured not only on radiographs but also accurately predicted from depth images before radiographs are acquired. This extends the usefulness of the diagnostic quality metric from reacting to suboptimal radiographs to actively preventing them. The findings show that prediction with an accuracy of 90.45% is possible, which is comparable to quality assessment on the radiographs themselves. While this research is at an early stage, it already works reliably for ankle radiographs.

Overall, the findings demonstrate that diagnostic quality can be automatically and accurately assessed on radiographs, matching the performance of radiologists. Furthermore, this automated assessment can be practically implemented as a feedback system that is valued by radiographers. The measured labeling effort reveals that the number of integrated body parts can be easily scaled for further research or deployment. Finally, diagnostic quality can be predicted from depth images, enabling quality assurance earlier in the examination process and thereby further protecting patients from unnecessary radiation exposure.

Depth Imaging in Clinical Routine This work systematically investigated the integration of depth cameras into the radiograph acquisition workflow, demonstrating multiple use cases to support radiographers and improve image quality. Two applications

were developed and evaluated. First, in Chapter 4, Time-of-Flight (ToF) depth cameras were used to predict the quality of radiographs to be acquired, achieving a prediction accuracy of 90.45%. Second, in Chapter 5, stereo depth cameras were used to predict optimal collimation areas. Even with limited data, this prediction achieved an IoU of 82.92%, approaching the radiographers' IoU of 86.75% for the best practically achievable collimation. As both use cases employed different camera technologies with varying depth accuracies, successful integration does not depend on specific depth camera characteristics.

Regarding camera positioning, mounting cameras directly on the X-ray source or collimator proved advantageous. This configuration enables easy alignment with the X-ray beam and ensures consistent positioning relative to the patient, regardless of where the examination takes place in the room. While multiple cameras can be used to avoid occlusions, using two cameras instead of one provides only limited benefit. A single camera achieved an accuracy of 86.63%, while two cameras improved this to 89.53%. Nevertheless, camera positioning proved to be critical, as one viewpoint yielded an accuracy of 86.63% while the other achieved only 75.06%. This indicates that multiple camera viewpoints contain complementary information for pose quality assessment.

Overall, the developed systems provide immediate feedback during acquisition while the patient is present, preventing delayed repeat imaging and supporting radiographers with expert-level guidance without disrupting established workflows. The results from both experiments demonstrate that depth cameras can be effectively integrated to solve practical tasks in clinical routine.

However, regulatory hurdles, not technical limitations, represent the main barrier to clinical deployment. Technical aspects such as depth accuracy, labeling effort, and network training were less constraining than the ability to deploy such systems in clinical routine. This limitation constrains academic research and results in X-ray equipment manufacturers having exclusive access to large datasets through their built-in cameras. Privacy concerns and ethical constraints further limited data collection from clinical routine, necessitating alternative data acquisition strategies such as the synthetic data generation framework developed in this work.

Limited Data While deep learning tasks typically rely on large datasets, such data are not always available. In the work presented in this thesis, only small datasets could be acquired for training. For quality assessment on radiographs, 3,500 images were labeled in total; however, datasets for a single view of a single body part contained only approximately 400 radiographs. Even with this limited data, precise and reliable predictions with a mean accuracy of 92.24% were achieved. Image standardization played a key role in this success. Splitting the quality assessment problem into smaller, view-specific problems allowed the model to focus on quality-relevant features, improving

accuracy by 1.7 percentage points. Although dataset sizes were halved by splitting into *AP* and *LAT* views, the image content became more homogeneous, as radiographs of the same body part and view are highly similar. Furthermore, restricting training to the relevant part of the image prevented the model from using irrelevant features, further improving accuracy by 9.0 percentage points. These design choices exploited the inherent structure and consistency of radiographic imaging, demonstrating that medical deep learning tasks need not require massive datasets when properly constrained. By systematically reducing the training set size, we found that as few as 150 labeled radiographs suffice to achieve 90% accuracy.

A similar approach applied to depth images demonstrated that quality-relevant features could be learned from limited data, even using only a single anatomical preparation. To overcome regulatory and ethical barriers in depth image collection, a novel framework was developed to synthetically generate paired depth images and radiographs from CT scans. This approach enables retrospective use of existing CT data and intentional generation of non-diagnostic poses that would be impossible to ethically acquire from real patients. Pretraining on synthetic data improved pose assessment accuracy by up to 11 percentage points, validating the transfer learning approach and providing a scalable solution to data acquisition challenges.

The capability of extremely small datasets to solve medical tasks was further demonstrated in Chapter 5. Training two deep networks using only 59 radiographs (29 in *LL* view and 30 in *PA* view) achieved collimation prediction accuracy comparable to that of radiographers.

Despite these small datasets, all frameworks demonstrated robust generalization: radiograph quality assessment across three body parts and multiple views; pose quality across different anatomical preparations and individuals; and collimation optimization across four anatomical sites and varied patient characteristics. These results demonstrate that deep learning can be effectively applied to medical tasks even with minimal data, overcoming data acquisition challenges through careful problem formulation and use of retrospective available data.

6.3 Limitations

While promising results were demonstrated for each problem, each also comes with limitations. The quality assessment on radiographs works reliably for different body parts, but all these radiographs were collected from a single hospital and labeled by a small group of radiologists. A multi-center study could further improve the demonstrated generalizability. Furthermore, the research is limited to three body parts. While particularly hard-to-position body parts with high reject rates were chosen, it is still possible that quality assessment is harder to learn for body parts not considered in this study. Finally, the designed framework requires two neural networks to be

trained for each view, plus a view recognition network for each body part. While this approach offers modularity and allows small networks to be retrained for specific views, integrating an increasing number of body parts can quickly result in a high number of trained networks and complicate inference.

Since radiographs for quality assessment could be collected retrospectively, raw data availability was virtually unlimited, leaving the manual labeling effort as the primary constraint. This was in strong contrast to the quality prediction on the depth images and the collimation prediction. A strong limitation on both is that only limited data could be acquired because of regulatory and privacy concerns. While prediction still worked well on the unseen data, an evaluation with larger datasets is absolutely necessary to demonstrate real generalization. Larger clinical studies with real patient data are needed to show that it works also in practice.

For both applications, we mounted the 3D cameras directly on the X-ray machine. While this is a reasonable position because the area to be X-rayed stays in the field of view of the camera, no systematic research was carried out regarding positioning or orientation, leaving this open for future work. Similarly, for both applications, we used only depth images, while color images could also be captured. Since data protection concerns are less relevant with pure depth images, no color images were used in order to facilitate data collection. However, from a research point of view, the question of whether color information helps or is even sufficient remains open.

References

[Abozeed et al., 2024] Abozeed, M., Junck, K., Lirette, S., Kimpe, T., Xthona, A., Tridandapani, S., and Perchik, J. "Interpretation Time Efficiency with Radiographs: A Comparison Study between Standard 6 and 12 MP High-Resolution Display Monitors". *Journal of Medical Imaging* 11 (3), 2024, p. 035502.

[Agunwa et al., 2019] Agunwa, C., Moradi, M., Wong, K. C. L., and Syeda-Mahmood, T. "Body Part and Imaging Modality Classification for a General Radiology Cognitive Assistant". In: *Medical Imaging 2019: Image Processing*. Vol. 10949. International Society for Optics and Photonics, 2019, p. 1094910.

[Ameli et al., 2025] Ameli, N., Miri Moghaddam, M., Lai, H., and Pacheco-Pereira, C. "Automated Quality Evaluation of Dental Panoramic Radiographs Using Deep Learning". *Imaging Science in Dentistry* 55 (2), 2025, pp. 175–188.

[Atkinson et al., 2020] Atkinson, S., Neep, M., and Starkey, D. "Reject Rate Analysis in Digital Radiography: An Australian Emergency Imaging Department Case Study". *Journal of Medical Radiation Sciences* 67 (1), 2020, pp. 72–79.

[Badal et al., 2009] Badal, A. and Badano, A. "Accelerating Monte Carlo Simulations of Photon Transport in a Voxelized Geometry Using a Massively Parallel Graphics Processing Unit". *Medical Physics* 36 (11), 2009, pp. 4878–4880.

[Badal et al., 2011] Badal, A. and Badano., A. "Chapter 50 - Fast Simulation of Radiographic Images Using a Monte Carlo x-Ray Transport Algorithm Implemented in CUDA". In: *GPU Computing Gems Emerald Edition*. Applications of GPU Computing Series. Morgan Kaufmann, 2011, pp. 813–829.

[Barile et al., 2017] Barile, A., Bruno, F., Arrigoni, F., Splendiani, A., Cesare, E. D., Zappia, M., Guglielmi, G., and Masciocchi, C. "Emergency and Trauma of the Ankle". *Seminars in Musculoskeletal Radiology* 21 (03), 2017, pp. 282–289.

[Becht et al., 2019] . *Lehrbuch der radiologischen Einstelltechnik*. Springer, 2019.

[Berg et al., 2020] Berg, J., Krönke, S., Gooßen, A., Bystrov, D., Brück, M., Harder, T., Wieberneit, N., and Young, S. "Robust Chest X-Ray Quality Assessment Using Convolutional Neural Networks and Atlas Regularization". In: *Medical Imaging 2020: Image Processing*. Vol. 11313. SPIE, 2020, pp. 391–398.

[Bigalke et al., 2021] Bigalke, A., Hansen, L., Diesel, J., and Heinrich, M. P. "Seeing under the Cover with a 3D U-Net: Point Cloud-Based Weight Estimation of

Covered Patients". *International Journal of Computer Assisted Radiology and Surgery* 16 (12), 2021, pp. 2079–2087.

[Breitwieser et al., 2025] Breitwieser, M., Wiesner, T., Moore, V., Wichlas, F., and Deininger, C. "Cost-Effectiveness of Routine X-Rays After Central Venous Catheter Removal: A Value-Based Analysis of Post-Removal Complications". *Journal of Clinical Medicine* 14 (4), 2025, p. 1397.

[Chen et al., 2017] Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H. "Rethinking Atrous Convolution for Semantic Image Segmentation". *arXiv:1706.05587 [cs]*, 2017. arXiv: `1706.05587 [cs]`.

[Chen et al., 2022] Chen, X., Deng, Q., Wang, Q., Liu, X., Chen, L., Liu, J., Li, S., Wang, M., and Cao, G. "Image Quality Control in Lumbar Spine Radiography Using Enhanced U-Net Neural Networks". *Frontiers in Public Health* 10, 2022, p. 891766.

[Choy et al., 2018] Choy, G., Khalilzadeh, O., Michalski, M., Do, S., Samir, A. E., Pianykh, O. S., Geis, J. R., Pandharipande, P. V., Brink, J. A., and Dreyer, K. J. "Current Applications and Future Impact of Machine Learning in Radiology". *Radiology* 288 (2), 2018, pp. 318–328.

[Decoster et al., 2023] Decoster, R., Toomey, R., Smits, D., Haygood, T. M., and Ryan, M.-L. "Understanding Reasons for Image Rejection by Radiologists and Radiographers". *Journal of Medical Radiation Sciences* 70 (2), 2023, pp. 127–136.

[Decoster et al., 2025] Decoster, R., Ryan, M.-L., and Toomey, R. "The Difference in Image Quality Assessment between Radiographers and Radiologists and Its Relationship with Diagnostic Accuracy". *Radiography (London, England: 1995)* 31 (1), 2025, pp. 89–96.

[Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. "ImageNet: A Large-Scale Hierarchical Image Database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255.

[Deutschland, 1988] Deutschland, *Sozialgesetzbuch Fünftes Buch – Gesetzliche Krankenversicherung (SGB V), § 135b Förderung Der Qualität Durch Die Kassenärztlichen Vereinigungen*. 1988.

[Elgaard et al., 2022] Elgaard, P. A., Simon, L., and Kusk, M. W. "Collimation Border with U-net Segmentation on Chest Radiographs Compared to Radiologists". *Journal of Medical Imaging and Radiation Sciences*. ISRRT Conference Proceedings 53 (4, Supplement 1), 2022, S43–S44.

[Esses et al., 2018] Esses, S. J., Lu, X., Zhao, T., Shanbhogue, K., Dane, B., Bruno, M., and Chandarana, H. "Automated Image Quality Evaluation of T2-weighted Liver MRI Utilizing Deep Learning Architecture". *Journal of Magnetic Resonance Imaging* 47 (3), 2018, pp. 723–728.

[Ester et al., 1996] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. KDD’96. AAAI Press, 1996, pp. 226–231.

[Fang et al., 2020] Fang, X., Harris, L., Zhou, W., and Huo, D. “Generalized Radiographic View Identification with Deep Learning”. *Journal of Digital Imaging*, 2020.

[Farzanegan et al., 2020] Farzanegan, Z., Tahmasbi, M., Cheki, M., Yousefvand, F., and Rajabi, M. “Evaluating the Principles of Radiation Protection in Diagnostic Radiologic Examinations: Collimation, Exposure Factors and Use of Protective Equipment for the Patients and Their Companions”. *Journal of Medical Radiation Sciences* 67 (2), 2020, pp. 119–127.

[Fillingim et al., 2002] Fillingim, R. B., Browning, A. D., Powell, T., and Wright, R. A. “Sex Differences in Perceptual and Cardiovascular Responses to Pain: The Influence of a Perceived Ability Manipulation”. *The Journal of Pain* 3 (6), 2002, pp. 439–445.

[Foos et al., 2009] Foos, D. H., Sehnert, W. J., Reiner, B., Siegel, E. L., Segal, A., and Waldman, D. L. “Digital Radiography Reject Analysis: Data Collection Methodology, Results, and Recommendations from an In-depth Investigation at Two Hospitals”. *Journal of Digital Imaging* 22 (1), 2009, pp. 89–98.

[Gabryś et al., 2025] Gabryś, P. D., Łapińska, N., Mendyk, A., and Tatoń, G. “Assessment of X-ray Ankle Joint Image Projection Correctness with the Use of Machine Learning Algorithms”. *Polish Journal of Radiology* 90, 2025, pp. 451–457.

[Gemeinsamer Bundesausschuss, 2023] Gemeinsamer Bundesausschuss, *Richtlinie Des Gemeinsamen Bundesausschusses Über Kriterien Zur Qualitätsbeurteilung in Der Radiologischen Diagnostik Nach § 135b Absatz 2 SGB V (Qualitätsbeurteilungs-Richtlinie Radiologie/QBR-RL)*. Richtlinie BAnz AT 24.07.2023 B2. Gemeinsamer Bundesausschuss, 2023.

[Gerdes et al., 2026] Gerdes, H., Mairhöfer, D., Laufer, M., Reis, F. L., Bischof, A., Wegner, F., Käster, T., Barth, E., Barkhausen, J., Martinetz, T., and Sieren, M. *Generalizable Deep Learning Framework for Diagnostic Quality Assessment of Musculoskeletal Radiographs*. 2026.

[Hansen et al., 2019] Hansen, L., Siebert, M., Diesel, J., and Heinrich, M. P. “Fusing Information from Multiple 2D Depth Cameras for 3D Human Pose Estimation in the Operating Room”. *International Journal of Computer Assisted Radiology and Surgery* 14 (11), 2019, pp. 1871–1879.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. arXiv: 1512.03385.

[Hinton, 1987] Hinton, G. E. "Learning Translation Invariant Recognition in a Massively Parallel Networks". In: *PARLE Parallel Architectures and Languages Europe*. Springer, 1987, pp. 1–13.

[Hofmann et al., 2015] Hofmann, B., Rosanowsky, T. B., Jensen, C., and Wah, K. H. C. "Image Rejects in General Direct Digital Radiography". *Acta Radiologica Open* 4 (10), 2015, p. 2058460115604339.

[Huang et al., 2016] Huang, G., Sun, Y., Liu, Z., Sedra, D., and Weinberger, K. Q. "Deep Networks with Stochastic Depth". In: *Computer Vision – ECCV 2016*. Springer International Publishing, 2016, pp. 646–661.

[Jeong et al., 2025] Jeong, S., Han, K., Kang, Y., Kim, E.-K., Song, K., Vasanawala, S., and Shin, H. J. "The Impact of Artificial Intelligence on Radiologists' Reading Time in Bone Age Radiograph Assessment: A Preliminary Retrospective Observational Study". *Journal of Imaging Informatics in Medicine* 38 (4), 2025, pp. 1915–1923.

[Kadkhodamohammadi et al., 2017] Kadkhodamohammadi, A., Gangi, A., de Mathelin, M., and Padoy, N. "A Multi-view RGB-D Approach for Human Pose Estimation in Operating Rooms". In: *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2017, pp. 363–372.

[Karami et al., 2016] Karami, V., Zabihzadeh, M., Gilavand, A., and Shams, N. "Survey of the Use of X-ray Beam Collimator and Shielding Tools during Infant Chest Radiography". *International Journal of Pediatrics* 4 (4), 2016, pp. 1637–1642.

[Karami et al., 2017] Karami, V. and Zabihzadeh, M. "Beam Collimation during Lumbar Spine Radiography: A Retrospective Study". *Journal of Biomedical Physics & Engineering* 7 (2), 2017, pp. 101–106.

[Kashyap et al., 2019] Kashyap, S., Moradi, M., Karargyris, A., Wu, J. T., Morris, M., Saboury, B., Siegel, E., and Syeda-Mahmood, T. "Artificial Intelligence for Point of Care Radiograph Quality Assessment". In: *Medical Imaging 2019: Computer-Aided Diagnosis*. Vol. 10950. SPIE, 2019, pp. 893–899.

[Kingma et al., 2017] Kingma, D. P. and Ba, J. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: `1412.6980 [cs]`.

[Kjelle et al., 2024] Kjelle, E., Brandsæter, I. Ø., Andersen, E. R., and Hofmann, B. M. "Cost of Low-Value Imaging Worldwide: A Systematic Review". *Applied Health Economics and Health Policy* 22 (4), 2024, pp. 485–501.

[Koo et al., 2016] Koo, T. K. and Li, M. Y. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research". *Journal of Chiropractic Medicine* 15 (2), 2016, pp. 155–163.

[Köpnick et al., 2023] Köpnick, J., May, J. M., Lundt, B., Brück, M., and Wülker, C. "Estimation of the Ankle-Joint Space Visibility in x-Ray Images Using

Convolutional Neural Networks". In: *Medical Imaging 2023: Image Processing*. Vol. 12464. SPIE, 2023, pp. 61–65.

[Krönke et al., 2022] Krönke, S., Berg, J., Brueck, M., Bystrov, D., Gooßen, A., Harder, T., Lundt, B., May, J. M., Wieberneit, N., Wissel, T., Hertgers, O., Lamb, H. J., and Young, S. "CNN-based Pose Estimation for Assessing Quality of Ankle-Joint X-ray Images". In: *Medical Imaging 2022: Image Processing*. Vol. 12032. SPIE, 2022, pp. 344–352.

[Lalone et al., 2015] Lalone, E. A., Grewal, R., King, G. J. W., and MacDermid, J. C. "A Structured Review Addressing the Use of Radiographic Measures of Alignment and the Definition of Acceptability in Patients with Distal Radius Fractures". *HAND* 10 (4), 2015, pp. 621–638.

[Laufer et al., 2024a] Laufer, M., Mairhöfer, D., Bischof, A., Käster, T., Sieren, M., Reis, F. L., Gerdes, H. W., and Simon, P. "Verfahren zur Erzeugung von Trainingsdaten für ein KI-basiertes Assistenzsystem und Vorrichtung zur Unterstützung der Röntgendiagnostik". DE102022133272A1. 2024.

[Laufer et al., 2024b] Laufer, M., Mairhöfer, D., Sieren, M., Gerdes, H., Reis, F. L., Bischof, A., Käster, T., Barth, E., Barkhausen, J., and Martinetz, T. "Patient Pose Assessment in Radiography Using Time-of-Flight Cameras". In: *Medical Imaging 2024: Image Processing*. Vol. 12926. SPIE, 2024, pp. 385–393.

[Laufer et al., 2025] Laufer, M., Mairhöfer, D., Sieren, M., Gerdes, H., Leal dos Reis, F., Bischof, A., Käster, T., Barth, E., Barkhausen, J., and Martinetz, T. "Synthetic Data Generated from CT Scans for Patient Pose Assessment". In: *Proceedings of the Eighth Conference on Medical Imaging with Deep Learning (MIDL 2025)*. PMLR, 2025.

[Lee et al., 2013] Lee, C. S., Nagy, P. G., Weaver, S. J., and Newman-Toker, D. E. "Cognitive and System Factors Contributing to Diagnostic Errors in Radiology". *American Journal of Roentgenology* 201 (3), 2013, pp. 611–617.

[Lichtman et al., 2011] Lichtman, D. M., Bindra, R. R., Boyer, M. I., Putnam, M. D., Ring, D., Slutsky, D. J., Taras, J. S., Watters, W. C. I., Goldberg, M. J., Keith, M., Turkelson, C. M., Wies, J. L., Haralson, R. H. I., Boyer, K. M., Hitchcock, K., and Raymond, L. "American Academy of Orthopaedic Surgeons Clinical Practice Guideline on: The Treatment of Distal Radius Fractures". *JBJS* 93 (8), 2011, p. 775.

[Lin et al., 2012] Lin, Y., Luo, H., Dobbins III, J. T., Page McAdams, H., Wang, X., Sehnert, W. J., Barski, L., Foos, D. H., and Samei, E. "An Image-Based Technique to Assess the Perceptual Quality of Clinical Chest Radiographs". *Medical Physics* 39 (11), 2012, pp. 7019–7031.

[Lindsey et al., 2018] Lindsey, R., Daluiski, A., Chopra, S., Lachapelle, A., Mozer, M., Sicular, S., Hanel, D., Gardner, M., Gupta, A., Hotchkiss, R., and Potter, H.

"Deep Neural Network Improves Fracture Detection by Clinicians". *Proceedings of the National Academy of Sciences* 115 (45), 2018, pp. 11591–11596.

[Little et al., 2017] Little, K. J., Reiser, I., Liu, L., Kinsey, T., Sánchez, A. A., Haas, K., Mallory, F., Froman, C., and Lu, Z. F. "Unified Database for Rejected Image Analysis Across Multiple Vendors in Radiography". *Journal of the American College of Radiology* 14 (2), 2017, pp. 208–216.

[Liu et al., 2019] Liu, Z., Tang, H., Lin, Y., and Han, S. "Point-Voxel CNN for Efficient 3D Deep Learning". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019.

[Lorensen et al., 1987] Lorensen, W. E. and Cline, H. E. "Marching Cubes: A High Resolution 3D Surface Construction Algorithm". In: *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*. ACM, 1987, pp. 163–169.

[Lysdahlgaard et al., 2023] Lysdahlgaard, S., Baressi Šegota, S., Hess, S., Antulov, R., Weber Kusk, M., and Car, Z. "Quality Assessment Assistance of Lateral Knee X-rays: A Hybrid Convolutional Neural Network Approach". *Mathematics* 11 (10), 2023, p. 2392.

[Mairhöfer et al., 2021] Mairhöfer, D., Laufer, M., Simon, P. M., Sieren, M., Bischof, A., Käster, T., Barth, E., Barkhausen, J., and Martinetz, T. "An AI-based Framework for Diagnostic Quality Assessment of Ankle Radiographs". In: *Proceedings of the Fourth Conference on Medical Imaging with Deep Learning (MIDL 2021)*. PMLR, 2021, pp. 484–496.

[Mairhöfer et al., 2024] Mairhöfer, D., Laufer, M., Berkel, L., Bischof, A., Barth, E., Barkhausen, J., and Martinetz, T. "AI-based Collimation Optimization for X-Ray Imaging Using Time-of-Flight Cameras". In: *ESANN 2024 Proceedings*. Ciaco - i6doc.com, 2024, pp. 703–708.

[Mairhöfer et al., 2026] Mairhöfer, D., Laufer, M., Berkel, L., Sieren, M., Bischof, A., Barth, E., Barkhausen, J., and Martinetz, T. "AI-based Collimation Optimization for X-ray Imaging Using Depth Cameras". *Neurocomputing* 661, 2026, p. 131881.

[Mann et al., 1947] Mann, H. B. and Whitney, D. R. "On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other". *The Annals of Mathematical Statistics* 18 (1), 1947, pp. 50–60.

[Meng et al., 2022] Meng, Y., Ruan, J., Yang, B., Gao, Y., Jin, J., Dong, F., Ji, H., He, L., Cheng, G., and Gong, X. "Automated Quality Assessment of Chest Radiographs Based on Deep Learning and Linear Regression Cascade Algorithms". *European Radiology* 32 (11), 2022, pp. 7680–7690.

[NEMA, 2025] NEMA, *Digital Imaging and Communications in Medicine (DICOM) Standard*. 2025.

[Nguyen et al., 2020] Nguyen, T. P., Chae, D.-S., Park, S.-J., Kang, K.-Y., Lee, W.-S., and Yoon, J. “Intelligent Analysis of Coronal Alignment in Lower Limbs Based on Radiographic Image with Convolutional Neural Network”. *Computers in Biology and Medicine* 120, 2020, p. 103732.

[NHS England, 2024] NHS England, *Diagnostic Imaging Dataset 2024-25 Data: DID Table 13 Modality Provider Report Turnaround*. 2024.

[Paszke et al., 2019] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 721. Curran Associates Inc., 2019, pp. 8026–8037.

[Pinto et al., 2018] Pinto, A., Berritto, D., Russo, A., Riccitiello, F., Caruso, M., Belfiore, M. P., Papapietro, V. R., Carotti, M., Pinto, F., Giovagnoni, A., Romano, L., and Grassi, R. “Traumatic Fractures in Adults: Missed Diagnosis on Plain Radiographs in the Emergency Department”. *Acta Biomedica Atenei Parmensis* 89 (1-S), 2018, pp. 111–123.

[Placht et al., 2012] Placht, S., Stancanello, J., Schaller, C., Balda, M., and Angelopoulou, E. “Fast Time-of-Flight Camera Based Surface Registration for Radiotherapy Patient Positioning”. *Medical Physics* 39 (1), 2012, pp. 4–17.

[Poggenborg et al., 2021] Poggenborg, J., Yaroshenko, A., Wieberneit, N., Harder, T., and Gossmann, A. *Impact of AI-based Real Time Image Quality Feedback for Chest Radiographs in the Clinical Routine*. 2021.

[Qi et al., 2017a] Qi, C. R., Su, H., Mo, K., and Guibas, L. J. “PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 652–660.

[Qi et al., 2017b] Qi, C. R., Yi, L., Su, H., and Guibas, L. J. “PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space”. In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017.

[Rasche et al., 2024] Rasche, A., Brader, P., Borggrefe, J., Seuss, H., Carr, Z., Hebecker, A., and ten Cate, G. “Impact of Intelligent Virtual and AI-based Automated Collimation Functionalities on the Efficiency of Radiographic Acquisitions”. *Radiography* 30 (4), 2024, pp. 1073–1079.

[Ritchie et al., 2025] Ritchie, B., Summerville, L., Sheng, M., Choi, M., Tirumani, S., and Ramaiya, N. “Impact of Turnaround Time in Radiology: The Good, the Bad, and the Ugly”. *Current Problems in Diagnostic Radiology*, 2025.

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer International Publishing, 2015, pp. 234–241.

[Russakoff et al., 2005] Russakoff, D., Rohlfing, T., Mori, K., Rueckert, D., Ho, A., Adler, J., and Maurer, C. "Fast Generation of Digitally Reconstructed Radiographs Using Attenuation Fields with Application to 2D-3D Image Registration". *IEEE Transactions on Medical Imaging* 24 (11), 2005, pp. 1441–1454.

[Saba et al., 2019] Saba, L., Biswas, M., Kuppili, V., Cuadrado Godia, E., Suri, H. S., Edla, D. R., Omerzu, T., Laird, J. R., Khanna, N. N., Mavrogeni, S., Protogerou, A., Sfikakis, P. P., Viswanathan, V., Kitas, G. D., Nicolaides, A., Gupta, A., and Suri, J. S. "The Present and Future of Deep Learning in Radiology". *European Journal of Radiology* 114, 2019, pp. 14–24.

[Saiti et al., 2022] Saiti, E. and Theoharis, T. "Multimodal Registration across 3D Point Clouds and CT-volumes". *Computers & Graphics* 106, 2022, pp. 259–266.

[Salvat et al., 2006] Salvat, F., Fernández-Varea, J. M., and Sempau, J. "PENELOPE-2006: A Code System for Monte Carlo Simulation of Electron and Photon Transport". In: *Workshop Proceedings*. Vol. 4. 2006, p. 7.

[Samei et al., 2014] Samei, E., Lin, Y., Choudhury, K. R., and Page McAdams, H. "Automated Characterization of Perceptual Quality of Clinical Chest Radiographs: Validation and Calibration to Observer Preference". *Medical Physics* 41 (11), 2014, p. 111918.

[Sandler et al., 2018] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. "MobileNetV2: Inverted Residuals and Linear Bottlenecks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 4510–4520.

[Seeland et al., 2021] Seeland, M. and Mäder, P. "Multi-View Classification with Convolutional Neural Networks". *PLOS ONE* 16 (1), 2021, e0245230.

[Sénégas et al., 2018] Sénégas, J., Saalbach, A., Bergtholdt, M., Jockel, S., Mentrup, D., and Fischbach, R. "Evaluation of Collimation Prediction Based on Depth Images and Automated Landmark Detection for Routine Clinical Chest X-Ray Exams". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Lecture Notes in Computer Science. Springer International Publishing, 2018, pp. 571–579.

[Serra et al., 2024] Serra, D., Neep, M. J., and Ryan, E. "Multi-Centre Digital Radiography Reject Analysis for Different Clinical Room Use Types: The Establishment of Local Reject Reference Levels for Public Hospital Departments". *Journal of Medical Radiation Sciences* 71 (3), 2024, pp. 412–420.

[Srivastava et al., 2014] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". *Journal of Machine Learning Research* 15 (56), 2014, pp. 1929–1958.

[Srivastav et al., 2021] Srivastav, V., Issenhuth, T., Kadkhodamohammadi, A., de Mathelin, M., Gangi, A., and Padoy, N. *MVOR: A Multi-view RGB-D Operating Room Dataset for 2D and 3D Human Pose Estimation*. 2021. arXiv: `1808.08180 [cs]`.

[Strash et al., 2004] Strash, W. W. and Berardo, P. "Radiographic Assessment of the Hindfoot and Ankle". *Clinics in Podiatric Medicine and Surgery*. Reconstruction of Failed Hindfoot, Ankle and Lower Leg Surgery 21 (3), 2004, pp. 295–304.

[Takaki et al., 2020] Takaki, T., Murakami, S., Watanabe, R., Aoki, T., and Fujibuchi, T. "Calculating the Target Exposure Index Using a Deep Convolutional Neural Network and a Rule Base". *Physica Medica* 71, 2020, pp. 108–114.

[Takamoto et al., 2020] Takamoto, H., Nishine, H., Sato, S., Sun, G., Watanabe, S., Seokjin, K., Asai, M., Mineshita, M., and Matsui, T. "Development and Clinical Application of a Novel Non-contact Early Airflow Limitation Screening System Using an Infrared Time-of-Flight Depth Image Sensor". *Frontiers in Physiology* 11, 2020.

[Tan et al., 2019a] Tan, M., Chen, B., Pang, R., Vasudevan, V., Sandler, M., Howard, A., and Le, Q. V. "MnasNet: Platform-aware Neural Architecture Search for Mobile". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[Tan et al., 2019b] Tan, M. and Le, Q. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks". In: *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[The Royal College of Radiologists, 2025a] The Royal College of Radiologists, *Clinical Radiology: Workforce Census 2024*. Workforce Census Report. The Royal College of Radiologists, 2025.

[The Royal College of Radiologists, 2025b] The Royal College of Radiologists, *Radiology Delays Worst on Record despite Spend on Private Providers Soaring*. 2025.

[Thian et al., 2019] Thian, Y. L., Li, Y., Jagmohan, P., Sia, D., Chan, V. E. Y., and Tan, R. T. "Convolutional Neural Networks for Automated Fracture Detection and Localization on Wrist Radiographs". *Radiology: Artificial Intelligence* 1 (1), 2019, e180001.

[Unberath et al., 2018] Unberath, M., Zaech, J.-N., Lee, S. C., Bier, B., Fotouhi, J., Armand, M., and Navab, N. "DeepDRR – A Catalyst for Machine Learning in Fluoroscopy-Guided Procedures". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Springer International Publishing, 2018, pp. 98–106.

[Waite et al., 2017] Waite, S., Scott, J. M., Legasto, A., Kolla, S., Gale, B., and Krupinski, E. A. "Systemic Error in Radiology". *American Journal of Roentgenology* 209 (3), 2017, pp. 629–639.

[Wang et al., 2020] Wang, Y., Song, Y., Wang, F., Sun, J., Gao, X., Han, Z., Shi, L., Shao, G., Fan, M., and Yang, G. "A Two-Step Automated Quality Assessment for Liver MR Images Based on Convolutional Neural Network". *European Journal of Radiology* 124, 2020.

[Willis et al., 2018] Willis, C. E., Nishino, T. K., Wells, J. R., Ai, H. A., Wilson, J. M., and Samei, E. "Automated Quality Control Assessment of Clinical Chest Images". *Medical Physics* 45 (10), 2018, pp. 4377–4391.

[Winberg et al., 2025] Winberg, C., Prager, R., Kim, C. S., Meyer, M., and Arntfield, R. "Comparative Evaluation of Lung Ultrasound versus Chest X-ray for Pneumothorax Assessment Post-Invasive Intrathoracic Procedures: A Case-Costing Evaluation". *Medicine* 104 (17), 2025, e41959.

[Zhou et al., 2018] Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J. *UNet++: A Nested U-Net Architecture for Medical Image Segmentation*. 2018. arXiv: `1807.10165 [cs, eess, stat]`.

List of Publications

This list contains journal articles, conference papers, and patents published or currently under review.

Journal Articles as First Author

- Gerdes, H.*, Mairhöfer, D.*, Laufer, M., Reis, F. L., Bischof, A., Wegner, F., Käster, T., Barth, E., Barkhausen, J., Martinetz, T., and Sieren, M. *Generalizable Deep Learning Framework for Diagnostic Quality Assessment of Musculoskeletal Radiographs*. *Authors contributed equally. 2026. **Under Review at Radiology: Artificial Intelligence**.
- Mairhöfer, D.*, Laufer, M.*, Berkel, L., Sieren, M., Bischof, A., Barth, E., Barkhausen, J., and Martinetz, T. "AI-based Collimation Optimization for X-ray Imaging Using Depth Cameras". *Neurocomputing* 661, 2026. *Authors contributed equally., p. 131881.

Conference Papers as First Author

- Mairhöfer, D.*, Laufer, M.*, Berkel, L., Bischof, A., Barth, E., Barkhausen, J., and Martinetz, T. "AI-based Collimation Optimization for X-Ray Imaging Using Time-of-Flight Cameras". In: *ESANN 2024 Proceedings*. *Authors contributed equally. Ciaco - i6doc.com, 2024, pp. 703–708.
- Laufer, M.*, Mairhöfer, D.*, Sieren, M., Gerdes, H., Reis, F. L., Bischof, A., Käster, T., Barth, E., Barkhausen, J., and Martinetz, T. "Patient Pose Assessment in Radiography Using Time-of-Flight Cameras". In: *Medical Imaging 2024: Image Processing*. Vol. 12926. *Authors contributed equally. SPIE, 2024, pp. 385–393.
- Mairhöfer, D., Laufer, M., Simon, P. M., Sieren, M., Bischof, A., Käster, T., Barth, E., Barkhausen, J., and Martinetz, T. "An AI-based Framework for Diagnostic Quality Assessment of Ankle Radiographs". In: *Proceedings of the Fourth Conference on Medical Imaging with Deep Learning (MIDL 2021)*. PMLR, 2021, pp. 484–496.

Patents

- Laufer, M., Mairhöfer, D., Bischof, A., Käster, T., Sieren, M., Reis, F. L., Gerdes, H. W., and Simon, P. "Verfahren zur Erzeugung von Trainingsdaten für ein KI-basiertes Assistenzsystem und Vorrichtung zur Unterstützung der Röntgendiagnostik". DE102022133272A1. 2024.

Journal Articles and Conference Papers as Co-Author

- Laufer, M., Mairhöfer, D., Sieren, M., Gerdes, H., Reis, F. L., Bischof, A., Käster, T., Barth, E., Barkhausen, J., and Martinetz, T. *Synthetic Data Generated from CT Scans for Patient Pose Assessment*. **Under Review at Machine Learning for Biomedical Imaging**.
- Laufer, M., Haas, J., Mairhöfer, D., Sieren, M., Gerdes, H., Reis, F. L., Bischof, A., Käster, T., Barth, E., Barkhausen, J., and Martinetz, T. "Evaluation of Time-of-Flight Camera Positioning for AI-based Patient Pose Assessment in Radiography". In: *Bildverarbeitung für die Medizin 2026*. Proceedings not yet published. Springer Fachmedien, 2026.
- Laufer, M., Mairhöfer, D., Sieren, M., Barth, E., Barkhausen, J., and Martinetz, T. "From Body Surface to Bones: Bone Segmentation Using Depth Images". In: *Medical Imaging 2026: Image Processing*. Proceedings not yet published. SPIE, 2026.
- Laufer, M., Brokmann, F., Mairhöfer, D., Barth, E., and Martinetz, T. "Investigating the Impact of Imbalanced Medical Data on the Performance of Self-Supervised Learning Approaches". In: *ESANN 2025 Proceedings*. Ciaco - i6doc.com, 2025, pp. 401–406.
- Laufer, M., Mairhöfer, D., Sieren, M., Gerdes, H., Leal dos Reis, F., Bischof, A., Käster, T., Barth, E., Barkhausen, J., and Martinetz, T. "Synthetic Data Generated from CT Scans for Patient Pose Assessment". In: *Proceedings of the Eighth Conference on Medical Imaging with Deep Learning (MIDL 2025)*. Proceedings not yet published. PMLR, 2025.
- Mamlouk, A. M. and Mairhöfer, D. "Gameful Learning as a Safe Haven: Embracing Diversity in Student-Centered Education". In: *2025 IEEE 4rd German Education Conference (GECon)*. Proceedings not yet published. 2025.
- Lämmermann, K., Mairhöfer, D., and Mamlouk, A. M. "GradebookXP: A Moodle Plugin for Competency-Oriented and Student-Centered Feedback on Learning Objectives". In: *2024 IEEE 3rd German Education Conference (GECon)*. 2024, pp. 1–6.

- Gerdes, H., Mairhöfer, D., Laufer, M., Reis, F. L., Käster, T., Barth, E., Martinetz, T., Barkhausen, J., Bischof, A., and Sieren, M. "Anwendung eines KI-basierten Algorithmus zur Qualitätsbewertung von Röntgenaufnahmen des Kniegelenks". In: *RöFo - Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*. Vol. 195. Georg Thieme Verlag, 2023, ab101.
- Gerdes, H., Mairhöfer, D., Laufer, M., Reis, F. L., Preuss, J., Käster, T., Barth, E., Martinetz, T., Barkhausen, J., Bischof, A., and Sieren, M. "Automatisierte Qualitätsbewertung von Röntgenaufnahmen des oberen Sprunggelenks mittels künstlicher Intelligenz". In: *RöFo - Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*. Vol. 194. Georg Thieme Verlag, 2022, ab18.
- Reis, F. L., Laufer, M., Mairhöfer, D., Gerdes, H., Preuss, J., Käster, T., Barth, E., Martinetz, T., Barkhausen, J., Bischof, A., and Sieren, M. "Vorhersage der Röntgenbildqualität des oberen Sprunggelenks mittels Tiefenbildtechnik und künstlicher Intelligenz – eine Kadaverstudie". In: *RöFo - Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*. Vol. 194. Georg Thieme Verlag, 2022, ab28.