


ORIGINAL ARTICLE

Open Access



A deep learning approach for projection and body-side classification in musculoskeletal radiographs

Anna Fink^{1*} , Hien Tran¹, Marco Reiser^{2,3}, Alexander Rau^{1,4}, Jörg Bayer⁵, Elmar Kotter¹, Fabian Bamberg¹ and Maximilian F. Russe¹

Abstract

Background The growing prevalence of musculoskeletal diseases increases radiologic workload, highlighting the need for optimized workflow management and automated metadata classification systems. We developed a large-scale, well-characterized dataset of musculoskeletal radiographs and trained deep learning neural networks to classify radiographic projection and body side.

Methods In this IRB-approved retrospective single-center study, a dataset of musculoskeletal radiographs from 2011 to 2019 was retrieved and manually labeled for one of 45 possible radiographic projections and the depicted body side. Two classification networks were trained for the respective tasks using the Xception architecture with a custom network top and pretrained weights. Performance was evaluated on a hold-out test sample, and gradient-weighted class activation mapping (Grad-CAM) heatmaps were computed to visualize the influential image regions for network predictions.

Results A total of 13,098 studies comprising 23,663 radiographs were included with a patient-level dataset split, resulting in 19,183 training, 2,145 validation, and 2,335 test images. Focusing on paired body regions, training for side detection included 16,319 radiographs (13,284 training, 1,443 validation, and 1,592 test images). The models achieved an overall accuracy of 0.975 for projection and 0.976 for body-side classification on the respective hold-out test sample. Errors were primarily observed in projections with seamless anatomical transitions or non-orthograde adjustment techniques.

Conclusions The deep learning neural networks demonstrated excellent performance in classifying radiographic projection and body side across a wide range of musculoskeletal radiographs. These networks have the potential to serve as presorting algorithms, optimizing radiologic workflow and enhancing patient care.

Relevance statement The developed networks excel at classifying musculoskeletal radiographs, providing valuable tools for research data extraction, standardized image sorting, and minimizing misclassifications in artificial intelligence systems, ultimately enhancing radiology workflow efficiency and patient care.

Key points

- A large-scale, well-characterized dataset was developed, covering a broad spectrum of musculoskeletal radiographs.
- Deep learning neural networks achieved high accuracy in classifying radiographic projection and body side.

*Correspondence:

Anna Fink

anna.fink@uniklinik-freiburg.de

Full list of author information is available at the end of the article



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

- Grad-CAM heatmaps provided insight into network decisions, contributing to their interpretability and trustworthiness.
- The trained models can help optimize radiologic workflow and manage large amounts of data.

Keywords Artificial intelligence, Bone and bones, Deep learning, Musculoskeletal diseases, Radiography

Graphical Abstract

A deep learning approach for projection and body side classification in musculoskeletal radiographs

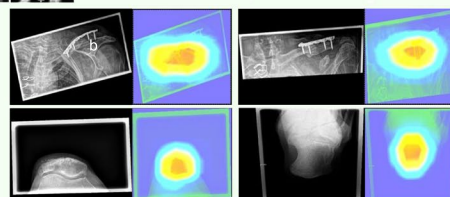
ESR[®] EUROPEAN SOCIETY OF RADIOLOGY

- A large-scale, well-characterized dataset was developed, covering a broad spectrum of MSK radiographs.
- Deep learning neural networks achieved high accuracy in classifying radiographic projection and body side.
- Grad-CAM heatmaps provided insight into network decisions, contributing to their interpretability and trustworthiness.
- The trained models can help optimize radiologic workflow and manage large amounts of data.



Exemplary cases of the dataset representing the broad variability of body parts, radiographic projections and pathologies.

Input radiographs resized to 256×256 pixels with corresponding Grad-CAM overlay of correctly classified radiographs for projection (a-d) and body side (e-h) demonstrating the influential image regions (red overlay). a, b Clavicle anterior-posterior. c, d Clavicle oblique. e, f Right patella. g, h Left calcaneus.



These networks excel at classifying MSK radiographs, providing tools for research data extraction, standardized image sorting, and minimizing misclassifications, enhancing workflow efficiency and patient care.



**Eur Radiol Exp (2024) Fink A, Tran H, Reisert M et al.
DOI: 10.1186/s41747-023-00417-x**

Background

Musculoskeletal diseases impose a high burden on healthcare systems worldwide. The high prevalence of these conditions, combined with the long-term impact of chronic pain and disability after acute treatment, not only diminishes patient well-being but also places a substantial financial load on societies [1]. Customized and appropriate therapy relies on accurate diagnoses and is crucial for the prevention of chronic conditions. Despite the increasing number of cross-sectional computed tomography and magnetic resonance examinations, conventional radiographs still play an indispensable role in the workup of musculoskeletal diseases [2].

Given the rapidly aging population, the prevalence of musculoskeletal conditions is on the rise, leading to a surge in radiological examinations [1, 3]. Consequently, optimizing radiologic workflows becomes paramount, paving the way for supporting artificial intelligence (AI) systems. Numerous models have been developed for the

automated identification of pathologies in radiographs, including fracture detection [4, 5], osteoarthritis grading [6], or skeletal maturity assessment [7, 8].

The performance of automated algorithms in pathology detection is significantly enhanced by utilizing larger training datasets [9]. While the Digital Imaging and Communications in Medicine (DICOM) format offers the opportunity to store metadata such as image modality, projection, or side, this information is often inconsistent or missing altogether [10].

To address these constraints and harness image data more effectively, automated metadata classification systems have been proposed. However, existing algorithms primarily focus on classifying body regions [11, 12] or differentiating two singular projections [10, 13].

Operating a multi-classification task, these networks require a substantial amount of training data. While publicly available musculoskeletal datasets exist for singular body regions such as hands [14], knees [15], or upper

[16] and lower extremities [17], an open-access dataset encompassing a broad spectrum of all relevant musculoskeletal projections and body regions is currently lacking.

We therefore sought to create a large-scale, well-characterized musculoskeletal radiograph dataset and utilize this training foundation to develop neural networks for the automatic classification of radiographic projection and body side.

Methods

Dataset

This retrospective, monocentric study was approved by the local institutional review board (Ethics Committee University of Freiburg; EK:570/19). Informed written consent was waived due to the retrospective study design and patient pseudonymization.

We retrieved all musculoskeletal radiographic studies performed on adult patients between 2018 and 2019 from our institution's Picture Archiving and Communication System (PACS). To ensure an adequate amount of data for each class, radiographs of rarely examined body regions were also included from the period of 2011 to 2017. These additional body regions comprised the nasal bone, dens, thoracic spine, clavicle, acromioclavicular joint, elbow (radial head), hand, hip, patella, and foot (forefoot, calcaneus, toe). Images of particularly poor quality (not attributable to a radiographic projection, joints destroyed beyond recognition, and incorrectly transferred images) were manually marked and excluded from the dataset.

As a result, a total of 13,098 studies encompassing 23,663 radiographs were included, covering a wide range of

musculoskeletal radiology fields with diverse body regions and pathologies as well as radiographs with and without orthopedic implants. The project workflow is depicted in Fig. 1. Figure 2 illustrates a sample selection of the dataset.

To prevent data leakage between training, validation, and test datasets, we only used the first obtained study for each patient within the period of 2011–2019. As some patient studies consisted of multiple individual radiographic projections, a randomized split was performed at the patient level. This resulted in three independent datasets, comprising 19,183 training, 2,145 validation, and 2,335 test images. For side detection, we only included images of paired body regions, leaving a total of 16,319 radiographs and a division into 13,284 training, 1,443 validation, and 1,592 test images.

Data annotation

Annotation for both network tasks was initially performed by a junior resident (first year of training, A.F.), followed by a consensus reading of uncertain cases with a senior resident (last year of training, H.T.) and a board-certified radiologist (M.F.R.), employing a local instance of the imaging platform Nora [18].

Each x-ray was manually classified according to the represented projection, allocating one of 45 possible machine-readable text labels, a list of which can be found in the supplementary materials (Suppl. 1). Additionally, two labels were assigned to indicate the body side (left or right) on radiographs of paired body regions only. Laterality ground truth was established based on examination notes. This manual classification process, involving initial labeling by a resident followed by a joint evaluation of

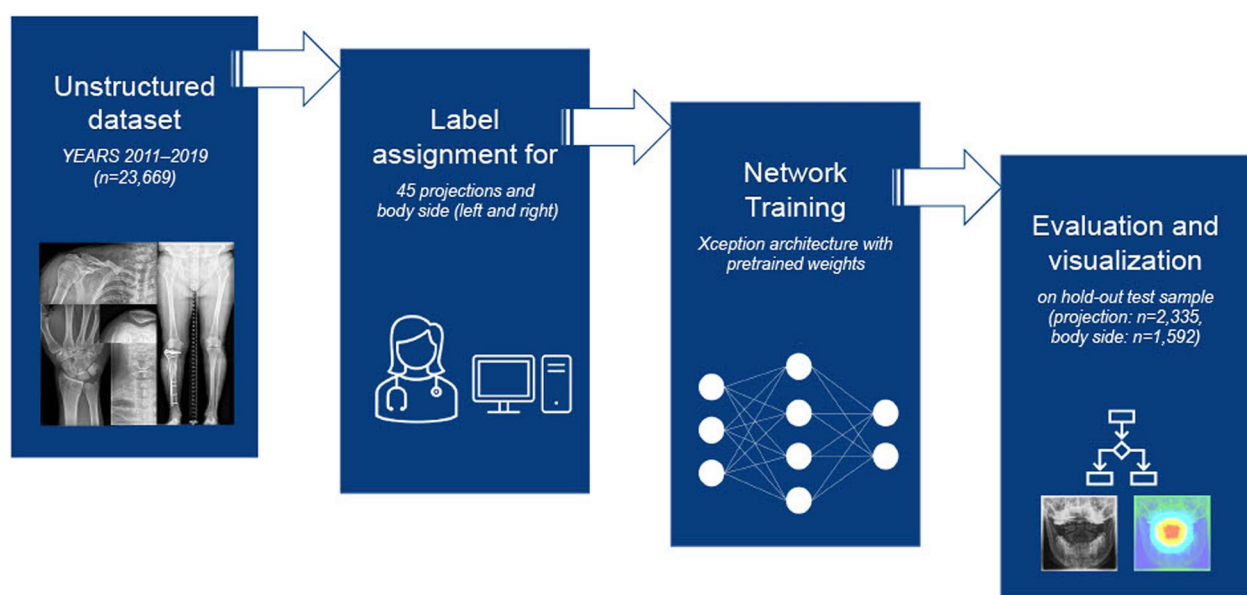


Fig. 1 Project workflow from dataset composition, annotation, and network training to final evaluation



Fig. 2 Exemplary cases of the dataset representing the broad variability of body parts, radiographic projections, and pathologies

indeterminate cases with an experienced and a subspecialized trained radiologist, ensured accurate labeling for the subsequent network training.

Table 1 presents an overview of the final dataset for the classification of radiographic projection, displaying the unbalanced label distribution within the dataset with a range from 189 images (toe anterior–posterior [AP]) to 1,267 images (patella tangential). For body-side classification, the dataset was split up into 9,028 images for the left and 7,291 images for the right side, utilizing all available radiographs despite the uneven distribution of examinations for both sides.

Network training

Based on this large-scale labeled dataset, we trained two separate neural networks for the classification of radiographic projection and body side, respectively. Network training was conducted on a standard server graphics processing unit (GPU, Nvidia Tesla RTX A6000). As a

deep learning framework, we used the open-source Python library TensorFlow 2.6 [19] and its programming interface Keras [20]. The established network architecture Xception by Chollet et al. [21], originally designed for the classification of multi-colored images with three input channels for the basic colors red, blue, and green, acted as Convolutional Neural Network base. Leveraging this feature, we utilized the original three input channels for each basic color to process our augmented training data.

To optimize the network architecture, adjusting for the reduced number of classes in comparison to the initial network configuration, we removed the top layer and replaced it with a global average pooling layer, a dropout layer to prevent overfitting during training, a dense layer with a rectified linear unit activation function to capture nonlinear dependencies between features and learn complex patterns from the data, and a dense layer with output neurons adapted to the number of classes. The

Table 1 Overview of every depicted projection in the dataset and its frequency of representation

Head/spine		<i>n</i>	Arm		<i>n</i>	Hand		<i>n</i>	Leg		<i>n</i>	Foot		<i>n</i>				
Nasal bone	Lateral	268	AC-joint	AP	445	Hand	AP	487	Pelvis	Pelvis AP	833	Foot	AP	864				
							Oblique	454			Hip AP		355		Oblique	771		
											Lauenstein		349		Lateral	392		
Cervical spine	AP	367	Shoulder	AP	1,132	Wrist	AP	404	Whole leg	AP	359	Forefoot	AP	306				
	Lateral	395			Axial		658			Lateral	443				Oblique	560		
	Dens	195			Outlet		977											
Thoracic spine	AP	659	Clavicle	AP	574	Finger	AP	333	Knee	AP	1,099	Calcaneus	Lateral	299				
	Lateral	620			Oblique		946			Lateral	346			Lateral	1,157		Axial	435
										Tangential	1,267							
Lumbar spine	AP	604	Elbow	AP	358	Thumb	AP	208	Ankle	AP	720	Toe	AP	189				
					Lateral		384								Lateral	168		
	Lateral	608			Radial head		329			Lateral	208			Lateral	689	Big toe	AP	224
															Lateral		230	

AP Anterior–posterior, *n* Number of radiographs in the dataset

final output decision was determined using a softmax function.

To improve overall network performance and shorten training time, we applied pretrained network weights using the open-access ImageNet database [22]. For training input, we rescaled the variably sized radiographs to a standard network input size of 256×256 pixels. To utilize the three input channels of the Xception network, the radiographs were transformed into a three-channel image by incorporating a derived inversion and an edge enhancement image. This approach can improve network performance compared to only using original input radiographs, as shown by Rahman et al. [23]. Edge enhancement was achieved by applying the medianBlur and adaptiveThreshold operations. Training data was augmented using lateral flip and rotation up to 10° for projection training.

To enable body-side detection, the corresponding training process did not involve lateral flip.

We trained both networks for a total of 400 epochs with 300 steps per epoch and a batch size of 15. The initial learning rate started from 0.1 and was gradually reduced to 0.005 using a polynomial decay function.

Evaluation metrics

We calculated outcome statistics using the Scikit-Learn software library [24]. For statistical analysis, each network output was compared to the manually assigned text label, thus determining model accuracy, precision, and recall. We additionally calculated the Matthews Correlation Coefficient (MCC), which provides a balanced assessment of model accuracy, particularly for unbalanced class distributions. Bootstrapping was used to calculate 95% confidence intervals, which are presented in brackets alongside each metric in the results section.

To address the potential issue of intransparent network predictions, we employed Gradient-weighted Class Activation Mapping (Grad-CAM) [25]. Heatmaps were computed based on the final convolutional layer, providing insight into the specific image regions that influenced the network's classification decision for every radiograph in the test dataset.

Code and dataset availability

The model code will be openly accessible as an interactive Jupyter notebook on GitHub. This codebase was created using Python 3.10.12 and leverages framework of TensorFlow 2.13.0, tf-explain 0.3.1, nibabel 4.0.2, cv2 4.8.0, and numpy 1.23.5. It is openly available under the MIT License and can be retrieved from the project's home page, the XraySorterAI Project (<https://github.com/maxrusse/XraySorterAI>).

The dataset generated in this study will be provided upon reasonable request, taking into consideration compliance with European data protection regulations and laws.

Results

Dataset

The dataset consisted of musculoskeletal radiographs with a mean age of 51.6 years (standard deviation 19.8). The distribution of files by gender was 56% for males and 44% for females.

The x-ray machines used were mainly manufactured by Philips Medical Systems (Hamburg, Germany), to a lesser extent from Samsung Electronics. The datasets are comparable across acquisition technology, x-ray machine manufacturer, spatial resolution, and exposure dosage. A detailed breakdown of the corresponding metadata can be found in the supplementary materials (Suppl. 2–5).

Radiographic projections

The DICOM-headers used in clinical routine did not contain information on the projection in 28.4% of the 2335 radiographs in the test dataset, emphasizing the necessity of manual labeling for accurate classification within this study. Processing all test images using a single-core server central processing unit (CPU) and no GPU took 139 s, resulting in a classification rate of 16 images/s. The model achieved an overall accuracy of 0.975 (95% confidence interval 0.968–0.981) on the hold-out test sample. Precision measured 0.978 (0.970–0.982), recall 0.973 (0.969–0.981), and MCC 0.974 (0.967–0.981).

Table 2 displays the radiographic projections in which incorrect predictions occurred, along with the corresponding proportion of misclassified radiographs within the overall test dataset. The remaining portion of the test dataset was correctly classified. Among the projections, performance was comparatively lower for the AP view of the clavicle (true positive rate of 0.822) and radial head (true positive rate of 0.800). For a detailed and comprehensive analysis of all network predictions, including true and false positives, the complete confusion matrix can be found in the supplementary materials (Suppl. 6).

Grad-CAM heatmaps provided visual evidence of the image regions that influenced network output decisions. Among the misclassified test images, the most common errors arose from smooth transitions between different projection angles (56%), such as AP and oblique views of the clavicle. Challenges also arose from collimation, mainly making the choice between AP views of the acromioclavicular joint, shoulder, and clavicle (34%) difficult. Metal-dense implant overlay also contributed to classification errors in some cases (5%). In 4% of cases, the exact reason for misclassification remained unclear.

Table 2 Overview of the radiographic projections in which incorrect network predictions were observed

True label	Prediction	r	True label	Prediction	r	True label	Prediction	r
Clavicle AP	Clavicle oblique	0.16	Hip AP	Lauenstein	0.04	Foot AP	Foot oblique	0.06
	Shoulder AP	0.02						
Clavicle oblique	Clavicle AP	0.07	Lauenstein	Hip AP	0.02	Foot oblique	Foot AP	0.01
	AC-joint	0.01					Forefoot oblique	0.03
AC-joint	Clavicle oblique	0.02	Knee AP	Knee lateral	0.01	Forefoot AP	Foot AP	0.04
	Shoulder AP	0.02						
Shoulder AP	AC-joint	0.04	Knee lateral	Elbow AP	0.01	Forefoot oblique	Foot oblique	0.09
	Shoulder outlet	0.03						
Radial head	Elbow AP	0.17	Ankle lateral	Calcaneus lateral	0.02	Big toe AP	Toe AP	0.05
	Elbow lateral	0.03						
Hand oblique	Hand AP	0.05	Calcaneus lateral	Ankle lateral	0.07			
Thumb lateral	Thumb AP	0.05						

AP Anterior–posterior, r Relative proportion of misclassified radiographs within the test dataset

Across all the incorrectly classified test images and 50 randomly selected correctly classified test images, heatmaps consistently highlighted that the image regions influencing network predictions were central parts of the radiograph, such as joint regions or large bone structures.

Figure 3 depicts the heatmaps of two correctly classified radiographs of the clavicle. Exemplary heatmaps illustrating the regions of influence for misclassified projections are provided in the supplementary materials (Suppl. 7).

Body side

Processing all 1,592 test images using a single-core CPU and no GPU took 48 s, resulting in a classification rate of 33 images per second. The model achieved an overall accuracy of 0.976 (95% confidence interval 0.969–0.983) on the hold-out test sample. Precision measured 0.976 (0.969–0.983), recall 0.976 (0.969–0.983), and MCC 0.973 (0.965–0.981).

Grad-CAM heatmaps were also computed for this task to illustrate which image regions influenced the network's output decision. Among the misclassified test images, the most common errors were observed

in lateral views of single fingers and knees (18% each), followed by AP view of thumb and knee (12% each), lateral view of the foot (9%), and AP view of single fingers and toes (6% each). Closer examination of the misclassifications revealed prominent problems arising from a projection technique inconsistent with our clinic's SOP, such as inverted radiation beam path or body part position (48%), alongside challenges posed by metal-dense implants (15%) and unusual pathologies such as foot amputation (6%). In 24% of cases, the exact reason for misclassification remained unclear, mainly involving lateral views of individual fingers.

Across all the incorrectly classified test images and 50 randomly selected correctly classified test images, the heatmaps consistently highlighted that the network's output decision was centered on crucial image areas, particularly joint gaps. Notably, none of the heatmaps focused on the sometimes visually displayed side labels "L" and "R," as visualized in the sample heatmaps provided in the supplementary materials (Suppl. 8).

Figure 4 provides two examples of heatmaps representing correctly classified radiographs, highlighting

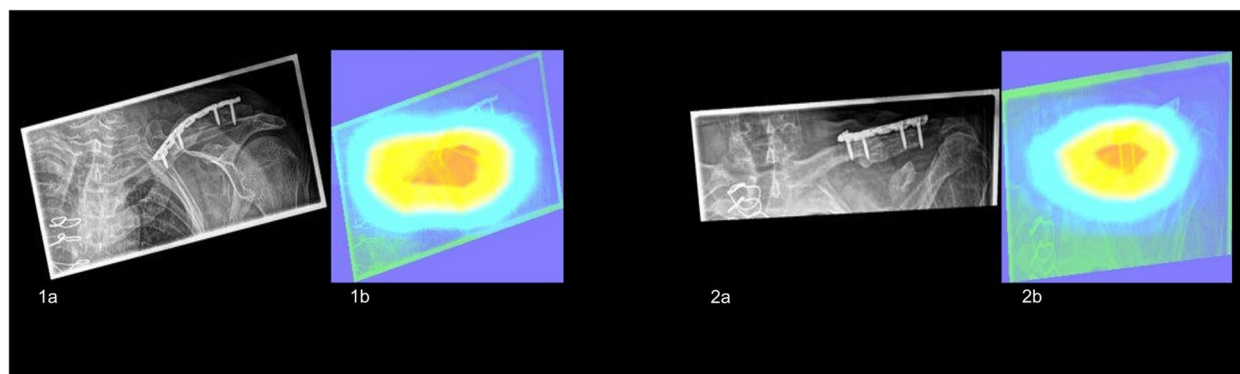


Fig. 3 Input radiographs resized to 256×256 pixels with corresponding Grad-CAM overlay of two correctly classified projections demonstrating the influential image regions (red overlay). **1a, 1b** Clavicle anterior–posterior. **2a, 2b** Clavicle oblique

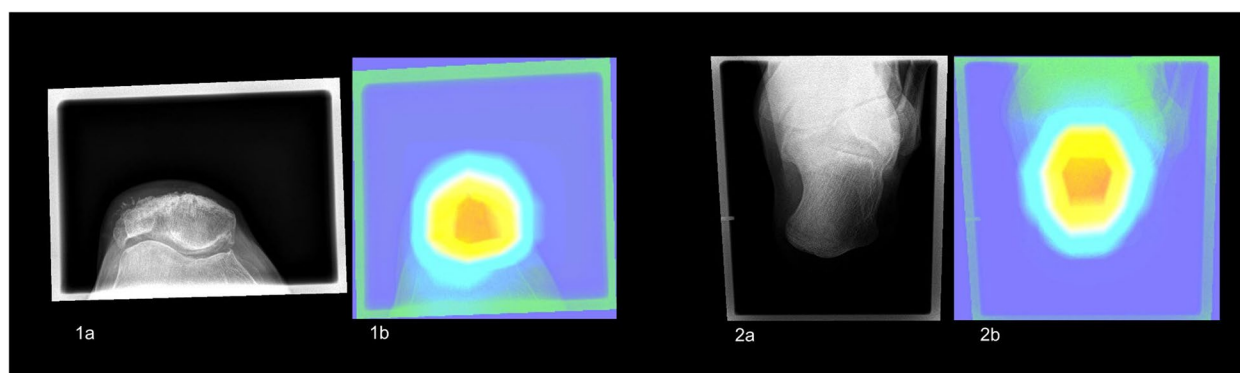


Fig. 4 Input radiographs resized to 256×256 pixels with corresponding Grad-CAM overlay of two correctly classified radiographs for the body side demonstrating the influential image regions (red overlay). **1a, 1b** Right patella. **2a, 2b** Left calcaneus

the influential regions. Supplementary materials contain additional heatmaps showcasing instances of incorrect classifications (Suppl. 9).

Discussion

We developed a large-scale, well-characterized dataset of musculoskeletal radiographs and trained corresponding networks for the classification of radiographic projection and body side. The models exhibited excellent and fast performance, achieving an accuracy of 0.975 for projection and 0.976 for body-side classification. The models' robustness was further highlighted by their performance on an unknown test dataset containing radiographs with various underlying pathologies and orthopedic implants. Moreover, the utilization of Grad-CAM heatmaps provided an additional layer of interpretability by visualizing the image regions that influenced the model's output decisions.

In the context of rapidly increasing examination numbers, it is crucial to organize and validate both radiographs and their associated metadata, particularly considering the prevalent inconsistencies or lack of image-related metadata in DICOM-headers. Previous studies have emphasized the importance of large labeled datasets for neural network training, such as the MURA dataset for the upper extremity (40,561 images [16]) and the LERA dataset for the lower extremity (93,455 images [17]) provided by the Stanford Machine Learning Group. The release of both datasets each prompted multiple subsequent projects focusing on abnormality detection in musculoskeletal radiographs [26–28]. However, these datasets primarily focused on presorting body regions, assigning labels at study and patient levels, respectively. Our dataset stands out for its comprehensive coverage of musculoskeletal radiographs, encompassing a broader spectrum of images than previously available datasets. This breadth allows our models to handle multiclassification tasks across a wide range of body regions, with 45 distinct labels for radiographic projection and additional differentiation of body side. The dataset's high quality was further ensured by involving three distinct labelers, including a resident and two experienced radiologists, in the manual classification process.

Previous studies on sorting networks primarily focused on classifying musculoskeletal radiographs based on broader body regions [11, 12]. In contrast, our approach takes a step further by classifying radiographs based on their precise projection and body side. Compared to related studies that primarily focused on distinguishing two chest x-ray projections [10, 13] or classifying radiographs into 30 categories [29], our models demonstrate the ability to classify radiographs across a wide range of 45 different projections while also incorporating body

side detection, outperforming the previous research in terms of accuracy and scope, respectively.

In our study, projections with unique features, such as nasal bone or whole leg AP, achieved excellent classification rates. Errors were infrequent and occurred primarily in projections such as the AP view of the clavicle (often misclassified as clavicle oblique) and the radial head (often misclassified as AP elbow). In clinical practice, these projections are often affected by non-orthograde adjustment techniques and show a seamless anatomical transition to other views. Similarly, body-side detection errors were more prevalent in radiographs of single fingers and toes or the tangential view of the patella, where distinguishing the body side is subjectively challenging. Nonetheless, our models demonstrated success in accurately distinguishing even these challenging classes, resulting in overall accuracies comparable to previous studies [13].

The incorporation of Grad-CAM heatmaps in our analysis enhanced the interpretability and transparency of the network's outputs, addressing the inherent "black box" nature of neural networks with multiple hidden layers. By visualizing the image regions that played a decisive role in the output, we showed that the network's decisions aligned with human viewers' interpretations. Even for the majority of incorrect predictions, we managed to make network decisions understandable. The influential regions identified by the heatmaps often corresponded to clinically relevant areas such as the joint space or prominent bone structures.

Furthermore, our findings demonstrated that the network's body-side classification was not reliant on the visually depicted side labels "L" and "R", as a human viewer would interpret. Instead, the classification was primarily based on bone structures within the radiographs. It is noteworthy that the side label was not always a physical opaque marker added by the technologist prior to imaging but rather often a digital overlay within the PACS, and thus not directly encoded in the raw data accessible to the network.

Despite these promising results, our study has limitations. Given the large number of classes in the projection training, class balancing was not feasible. Nevertheless, the substantial number of radiographs per class allowed for an excellent classification accuracy. This finding is consistent with previous studies, where increasing data volume significantly improved precision and recall, while balancing techniques barely showed any improvement [9].

As the study was monocentric and retrospective in nature, we did not have the opportunity to validate the trained models on radiographs from external institutions. To mitigate this, we implemented a randomized dataset split on a patient level, creating a hold-out test sample that was unknown to the models. Furthermore,

we took measures to create a highly heterogeneous dataset that encompasses radiographs from everyday clinical practice. This dataset was obtained from various examiners, captured using different devices, and depicted a wide range of pathologies and orthopedic implants. We believe that the excellent performance of our models on such a diverse dataset suggests their applicability to external datasets, but further validation through external studies is warranted.

In summary, the developed networks exhibited exceptional performance in classifying a wide range of musculoskeletal radiographs, enabling precise data extraction in research and automated image sorting for standardized reporting. Implementing them as pre-sorting algorithms for end-to-end solutions targeted on specific body regions showcases the great potential for minimizing misclassifications, ultimately enhancing radiology workflow efficiency and patient care.

Abbreviations

AI	Artificial intelligence
AP	Anterior–posterior
CPU	Central processing unit
DICOM	Digital Imaging and Communications in Medicine
GPU	Graphics processing unit
Grad-CAM	Gradient-weighted class activation mapping
MCC	Matthews correlation coefficient
PACS	Picture Archiving and Communication System

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s41747-023-00417-x>.

Additional file 1: Suppl. 1. List of all labels used in annotating training data for projection classification (AP: anterior-posterior). **Suppl. 2.** Image metadata on acquisition technology, x-ray machine manufacturer, and spatial resolution, for the training dataset (CR: computed radiography, DX: digital x-ray). **Suppl. 3.** Image metadata on acquisition technology, x-ray machine manufacturer, and spatial resolution for the validation dataset (CR: computed radiography, DX: digital x-ray). **Suppl. 4.** Image metadata on acquisition technology, x-ray machine manufacturer, and spatial resolution for the test dataset (CR: computed radiography, DX: digital x-ray). **Suppl. 5.** Image metadata on exposure dose in kVp and mAS for the training, validation, and test dataset (kVp: kilovoltage peak, mAS: milliampere-seconds). **Suppl. 6.** Normalized confusion matrix for the classification of 45 distinct radiographic projections. **Suppl. 7.** Input radiographs resized to 256 x 256 pixels with corresponding Grad-CAM overlay of wrongly classified projections, demonstrating the influential image regions (red overlay). 1a/b: clavicle AP (prediction: shoulder AP), 2a/b: dens (prediction: c-spine AP), 3a/b: knee lateral (prediction: elbow lateral), 4a/b: l-spine AP (prediction: t-spine AP). **Suppl. 8.** Input radiographs resized to 256 x 256 pixels with corresponding Grad-CAM overlay of correctly classified radiographs for body side with a visually displayed radiopaque side marker (1a/b: right hand, 2a/b: right thumb, 3a/b: right shoulder AP, 4a/b: right toe), demonstrating the influential image regions (red overlay). **Suppl. 9.** Input radiographs resized to 256 x 256 pixels with corresponding Grad-CAM overlay of left body parts wrongly classified as right body side (1a/b: foot, 2a/b: knee, 3a/b: thumb, 4a/b: knee), demonstrating the influential image regions (red overlay).

Acknowledgements

AR was supported by the Berta-Ottenstein-Program for Clinician Scientists, Faculty of Medicine, University of Freiburg.

The authors did not use any generative AI tools during the preparation of this paper.

Authors' contributions

MFR and EK conceptualized the study. Data annotation was initially performed by AF, followed by a consensus reading with HT and MFR. Network training was conducted by AF, MFR, and MR. AF and MFR wrote the main manuscript text. All authors reviewed and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials

The model code will be openly accessible as an interactive Jupyter notebook on GitHub. This codebase was created using Python 3.10.12 and leverages framework of TensorFlow 2.13.0, tf-explain 0.3.1, nibabel 4.0.2, cv2 4.8.0, and numpy 1.23.5. It is openly available under the MIT License and can be retrieved from the project's home page, the XraySorterAI Project (<https://github.com/maxrusse/XraySorterAI>).

Declarations

Ethics approval and consent to participate

This study was approved by the local institutional review board (Ethics Committee University of Freiburg: EK:570/19). Informed written consent was waived due to the retrospective study design and patient pseudonymization.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Department of Diagnostic and Interventional Radiology, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Breisacher Str. 64, 79106 Freiburg, Germany. ²Department of Stereotactic and Functional Neurosurgery, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany. ³Medical Physics, Department of Diagnostic and Interventional Radiology, Medical Center, University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany. ⁴Department of Neuroradiology, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany. ⁵Department of Trauma and Orthopaedic Surgery, Schwarzwald-Baar Hospital, Villingen-Schwenningen, Germany.

Received: 18 September 2023 Accepted: 29 November 2023

Published online: 14 February 2024

References

- Vos T, Lim SS, Abbafati C et al (2020) Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet* 396:1204–1222. [https://doi.org/10.1016/S0140-6736\(20\)30925-9](https://doi.org/10.1016/S0140-6736(20)30925-9)
- Nekolla EA, Schegerer AA, Griebel J, Brix G (2017) Häufigkeit und Dosis diagnostischer und interventioneller Röntgenanwendungen. *Radiologe* 57:555–562. <https://doi.org/10.1007/s00117-017-0242-y>
- Bhargavan M, Kaye AH, Forman HP, Sunshine JH (2009) Workload of radiologists in United States in 2006–2007 and trends since 1991–1992. *Radiology* 252:458–467. <https://doi.org/10.1148/radiol.2522081895>
- Jiménez-Sánchez A, Kazi A, Albarqouni S et al (2020) Precise proximal femur fracture classification for interactive training and surgical planning. *Int J Comput Assist Radiol Surg* 15:847–857. <https://doi.org/10.1007/s11548-020-02150-x>

5. Thian YL, Li Y, Jagmohan P et al (2019) Convolutional neural networks for automated fracture detection and localization on wrist radiographs. *Radiol Artif Intell* 1:e180001. <https://doi.org/10.1148/ryai.2019180001>
6. Norman B, Padoia V, Noworolski A et al (2019) Applying densely connected convolutional neural networks for staging osteoarthritis severity from plain radiographs. *J Digit Imaging* 32:471–477. <https://doi.org/10.1007/s10278-018-0098-3>
7. Larson DB, Chen MC, Lungren MP et al (2017) Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. *Radiology*. <https://doi.org/10.1148/radiol.2017170236>
8. Gao Y, Zhu T, Xu X (2020) Bone age assessment based on deep convolution neural network incorporated with segmentation. *Int J Comput Assist Radiol Surg* 15:1951–1962. <https://doi.org/10.1007/s11548-020-02266-0>
9. Juba B, Le HS (2019) Precision-recall versus accuracy and the role of large data sets. *Proc AAAI Conf Artif Intell* 33:4039–4048. <https://doi.org/10.1609/aaai.v33i01.33014039>
10. Rajkomar A, Lingam S, Taylor AG et al (2017) High-throughput classification of radiographs using deep convolutional neural networks. *J Digit Imaging* 30:95–101. <https://doi.org/10.1007/s10278-016-9914-9>
11. Yi PH, Kim TK, Wei J et al (2019) Automated semantic labeling of pediatric musculoskeletal radiographs using deep learning. *Pediatr Radiol* 49:1066–1070. <https://doi.org/10.1007/s00247-019-04408-2>
12. Hinterwimmer F, Consalvo S, Wilhelm N et al (2023) SAM-X: sorting algorithm for musculoskeletal x-ray radiography. *Eur Radiol* 33:1537–1544. <https://doi.org/10.1007/s00330-022-09184-6>
13. Kim TK, Yi PH, Wei J et al (2019) Deep learning method for automated classification of anteroposterior and posteroanterior chest radiographs. *J Digit Imaging* 32:925–930. <https://doi.org/10.1007/s10278-019-00208-0>
14. Cao F, Huang HK, Pietka E, et al (2003) Image database for digital hand atlas. *Proc. SPIE 5033, Medical Imaging 2003: PACS and Integrated Medical Information Systems: Design and Evaluation*. 5033:461–470. <https://doi.org/10.1117/12.480681>
15. Eckstein F, Wirth W, Nevitt MC (2012) Recent advances in osteoarthritis imaging—the Osteoarthritis Initiative. *Nat Rev Rheumatol* 8:622–630. <https://doi.org/10.1038/nrrheum.2012.113>
16. Rajpurkar P, Irvin J, Bagul A et al (2018) MURA: Large dataset for abnormality detection in musculoskeletal radiographs. 1st Conference on Medical Imaging with Deep Learning (MIDL 2018), Amsterdam, The Netherlands
17. Varma M, Lu M, Gardner R et al (2019) Automated abnormality detection in lower extremity radiographs using deep learning. *Nat Mach Intell* 1:578–583. <https://doi.org/10.1038/s42256-019-0126-0>
18. Nora - The medical imaging platform. <https://www.nora-imaging.com/>. Accessed 1 Sep 2023
19. Abadi M, Agarwal A, Barham P, et al (2016) TensorFlow: Large-scale machine learning on heterogeneous distributed systems. *arXiv:1603.04467* <https://doi.org/10.48550/arXiv.1603.04467>
20. Keras: Deep learning for humans. <https://keras.io/>. Accessed 1 Sep 2023
21. Chollet F (2017) Xception: Deep learning with depthwise separable convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Honolulu, HI, pp 1800–1807
22. Deng J, Dong W, Socher R, et al (2009) ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 2009, pp 248–255. <https://doi.org/10.1109/CVPR.2009.5206848>
23. Rahman T, Khandakar A, Qiblawey Y et al (2021) Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. *Comput Biol Med* 132:104319. <https://doi.org/10.1016/j.combiomed.2021.104319>
24. Pedregosa F, Varoquaux G, Gramfort A, et al (2018) Scikit-learn: machine learning in python. *arXiv:1201.0490* <https://doi.org/10.48550/arXiv.1201.0490>
25. Selvaraju RR, Cogswell M, Das A et al (2020) Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Int J Comput Vis* 128:336–359. <https://doi.org/10.1007/s11263-019-01228-7>
26. Guan B, Zhang G, Yao J et al (2020) Arm fracture detection in X-rays based on improved deep convolutional neural network. *Comput Electr Eng* 81:106530. <https://doi.org/10.1016/j.compeleceng.2019.106530>
27. Liang S, Gu Y (2020) Towards robust and accurate detection of abnormalities in musculoskeletal radiographs with a multi-network model. *Sensors (Basel)* 20:3153. <https://doi.org/10.3390/s20113153>
28. Urinbayev K, Orazbek Y, Nurambek Y et al (2020) End-to-end deep diagnosis of x-ray images. 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, Montreal, QC, Canada, pp 2182–2185
29. Dratsch T, Korenkov M, Zopf D et al (2021) Practical applications of deep learning: classifying the most common categories of plain radiographs in a PACS using a neural network. *Eur Radiol* 31:1812–1818. <https://doi.org/10.1007/s00330-020-07241-6>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.