

Machine Learning for Analysis and Diagnosis of Musculoskeletal Tumours

Florian Georg Maximilian Hinterwimmer

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology
der Technischen Universität München zur Erlangung eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
genehmigten Dissertation.

Vorsitz: Prof. Dr. Nassir Navab

Prüfende der Dissertation:

1. Prof. Dr. Daniel Rückert
2. Prof. Dr. Ferdinando Rodriguez y Baena

Die Dissertation wurde am 20.02.2024 bei der Technischen Universität München eingereicht
und durch die TUM School of Computation, Information and Technology am 24.06.2024
angenommen.

Abstract

Musculoskeletal tumours represent a rare and heterogeneous group of diseases, posing significant challenges for early and precise diagnosis. Current diagnostic procedures necessitate interdisciplinary examinations, that require the expertise of specialists and advanced imaging technologies. The rarity and intricate variability of these conditions, combined with a lack of standardized, digital data repositories, significantly degrade the quality of data available for analysis. This situation impedes the effective deployment of machine learning methods. However, the timely aggregation of comprehensive, prospective datasets is impractical owing to their low incidence. This elevates the significance of leveraging retrospective data as indispensable resource. This dissertation endeavours to develop innovative approaches that exploit retrospective data for the accurate assessment of musculoskeletal tumours, addressing these challenges. It focuses on employing state-of-the-art machine learning models to enhance diagnostic accuracy, optimise radiological workflows, and integrate multimodal datasets to unlock the latent potential within clinical systems. The research analyzed clinical and imaging data from 1962 to 2021, primarily unstructured, from Klinikum rechts der Isar. The data preparation process involved meticulous cleaning and verification, with the collaboration of radiologists and surgeons. The dissertation is based on four core publications. An initial review article highlights the challenges and limitations of applying machine learning in diagnosing musculoskeletal malignancies, advocating for improved data collection frameworks and international networks to advance orthopaedic oncology. Key methodological and application-related developments include the successful utilization of transfer learning with a dataset of 42,608 tumour-associated, unstructured X-ray images, addressing data insufficiency; the development of a novel sorting algorithm with a 96.6% accuracy in classifying radiographs into 28 anatomical regions, thus potentially optimising workflows; and a recommender-based approach that incorporates extensive experience lying dormant in clinical systems to link new patients to previous patients while classifying tumours with a mean accuracy of 92.86% across ten entities, outperforming current state-of-the-art models by over 30 percentage points. By leveraging real-world clinical data, this work navigates the constraints of small datasets and data quality, paving new avenues for the early and precise diagnosis of musculoskeletal tumours, optimizing radiological practices, aiding non-specialist clinicians, and ultimately promoting more effective, patient-centred care.

Zusammenfassung

Muskuloskelettale Tumoren stellen eine seltene und heterogene Gruppe von Krankheiten dar, die erhebliche Herausforderungen für eine frühe und präzise Diagnose mit sich bringen. Aktuelle Diagnoseverfahren erfordern interdisziplinäre Untersuchungen, die das Fachwissen von Spezialisten und fortgeschrittene Bildgebungstechnologien benötigen. Die Seltenheit und komplexe Variabilität dieser Zustände, kombiniert mit einem Mangel an standardisierten, digitalen Datenbanken, mindern signifikant die Qualität der verfügbaren Daten für Analysen. Diese Situation behindert den effektiven Einsatz von Methoden des maschinellen Lernens. Jedoch ist die zeitnahe Sammlung umfassender, prospektiver Datensätze aufgrund ihrer geringen Inzidenz nicht praktikabel. Dies hebt die Bedeutung der Nutzung retrospektiver Daten als unverzichtbare Ressource hervor. Diese Dissertation strebt die Entwicklung innovativer Ansätze an, die retrospektive Daten für die genaue Bewertung muskuloskelettaler Tumoren nutzen und diese Herausforderungen angehen. Sie konzentriert sich auf den Einsatz modernster maschineller Lernmodelle, um die diagnostische Genauigkeit zu erhöhen, radiologische Arbeitsabläufe zu optimieren und multimodale Datensätze zu integrieren, um das latente Potenzial innerhalb klinischer Systeme zu erschließen. Die Forschung analysierte klinische und bildgebende Daten von 1962 bis 2021, überwiegend unstrukturiert, vom Klinikum rechts der Isar. Der Datenbearbeitungsprozess umfasste eine sorgfältige Reinigung und Überprüfung in Zusammenarbeit mit Radiologen und Chirurgen. Die Dissertation basiert auf vier Kernpublikationen. Ein einleitender Übersichtsartikel beleuchtet die Herausforderungen und Einschränkungen des Einsatzes maschinellen Lernens bei der Diagnose muskuloskelettaler Malignome und spricht sich für verbesserte Datenerfassungsrahmen und internationale Netzwerke aus, um die orthopädische Onkologie voranzubringen. Zu den wichtigsten methodischen und anwendungsbezogenen Entwicklungen gehören die erfolgreiche Nutzung von Transferlernen mit einem Datensatz von 42.608 tumorassoziierten, unstrukturierten Röntgenbildern, die Datenknappheit adressieren; die Entwicklung eines neuartigen Sortieralgorithmus mit einer Genauigkeit von 96,6% bei der Klassifizierung von Röntgenbildern in 28 anatomische Regionen und somit eine potenzielle Optimierung der Arbeitsabläufe; sowie ein empfehlungsbasierter Ansatz, der umfangreiche Erfahrungen, die in klinischen Systemen ruhen, nutzt, um neue Patienten mit vorherigen Patienten zu verknüpfen, während Tumoren mit einer durchschnittlichen Genauigkeit von 92,86%

über zehn Entitäten klassifiziert werden und damit aktuelle Modelle um mehr als 30 Prozentpunkte übertreffen. Durch die Nutzung realer klinischer Daten navigiert diese Arbeit durch die Einschränkungen kleiner Datensätze und Datenqualität und ebnet neue Wege für die frühe und präzise Diagnose muskuloskelettaler Tumoren, optimiert radiologische Praktiken, unterstützt Nichtspezialisten und fördert letztendlich eine effektivere, patientenzentrierte Versorgung.

Acknowledgments

My journey into science began at the age of nine, when a birthday present - a microscope - led me to cut my finger and examine my own blood, a curiosity-driven experiment that marked the beginning of a lifelong fascination with science. This early experience laid the foundation for a dream that lay dormant for many years, but has now become a reality with the completion of my doctoral studies. Looking back over the past four years, I am very satisfied with my decision to embark on this journey.

At the forefront of my gratitude is Daniel, my supervisor, whose guidance has been invaluable. Daniel's ability to elegantly guide my research, often with just one insightful sentence, has had a decisive influence on my work. His mentorship has not only enriched my research experience but also significantly contributed to my personal and professional growth. I am immensely proud to be the first doctoral graduate from TUM under his supervision, an honor I cherish greatly. I am equally grateful to my mentor, Rainer, for the profound discussions we've had about the essence of medicine, the potential and impactful role of AI, and the importance of thinking outside the box. Rainer's insights and enthusiasm have been a constant source of inspiration, pushing me to explore innovative approaches in my research. My gratitude extends to Rüdiger, my boss, who has supported me on numerous occasions and created an environment that has been crucial to my growth and success. His encouragement and belief in my abilities have played a significant role in my successes. To my students, who have journeyed with me through this doctoral process, your eagerness to learn and contribute has enriched our projects and my own learning. Collaborating with you has been one of the highlights of my doctoral experience, and I am thankful for the opportunity to have worked alongside such dedicated individuals. I also wish to thank my colleagues for the endless discussions on navigating the challenges of completing a doctorate and for the countless late-night sessions that have bonded us. Your camaraderie and shared insights have made this journey less daunting and more enjoyable. Lastly, my family - Sabine, Georg, Sebastian, SabineW, Lukas, Mona, and Freya - your unshakable support and belief in my potential have been a constant source of motivation. Even when doubts arose about pursuing a doctorate, your confidence

Acknowledgments

in this path was unbroken for me. Your encouragement has been a guiding light, steering me towards this achievement. For that, and for your endless support, I am truly grateful.

To everyone who has been a part of my doctoral journey, thank you. Your support, guidance, and encouragement have been indispensable to my success.

List of Publications

The following publications constitute the core of my cumulative doctoral thesis and are printed in the chapter *Publications*.

1. F. Hinterwimmer, S. Consalvo, J. Neumann, D. Rueckert, R. von Eisenhart-Rothe, and R. Burgkart. “Applications of machine learning for imaging-driven diagnosis of musculoskeletal malignancies - a scoping review.” In: *European Radiology* 32.10 (2022), pp. 7173–7184
2. F. Hinterwimmer, S. Consalvo, J. Neumann, C. Micheler, N. Wilhelm, J. Lang, R. v. Eisenhart-Rothe, R. Burgkart, and D. Rueckert. “From Self-supervised Learning to Transfer Learning with Musculoskeletal Radiographs.” In: *Current Directions in Biomedical Engineering*. Vol. 8. 2. De Gruyter. 2022, pp. 9–12
3. F. Hinterwimmer, S. Consalvo, N. Wilhelm, F. Seidl, R. H. Burgkart, R. von Eisenhart-Rothe, D. Rueckert, and J. Neumann. “SAM-X: sorting algorithm for musculoskeletal x-ray radiography.” In: *European Radiology* 33.3 (2023), pp. 1537–1544
4. F. Hinterwimmer, R. S. Serena, N. Wilhelm, S. Breden, S. Consalvo, F. Seidl, D. Juestel, R. H. Burgkart, K. Woertler, R. von Eisenhart-Rothe, et al. “Recommender-based bone tumour classification with radiographs—a link to the past.” In: *European Radiology* (2024), pp. 1–10

The following publications are related to the content of this thesis but not part of the cumulative thesis; the publications are printed in the *Appendix*.

1. S. Consalvo, F. Hinterwimmer, J. Neumann, M. Steinborn, M. Salzmann, F. Seidl, U. Lenze, C. Knebel, D. Rueckert, and R. H. Burgkart. “Two-Phase Deep Learning Algorithm for Detection and Differentiation of Ewing Sarcoma and Acute Osteomyelitis in Paediatric Radiographs.” In: *Anticancer Research* 42.9 (2022), pp. 4371–4380
2. M. Bloier, F. Hinterwimmer, S. Breden, S. Consalvo, J. Neumann, N. Wilhelm, R. v. Eisenhart-Rothe, D. Rueckert, and R. Burgkart. “Detection and Segmentation

- of Heterogeneous Bone Tumours in Limited Radiographs." In: *Current Directions in Biomedical Engineering*. Vol. 8. 2. De Gruyter. 2022, pp. 69–72
3. F. Hinterwimmer, M. Günther, S. Consalvo, A. Gersing, K. Woertler, R. von Eisenhart-Rothe, J. Neumann, R. Burgkart, and D. Rueckert. "Impact of meta-data in multimodal bone tumour classification." In: *PLOS Digital Health* (2024). submitted 01/2024

The following additional publications were published during the time of my doctoral thesis, but have no significant connection to the topic of this thesis.

1. F. Lenze, F. Hinterwimmer, L. Fleckenstein, I. Lazic, D. Dammerer, R. VON Eisenhart-Rothe, N. Harrasser, and F. Pohlig. "Minimally invasive total hip arthroplasty: a comparison of restoring hip biomechanics with and without a traction table." In: *in vivo* 36.1 (2022), pp. 424–429
2. F. Hinterwimmer, I. Lazic, C. Suren, M. T. Hirschmann, F. Pohlig, D. Rueckert, R. Burgkart, and R. von Eisenhart-Rothe. "Machine learning in knee arthroplasty: specific data are key - a systematic review." In: *Knee Surgery, Sports Traumatology, Arthroscopy* 30.2 (2022), pp. 376–388
3. C. Zanzinger, N. Harrasser, O. Gottschalk, P. Dolp, F. Hinterwimmer, H. Hoerterer, and M. Walther. "One-year Follow-Up Results with Hydrogel Implant in Therapy of Hallux Rigidus: Case Series with 44 Patients." In: *Zeitschrift für Orthopädie und Unfallchirurgie* 160.04 (2022), pp. 414–421
4. D. M. Hedderich, M. Keicher, B. Wiestler, M. J. Gruber, H. Burwinkel, F. Hinterwimmer, T. Czempel, J. E. Spiro, D. Pinto dos Santos, D. Heim, et al. "AI for Doctors - a course to educate medical professionals in artificial intelligence for medical imaging." In: *Healthcare*. Vol. 9. 10. MDPI. 2021, p. 1278
5. F. Hinterwimmer, I. Lazic, S. Langer, C. Suren, F. Charitou, M. T. Hirschmann, G. Matziolis, F. Seidl, F. Pohlig, D. Rueckert, et al. "Prediction of complications and surgery duration in primary TKA with high accuracy using machine learning with arthroplasty-specific data." In: *Knee Surgery, Sports Traumatology, Arthroscopy* 31.4 (2023), pp. 1323–1333
6. I. Lazic, F. Hinterwimmer, S. Langer, F. Pohlig, C. Suren, F. Seidl, D. Rückert, R. Burgkart, and R. von Eisenhart-Rothe. "Prediction of Complications and Surgery Duration in Primary Total Hip Arthroplasty Using Machine Learning: The Necessity of Modified Algorithms and Specific Data." In: *Journal of Clinical Medicine* 11.8 (2022), p. 2147

7. R. von Eisenhart-Rothe, F. Hinterwimmer, H. Graichen, and M. T. Hirschmann. "Artificial intelligence and robotics in TKA surgery: promising options for improved outcomes?" In: *Knee Surgery, Sports Traumatology, Arthroscopy* 30.8 (2022), pp. 2535–2537
8. N. J. Wilhelm, S. Haddadin, J. J. Lang, C. Micheler, F. Hinterwimmer, A. Reiners, R. Burgkart, and C. Glowalla. "Development of an Exoskeleton Platform of the Finger for Objective Patient Monitoring in Rehabilitation." In: *Sensors* 22.13 (2022), p. 4804
9. I. Lazic, F. Hinterwimmer, and R. von Eisenhart-Rothe. "Vorhersage von irregulären Operationsdauern bei Knieendoprothesen mit Daten aus dem Endoprothesenregister Deutschland und EndoCert." In: *Knie Journal* 4.4 (2022), pp. 224–229
10. C. M. Micheler, J. J. Lang, N. J. Wilhelm, I. Lazic, F. Hinterwimmer, C. Fritz, R. v. Eisenhart-Rothe, M. F. Zäh, and R. H. Burgkart. "Scaling Methods of the Pelvis without Distortion for the Analysis of Bone Defects." In: *Current Directions in Biomedical Engineering*. Vol. 8. 2. De Gruyter. 2022, pp. 797–800
11. J. J. Lang, V. Baylacher, C. M. Micheler, N. J. Wilhelm, F. Hinterwimmer, B. Schwaiger, D. Barnewitz, R. v. Eisenhart-Rothe, C. U. Grosse, and R. Burgkart. "Improving Equine Intramedullary Nail Osteosynthesis via Fracture Adjacent Polymer Reinforcement." In: *Current Directions in Biomedical Engineering*. Vol. 8. 2. De Gruyter. 2022, pp. 129–132
12. S. Consalvo, F. Hinterwimmer, N. Harrasser, U. Lenze, G. Matziolis, R. von Eisenhart-Rothe, and C. Knebel. "C-Reactive Protein Pretreatment-Level Evaluation for Ewing's Sarcoma Prognosis Assessment - A 15-Year Retrospective Single-Centre Study." In: *Cancers* 14.23 (2022), p. 5898
13. N. Harrasser, F. Hinterwimmer, S. Baumbach, K. Pfahl, C. Glowalla, M. Walther, and H. Hörterer. "The distal metatarsal screw is not always necessary in third-generation MICA: a case-control study." In: *Archives of Orthopaedic and Trauma Surgery* (2022), pp. 1–7
14. C. Glowalla, S. Langer, U. Lenze, I. Lazic, M. T. Hirschmann, F. Hinterwimmer, R. von Eisenhart-Rothe, and F. Pohlig. "Postoperative full leg radiographs exhibit less residual coronal varus deformity compared to intraoperative measurements in robotic arm-assisted total knee arthroplasty with the MAKO™ system." In: *Knee Surgery, Sports Traumatology, Arthroscopy* (2023), pp. 1–7

15. D. Jüstel, H. Irl, F. Hinterwimmer, C. Dehner, W. Simson, N. Navab, G. Schneider, and V. Ntziachristos. "Spotlight on nerves: Portable multispectral optoacoustic imaging of peripheral nerve vascularization and morphology." In: *arXiv preprint arXiv:2207.13978* (2022)
16. S. Breden, F. Hinterwimmer, S. Consalvo, J. Neumann, C. Knebel, R. von Eisenhart-Rothe, R. H. Burgkart, and U. Lenze. "Deep learning-based detection of bone tumors around the knee in X-rays of children." In: *Journal of Clinical Medicine* 12.18 (2023), p. 5960
17. S. Breden, F. Hinterwimmer, S. Beischl, S. Consalvo, A. S. Gersing, U. Lenze, R. von Eisenhart-Rothe, and C. Knebel. "A New Method for Assessing Patients' Obesity-Associated Infection Risk Using X-rays in Hip Arthroplasties." In: *Journal of Clinical Medicine* 12.23 (2023), p. 7277
18. N. Wilhelm, C. M. Micheler, J. J. Lang, F. Hinterwimmer, V. Schaack, R. Smits, S. Haddadin, and R. Burgkart. "Development and Evaluation of a Cost-effective IMU System for Gait Analysis: Comparison with Vicon and VideoPose3D Algorithms." In: *Current Directions in Biomedical Engineering*. Vol. 9. 1. De Gruyter. 2023, pp. 254–257

Contents

Abstract	iii
Zusammenfassung	iv
Acknowledgments	vi
List of Publications	viii
1. Introduction	1
1.1. Orthopaedic Oncology	1
1.1.1. Clinical Features of Musculoskeletal Tumours	2
1.1.2. Diagnostic Workflow at Tumour Centre	3
1.1.3. Obstacles in Tumour Assessment	5
1.2. Deep Learning for Image Analysis	7
1.2.1. A Brief History of Deep Learning: Pioneers, Milestones, and Evolution	7
1.2.2. State-of-the-art in Deep Learning	8
1.2.3. Current Limitations and Implications for Medicine	10
1.3. Subject of this Dissertation	13
2. Materials and Methods	15
2.1. Patient Cohort and Data	15
2.2. Current State of Machine Learning for Musculoskeletal Tumour Diagnostics	20
2.3. Handling Limited Datasets	23
2.4. Leveraging Unstructured Data for Workflow Optimization	26
2.5. Multimodal Data for Diagnostic Decision Support Tools	28
3. Discussion	31
3.1. Cross-thematic discussion	31
3.2. Future Work	35
3.3. Conclusion	36

4. Publications	38
4.1. Applications of machine learning for imaging-driven diagnosis of musculoskeletal malignancies - a scoping review	38
4.2. From Self-supervised Learning to Transfer Learning with Musculoskeletal Radiographs	54
4.3. SAM-X: sorting algorithm for musculoskeletal x-ray radiography	61
4.4. Recommender-based Bone tumour Classification – a Link to the Past . .	73
Abbreviations	87
List of Figures	89
List of Tables	90
A. Related publications	91
Bibliography	128

1. Introduction

1.1. Orthopaedic Oncology

Orthopaedic oncology is a speciality of orthopaedic surgery that focuses on the diagnosis, treatment and care of bone and soft tissue tumours. These tumours can arise in various anatomical sites, including bone, muscle, tendon, ligament and other connective tissues [26]. The primary goal of orthopaedic oncology is to provide optimal care for patients with musculoskeletal (MSK) tumours, aiming for both limb salvage and optimal long-term functional outcomes [27]. This requires a multidisciplinary approach involving collaboration between orthopaedic surgeons, medical oncologists, radiologists, pathologists and other medical professionals [1]. The specialty encompasses a wide range of surgical techniques, including tumour resection, reconstruction and limb-sparing procedures tailored to the individual needs of the patient and the characteristics of the tumour [27].

Imaging plays an important role in the diagnosis of bone and soft tissue tumours. For the assessment of bone lesions for example, the most common imaging modalities used are standard radiographs, computed tomography (CT) and magnetic resonance imaging [28]. Even recently, the Musculoskeletal Tumour Society and American Academy of Orthopedic Surgeons working group affirmed plain radiography as the initial screening for possible bone tumours [28, 29]. Patients with a suspected malignant lesion should be referred to an MSK tumour centre based on their radiograph only to avoid delaying treatment. The additional diagnostic information provided by CT and magnetic resonance imaging (MRI) is moderate (for early assessment!) and should not delay medical care. The relevance of further imaging studies should be determined at a referral centre before a biopsy is performed [28].

Experts in tumour centres are trained to make precise diagnoses through a combination of clinical assessment, imaging studies and histopathological analysis. They use advanced imaging techniques such as MRI, CT and positron emission tomography (PET) to accurately locate and characterise the tumours. They also play a crucial role in post-operative management by monitoring patients for signs of therapy effectiveness, recurrence or metastasis. Through continuous research and technological advances, orthopaedic oncology is constantly evolving and improving outcomes such as survival rate and quality of life of patients affected by bone and soft tissue tu-

mours. Nevertheless, for several reasons (discussed in the following sections), more advanced approaches for e.g. image analysis or processing of complex multimodal data is required.

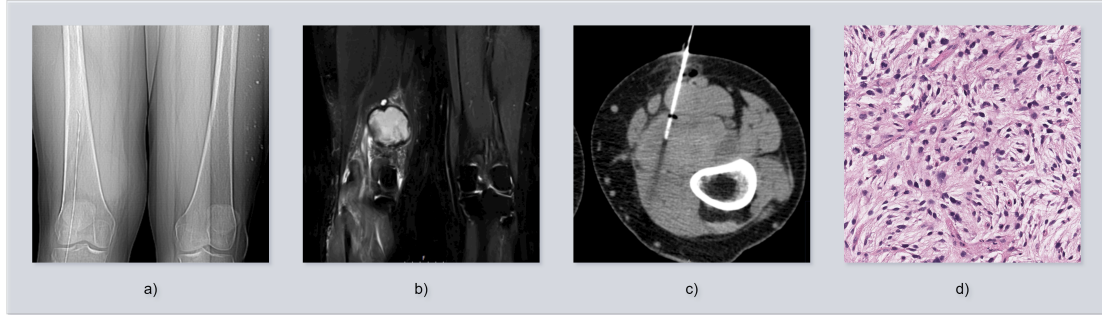


Figure 1.1.: Exemplary radiological (a-c) and pathological (d) imaging data of a musculoskeletal tumour patient: a) preoperative x-ray, b) preoperative MRI, c) CT-guided needle biopsy, d) histology.

1.1.1. Clinical Features of Musculoskeletal Tumours

MSK tumours have a variety of clinical features that can vary depending on tumour type, location, and stage [26]. Recognising these clinical features is critical for early detection, precise diagnosis, and appropriate treatment of MSK tumours. The following chapter discusses the clinical manifestations commonly seen in patients with MSK tumours.

Picci et al. [26] describe the symptoms in their book *Diagnosis of Musculoskeletal tumours and tumour-like Conditions* with the first case from 1900 and more than 47,000 cases overall as follows: Pain and swelling are among the most common symptoms of MSK tumours. The pain can range from mild discomfort to severe, persistent pain and may be limited to the tumour site or radiate to neighbouring tissue. Swelling is another common feature, often associated with increased vascularity and oedema within the tumour. Patients with MSK tumours often report the presence of a palpable mass or lump. The mass may be firm and immobile and gradually increase in size. Depending on the origin of the tumour, it may be localised in the bones, soft tissues, or both. MSK tumours may feel painful, especially when pressure is applied directly to the tumour or during movement. Tenderness may be accompanied by localised warmth or erythema in some cases. Tumours affecting the bones often cause bone pain, which can be severe and worsen at night. This pain can be localised or radiate along the affected bone. In addition, the weakened bone structures due to tumour infiltration can lead to

pathological fractures even with minor trauma. As MSK tumours grow and progress, they can lead to functional impairment in the affected area. Patients may experience restricted joint movement, muscle weakness, and difficulty in performing activities of daily living. In some cases, MSK tumours may be associated with constitutional symptoms such as fatigue, unwanted weight loss, and night sweats. These systemic symptoms are more common in aggressive or metastatic tumours. Tumours that are close to nerves or compress neural structures can cause neurological deficits. Depending on the location and size of the tumour, patients may experience sensory disturbances, muscle weakness, loss of reflexes, or even paralysis.

These entities are often detected incidentally due to lack of specific symptoms and their non-specific presentation [30, 31]. A typical scenario is that a child has injured himself/herself while playing and a possible mass is discovered during the subsequent routine X-ray examination to check the bone structure. Certainly, this is not an ideal way to identify MSK malignancies.



1.1.2. Diagnostic Workflow at Tumour Centre

The diagnostic workflow for MSK tumours involves a comprehensive evaluation that includes clinical assessment, imaging studies, biopsy, histopathological analysis and multidisciplinary discussions (=weekly tumour board). This systematic approach ensures precise diagnosis, staging and prognostic evaluation and facilitates appropriate treatment planning for patients with MSK tumours [26, 27, 32].

1. **Clinical assessment:** The diagnostic workflow begins with a thorough clinical examination, including a detailed history and physical examination. The doctor assesses the patient's symptoms, duration of symptoms, associated risk factors and any history of cancer or genetic syndromes. The physical examination is primarily concerned with determining the location, size, consistency, painfulness and patient mobility with tumour.
2. **Laboratory tests:** Laboratory tests complement the diagnostic process by providing additional information. Blood tests, including complete blood count, liver function tests, renal function tests and tumour markers (such as alkaline phosphatase, lactate dehydrogenase or prostate specific antigen) can be performed to assess general health, organ function and potential markers of tumour activity. Further evaluation through laboratory tests is still under research [19].
3. **Imaging examinations:** Imaging plays a crucial role in the diagnosis of MSK tumours. Various imaging techniques are used to assess the location, size and morphology of the tumour and its relationship to neighbouring tissue. Commonly

1. Introduction

**INTERDISZIPLINÄRES ZENTRUM
für Knochen- und Weichteiltumore
an der
Klinik für Orthopädie und Sportorthopädie
Klinikum rechts der Isar
der Technischen Universität München
gegründet 1972**



TUMORKONFERENZ (Protokoll)
**Freitag, den 07.01.2022, 7:30 Uhr im Hörsaal der Pathologie im Institut für Pathologie
und pathologische Anatomie, Klinikum rechts der Isar der TUM**

Teil I Malignome und malignitätsverdächtige Tumoren

23.11.1940, ORPOL, M, 81 **OP geplant 12.01.2021**

A: Seit ca. November 2021 Schwellung an der lateralen linken Schulter.
12.11.2021 MRT allo loco: großen Lipom DD ALT zwischen M. infraspinatus und M. deltoides, kleines Lipom im proximalen M. trizeps brachii

D: V.a. intramuskuläres Lipom DD ALT linke Schulter zwischen M. infraspinatus und M. deltoides und M. trizeps brachii
Fragestellung: Prä-operative Vorstellung. Weiteres Procedere?
E: Bildgebend V.a. Lipom linke Schulter
Procedere: Marginale Resektion

Seite 1 von 10

12.12.1977, ORAMB, W, 44

A: Im April 2021 frustrierte Resektion eines Chondrosarkoms Hüfte rechts in Ägypten ohne vorherige Biopsie.
Bei intraoperativ nicht möglicher Resektion des Tumors und starker Blutung erfolgte nach partieller Curettage eine Defektfüllung mit Zement (histologisch Chondrosarkom)
Adjuvante RTx 25 x 2 Gy

Seit der OP bestehen starke Schmerzen.
Vorstellung in Deutschland bei Prof. Hempel
Ägyptische Proben zur Referenzpathologie nach Kaulbeuren
(Differenzierung eines Chondrosarkoms und eines chondroblastischen Osteosarkoms kann allerdings nicht allein auf Basis der histologischen Schnittpräparate erfolgen. Es müssen hier auch entsprechende radiologische Befunde vorliegen, die dann eine Unterscheidung dieser beiden Entitäten ermöglichen.)
10.12.2021 Sonographische Probenentnahme durch Prof. Hempel.
Im PET-CT kein Hinweis für Metastasen mit osteodestruktiven/postoperativen Veränderungen des Acetabulum rechts.

D: Chondrosarkom G2 Hüfte rechts bei partieller intraläsionaler Resektion und Defektfüllung mit Zement in Ägypten 04/21
Z.n. Adjuvanter Strahlentherapie
Fragestellung: Externe Histologie. Schnitte. Bildgebung. Weiteres Procedere?
E: Bildgebend V.a. Chondrosarkom Acetabulum bis Schambeinast und vorderer Hüftpfeller rechts.
V.a. Großen intraartikulären Resttumor (Iliopsoas)
Histologisch Chondrosarkom G2
Procedere: 1. Prä-operatives MRT (Versuch bei adipsitas per magna)
2. externe Hemipelvectomy

11.02.1986, ORPOL, M, 35 **OP geplant 07.02.2021**

A: Seit ca. 2016 intermittierende, belastungsabhängige Schmerzen, Schwellung und Druckgefühl an der distalen Wadenmuskulatur links
15.09.2021 MRT Unterschenkel links: V.a. Adamantinom dist. Linker Tibiaschaft
16.09.2021 Offene Biopsie (histologisch atypisches Ewing Sarkom)

20220107_07. Januar 2022 Tumorkonferenz Seite 2 von 10

Figure 1.2.: Example protocol from weekly MSK tumour board meeting.

used imaging techniques include X-rays, CT, MRI and PET. X-rays provide valuable information about bone lesions, while CT and MRI provide detailed anatomical imaging of both bone and soft tissue tumours. PET scans help assess metabolic activity and detect possible metastases. Nuclear imaging techniques are also used in some specialised centres. However, their diagnostic value has yet to be proven [26].

4. **Staging and prognostic evaluation:** After diagnosis, the tumour is staged to determine its extent and possible spread. Staging helps to make treatment decisions and provides prognostic information. Different staging systems are used for MSK tumours, e.g. the TNM classification system (tumour, node, metastases). Prognostic factors such as tumour grade, size, histological subtype and molecular markers are also taken into account to estimate the patient's overall prognosis.
5. **Biopsy and histopathological analysis:** A definitive diagnosis of MSK tumours is made by biopsy, where tissue samples are taken for histopathological analysis. Depending on the nature and location of the tumour, different biopsy techniques such as needle biopsy, CT guided needle biopsy, incisional biopsy or excisional biopsy may be used. The tissue samples taken are sent to a pathologist who examines them under the microscope to determine the histological type, grade and other important features of the tumour.
6. **Genetic testing:** In some cases, genetic testing may be warranted to identify specific genetic alterations or mutations associated with certain MSK tumours. This information can provide valuable insight into tumour behaviour, prognosis and possible treatment options.
7. **Multidisciplinary team discussion:** The complexity of diagnosing and especially treating MSK tumours often requires a multidisciplinary approach. A team of orthopaedic surgeons, oncologists, radiologists, pathologists and other specialists review the patient's clinical data, imaging results and histopathological findings. This joint discussion helps formulate an precise diagnosis, determine the stage of the tumour and develop an individualised treatment plan.

1.1.3. Obstacles in Tumour Assessment

The diagnosis of MSK tumours faces obstacles related to late detection, heterogeneous subtypes (>100 subtypes according to WHO [33]), need for multimodality assessment [1], susceptibility to diagnostic errors [34], limited experience of general practitioners [35, 29], low incidence [26] and specific challenges in paediatric cases [36, 5].

Overcoming these barriers requires increased awareness, specialised training, multi-disciplinary collaboration, access to advanced diagnostic resources and **development of new methodologies for precise diagnosis**, ultimately leading to better patient outcomes in the diagnosis and treatment of MSK tumours.

- **Late detection and delayed referral:** MSK tumours often present with non-specific symptoms, leading to delayed recognition and up to 12 months delayed referral to specialised centres [35]. Patients may initially associate their symptoms with more general MSK conditions, which leads to late consultation with a doctor. Late recognition and referral to a specialised centre can lead to advanced disease stage and limited treatment options, negatively impacting patient outcomes [29].
- **Heterogeneous subtypes and clinical presentations:** MSK tumours encompass a wide range of histological subtypes with different clinical presentations. The variability of tumour types and their presentations can complicate diagnosis, as symptoms may overlap with benign disease. Differentiation between malignant and benign lesions requires specialised expertise and accurate histopathological analysis [33].
- **Multimodal assessment:** Accurate diagnosis of MSK tumours often requires a multimodal assessment that includes clinical data, imaging studies and histopathological analysis. Interpretation and correlation of data from different modalities such as radiographs, CT scans, MRI, PET scans and biopsy results requires extensive expertise and access to sophisticated imaging techniques that only specialised tumour centres have, but that general practitioners and hospitals in less developed and wealthy countries, for example, usually do not [1, 26, 34].
- **Error-prone diagnostic process:** The diagnostic process for MSK tumours is error-prone and can lead to misinterpretation. False-negative or false-positive results can occur, leading to incorrect diagnoses and inappropriate treatment decisions. The complexity of diagnosing MSK tumours, combined with the rarity of these tumours, increases the risk of diagnostic errors [34].
- **Limited experience of general practitioners:** General practitioners may have limited experience in diagnosing MSK tumours due to the low incidence and specialisation of these tumours. As a result, early detection and precise diagnosis of these tumours can be challenging in general care. Referral to specialised tumour centres is critical to ensure precise diagnosis and appropriate treatment [37, 1, 4, 7].
- **Low incidence and familiarity:** MSK tumours are relatively rare compared to other cancers, making them less familiar to healthcare providers. The low

incidence can lead to a lack of awareness and expertise, resulting in delayed or overlooked diagnoses. Better education and awareness, as well as access to specialised tumour centres, are essential to address this challenge [26, 1].

- **MSK tumours in children:** MSK tumours are one of the most common cancers in children. Diagnosing tumours in paediatric patients requires additional considerations as tumour types, presentation and treatment methods differ from those in adults. Specialised paediatric oncology and age-specific therapeutic protocols are necessary to provide optimal care for paediatric MSK tumours [5, 36].

1.2. Deep Learning for Image Analysis

1.2.1. A Brief History of Deep Learning: Pioneers, Milestones, and Evolution

Deep learning (DL), a subset of machine learning (ML), has gained tremendous popularity and attention in the 21st century due to its ability to solve complex tasks across various domains, such as image and speech recognition, natural language processing (NLP), and reinforcement learning [38]. DL is based on artificial neural networks (ANNs), computational models inspired by the structure and functioning of the human brain [39]. The concept of ANNs dates back to the 1940s, with the pioneering work of Warren McCulloch and Walter Pitts. In their seminal 1943 paper [40], they proposed a simplified model of biological neurons and demonstrated that these artificial neurons could perform logical computations. This work laid the foundation for further research in the field of artificial intelligence (AI) and neural networks. In 1958, Frank Rosenblatt introduced the perceptron [41], an early ANNs that could learn to classify linearly separable patterns through supervised learning. Despite the perceptron's initial promise, Marvin Minsky and Seymour Papert's 1969 book, "Perceptrons" [42], demonstrated its limitations and criticized its inability to solve more complex, non-linear problems. This critique contributed to the decline in neural network research for a period of time, known as the "AI Winter."

Interest in neural networks was revived in the 1980s, fueled by the development of the backpropagation algorithm by Geoffrey Hinton, David Rumelhart, and Ronald Williams [43]. This algorithm allowed for efficient training of multi-layer perceptrons, overcoming the limitations highlighted by Minsky and Papert. Yann LeCun's work in the late 1980s and early 1990s led to the development of the convolutional neural networks (CNNs), a key milestone in DL. CNNs are characterized by their ability to process grid-like data, such as images, by incorporating convolutional layers that can learn local features, thus enabling the automatic extraction of hierarchical features. LeCun's LeNet-5 architecture, developed for handwriting recognition, exemplified the

potential of CNNs [44].

Recurrent neural network (RNNs) [45], which possess the ability to process sequences of data, were introduced by John Hopfield and David Rumelhart in the 1980s. Later, in 1997, Sepp Hochreiter and Jürgen Schmidhuber proposed the long short-term memory (LSTM) architecture [46], addressing the vanishing gradient problem in RNNs and enabling the processing of longer sequences. The advent of powerful graphical processing units (GPUs), big data, and improved algorithms spurred the resurgence of DL in the 21st century. In 2012, Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton achieved a groundbreaking result in the *ImageNet Large Scale Visual Recognition Challenge* using a deep CNNs called AlexNet [47]. This event marked the beginning of the modern DL era. Recent years have witnessed the development of more advanced DL architectures, such as residual neural networks (ResNets) [48], generative adversarial networks (GANs) [49], and transformers [50], which have driven significant breakthroughs in various domains.

1.2.2. State-of-the-art in Deep Learning

Deep Learning has made significant advances in image analysis, revolutionising computer vision tasks and achieving excellence in several areas. In this section, the current state of the art in DL for image analysis will be described and the key techniques and models that have advanced the field will be highlighted.

CNNs have been instrumental in making breakthroughs in image analysis. CNNs use convolutional layers that enable hierarchical feature extraction, allowing the network to capture spatial and compositional information in images. Architectures such as ResNet [48], Inception [51] and EfficientNet [52] have pushed the boundaries of CNNs performance and achieved high accuracy in tasks such as image classification.

Transfer learning and pre-trained models have played a crucial role in image analysis tasks, especially when there is limited labelled data. Pre-trained models trained on large image datasets such as ImageNet [53, 47] have learned extensive feature representations that can be fine-tuned for specific tasks. This approach has greatly reduced the need for extensive labelled data and computational resources, and allows researchers and practitioners to efficiently build powerful models. Common pre-trained models include VGGNet [54], ResNet [48] and EfficientNet [52]. Pre-training followed by fine-tuning has proven effective in improving performance also on various language tasks.

DL has led to remarkable advances in object detection and localisation, enabling accurate identification of objects in images in real time. Architectures such as Faster R-CNN [55] and YOLO [56] have pioneered object recognition by combining Region Proposal Networks and convolutional features to recognise and classify objects. These models have applications in autonomous driving, surveillance systems and object

recognition.

Semantic segmentation is about labelling objects and their boundaries at the pixel level within an image. DL has made significant progress in this area, with models such as U-Net [57], SegNet [58] and DeepLab [59] achieving remarkable performance. These models use fully convolutional architectures and skip connections to capture fine-grained details and spatial relationships, enabling accurate segmentation of objects in images. Semantic segmentation has applications in medical imaging, autonomous systems and augmented reality.

Generative models, especially GANs and variational autoencoders (VAEs), have demonstrated their capabilities in image synthesis tasks. GANs produce realistic images by training a generator network to produce patterns that fool a discriminator network, leading to visually compelling results. Models such as DCGAN [60] and StyleGAN [61] have achieved impressive results in image synthesis and enable applications for artistic style transfer, image-to-image translation and data augmentation.

DL models have also made progress in tasks that involve the fusion of image and language. Image labelling models generate natural language descriptions of images, while visual question answering models answer questions based on visual content. Approaches such as *Show and Tell*, *Show, Attend and Tell*, and bottom-up and top-down attention mechanisms have improved the quality and accuracy of generated captions. Visual question answering models combining CNNs with recurrent networks have achieved impressive results in answering questions about images.

BERT [62] is an example of a pre-trained language model that can be fine-tuned for tasks such as answering questions and classifying sentences. Large language models (LLMs) such as generative pre-trained transformer (GPT) trained on large web corpora have achieved peak performance in translation, question answering, essay writing and program generation. Research has also focused on fine-tuning and transfer learning to improve performance on specific tasks with smaller datasets. LLMs are not the focus of this dissertation, but are still worth mentioning due to the enormous impact in society and research at this time.

As DL models are used in critical applications, it is crucial that they are robust against outside attacks and ensure interpretability of the models. Researchers are actively working to develop robust models that are resistant to interference from attackers in order to improve the reliability and security of the models. Efforts are also being made to improve the interpretability of DL models to enable better understanding and confidence in the decisions made by these models.

1.2.3. Current Limitations and Implications for Medicine

While it is crucial for research in this field to understand the extensive possibilities of modern DL methods, it is at least as crucial to know their limitations. In a field as specific and highly sensitive as medicine, certain limitations come with specific implications. The following section discusses the current limitations and specific implications for the field of medicine.

Data dependency and labelling

Models require large amounts of labelled data for training in order to effectively learn and generalise from underlying patterns in the data [54, 53, 48]. However, in medical domains, such as rare diseases (e.g. MSK tumours) or special patient populations, it can be particularly difficult to obtain large annotated datasets due to limited data availability. This scarcity of labelled data makes it difficult to develop and deploy accurate and reliable models in various medical applications. The implications of this limitation are multi-faceted. First, the lack of sufficiently labelled data may limit the performance and generalisability of DL models in medical scenarios [63]. Without access to representative and diverse datasets, models may not be able to capture the full spectrum of disease conditions, resulting in suboptimal diagnostic accuracy and treatment recommendations. Second, the process of labelling medical data requires a significant amount of expertise and time on the part of professionals [3]. Manual annotation is often required to ensure accuracy and reliability of the labels. The labour-intensive nature of this task can slow down the development and use of AI models in medicine, making it difficult to keep up with rapidly evolving medical knowledge and new healthcare demands. In addition, the limited availability of labelled data in certain medical fields can lead to unbalanced datasets in which certain classes or conditions are underrepresented [1, 3]. This can lead to biased model performance, potentially aggravating inequalities in healthcare and jeopardising the equitable provision of health services. Constraining data dependency and labelling in DL for medicine is critical. Innovative approaches such as transfer learning, active learning and data augmentation can help alleviate the problem of data scarcity by effectively using existing labelled data and generating synthetic data to complement the training process [6, 5, 2]. Collaborative efforts between healthcare institutions, researchers and regulators are essential to facilitate data sharing and create standardised datasets that benefit the development of AI models in medicine.

Lack of explainability

DL models are characterised by their complex and hierarchical structures, consisting of numerous interconnected layers. While these models can achieve remarkable accuracy in various medical tasks, such as diagnosing diseases or predicting treatments, the lack of explainability poses a challenge when it comes to understanding how and why the models arrive at their decisions [64, 65]. This limitation has several implications for the field of medicine. In healthcare, trust and acceptance of computer models are of utmost importance. Medical professionals and patients must have confidence in the decisions made by AI models. The lack of explanation can affect the trustworthiness of these models, as their internal workings remain opaque, even for experts. Without clear explanations, medical professionals may continue to be reluctant to rely on DL models for critical decision-making processes, potentially limiting their use in clinical practice. Furthermore, models have the potential to provide valuable decision support in the clinical setting. However, without clear explanations of their predictions, it becomes difficult for healthcare professionals to assess the validity and reliability of the model recommendations [66]. Explainability is particularly important when it comes to patient-specific medical interventions, as it helps clinicians understand the rationale behind the model's suggestions and adjust treatment accordingly. The lack of explainability raises ethical concerns, especially in sensitive medical scenarios. When DL models are used to make decisions with significant consequences for patients, such as treatment plans or allocation of healthcare resources, the ability to provide transparent explanations becomes essential. Explainability is crucial to ensure fairness, avoid bias and meet legal and ethical requirements related to accountability and transparency in medical decision-making. In medicine, it is important to understand the factors and features that contribute to the predictions of an AI model. Without explainability, it becomes difficult to assess the safety and reliability of the model's results. In cases where DL models are used for critical tasks such as detecting adverse events or predicting patient outcomes, the inability to explain their decision-making process can hinder the identification of potential risks or errors.

Generalisation to unseen or out-of-distribution data

DL models excel at learning patterns and making accurate predictions within the constraints of the training data they have been given. However, when faced with data that deviates significantly from the training distribution, such as new disease manifestations or patient populations that were not adequately represented during training, these models may have difficulty generalising effectively. This limitation has several implications for the field of medicine. Models used for disease diagnosis are

highly dependent on their ability to generalise well to unseen data. However, if a model has not encountered certain disease variants or rare conditions during training, it may not be able to accurately diagnose real cases. This limitation may affect the reliability and effectiveness of DL models in clinical practice, leading to misdiagnosis or delayed treatments. DL is increasingly used to assist in treatment planning and decision-making. However, when patient-specific data are available that differ significantly from the training distribution, such as comorbidities or unique genetic profiles, the model's recommendations may not be reliable or optimal. This limitation may hinder the potential benefits of DL in personalised medicine, where tailored treatment strategies are critical. The limited generalisability of DL models can affect their performance and applicability in the real world [67]. Models that demonstrate high accuracy when evaluated on benchmark datasets may underperform or exhibit unexpected behaviour when used in real-world clinical settings. This limitation may hinder the adoption and acceptance by healthcare professionals who need robust and reliable performance for decision-making. Limiting generalisation to unseen or undistributed data in DL for medicine is an active area of research. Efforts are being made to develop strategies that improve model generalisation, such as incorporating diverse and representative training data, data augmentation techniques and transfer learning approaches [6]. In addition, advances in domain adaptation and robustness techniques aim to improve the ability of DL models to deal with data distributions that deviate from the training data.

Integration into clinical workflow

Integrating DL models into the clinical workflow presents challenges, including integration with electronic health records, real-time decision making and clinical validation. Integration should be seamless and practical to ensure that the models provide meaningful and actionable insights for healthcare providers without disrupting established clinical processes [68, 69].

Hardware resources

Training DL models, especially large-scale architectures, requires significant computational resources, including high-performance GPUs or specialised hardware [70]. This need for computing power can be a significant barrier for researchers and organisations with limited access to such resources. Efficient model architectures and training techniques are being explored to reduce the computational demands.

Ethical and bias concerns

DL models, while powerful and versatile, are not immune to ethical concerns and biases, particularly in the field of medicine [71]. These concerns stem from several factors, including biased training data, opacity of model decisions, and the potential for discriminatory outcomes [72]. These limitations have significant implications for the responsible and equitable application of DL in medicine. AI can lead to discriminatory results in medical applications if they are not properly screened for bias. For example, a model trained on biased data may disproportionately misdiagnose or underdiagnose certain groups of patients based on factors such as race, gender or socioeconomic status. This can lead to unequal access to health resources, unequal treatment outcomes and the perpetuation of systemic biases in healthcare [73].

In medicine, DL models often rely on sensitive patient data for training and inference. Privacy concerns [74] and the need for informed consent are critical when using these models. Appropriate measures must be taken to ensure patient privacy and data protection, as well as transparent communication about the use of patient data for model development and evaluation. Limiting ethical concerns and bias in DL for medicine is imperative to address. This requires a multi-faceted approach that includes data collection practices that prioritise diversity and fairness, robust strategies to detect and mitigate bias in training data, transparency and explainability techniques to improve model interpretability, and ethical guidelines for the development and use of DL models in medicine [75]. In addition, interdisciplinary collaboration between computer scientists, healthcare professionals, ethicists and policy makers is essential to establish legal frameworks, standards and guidelines that promote fairness, transparency and accountability in the development and use of DL models in medicine. Responsible use of DL models can help mitigate biases, ensure equitable access to healthcare and improve patient outcomes.

1.3. Subject of this Dissertation

The main topic of the present work was to find ways to cope with the data of MSK tumour patients from several decades at Klinikum rechts der Isar and to develop analysis and support tools mainly for diagnostic purposes based on state-of-the-art ML methods for the purpose of decreasing the *time to diagnosis*. Due to the low incidence of MSK tumours in general, very little data is available. In addition, the workflow for data acquisition and processing for ML and other research areas require extensive domain knowledge from different medical fields such as orthopaedics, radiology, and pathology. This implies that data is heterogeneous and multimodal, which is another challenge for ML applications and that interdisciplinary collaboration is crucial to develop clinically

1. Introduction

relevant research questions. Therefore, the four main chapters of this thesis contain the following major contributions:

- Current State of Machine Learning for Musculoskeletal Tumour Diagnostics
- Handling Limited Datasets
- Leveraging Unstructured Data for Workflow Optimization
- Multimodal Data for Diagnostic Decision Support Tools

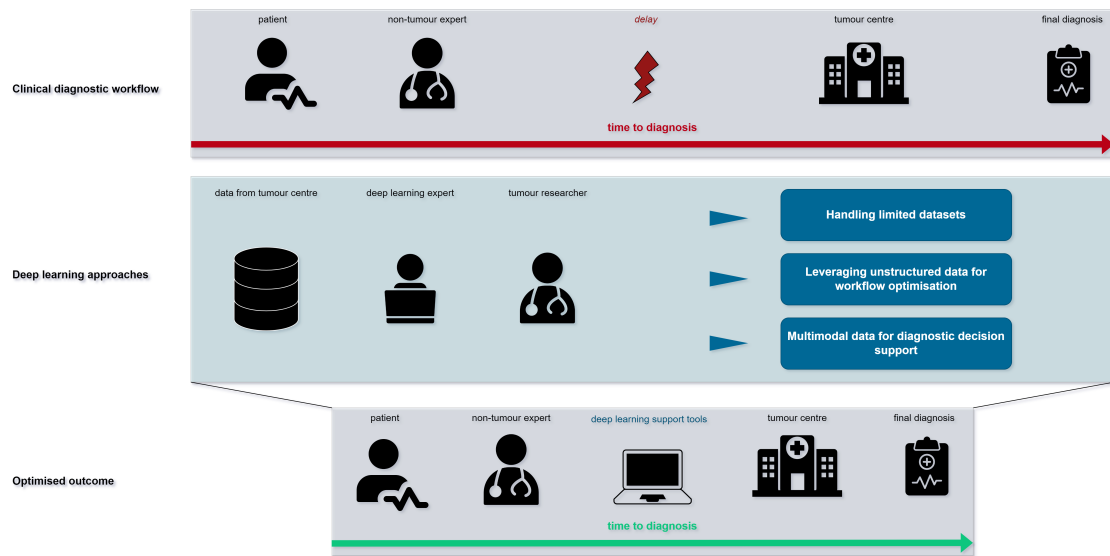


Figure 1.3.: proposed approach of deep learning methods for musculoskeletal tumour diagnostics: the overall problem of musculoskeletal tumour diagnostic is that early diagnostics is complex and delays occur. Deep learning models can help based on retrospective data to handle limited data, optimise workflows and finally build diagnostic support tools.

2. Materials and Methods

2.1. Patient Cohort and Data

Source of data

The present dissertation relies solely on retrospective data; however, this data was derived from real clinical sources. The clinical data utilized in this study was obtained from our hospital information system (HIS), while the imaging data was retrieved from our picture archiving and communication system (PACS). The dataset spans from the initial entries dating back to 1962, although the quality and completeness of the data significantly improved starting in 2004. This enhancement coincided with the implementation of a new PACS at Klinikum rechts der Isar. The dataset encompasses information until the year April 2021. It is important to note that follow-up data is only available for patients who sought further treatment or passed away within our hospital. Approximately 10% of the patients included in this study were referred to us by general practitioners or originated from foreign countries, thus, these patients' data can be referred to as "external data".

All following studies were approved by the local ethics board and conducted as per national and international guidelines. Informed consent was waived due to the studies' retrospective and anonymised nature.

Patients

The patient cohort encompasses 8,377 patients with 8,922 cases (as of April 2021) diagnosed with various types of bone and soft tissue tumours, comprising benign, intermediate, and malignant tumours. Here a case is representing a economic parameter and is defined, among other things, by whether a patient was admitted on an outpatient or inpatient basis. This potentially results in multiple cases being associated with one patient if the admission status changes or a patient presents again at the hospital with a new tumour entity. These patients were treated at Klinikum rechts der Isar. Because it is a maximum care centre, it provides comprehensive care at various stages of the patient's life, including general population, primary care, secondary care and palliative care. Detailed information regarding the patient population and tumour entities can be

2. Materials and Methods

found in Table 2.1 and Table 2.2, providing a comprehensive overview of the dataset and supporting the subsequent analyses conducted in this research. It is important to note that due to the nature of our hospital as a centre for musculoskeletal tumours, our dataset does not reflect the real ratio of benign, intermediate, and malignant tumours. We have a significantly higher proportion of extremely rare malignant tumours, which again reflects the great value of this dataset.

Entity	incidence	%
adamantinoma	8	0,09 %
angiosarcoma	37	0,41 %
aneurysmatic bone cyst	196	2,2 %
juvenile bone cyst	128	1,43 %
solitary bone cyst	62	0,69 %
breast carcinoma	1	0,01 %
bronchial carcinoma	1	0,01 %
bursitis	19	0,21 %
chondroblastoma	50	0,56 %
chondromatosis, synovial	95	1,06 %
chondrosarcoma	317	3,55 %
chordoma	28	0,31 %
clear cell sarcoma	11	0,12 %
cancer of unknown primary	2	0,02 %
dermatofibrosarcoma protuberans	24	0,27 %
desmoid (aggressive fibromatosis)	110	1,23 %
fibrous dysplasia	142	1,59 %
enchondroma	440	4,93 %
epithelioid sarcoma	21	0,24 %
Ewing's sarcoma	150	1,68 %
cartilaginous exostosis	17	0,19 %
nodular fasciitis	50	0,56 %
fibroma	73	0,82 %
fibrosarcoma	26	0,29 %
ganglion	261	2,93 %
giant cell tumour	145	1,63 %
eosinophilic granuloma	72	0,81 %
hemangioendothelioma	5	0,06 %
malignant hemangioendothelioma	15	0,17 %
hemangioma	272	3,05 %

2. Materials and Methods

hemangiopericytoma	8	0,09 %
epithelioid hemangiosarcoma	3	0,03 %
hematoma	28	0,31 %
Hodgkin's lymphoma	11	0,12 %
leiomyosarcoma	124	1,39 %
leukemia	11	0,12 %
lipoma	537	6,02 %
lipoma arborescens	8	0,09 %
liposarcoma	357	4 %
lymphangioma	9	0,1 %
malignant fibrous histiocytoma	273	3,06 %
morbus Ledderhose	30	0,34 %
malignant peripheral nerve sheath	44	0,49 %
myofibblas	1	0,01 %
myositis ossificans	50	0,56 %
myxofibrosarcoma	115	1,29 %
myxoma	58	0,65 %
neurinoma	168	1,88 %
neurofibroma / neurofibromatosis	13	0,15 %
NHL of T-cell type	10	0,11 %
NHL of the B-cell type	161	1,8 %
non ossifying fibroma	81	0,91 %
osteochondroma	495	5,55 %
osteoid osteoma	91	1,02 %
osteomyelitis	195	2,19 %
osteosarcoma	259	2,9 %
other	692	7,76 %
Paget's disease	4	0,04 %
plasmocytoma / multiple myeloma	190	2,13 %
primitive neuroectodermal tumour	15	0,17 %
prostate carcinoma	1	0,01 %
pseudotumour	34	0,38 %
pigmented villonodular synovitis	325	3,64 %
renal cell carcinoma	1	0,01 %
rhabdomyosarcoma (alveolar / embryonal)	39	0,44 %
not otherwise specified sarcoma	64	0,72 %
myofibroblastic sarcoma	13	0,15 %
alveolar soft tissue sarcoma	16	0,18 %
synovial sarcoma (monophasic / biphasic)	118	1,32 %

2. Materials and Methods

solitary, fibrous tumour,	20	0,22 %
unknown	1472	16,5 %

Gender	incidence	%
female	2943	32,99 %
male	3114	34,9 %
other	0	0 %
unknown	2865	32,11 %

Localisation	incidence	%
abdominal wall	114	1,28 %
thoracic spine	62	0,69 %
cervical spine	26	0,29 %
dorsum	56	0,63 %
elbow	96	1,08 %
foot	602	6,75 %
forearm	234	2,62 %
hand	250	2,8 %
head-neck region	94	1,05 %
hips	160	1,79 %
knee	696	7,24 %
lower leg	980	10,98 %
os coccygeum	6	0,07 %
os sacrum	114	1,28 %
pelvis	538	6,03 %
shoulder	476	5,34 %
spine	114	1,28 %
thigh	1840	20,62 %
thoracic wall	203	2,28 %
unknown	1701	19,07 %
upper arm	610	6,84 %

Malignancy	incidence	%
benign	2608	29.23 %
intermediate	675	7.57 %
malignant	5639	63.20 %

Table 2.1.: Distribution of discrete parameters with incidence and percentage ratio.

	median	interquartile range
age	46.00	35.00
year	2011.00	11.00

Table 2.2.: Distribution of continuous parameters with median and interquartile range.

Dataset

The underlying dataset of this dissertation consisted of clinical data in tabular form and images in Digital Imaging and Communications in Medicine (DICOM) format. The clinical data was unstructured and unlabelled, indicating that no verification had been conducted to confirm the accuracy of e.g. associated diagnoses or TNM classifications. Preliminary values for these clinical parameters were often based on radiological assessments alone, with final values recorded separately. The tabular data encompassed over 1,000,000 data points but had not been validated and was prone to errors. Due to the prior mapping process of this data, which was conducted without proper database methodology and by non-experts, we were not able to directly identify the stage (e.g. pre-op or post-op) of the acquired parameters. Consequently, extensive data cleaning and verification was necessary before utilizing the data. Furthermore, due to the heterogeneity of MSK tumours, a more detailed analysis of each case was required. Differentiating between primary tumours and recurrent tumours, as well as determining the tumour location (e.g., spine or forearm), had important clinical implications. This information was solely available in textual formats, such as radiology and pathology reports or tumour board protocols. The preparation of the data demanded significant domain knowledge acquired over the years, with essential support from MSK radiologists and tumour surgeons. The dataset comprised more than 250,000 x-ray images. However, crucial information such as the stage at which the imaging examinations were conducted (e.g., pre-chemotherapy or preoperative) could only be found in reports, necessitating the sorting of data from a clinical standpoint prior to using it for ML purposes.

As already mentioned, more information was available in the form of medical reports in the HIS. However, the interfaces to the clinical system were not accessible (for privacy and internal compliance reasons), so each patient had to be manually reviewed in the clinical system and the text analysed.

2.2. Current State of Machine Learning for Musculoskeletal Tumour Diagnostics

Methods

To achieve an overview of the current situation of machine learning applications in the field of MSK research, especially in a diagnostic context, we conducted a scoping literature review according to the guidelines of the PRISMA statement [76]. The PRISMA statement refers to the *Preferred Reporting Items for Systematic Reviews and Meta-Analyses*. It is a set of guidelines that provides a structured approach for reporting systematic reviews and meta-analyses of healthcare interventions. The PRISMA statement was developed to enhance the transparency, completeness, and accuracy of reporting in systematic reviews, thereby improving the quality of evidence synthesis. Studies meeting the following criteria were included in this review:

- Primary malignant musculoskeletal tumours
- Application of Machine Learning or DL
- Imaging data or data retrieved from images
- Human or preclinical
- Written in English
- Original research articles

The following focus led to the exclusion of articles for this review:

- Metastases
- Histological data
- Secondary bone/soft tissue tumours
- Lymphoma
- Myeloma
- Benign, intermediate
- Review articles

Articles that contained benign or intermediate lesions but focused primarily on e.g. the detection of malignant lesions were included. In contrast, articles that did not contain data on malignant lesions were excluded. The focus was on malignant lesions because of their clinical relevance and difficulty in accurate assessment. In December 2021, a thorough literature search through MEDLINE (PubMed), CENTRAL (Cochrane Library) and LISTA (EBSCO) was conducted. Grey literature was not considered. For the systematic search, the following search terms were used without any filters or limits:

((Artificial Intelligence) OR (Deep Learning) OR (Machine Learning)) AND (malignant) AND (tumour OR neoplasm OR cancer) AND (musculoskeletal OR sarcoma OR bone OR (soft tissue)) AND (imaging OR radiographic OR (computer-assisted) OR (image interpretation)).

Study titles were reviewed and evaluated by an MSK radiologist, an orthopaedic surgeon, and a data scientist at our institution using the above selection criteria (Figure 2.1). All discrepancies were resolved by consensus. The results were summarised, and duplicates were discarded. All articles were initially screened for relevance by title and abstract to assess the inclusion criteria. The three authors independently performed a careful reading of the studies and extracted the data. The following information was extracted from each article: title, author, year of publication, tumour entity group, number of patients, malignancy, imaging modality, algorithm, model, task, applied metric, outcome label and if or if not focused on diagnosis. For the synthesis, studies with diagnosis-oriented tasks were further examined by retrieving the scores of the most common metrics and the number of class labels to assess the number of samples per class and illustrate a potential relationship between these parameters through linear analysis and a correlation coefficient.

Scientific contribution

This review highlights that ML applications have yet to make a substantial impact on imaging-guided diagnosis of MSK malignancies, primarily due to various data-related challenges. Quality data availability is hindered by the lack of systematic and structured data collection by research institutes, resulting in small and non-increasing datasets for musculoskeletal malignancies. Even when patient data is available, it is often not in a format suitable for data science. Additionally, the rarity of MSK malignancies complicates the collection of adequate prospective data.

Interestingly, the review indicates that studies with fewer samples per class tend to have slightly higher metric scores, contradicting the common assumption that more data leads to better model performance. This anomaly may be influenced by a class

2. Materials and Methods

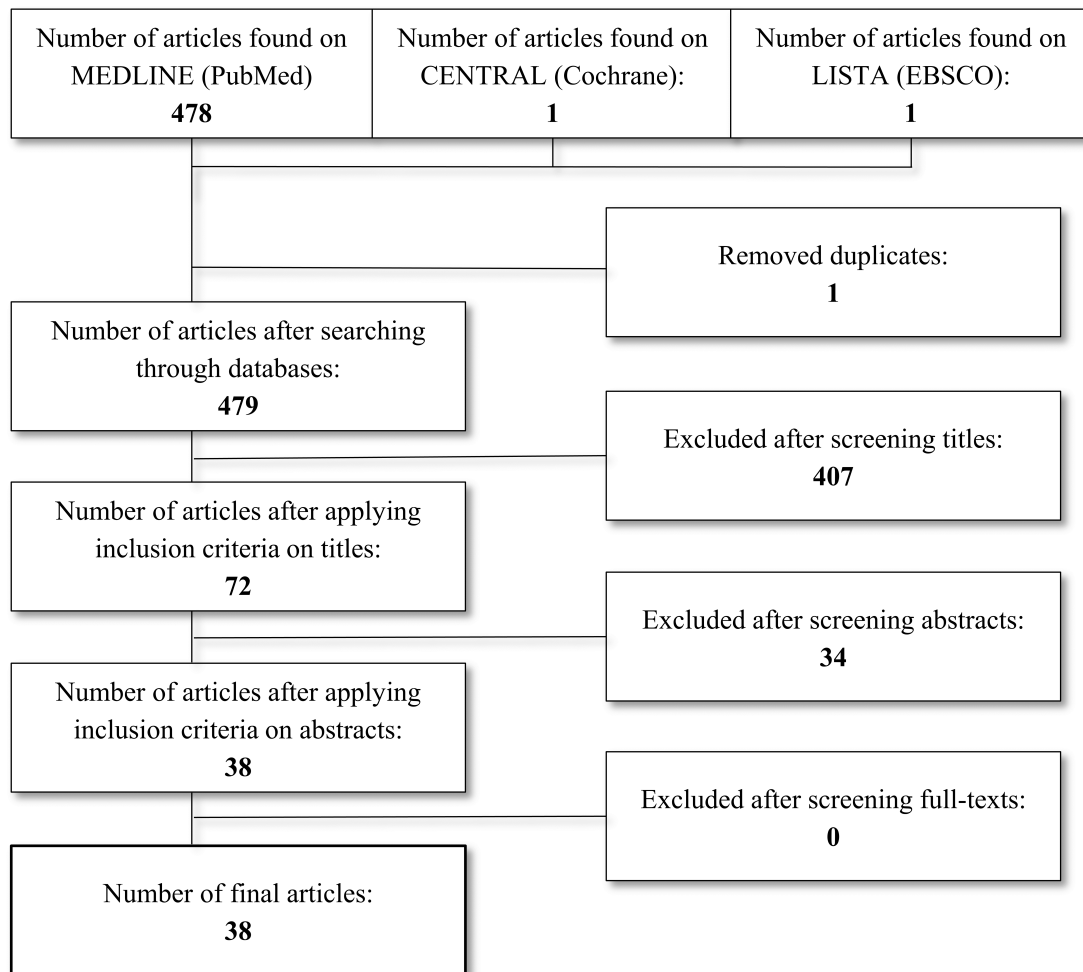


Figure 2.1.: Selection process for final references [1].

imbalance in the dataset, leading to falsely high metric values, and can also cause overfitting and suboptimal results. Most studies' problem definitions do not mirror real-world clinical scenarios, and the limited data availability hinders the development of more generalized models that address clinical needs. Although differentiating tumour entities and assessing tumour malignancy are crucial steps, identifying potential sarcomas poses the greatest challenge for MSK radiologists and orthopedic surgeons. MRI is the preferred imaging modality for ML analysis of MSK malignancies due to its superior soft tissue contrast and minimized radiation exposure. Radiomics, which extracts quantitative features from imaging data, is also popular as it can mitigate the limitations of small datasets and provide additional information. The review acknowledges the infancy of ML research in MSK malignant tumours, with most studies being in the early development stages, and highlights the heterogeneity in ML algorithms, models, and outcome designations used in the studies.

The major contribution of this study was to underscore the critical importance of data for the further development of ML for clinical image interpretation in MSK malignancies. It emphasizes the urgent need to establish national and international networks, perform systematic and structured data collection, and integrate multimodal data comparable to radiologists' practice.

2.3. Handling Limited Datasets

Methods

The dataset used in this study comprised 42,608 unstructured radiographs from a MSK tumour centre. These images, associated with sarcoma-related ICD codes, encompassed various regions of the MSK system, including extremities and joints typically affected by sarcomas, and images used for metastatic control and post-surgery or therapy monitoring. The DICOM images, sourced from the local PACS system at Klinikum rechts der Isar, spanned 25 years, exhibiting heterogeneity in quality, resolution, and data corruption. The header information of the DICOM images was fully anonymised, eliminating all meta-information for statistical analysis. A separate dataset of 63 images (22 acute osteomyelitis, 41 Ewing's sarcoma) of patients under 18 was used for evaluating the transfer learning approach, without additional restrictions on age, MSK characteristics, or sex.

To classify MSK radiographs, a two-step DL algorithm was developed (Figure 2.2). Initially, the unstructured dataset was clustered using DeepCluster, a self-supervised model proposed by Caron et al. [77]. This innovative and scalable clustering approach for unsupervised learning of CNNs alternates between clustering the features generated by the CNNs and updating the CNNs weights by using the cluster assign-

ments as pseudo-labels in a discriminative loss. DeepCluster was evaluated using different datasets, architectures, and tasks (classification, detection, segmentation, and instance-level image retrieval), demonstrating its superiority over current state-of-the-art methods in most applications, robustness to changes in image distribution, and significantly better performance with deeper architectures such as VGG-16 [54] compared to AlexNet [47]. The study also emphasized the importance of unsupervised pre-training for complex architectures when limited supervised data is available, and DeepCluster's efficiency in instance-level image retrieval tasks, highlighting the crucial role of pre-training in such applications. Overall, DeepCluster provides a robust and efficient solution for unsupervised learning of CNNs, especially in domains with scarce annotations.

In our study, k-means clustering was employed within DeepCluster to iteratively group image features and use the resulting mappings as pseudo-labels for network weight updates. The optimal number of clusters was determined through test runs based on the highest classification scores before training. Subsequently, the cluster assignments from the first step were used as "auxiliary" class labels for a classification task, in which a ResNet50 [48] model was pre-trained. The dataset was partitioned into training, validation, and hold-out test sets in 80%, 10%, and 10% proportions, respectively, for both the pretraining and transfer learning phases. A cross-validation approach was implemented to ensure robust results and prevent cross-contamination. The transfer learning approach evaluation involved a two-entity classification task with limited samples, and the dataset was divided in the same proportions as in the pretraining phase. The models' performance was evaluated based on accuracy scores. For the upstream task (pretraining), a stack size of 512, a learning rate of 0.05, and 500 epochs were selected. This phase had an approximate runtime of 7.5 hours. For the downstream task (transfer learning), a stack size of 4, a learning rate of 0.0001, and 100 epochs were chosen. The running time for all cross-validation folds was approximately 2 hours, and the inference step for all folds took about 7 minutes.

Scientific contribution

Our study represents a significant contribution to the field of DL applications in MSK image analysis, primarily by demonstrating the effective initiation of transfer learning using a state-of-the-art self-supervised model on a large dataset of 42,608 unstructured X-ray images. This strategy led to an improvement in downstream classification tasks, underscoring the potential of transfer learning to address the problem of insufficient data in medical applications. Moreover, we tackled the limitations associated with small datasets in orthopedic oncology by implementing data augmentation and transfer learning techniques. These methods showed promise in supporting various image

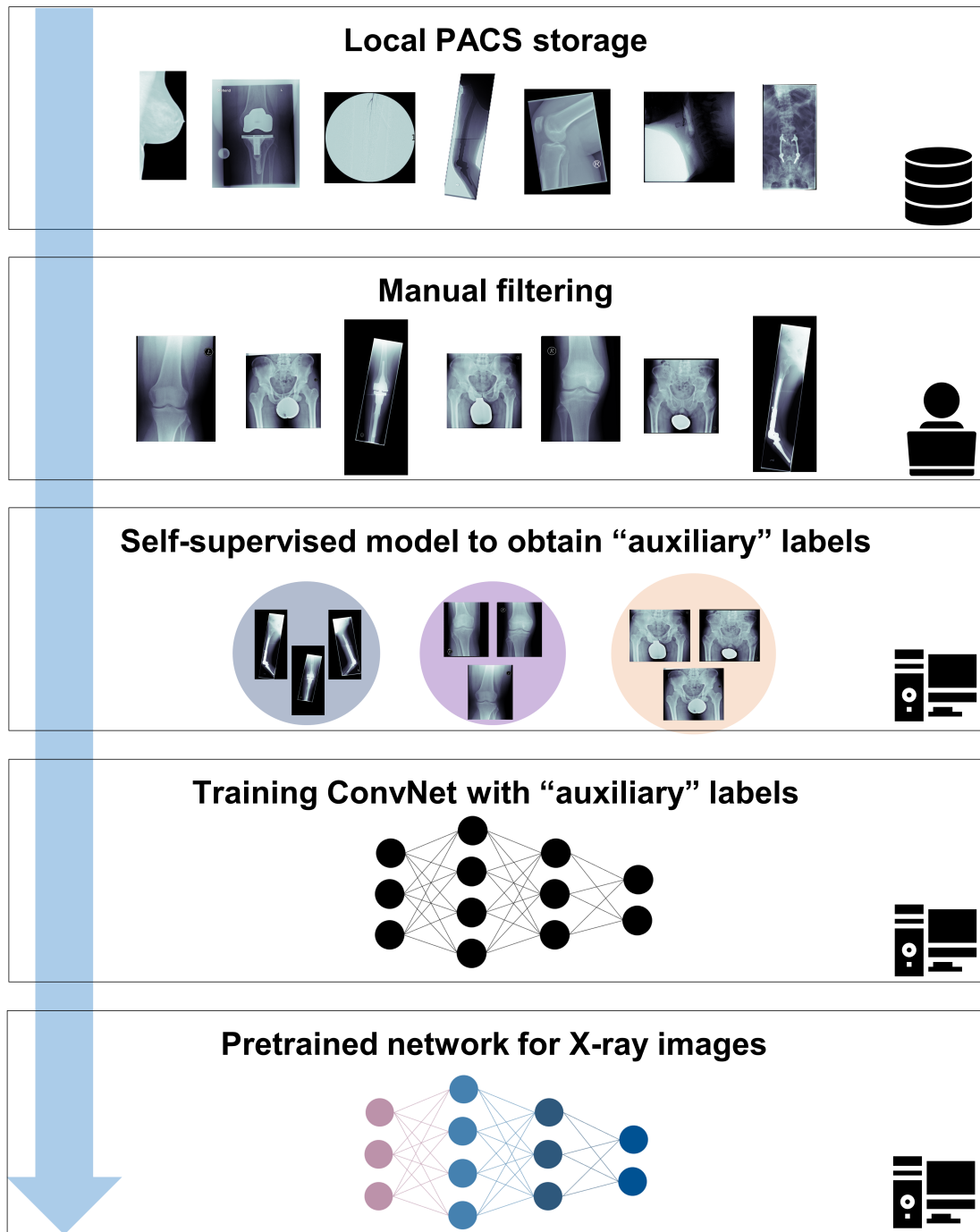


Figure 2.2.: Transfer learning approach through self-supervised pre-structuring of the data to obtain auxiliary labels [2].

interpretation tasks, thereby enhancing the performance of DL models. Our study underscores the necessity for ongoing developments to increase the quality and quantity of data for DL applications in medicine. Another key contribution is our observation on the impact of pre-training with larger datasets. We found that pre-trained networks effectively counteract the overfitting issue commonly encountered in non-pre-trained networks. This observation highlights the positive influence of pre-training with larger datasets and its potential to enhance the performance and generalizability of DL models.

The pretrained model developed in our study was also employed in a subsequent study by Consalvo et al.[5]. This study found that even very small datasets resulted in robust and stable results for image classification due to the domain-specific transfer learning approach.

The major contribution of this study was the advancement of DL in medical image analysis by showcasing the efficacy of transfer learning, addressing small dataset limitations through innovative techniques, and emphasizing the importance of systematic data collection for achieving clinically relevant results.

2.4. Leveraging Unstructured Data for Workflow Optimization

Methods

The study employed a two-phase DL framework to categorize a comprehensive dataset of 42,608 unstructured and pseudonymized radiographs into 28 distinct anatomical regions. The initial phase involved the utilization of DeepCluster [77], an innovative clustering approach for the unsupervised learning of CNNs, to cluster the entire dataset. This process involved alternating between clustering the features produced by the CNNs and updating the CNNs weights by predicting the cluster assignments as pseudo-labels in a discriminative loss. Subsequently, a senior radiologist identified 28 principal MSK classes from the clusters. This led to the exclusion of non-MSK images, resulting in a 'MSK subset' comprising 29,433 images retained for further training. In the second phase, a cross-validated classification of the MSK subset was performed. The classifier was trained to categorize images into one of the 28 pre-defined anatomical regions, with the calculation of accuracy scores for both the validation and hold-out test data. Two distinct accuracy scores were computed; one considering only the class with the highest prediction probability and another considering the two top predicted class labels. Additionally, Grad-CAMs were employed to visualize the algorithm's focus and predictions, indicating the regions of the images most pertinent for classification. Several limitations of the SAM-X model were also addressed, including the assumption that the input images correspond to one of the predefined radiographic classes, and the employment of weak supervision. Weak supervision

involves the use of imprecise (noisy) data for supervised learning, obviating the need for the laborious task of manually labeling the entire dataset. To mitigate this, a second accuracy score was calculated, considering the two predictions with the highest probability. This demonstrated that the model did not weakly label based on any incidental image features but labelled according to similar anatomical features.

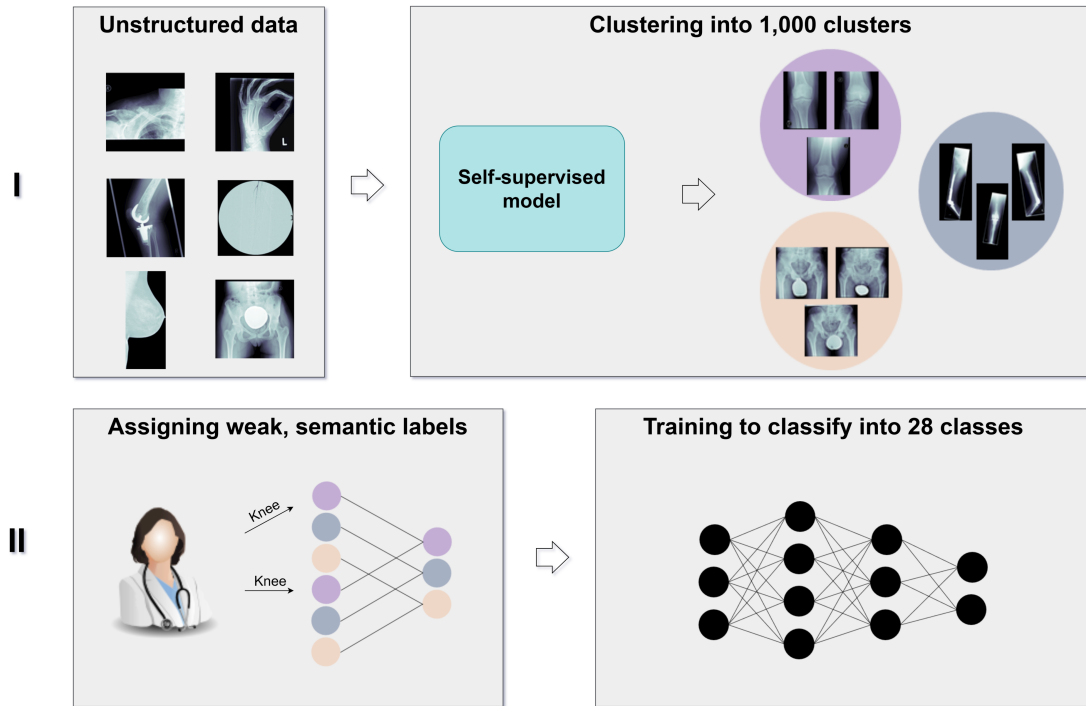


Figure 2.3.: Illustration of the presented framework in two phases: clustering data with a self-supervised model and training a network with human-annotated clusters [3].

Scientific contribution

The scientific contribution of SAM-X lies in its ability to structure and annotate vast amounts of images according to anatomical features, a task that is significantly more efficient than manual human annotation. Utilizing the strengths of DeepCluster, we developed a model that optimally organized images, demonstrating the potential to optimize workflows for radiologists and other clinicians working with images. This advancement represents a critical step toward managing the growing volume of radiographic data and, consequently, improving the efficiency of MSK disorder

assessments.

The major contribution of this study was the development of an innovative tool that can expedite the classification of musculoskeletal radiographs, thereby saving significant time for clinicians and contributing to the overall enhancement of radiological practice.

2.5. Multimodal Data for Diagnostic Decision Support Tools

Methods

Our single-center database was searched for patients treated for primary bone neoplasms from 2000 to 2021. Patients diagnosed with one of the ten most frequent tumours in the database were included. Data curation and validation were performed by two orthopedic residents, a senior MSK radiologist, and a data scientist. Data was presented following STROBE guidelines [78]. The precision-at-k metric was used to evaluate the recommender-systems clustering results. The dataset was divided into 80% training and 20% test data, with final metrics calculated three times with randomly shuffled data. Training and inference were conducted on a DGX Station A100 with four 80 GB graphical processing units. Preprocessing and model implementation were performed using Python 3.11.1 with PyTorch 1.13.1 and cuda toolkit 12.0. The proposed framework Figure 2.4 aimed to identify the most similar cases from previous patients based on radiographs relative to an undiagnosed image. Initially, baseline classification accuracies for bone tumour entities were calculated using a standard [48] and a state-of-the-art [50] DL model for multi-entity classification. Our approach involved two main steps: (I) Emphasizing tumorous tissue over background or non-relevant tissue by creating bounding boxes around the region of interest, either algorithmically [6] or through manual cropping by a domain expert. A CNN [48] was then trained with the training data in a supervised manner to extract meaningful image features for the respective bone lesion classes. The trained model and extracted features from the training data were saved, and image features of the test data were calculated by running the data through the trained CNNs model. (II) We created a hash table. Instead of comparing each set of new image features to the training data features, we used locality-sensitive hashing (LSH), an approximate nearest-neighbor algorithm that reduces the computational complexity from $O(N^2)$ to $O(\log N)$. LSH generates a hash value for image features by taking the spatiality of the data into account. Data elements that are similar in high dimensional space have a higher chance of obtaining the same hash value [79]. Based on a hamming distance function, we computed the k-nearest neighbors with respect to each target image. By assigning the k-nearest neighbors (train images) and the target image (test image) to one cluster, we established a link between the undiagnosed patient and past patients from our database. Since local patient identifiers from

the training data patients are known, this allowed us to potentially link to experiences from previous patients in our clinical systems, e.g. radiology reports, laboratory results, therapy results, etc. Furthermore, we obtained a classification of tumour entities by applying a majority vote to the entities of the images clustered to the target image.

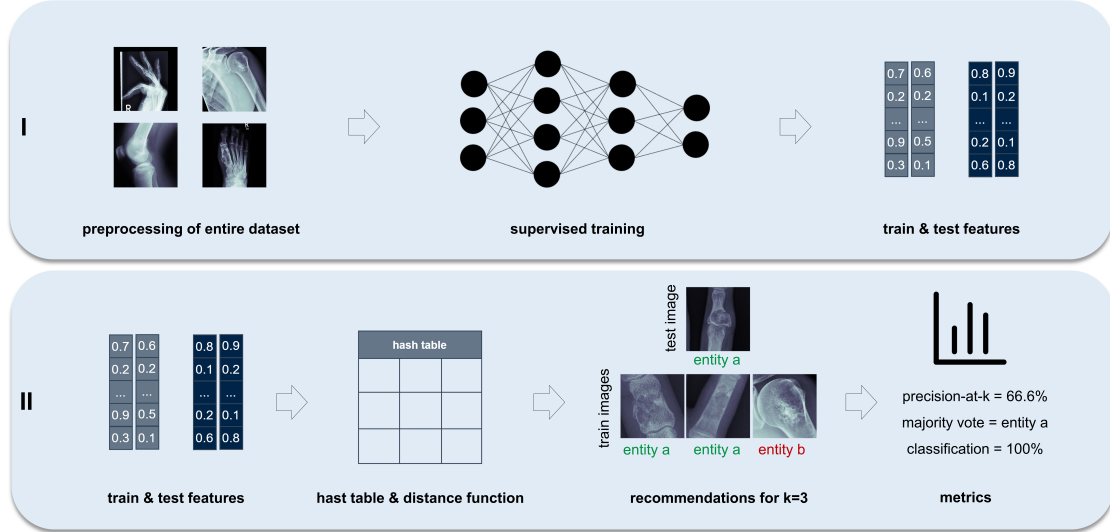


Figure 2.4.: Flow chart of the proposed model – (I) preparing the images, training of the convolutional neural network, saving the model and features, (II) calculating the high dimensional distances with a distance function, adding a hash tables, clustering of the most similar x-rays and calculating a precision-at-k and a tumour entity classification with a majority vote of the k-clustered images [4].

Scientific contribution

This study introduced a novel method for real-time classification of bone tumour entities, showcasing a significant advancement over conventional and state-of-the-art models. The innovative approach is grounded in the clustering of similar X-ray images and the application of majority voting for final classification. A pivotal aspect of this research is its ability to connect undiagnosed patients with the wealth of experience and knowledge encapsulated in clinical systems. By clustering the most similar cases, the algorithm leverages knowledge from past patient histories, thereby enabling precise diagnoses. This integration with previous patient data and histories not only augments the diagnostic process but also potentially revolutionizes the way physicians can harness dormant information. Furthermore, the study’s methodology is transferable to

other pathologies, indicating its potential versatility across various medical disciplines. The research also tackled the universal challenges associated with limited data and the diverse manifestations of tumours, proposing solutions that could influence the diagnosis of rare and intricate diseases.

The major contribution of this study was the development of a real-time classification method that leverages previous patient data and clinical knowledge to facilitate early and specific diagnosis, and notably, its capacity to link current cases with previous patient cases for comparison, potentially impacting the diagnosis of rare and complex diseases across various medical fields.

3. Discussion

3.1. Cross-thematic discussion

The following section is structured following the *Checklist for Artificial Intelligence in Medical Imaging (CLAIM)* [80] and *Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies* [78] and will address the key findings, limitations, interpretation, similar studies and generalisability.

The initial focus of this dissertation was to study MSK tumours, i.e., soft tissue and bone tumours. However, in the end, most studies focused only on bone tumours, as the clinical handling was different, and the potential within this subgroup was substantial.

The key findings from our studies are multifaceted and encompass several aspects. (I) the rarity of data in the MSK tumour field implies that not many studies have been conducted in this area [1]. As a result, most studies, including ours, have a "feasibility" character. Essentially, these studies ascertained that it is feasible to develop models or approaches that can aid in the classification and diagnosis of MSK tumours, even given the paucity of available data. (II) our studies, among others, have employed various technical approaches to address the challenges posed by limited, unstructured, and poorly documented data. For instance, our innovative algorithm utilizes a hash-based nearest-neighbour recommender approach and majority voting to classify bone tumours based on similar cases from previous patient data. This approach not only assists in managing limited data but also leverages dormant information in clinical systems to facilitate precise diagnosis [4]. Additionally, workflow optimization and structuring data are examples of technical approaches that can be employed to address the issues of unstructured and poorly documented data [3]. (III) early and precise diagnosis is of paramount importance for the effective treatment of MSK tumours. Our studies demonstrate that DL models in fact have the potential to become valuable support tools for non-tumour experts, young professionals, and general practitioners. For example, our proposed algorithm achieved a classification accuracy of 92.86% [4], significantly outperforming existing models and senior radiologists in a similar task [81]. This indicates the potential of DL models to assist in the early and precise diagnosis and analysis of MSK tumours. Moreover, while image-based approaches

have proven effective, our not yet published study titled "Impact of metadata in bone tumour classification" [7] underscores the importance of integrating multimodal data, similar to the approach of expert clinicians. This includes not only image data but also clinical and textual data, among others. Our studies contribute to ongoing efforts to develop technical approaches that can aid in the early and precise assessment of MSK tumours, despite the challenges posed by limited, unstructured, and poorly documented data. While our novel algorithms demonstrate promising results and outperform existing models, they also highlight the existing challenges, such as the need for a comprehensive solution to the issue of limited and unstructured data, and the necessity for further research and development to make DL models a valuable support tool for a broader range of medical professionals. Finally, the integration of multimodal data, as shown in our study, is crucial for improving the accuracy and effectiveness of these models.

The major limitation of our applied research is that we did not exactly follow the clinical gold standard for diagnosis, which might lead to limited acceptance by clinicians. Since DL applications in orthopaedic oncology are still in their infancy and there are several external limiting factors (time and political constraints, low incidence etc.) [1], we were forced to break down problems from holistic to simplified clinical questions. Therefore, we did focus solely on radiographs in [5] and [2]. Although plain radiographs are crucial for the initial screening for a possible bone tumour [28, 29, 82, 83], further classification requires the inclusion of clinical data (and possibly additional imaging) [84]. However, we hypothesize that some clinical information such as the patient's age, anatomical region, or tumour location is partially represented in the x-ray images and therefore indirectly integrated into our prediction models. Nevertheless, in [7] we intended to mimic the clinical workflow by including clinical metadata in addition to radiographs in the evaluation of tumour entities, or in [4] we build an unsupervised model and intended to rely on prior clinical knowledge and experience. With these approaches we increased performance of our models and came a step closer to clinical reality. Furthermore, in addition to technical progress, it should also be noted that AI tools will only find their way into the clinic if they are accepted by the medical professionals. Also MRI and CT play a crucial role in clinical diagnostics. These modalities were not investigated further within this dissertation, because metadata and radiography already held high potential. Nonetheless, these modalities will be investigated in future studies. Another limitation arises from the limited dataset. While a dataset with up to 10,000 patients and several thousand X-ray images represents a considerable amount for the very low incidence of MSK tumours, in context of DL the number is not particularly high. Further, to structure data according to clinical features such as entity, status of therapy, etc., the sample size per class was mostly lower than 500. In [6] we added a second, similar tumour entity for a segmentation

task to increase dataset size. Our clinical experts expressed doubts as this contradicts clinical approaches, where entities are treated and evaluated differently. However, we managed to increase overall performance of the model. We hypothesize, while research questions have to be designed carefully in close collaboration with clinical experts, in such cases the priority should rather be on addressing data science obstacles such as limited data than following the clinical guidelines to the last detail, especially in basic research and feasibility studies. This assessment should be done depending on the task at hand.

The proposed model's results do not yet show direct clinical relevance, but the increased accuracy achieved through state-of-the-art methodology shows promise. Enriching the imaging dataset with clinical metadata brings AI models closer to the approach of human experts. These promising results, along with other applications of AI models in medicine, could raise awareness among domain experts. Optimal AI model performance relies, amongst others, significantly on domain experts supporting the collection of complete, accurate, and comprehensive medical data, as data quality and quantity are vital factors.

Several prior studies have explored analysis and diagnostic tasks of MSK tumours using imaging data [5, 81] or have demonstrated multimodal approaches for integrating imaging and tabular data in medical classification [85, 86]. For instance, von Schacky et al. [81] developed a multitask DL model capable of simultaneously detecting, segmenting, and classifying bone lesions, comparing its performance against radiologists of varying experience levels. The overall task of classifying bone lesions and the specific entities examined in their study were similar to those in our research. While their model achieved a classification accuracy of 43.2%, a MSK radiologist achieved 58.6% accuracy in classifying bone lesions on an entity level. Although our performances in [7, 4] were significantly higher, von Schacky et al. had to contend with a lower sample ratio per class, fewer patients, and thus a smaller overall dataset. Their study primarily focused on a multitasking model and comparison with human experts, while our emphasis centered on integrating clinical metadata in conjunction with imaging data using state-of-the-art techniques. Nonetheless, their study underscores the intricate nature of accurately identifying bone neoplasms for both DL models and clinical professionals. In a similar study, Liu et al. [85] proposed a deep learning-machine learning model for classifying bone tumours using patient clinical metadata and radiographs. They collected 982 radiographs from 643 patients, incorporating clinical metadata such as age, gender, and location. Their approach involved using an Inception V3 model to process imaging data and fusing its output with clinical features to train an XGBoost model. Their fusion model achieved a top macro area under the curve of 0.872, outperforming five radiologists (0.819). The main difference between their study and ours is their focus on predicting tumour malignancy, while we aimed to classify two specific tumour

entities [5] or ten tumour entities [7, 4]. The classification task differs due to the number of classes and sample sizes. Our fusion approach [7] as well as our unsupervised clustering of similar cases [4] captures both image and metadata information, while Liu et al. combined DL model probabilities with metadata before using a secondary model. We hypothesize that our approaches better align with the clinical algorithm used by radiologists and surgeons, as it either distinguishes very specific entities (usually differential diagnoses) or simultaneously evaluates and links metadata and imaging data for comprehensive and accurate bone tumour assessment, leading to improved performance. Xu et al. [86] presented a notable study that employed multimodal data and a fusion approach for accurate differential diagnosis of skin tumours. They introduced a transformer model capable of leveraging multimodality imaging and non-imaging data to enhance diagnostic performance. Their approach involved integrating a cross-modality fusion module with a transformer-based multimodal classification system, enabling the fusion of data from multiple sources. The dataset used in their study encompassed dermoscopy, clinical imaging, and patient metadata. To evaluate the effectiveness of their proposed model, Xu et al. conducted experiments on both a public dataset (Derm7pt, 1,011 cases) and an in-house dataset (5,601 cases). The results were highly promising, surpassing the state-of-the-art performance with a 2.8% increase and achieving an impressive accuracy of 88.5%, respectively. In comparison to our models, the approach described by Xu et al. demonstrated the capability to incorporate multimodal imaging in addition to metadata. While "remixing" metadata within disease classes yielded positive results in their specific domain, we conjecture that in our case, metadata and image features are closely intertwined and should not be interchangeably treated. Nevertheless, it is important to note that no existing model, to the best of our knowledge, has proposed a multimodal approach that integrates both imaging and patient-specific metadata for bone tumour classification.

The generalizability of our studies is constrained by several key factors. Firstly, the dataset, while considerable in its number of radiographs, remains limited due to the rarity and heterogeneity of MSK tumours, with an average of 179 samples from the studies [7, 4] per class being relatively low considering the diversity of MSK lesions and the demands of DL applications. Secondly, the bulk of the data was collected from a single centre, leading to a model trained and tested on data with similar imaging devices and patient demographics, which could affect its performance when applied to data with different characteristics. Thirdly, there has been no external validation of the model, a crucial step in assessing its generalizability before it can be considered suitable for clinical use. Lastly, our models were developed and tested on a limited number of tumour entities (two in [5], ten in [7, 4]), leaving its performance on rarer tumour entities or other types of bone pathologies unassessed. While these limitations restrict the generalizability of the findings for clinical application, there is a positive aspect

to consider. The technical approaches employed in the studies, such as the fusion of a Multi-Layer Perceptron (MLP) and a transformer to cope with imaging and clinical data [7], or the combination of feature extraction, clustering, and majority vote [4], may be adaptable to other scenarios. This suggests that, although the findings may not be immediately generalizable for clinical use, the technical side of the models might be more broadly applicable, and therefore, may have a positive impact on other areas of medical imaging and diagnostics. Therefore, while further validation on larger, more diverse, and external datasets is necessary to assess the models' generalizability and suitability for clinical use, the technical approaches developed in our studies might still offer valuable contributions to the field.

3.2. Future Work

Future work in ML for the analysis and diagnosis of MSK tumours should aim to build on the foundation established in previous research while addressing identified limitations. This includes developing more sophisticated ML and DL models, integrating different types of data, improving data acquisition and management, and exploring innovative ways to validate models.

Developing more advanced models: Although significant progress has been made in using ML for tumour classification and diagnosis, there is still potential to develop more sophisticated models. In addition to classification tasks, models could be developed to predict prognosis, response to treatment, and risk of recurrence, which is especially interesting for clinicians. Techniques such as transfer learning and self-supervised learning can help overcome the challenge of limited datasets. Future work could also explore the potential of combining different AI techniques, such as integrating CNNs for image analysis with RNNs for sequential data analysis to analyse e.g. treatment progress.

Integration of multimodal data: Future studies could aim to integrate multimodal data, including radiographs, clinical metadata, histopathology data, and patient-reported outcomes. This would improve the predictive power and clinical relevance of ML models. As genomics plays an increasingly important role in cancer diagnosis and treatment, incorporating genomic data into these models could further enhance their ability to provide personalized diagnostic and therapeutic insights.

NLP applications: NLP has enormous potential in healthcare, particularly in extracting valuable information from unstructured textual data such as clinical notes, radiology and pathology reports, tumour board protocols (Figure 1.2) or surgery reports. For our future research, it could be useful to use NLP to extract and structure information such as whether an imaging study was performed pre- or post-operatively,

pre-chemo or post-chemo, etc.. In the context of MSK tumours, NLP could be used to automatically extract and structure relevant clinical information (e.g. tumour size, direction of tumour growth, visual appearance of the tumour, etc.), which could then be fed into ML models to support image-based diagnosis. This could significantly improve the efficiency and accuracy of the diagnostic process.

Improved data collection and management: A major challenge in this area is the scarcity and heterogeneity of data. Building comprehensive, structured, and systematic data collection systems and networks should be a priority for future work. This could include collaboration between different, national and international healthcare institutions and the development of standards for data collection and sharing.

Model validation and clinical translation: There is a need for rigorous validation of ML models to determine their performance, reliability, and clinical utility. Future work should aim to test models in real-world clinical scenarios and in different populations and settings. In addition, it is important to explore how these AI systems can be effectively integrated into clinical workflows and assess their impact on patient outcomes and healthcare costs. Close collaboration with clinical experts is considered absolutely crucial.

By addressing these areas, future work could significantly advance the application of ML in the diagnosis and treatment of MSK tumours, with the potential to improve patient outcomes and healthcare efficiency.

3.3. Conclusion

In conclusion, this dissertation has made significant progress in applying ML techniques to the analysis and diagnosis of MSK tumours. Our proposed models, which exploit methods of transfer learning, unsupervised learning and self-supervised learning have shown promising ability to accurately classify tumours and efficiently cope with small and manage large image datasets. In addition, our pioneering efforts to integrate clinical metadata with radiographic data have led to significant improvements in diagnostic performance, indicating the potential of multimodal learning approaches.

A major focus of our future work will be the integration of textual data. We see NLP as an important tool for analyzing unstructured data, such as radiology reports, clinical notes or tumour board protocols, and transforming them into valuable input for our ML models. By incorporating this rich source of data, we can create a more holistic picture of the patient, leading to an even more robust multimodal approach to diagnosis. Our models have contributed significantly to potentially shortening the *time to diagnosis* for MSK tumours. This progress is invaluable as earlier and more accurate diagnoses directly lead to a better prognosis for patients. Patients can receive timely

3. Discussion

and appropriate treatment, leading to better health outcomes and a higher quality of life through AI.

4. Publications

4.1. Applications of machine learning for imaging-driven diagnosis of musculoskeletal malignancies - a scoping review

Authors:

Florian Hinterwimmer, Sarah Consalvo, Jan Neumann, Daniel Rueckert, Rüdiger von Eisenhart-Rothe, Rainer Burgkart

Journal:

European Radiology

Synopsis:

Malignant tumours of the MSK system are extremely rare and diverse. They account for only 0.2% of all human malignancies, but are more common in children and adolescents. Early diagnosis is difficult due to their nonspecific clinical presentation and rarity. Timely referral to specialized sarcoma centers is critical for accurate diagnosis and improved patient prognosis. However, delays in diagnosis are common, due in part to the limited experience of general practitioners with such cases. The morphologic heterogeneity of MSK tumours complicates imaging and biopsy procedures. Biopsies can be difficult in certain types of lesions, leading to low success rates and potential complications. Adequate diagnostic biopsy is essential for proper treatment of MSK tumours and is considered the first step in therapy. ML and DL techniques have shown promise in various areas of medical research, but their application in orthopaedic oncology is still limited. The lack of structured and systematic data collection systems and the rarity and heterogeneity of sarcomas make it difficult to effectively apply AI methods. While some methods have been developed to address limited datasets, building appropriate structures and networks is critical. The aim of this review was to assess the extent to which ML can aid image interpretation of malignant MSK tumours,

particularly in diagnostic tasks. The review focused on identifying relevant studies that used ML or DL techniques with image data from primary malignant MSK tumours.

A scoping review of the literature was performed based on specific eligibility criteria. The search yielded 480 references, and after screening and assessment, 38 articles were included in the final analysis. These articles were published between 1994 and 2021 and covered a range of ML and DL applications in MSK malignancy imaging.

The majority of studies (71.1%) were diagnosis-oriented and focused on classification tasks. The median accuracy and area under the curve area under the curve (AUC) for these studies were 0.88 and 0.92, respectively, but there was no significant correlation between the metric values and the number of samples per class.

The review found that ML applications have not yet had a significant impact on the diagnosis of MSK malignancies. The limited availability of quality data, especially in structured and systematic formats, remains a major challenge. The rarity of sarcomas and the limited amount of research in orthopaedic oncology contribute to the lack of data. Imbalance of classes and limited dataset sizes may also affect the performance of ML models. In addition, many studies do not reflect real-world clinical scenarios because they focus on specific tumour entities rather than detection of potential sarcomas. The use of ML in imaging-based diagnosis of MSK malignancies is still at an early stage. The scarcity and heterogeneity of data and other challenges hinder the application of ML techniques in this field. Further efforts are needed to establish comprehensive data collection structures and networks to advance the application of ML in orthopaedic oncology.

Contribution of thesis author:

Florian Hinterwimmer was the principal investigator in this study and contributed at least 50%.

Copyright:

Applications of machine learning for imaging-driven diagnosis of musculoskeletal malignancies—a scoping review © 2022 by Florian Hinterwimmer is licensed under Creative Commons Attribution 4.0 International. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Applications of machine learning for imaging-driven diagnosis of musculoskeletal malignancies—a scoping review

SPRINGER NATURE

Author: Florian Hinterwimmer et al

Publication: European Radiology

Publisher: Springer Nature

Date: Jul 19, 2022

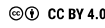
Copyright © 2022, The Author(s)

Creative Commons

This is an open access article distributed under the terms of the [Creative Commons CC BY](#) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.

To request permission for a type of use not listed, please contact [Springer Nature](#)



ATTRIBUTION 4.0 INTERNATIONAL

Deed

Canonical URL: <https://creativecommons.org/licenses/by/4.0/>

[See the legal code](#)

You are free to:

Share — copy and redistribute the material in any medium or format for any purpose, even commercially.

Adapt — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

Attribution — You must give **appropriate credit**, provide a link to the license, and **indicate if changes were made**. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

No additional restrictions — You may not apply legal terms or **technological measures** that legally restrict others from doing anything the license permits.

Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable **exception or limitation**.

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as **publicity, privacy, or moral rights** may limit how you use the material.

Notice

This deed highlights only some of the key features and terms of the actual license. It is not a license and has no legal value. You should carefully review all of the terms and conditions of the actual license before using the licensed material.

Creative Commons is not a law firm and does not provide legal services. Distributing, displaying, or linking to this deed or the license that it summarizes does not create a lawyer-client or any other relationship.

Creative Commons is the nonprofit behind the open licenses and other legal tools that allow creators to share their work. Our legal tools are free to use.

- [Learn more about our work](#)
- [Learn more about CC Licensing](#)
- [Support our work](#)
- [Use the license for your own material](#)
- [Licenses List](#)
- [Public Domain List](#)

Footnotes

appropriate credit — If supplied, you must provide the name of the creator and attribution parties, a copyright notice, a license notice, a disclaimer notice, and a link to the material. CC licenses prior to Version 4.0 also require you to provide the title of the material if supplied, and may have other slight differences.

- [More info](#)

indicate if changes were made — In 4.0, you must indicate if you modified the material and retain an indicator of previous modifications. In 3.0 and earlier license versions, the indication of changes is only required if you create a derivative.

- [Marking guide](#)
- [More info](#)

technological measures — The license prohibits application of effective technological measures, defined with reference to Article 11 of the WPO Copyright Treaty.

- [More info](#)

exception or limitation — The rights of users under exceptions and limitations, such as fair use and fair dealing, are not affected by the CC licenses.

- [More info](#)

publicity, privacy, or moral rights — You may need to get additional permissions before using the material as you intend.

- [More info](#)

[Contact](#)

[Newsletter](#)

[Privacy](#)

[Policies](#)

[Terms](#)

CONTACT US

Creative Commons PO Box 1396, Mountain View, CA 94042

info@creativecommons.org

+1 415 433 8200

SUBSCRIBE TO OUR NEWSLETTER

Your email

[SUBSCRIBE](#)

SUPPORT OUR WORK

Our work relies on you. Help us keep the Internet free and open.

[DONATE NOW](#)

Except where otherwise noted, content on this site is licensed under [Creative Commons Attribution 4.0 International License](#). Icons by [Font Awesome](#).



Applications of machine learning for imaging-driven diagnosis of musculoskeletal malignancies—a scoping review

Florian Hinterwimmer^{1,2} · Sarah Consalvo¹ · Jan Neumann³ · Daniel Rueckert² · Rüdiger von Eisenhart-Rothe¹ · Rainer Burgkart¹

Received: 16 May 2022 / Revised: 31 May 2022 / Accepted: 22 June 2022
© The Author(s) 2022

Abstract

Musculoskeletal malignancies are a rare type of cancer. Consequently, sufficient imaging data for machine learning (ML) applications is difficult to obtain. The main purpose of this review was to investigate whether ML is already having an impact on imaging-driven diagnosis of musculoskeletal malignancies and what the respective reasons for this might be. A scoping review was conducted by a radiologist, an orthopaedic surgeon and a data scientist to identify suitable articles based on the PRISMA statement. Studies meeting the following criteria were included: primary malignant musculoskeletal tumours, machine/deep learning application, imaging data or data retrieved from images, human/preclinical, English language and original research. Initially, 480 articles were found and 38 met the eligibility criteria. Several continuous and discrete parameters related to publication, patient distribution, tumour specificities, ML methods, data and metrics were extracted from the final articles. For the synthesis, diagnosis-oriented studies were further examined by retrieving the number of patients and labels and metric scores. No significant correlations between metrics and mean number of samples were found. Several studies presented that ML could support imaging-driven diagnosis of musculoskeletal malignancies in distinct cases. However, data quality and quantity must be increased to achieve clinically relevant results. Compared to the experience of an expert radiologist, the studies used small datasets and mostly included only one type of data. Key to critical advancement of ML models for rare diseases such as musculoskeletal malignancies is a systematic, structured data collection and the establishment of (inter)national networks to obtain substantial datasets in the future.

Key Points

- *Machine learning does not yet significantly impact imaging-driven diagnosis for musculoskeletal malignancies compared to other disciplines such as lung, breast or CNS cancer.*
- *Research in the area of musculoskeletal tumour imaging and machine learning is still very limited.*
- *Machine learning in musculoskeletal tumour imaging is impeded by insufficient availability of data and rarity of the disease.*

Keywords Primary musculoskeletal malignancies · Imaging-driven diagnosis · Diagnostic imaging · Machine learning · Deep learning

Abbreviations

Acc Accuracy
AI Artificial intelligence
AUC Area under the curve

DL Deep learning
IoU Intersection over union
IQR Interquartile range
ML Machine learning

✉ Florian Hinterwimmer
florian.hinterwimmer@tum.de

¹ Department of Orthopaedics and Sports Orthopaedics, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany

² Institute for AI and Informatics in Medicine, Technical University of Munich, Munich, Germany

³ Department of Diagnostic and Interventional Radiology, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany

MSK Musculoskeletal
 SD Standard deviation
 SVM Support vector machine

Introduction

Malignant tumours of the musculoskeletal system represent a group of extraordinarily rare and heterogeneous tumour entities. For example, malignant bone tumours account for only about 0.2% of all human malignancies, but they occur more frequently in children (sixth most common cancer) and adolescents (third most common cancer) [1–3]. In addition to the pronounced rarity, the mostly unspecific history or clinical presentation also complicates early diagnosis and often leads to significant delays [3]. However, undelayed diagnosis is of paramount importance in musculoskeletal tumours, as the diagnostic window also has a direct impact on resectability and patient survival prognosis [2]. Thus, prompt referral to a specialised sarcoma centre is crucial when a malignant musculoskeletal tumour is suspected. However, delays of more than 12 months sometimes occur in clinical care reality, which can be explained not least by the fact that a general medical practitioner encounters only about three malignant musculoskeletal tumours in his/her professional life [4].

Especially the morphologic heterogeneity within musculoskeletal tumours complicates imaging entity or malignancy assessment and even limits the informative value of a biopsy. In sclerotic, blastic or cartilaginous lesions, as well as in tumours with a large necrotic area, retrieving adequate material from a biopsy is extremely challenging and requires a high degree of experience [5]. The rate of biopsy-related complications that adversely affect biopsy outcome or prognosis is reported to be 15–20%, with up to 12 times higher rates in non-specialist institutions [6]. Therefore, the importance of adequate diagnostic biopsy cannot be overstated in musculoskeletal tumours, which is why biopsy is considered the “first step of therapy” by many experts.

Image interpretation as a part of precision medicine plays an increasingly important role in the future of orthopaedic oncology, and novel, more comprehensive and specific analysis tools are urgently needed, especially for outpatient clinics with limited experience and resources for detection and interpretation of rare bone and soft tissue malignancies. Machine learning (ML) and the subset deep learning (DL) represent distinct applications of artificial intelligence (AI), which evolved from pattern recognition and learning theory. ML is just in its early stages in orthopaedics, and standardised approaches are not yet established. While complex data analysis of cancerous tissue through AI and imaging data is already widely applied for research purposes in some cancers (e.g. lung, breast or CNS cancer) [7], the application of these methods in orthopaedic oncology research is still very limited

[8]. The fact that globally no far-reaching structures for systematic and structured data acquisition have yet been established (to the best of our knowledge) and that sarcomas are very rare and heterogeneous makes modern AI applications, for which a sufficient and qualitative amount of data is crucial, considerably more difficult. Although various methods for dealing with limited datasets have been developed (data augmentation [9], transfer learning [10], data simulation [11]), there is no way around building up appropriate structures and networks.

The main purpose of this review was to investigate whether ML can already substantially support image interpretation of musculoskeletal (MSK) malignancies with a focus on diagnostic tasks and what the respective reasons for this might be.

Materials and methods

Eligibility criteria

A scoping review of the literature was performed to identify ML applications in imaging of musculoskeletal malignancies based on the PRISMA statement [12]. Studies meeting the following criteria were included in this review:

- Primary malignant musculoskeletal tumours
- Application of machine learning or deep learning
- Imaging data or data retrieved from images
- Human or preclinical
- Written in English
- Original research articles

The following focus led to the exclusion of articles for this review:

- Metastases
- Histological data
- Secondary bone/soft tissue tumours
- Lymphoma
- Myeloma
- Benign, intermediate
- Review articles

Articles that contained benign or intermediate lesions but focused primarily on e.g. the detection of malignant lesions were included. In contrast, articles that did not contain data on malignant lesions were excluded. The focus was on malignant lesions because of their clinical relevance and difficulty in accurate assessment.

In December 2021, a thorough literature search through MEDLINE (PubMed), CENTRAL (Cochrane Library) and LISTA (EBSCO) was conducted. Grey literature was not

considered. For the systematic search, the following search terms were used without any filters or limits:

((Artificial Intelligence) OR (Deep Learning) OR (Machine Learning)) AND (malignant) AND (tumour OR neoplasm OR cancer) AND (musculoskeletal OR sarcoma OR bone OR (soft tissue)) AND (imaging OR radiographic OR (computer-assisted) OR (image interpretation))

Study titles were reviewed and evaluated by an MSK radiologist, an orthopaedic surgeon and a data scientist at our institution using the above selection criteria. All discrepancies were resolved by consensus. The results were summarised, and duplicates were discarded. All articles were initially screened for relevance by title and abstract to assess the inclusion criteria. The three authors independently performed a careful reading of the studies and extracted the data. The following information was extracted from each article: title, author, year of publication, tumour entity group, number of patients, malignancy, imaging modality, algorithm, model, task, applied metric, outcome label and if or if not focused on diagnosis. For the synthesis, studies with diagnosis-oriented tasks were further examined by retrieving the scores of the most common metrics and the number of class labels to assess

the number of samples per class and illustrate a potential relationship between these parameters through linear analysis and a correlation coefficient. The level of evidence is level V.

Statistical analysis

Continuous data is reported as mean with standard deviation (SD) or median with interquartile range (IQR), and the respective interval. Discrete data was reported as incidence and percentage share per entity. Due to the heterogeneous nature and the limited amount of data, a non-parametric test was chosen to calculate a correlation coefficient for metric values and number of samples per class label for the diagnosis-oriented studies.

Results

Selection and methodological characteristics

The first search resulted in 480 references in the databases mentioned above. One duplicate was discarded and 38 articles subsequently met the eligibility criteria (Fig. 1) [8, 10, 13–51]. Table 1 displays the final selection of articles with authors and continuous and discrete parameters. Final articles were published between 1994 and 2021. All 38 articles addressed an

Fig. 1 Selection process

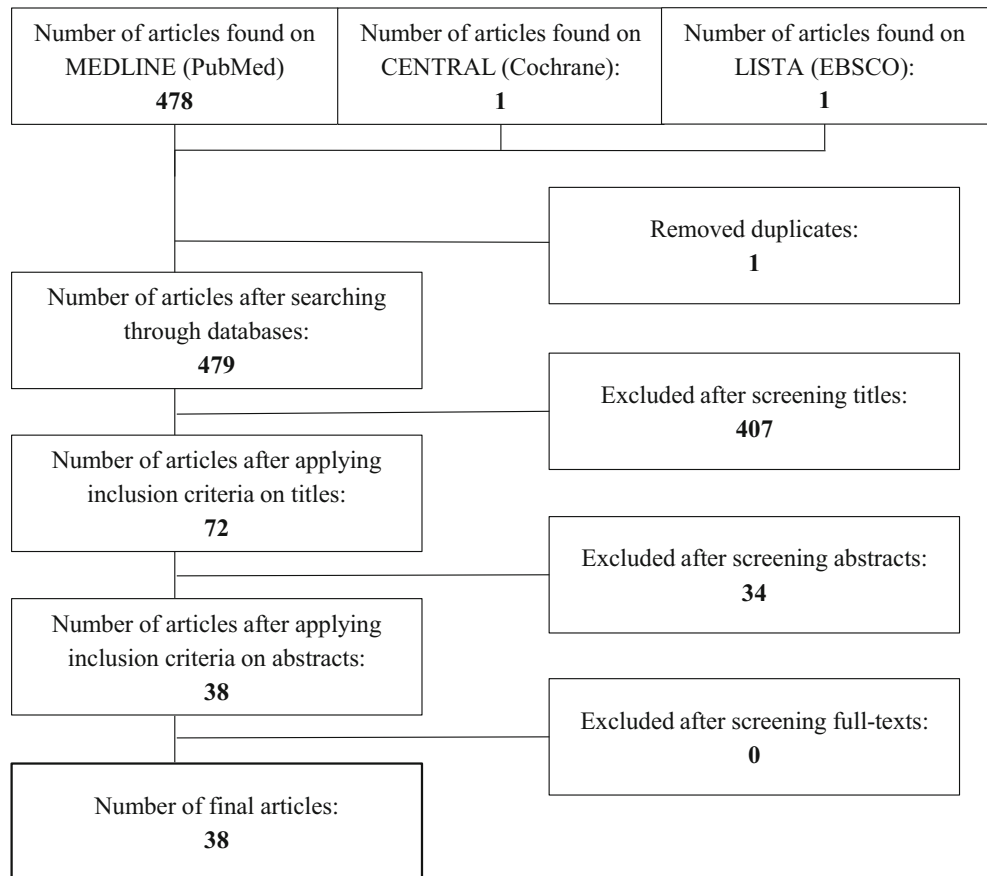


Table 1 Final articles with continuous and discrete parameters. Acc and AUC values as well as number of labels were further investigated for articles with diagnosis-oriented tasks

Author	Year	Number of patients / cases	Healthy cases	Benign cases	Intermediate cases	Malignant cases	Metastases cases	Study design	Tumour entity group	Imaging modality	Radiomic data
Bandyopadhyay et al	2019	150	0	0	0	150	0	Retrospective	Bone tumours	X-ray	No
Banerjee et al	2018	21	0	0	0	21	0	Retrospective	Soft tissue tumours	MRI	No
Chianca et al	2021	146	0	49	0	40	57	Retrospective	Bone tumours	MRI	Yes
Do et al	2021	1576	381	1061	0	134	0	Retrospective	Bone tumours	X-ray	No
Dufau et al	2019	69	0	0	0	69	0	Retrospective	Bone tumours	MRI	Yes
Eweje et al	2021	1060	0	582	0	478	0	Retrospective	Bone tumours	MRI	No
Fields et al	2021	128	0	36	0	92	0	Retrospective	Soft tissue tumours	MRI	Yes
Gao et al	2021	30	0	0	0	30	0	Prospective	Soft tissue tumours	MRI	No
Gao et al	2020	30	0	0	0	30	0	Prospective	Soft tissue tumours	MRI	Yes
García-Gómez et al	2004	430	0	267	0	163	0	Retrospective	Soft tissue tumours	MRI	No
Gitto et al	2020	58	0	0	0	58	0	Retrospective	Bone tumours	MRI	Yes
Glass et al	1998	43	0	0	0	43	0	Retrospective	Bone tumours	MRI	No
He et al	2020	1356	0	679	0	360	317	Retrospective	Bone tumours	X-ray	No
Holbrook et al	2020	79	0	0	0	79	0	Unknown	Soft tissue tumours	MRI	Yes
Hu et al	2021	160	0	90	0	70	0	Retrospective	Soft tissue tumours	MRI	Yes
Hu et al	2014	141	0	71	0	70	0	Unknown	Bone tumours	X-ray	No
Huang et al	2020	12	0	0	0	12	0	Prospective	Bone tumours	MRI	No
Huang et al	2017	23	0	0	0	23	0	Unknown	Bone tumours	CT	No
Juntu et al	2010	135	0	86	0	49	0	Unknown	Soft tissue tumours	MRI	No
Leporq et al	2020	81	0	40	0	41	0	Retrospective	Soft tissue tumours	MRI	Yes
Li et al	2019	210	0	154	0	56	0	Retrospective	Bone tumours	MRI	Yes
Liu et al	2021	643	0	392	93	158	0	Retrospective	Bone tumours	X-ray	No
Pan et al	2021	796	0	412	169	215	0	Retrospective	Bone tumours	X-ray	No
Peeken et al	2019	221	0	221	0	0	0	Retrospective	Soft tissue tumours	CT	Yes
Peeken et al	2018	136	0	0	0	136	0	Retrospective	Soft tissue tumours	MRI, CT	No
Reinus et al	1994	709	0	492	0	217	0	Retrospective	Bone tumours	X-ray	No
Shen et al	2018	36	0	15	0	21	0	Unknown	Bone tumours	X-ray	No
Terunuma et al	2018	1	N/A	N/A	N/A	N/A	N/A	Retrospective	Bone tumours	X-ray	No
von Schacky et al	2021	934	0	623	0	311	0	Retrospective	Bone tumours	X-ray	No

Table 1 (continued)

Author	Algorithm	Task	Model	Applied metric	Outcome label	Diagnosis-oriented	Acc	AUC	Number of labels	
Vos et al	2019	116	0	58	0	58	0		MRI	Yes
Wang et al	2021	227	0	147	0	80	0		US	No
Wang et al	2020	206	0	105	0	93	8		MRI	Yes
Yin et al	2019	120	0	0	30	54	36		MRI	Yes
Yin et al	2019	95	0	0	42	53	0		CT	Yes
Yin et al	2021	795	0	215	0	399	181		CT	Yes
Zhang et al	2020	51	N/A	N/A	N/A	N/A	N/A		MRI, CT	No
Zhang et al	2019	35	0	0	0	35	0		MRI	Yes
Zhang et al	2018	23	0	0	0	23	0		CT	No
Bandyopadhyay et al	Supervised	Classification	SVM, decision tree	acc, sens, Dice	Histopathological grading, staging	✓	0.85	2		
Banerjee et al	Supervised	Classification	AlexNet	acc, AUC, sens, spec	Tumour entities	✓	0.85	2		
Chianca et al	Supervised	Classification	LogitBoost, SVM	AUC, sens, spec, acc	Malignancy	✓	0.90	2		
Do et al	Supervised	Classification, segmentation	UNet	acc, IoU	Segmented tumour, tumour entities	✓	0.99	3		
Dufau et al	Supervised	Classification	SVM	AUC, sens, spec	Chemotherapy response	×				
Eweje et al	Supervised	Classification	Efficient-Net, logistic regression	acc, sens, spec, AUC	Malignancy	✓	0.79	2		
Fields et al	Supervised	Classification	AdaBoost, random forest	AUC, sens, spec	Malignancy	✓	0.77	2		
Gao et al	Supervised	Classification	VGG19	sens, spec, acc	Radiotherapy response	×				
Gao et al	Supervised	Classification	SVM, logistic regression	AUC	Radiotherapy response	×				
García-Gómez et al	Supervised	Classification	K-nearest neighbour, SVM	sens, spec	Malignancy	✓	0.90	2		
Gitto et al	Supervised	Classification	LogitBoost	acc, AUC	Histopathological grading	✓	0.75	0.78	2	
Glass et al	Unsupervised	Segmentation	Neural network	acc, sens, spec	Chemotherapy response	×				
He et al	Supervised	Classification	Efficient-Net	AUC, sens, spec, acc	Malignancy	✓	0.73	2		
Holbrook et al	Supervised	Segmentation	SVM, neural network	Dice, AUC	Segmented tumour	×				
Hu et al	Supervised	Classification	Least absolute shrinkage and selection operator	AUC, sens, spec, acc	Malignancy	✓	0.92	0.96	2	
Hu et al	Supervised	Classification	SVM	acc, AUC, sens, spec	Tumour occurrence	✓	0.96	2		
Huang et al	Supervised	Classification	Random forest	AUC, sens, spec, acc	Chemotherapy response	×				

Table 1 (continued)

Author	Supervised	Segmentation	VGG16	Dice score	Segmented tumour	×
Huang et al	Supervised	Classification	SVM, neural network, decision tree	AUC, sens, spec, acc	Malignancy	✓
Juntu et al	Supervised	Classification	SVM	AUC, sens, spec, acc	Malignancy	0.93
Leporq et al	Supervised	Classification	SVM	AUC, sens, spec, acc	Malignancy	0.95
Li et al	Supervised	Classification	SVM	AUC, sens, spec, acc	Tumour entities	0.96
Liu et al	Supervised	Classification	XGBoost, Inception V3	AUC, sens, spec, acc	Malignancy	0.87
Pan et al	Supervised	Classification	Random forest	AUC, acc	Malignancy	0.87
Peeken et al	Supervised	Classification	Random forest	AUC, Dice	Histopathological grading	0.95
Peeken et al	Supervised	Classification	Random forest	AUC, sens, spec, acc	Prognosis	0.64
Reinus et al	Supervised	Classification	Neural network	acc	Malignancy	0.85
Shen et al	Supervised	Classification	Random forest, SVM	AUC, sens, spec, acc	Malignancy	0.85
Terunuma et al	Supervised	Object detection, segmentation	SegNet	Jaccard index	Segmented tumour	×
von Schacky et al	Supervised	Object detection, segmentation, classification	Mask-RCNN	acc, sens, spec, IoU, Dice	Malignancy	×
Vos et al	Supervised	Classification	SVM, random forest	AUC, sens, spec	Tumour entities	0.89
Wang et al	Supervised	Classification	VGG16	acc, sens, spec, AUC	Malignancy	0.79
Wang et al	Supervised	Classification	SVM, generalised linear models, random forest	AUC, sens, spec, acc	Malignancy	0.86
Yin et al	Supervised	Classification	Random forest	AUC, acc	Segmented tumour, tumour entities	0.71
Yin et al	Supervised	Classification	Random forest	acc, AUC	Tumour entities	0.90
Yin et al	Supervised	Classification	Random forest	AUC, acc	Tumour entities	0.88
Zhang et al	Supervised	Classification	Inception-v3	acc, AUC	Histopathological grading	0.86
Zhang et al	Supervised	Classification	Random forest, SVM	AUC, sens, spec, acc	Histopathological grading	0.88
Zhang et al	Supervised	Segmentation	ResNet-50	Dice, sens	Segmented tumour	×

SVM support vector machine, *IoU* intersection over union *N/A* not assessed

application of ML or DL with imaging data of MSK malignancies. Three review articles were found and excluded from statistical analysis [8, 14, 25]. 75.7% (28) of the studies were conducted retrospectively, 8.1% (3) were conducted prospectively and 16.2% (6) did not clearly state the study design. 60.5% (23) of the studies focused on bone, while 39.5% (15) focused on soft tissue tumours. 50.3% of the cases included were from patients with benign tumours, 3.0% were from patients with intermediate tumours, 37.4% were from patients with malignant tumours, 5.4% were from patients with metastases, 3.6% were from patients without tumours (healthy) and 0.5% did not provide any information. Further details are reported in Tables 2 and 3.

Narrative review of best studies

Several studies have presented novel and interesting implementations. However, we would like to highlight two studies that, in our opinion, provide very intriguing frameworks. Liu et al [35] demonstrated a ML-DL fusion model that integrates not only imaging but also clinical data to assess the malignancy of tumours. This approach is similar to the diagnostic procedure a radiologist would use to diagnose MSK lesions. A second noticeable study was published by von Schacky et al [42]: they presented a multi-task DL model that shows the potential of state-of-the-art DL by simultaneously detecting, segmenting and classifying image data. To classify the DL results in the context of “man vs. machine,” they were also compared with the results of radiologists of different experience levels demonstrating strengths and limitations of DL with limited data.

In-depth investigation of diagnosis-oriented tasks

Twenty-seven (71.1%) of the studies were diagnosis-oriented and mainly aimed at classification tasks [10, 13, 15, 16, 18, 19, 22, 23, 26, 28, 29, 32–37, 39, 40, 43–49, 51]. A median

accuracy (Acc) of 0.88 with an interval of [0.71; 0.99] was found. For the area under the curve (AUC), the median resulted in 0.92 with a corresponding interval of [0.64; 0.98]. For the number of labels, a median of 2 with an interval of [2;3] was found. Further details are shown in Table 4.

Figure 2 demonstrates the findings of a linear analysis of the metric values Acc and AUC on the vertical axis and the quotient of total number of cases and number of labels per class (= mean number of samples per class). Further, a correlation coefficient for each metric and the mean number of samples per class was calculated. The number of studies examined is limited, and the data found show considerable heterogeneity. Subsequently, a Spearman’s rank-order correlation coefficient, which is a measure for linear correlation between two datasets and does not assume that both datasets are normally distributed, was applied. We chose $|\rho| > 0.5$ to infer a significant direct or indirect correlation between two parameters for this study. The correlation coefficient for Acc and AUC against the mean number of samples per class resulted in $\rho = -0.204$ / $\rho = -0.153$, respectively. Therefore, both results represent no significant correlation coefficient.

Discussion

The most important finding of the presented review was that imaging-driven diagnosis for MSK malignancies does not yet experience significant impact by ML applications and this has several reasons associated with data.

The main issue might be the availability of data. In most research institutes, a systematic and structured collection of quality data does not yet seem to take place or has only recently been introduced. This can be derived from the fact that datasets in general are comparably small and dataset size is not increasing yet. Consequently, even if according patient data is existing, this does not necessarily imply data is present in a format, validity, accessibility, consistency and completeness

Table 2 Continuous parameters with interval, median, mean IQR, and standard deviation

Continuous parameters					
Parameter	Interval	Median	IQR	Mean	Std
Year of publication	[1994; 2021]	2020	3	2018	6
Number of patients/cases	[1; 1565]	132.0	180.5	292.0	392.0
Healthy	[0; 381]	0.0	0.0	10.6	62.6
Benign	[0; 1061]	38.0	154.2	154.8	248.3
Intermediate	[0; 169]	0.0	4.6	9.3	32.0
Malignant	[12; 478]	69.5	79.5	115.1	113.4
Metastases	[0; 317]	0.0	4.3	17.1	60.4

IQR interquartile range, *std* standard deviation

Table 3 Discrete parameters with incidence and percentage share per entity

Discrete parameters			
Parameter	Entity	Σ	%
Study design	Retrospective	28	75.7%
	Prospective	3	8.1%
	Unknown	6	16.2%
Task	Classification	33	80.5%
	Segmentation	6	14.6%
	Object detection	2	4.9%
Model	AlexNet	1	1.9%
	LogitBoost	2	3.8%
	Support vector machine	14	26.4%
	U-Net	1	1.9%
	Efficient-Net	2	3.8%
	Logistic regression	2	3.8%
	Adaboost	1	1.9%
	Random forests	12	22.6%
	VGG19	1	1.9%
	k-nearest neighbour	1	1.9%
	Neural network	4	7.5%
	LASSO	1	1.9%
	VGG16	2	3.8%
	Decision tree	2	3.8%
	XGBoost	1	1.9%
	Inception v3	2	3.8%
	SegNet	1	1.9%
	Mask RCNN	1	1.9%
	Generalised linear model	1	1.9%
	ResNet-50	1	1.9%
Diagnosis oriented	Yes	27	71.1%
	No	11	28.9%
Outcome label	Segmented tumour	6	14.6%
	Tumour entities	7	17.1%
	Tumour occurrence	1	2.4%
	Histopathological grading	5	12.2%
	Radiotherapy response	2	4.9%
	Chemotherapy response	3	7.3%
	Malignancy	15	36.6%
	Staging	1	2.4%
	Prognosis	1	2.4%
	Tumour group	Bone tumour	23
Soft tissue tumour		15	39.5%

Table 3 (continued)

Discrete parameters			
Parameter	Entity	Σ	%
Imaging modality	MRI	22	55.0%
	CT	7	17.5%
	X-ray	10	25.0%
	US	1	2.5%
Radiomic data	Yes	16	42.1%
	No	22	57.9%
Algorithm	Supervised	37	97.4%
	Unsupervised	1	2.6%
	Reinforcement	0	0.0%
Applied metric	Accuracy	29	25.4%
	Sensitivity	25	21.9%
	Specificity	23	20.2%
	AUC	28	24.6%
	Jaccard index	1	0.9%
	Intersection over union	2	1.8%
	Dice score	6	5.3%

LASSO Least Absolute Shrinkage and Selection Operator

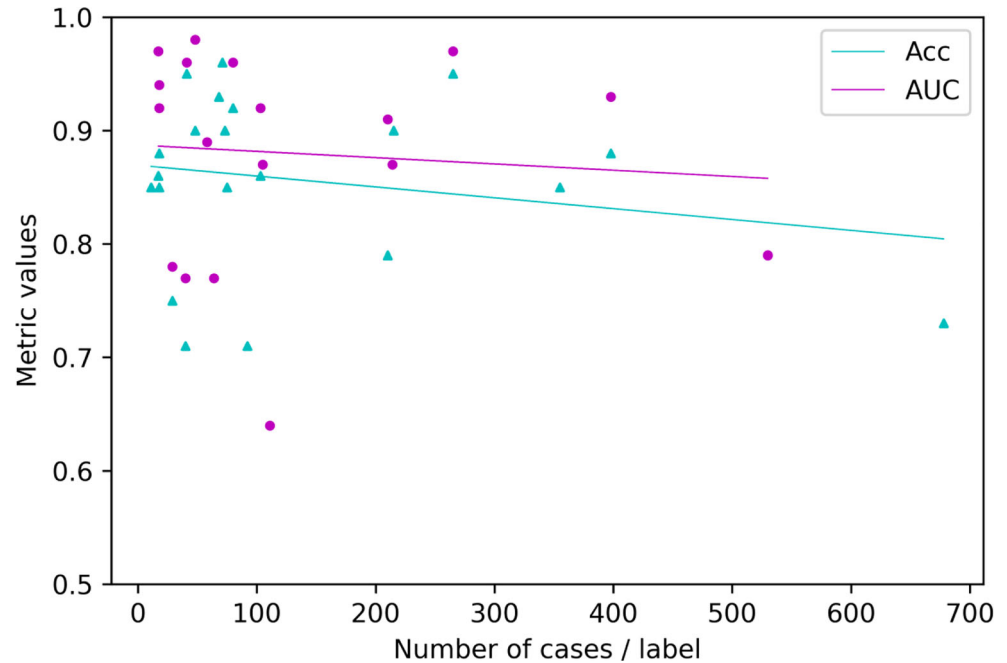
feasible for data science. In addition, sarcomas are a very rare entity of cancer, which does not allow for fast gathering of sufficient prospective data. Terenuma et al [41] developed a technique to obtain multiple images from a single patient, which is from a data science perspective very intriguing, but does not provide enough data for a clinical application and is not generally transferable to any other study. Several mathematical techniques to cope with limited data have emerged (e.g. transfer learning [10], data augmentation [9]). However, these techniques can at this point only support an AI task, but not solve the issue of limited data. For rare

Table 4 Continuous parameters of diagnosis-oriented studies with interval, median, mean and standard deviation

Continuous parameters of diagnosis-oriented parameters					
Parameter	Interval	Median	IQR	Mean	std
ACC	[0.71; 0.99]	0.88	0.07	0.87	0.07
AUC	[0.64; 0.98]	0.92	0.14	0.88	0.09
Number of labels	[2; 3]	2	0	2.19	0.39

IQR interquartile range, std standard deviation

Fig. 2 Distribution of final metric scores against the mean number of samples per class label



diseases, building networks and databases on a national or even international basis might be a future solution. Another reason might be the considerably limited amount of research in the field of orthopaedic oncology, which can again partly be explained by insufficient data. With the respectively adapted search term, more than 1300 articles can be found for lung malignancies and even more than 2200 articles for breast malignancies, while only 480 articles were detected for MSK malignancies (initial search, each in December 2021). ML in general is still in its infancy, but more so in MSK and orthopaedic oncology.

A further finding was presented by synthesising the relationship of number of cases and number of labels per class against the metric values. In the research field of AI, it is common knowledge that the amount of data has profound impact on the model performance [10, 11, 52]. Nonetheless, Fig. 2 tells a different story. The median number of samples per class resulted in 75 and 59.3% of the diagnosis-oriented studies had less than 100 samples per class. Further, the mean metric scores of studies with fewer than 100 samples per class (Acc 0.86, AUC 0.89) were slightly higher than those of studies with more than 100 samples per class (Acc 0.85, AUC 0.86), as indicated by the linear regression lines in Fig. 2. This would suggest that less data leads to higher results. One explanation for these unexpected results could be the class imbalance: several studies developed models to classify tumour malignancy, for example [15, 18, 19, 22, 26, 28, 32, 33, 35, 36, 39, 40, 44, 45]. Benign MSK tumours occur more often than malignant MSK tumours, which results in a class imbalance in the dataset. Such an imbalance can lead to spuriously high metric values, especially for AUC. A detailed and interdisciplinary interpretation of results with regard to

composition of data is crucial. Another issue associated with limited datasets and class imbalance is that specific classes of data might be sparse. Therefore, overfitting may occur, resulting in suboptimal results.

Yet another indication is that problem statements of most studies do not reflect real clinical scenarios. Most studies aim at distinguishing two to three specific tumour entities [10, 16, 34, 43, 46–48] or assessing tumour malignancy [15, 18, 19, 22, 26, 28, 32, 33, 35, 36, 39, 40, 42, 44, 45]. If one fed a third entity to a two-entity classifier, the model would try to fit the third entity into one of the first two entity classes. While confining a tumour entity from another is an imperative step in tumour assessment, nonetheless, most sarcoma diagnoses are incidental findings, and in daily practice, MSK radiologists and orthopaedic surgeons are first confronted with detecting a potential sarcoma at all [1, 4, 53]. Whereas von Schacky et al [42] aimed at differentiating various tumour entities, thus modelling a more realistic clinical scenario, the results were only moderate. More general models are needed to comply with clinical needs and difficulties. However, we hypothesise that this is again very difficult to achieve due to the very limited amount of data available and probably also closely related to the distribution of the data. Naturally, the quality and problems of AI models cannot be assessed by dataset size and data distribution alone, but data undoubtedly have major impact on the overall performance and clinical relevance.

No biopsy-focused studies

The most applied outcome labels among the 38 investigated original research articles were tumour malignancy (15, 36.6%) [15, 18, 19, 22, 26, 28, 32, 33, 35, 36, 39, 40, 42,

44, 45], tumour entities (7, 17.1%) [10, 16, 34, 43, 46–48] and segmented tumour (6, 14.6%) [16, 27, 31, 41, 46, 50]. A distinct finding of this review is that although a biopsy is a crucial step in the diagnostic process of MSK malignancies, there is no study focused on radiological images and biopsies. Retrieving relevant biopsy material—for example, via CT-guided needle biopsy—is a highly complex task and requires significant experience. From this, it could be derived that ML research in the field of MSK malignancies is currently not mainly oriented on medical needs, but models and research questions are built around available data. This underlines that ML is still in its very infancy in MSK tumour research.

MRI and radiomics

MRI is the most popular kind of imaging data for ML analysis at this point (55.0%, 22). This might be explained by the fact that MR imaging plays a fundamental role in the assessment of sarcomas due to superior soft tissue contrast and the desire to reduce unnecessary radiation dose. But also, from a data science perspective, this is comprehensible: with one patient, multiple 2D data samples (or one 3D data sample) are produced. Additionally, various image planes and weightings are possible. This suggests that less patients are necessary to acquire more data.

Likewise, radiomics appears to be on demand. 42.1% of articles (16) utilised radiomic data [15, 17, 19, 21, 23, 27, 28, 33, 34, 37, 43, 45–48, 51], while only 17.5% (7) integrated CT, 25.0% (10) X-ray and 2.5% (1) US. With radiomics, a large number of quantitative features can be extracted from imaging data. These are combined with other patient data and can be mined with modern techniques of e.g. bioinformatics and data science. In consequence, the popularity of radiomics might be associated with the capability to extract additional information from images and therefore tackle the issue of small datasets.

Limitations

This review article has several limitations. The major limitation is the early stage of the examined studies. Because ML in orthopaedic oncology is still in its infancy, most studies are also at an early stage, making it difficult to examine the impact of the studies presented and assess their quality. Most studies were not published until 2021. Further, the mean number of cases per study is 292. While a limited number of cases is related to the type of entities studied [53], the number is very small in the context of ML applications. These facts underline the early stage of the studies. Another limitation is the overall heterogeneity of the examined studies. We restricted the tumour entities and the type of data by the eligibility criteria. However, we did not impose any restrictions on ML algorithms, models, or tasks. Thus, the studies presented three

distinct algorithm types, 20 different models and nine groups of outcome labels for various tasks.

Conclusion

In conclusion, for a rare disease, there are very limited amounts of data and no established large-scale networks between multiple national and international facilities yet. The impact of imaging-driven ML research in other disciplines is already present [52]. Also, several studies presented in this review demonstrated that ML can selectively support imaging-driven diagnosis for MSK malignancies. However, until statistically robust results can be achieved and clinically relevant models to cope with heterogeneous cases an orthopaedic surgeon or MSK radiologist encounters on a regular basis can be developed, data quality and quantity have to be improved. An expert radiologist from a specialised centre has seen thousands of images in his/her professional life and incorporates meta data as well as other factors into his/her decision-making process. In contrast, the presented studies only worked with 1 [41] up to 1576 [16] cases mostly focusing on one single kind of data and imaging modality.

The key to bring ML to a level where it can substantially impact clinical image interpretation in the diagnosis of MSK malignancies is data: establishing national and international networks, implementing a systematic and structural data acquisition and finally integrating multimodal data comparable to expert radiologists.

Acknowledgements Many thanks to Fritz Seidl for his assistance with the language editing; looking forward to working with you again.

Funding Open Access funding enabled and organised by Projekt DEAL. The authors state that this work has not received any funding.

Declarations

Guarantor The scientific guarantor of this publication is Prof. MD Rainer Burgkart.

Conflict of interest The authors of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

Statistics and biometry One of the authors has significant statistical expertise (Florian Hinterwimmer).

Informed consent Not applicable

Ethical approval Not applicable

Methodology

- retrospective
- performed at one institution

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/>.

References

- Grimer RJ, Briggs TW (2010) Earlier diagnosis of bone and soft-tissue tumours. *J Bone Joint Surg Br* 92:1489–1492
- Grimer RJ, Carter SR, Pynsent PB (1997) The cost-effectiveness of limb salvage for bone tumours. *J Bone Joint Surg Br* 79:558–561
- Rechl H, Kirchhoff C, Wortler K, Lenze U, Topfer A, von Eisenhart-Rothe R (2011) Diagnosis of malignant bone and soft tissue tumors. *Orthopade* 40:931–94
quiz 942–933
- Clark MA, Thomas JM (2005) Delay in referral to a specialist soft-tissue sarcoma unit. *Eur J Surg Oncol* 31:443–448
- Ayala AG, Zomosa J (1983) Primary bone tumors: percutaneous needle biopsy. Radiologic-pathologic study of 222 biopsies. *Radiology* 149:675–679
- Mankin HJ, Mankin CJ, Simon MA (1996) The hazards of the biopsy, revisited. Members of the Musculoskeletal Tumor Society. *J Bone Joint Surg Am* 78:656–663
- Savage N (2020) How AI is improving cancer diagnostics. *Nature* 579:S14+
- Vogrin M, Trojner T, Kelc R (2020) Artificial intelligence in musculoskeletal oncological radiology. *Radiol Oncol* 55:1–6
- Zaman A, Park SH, Bang H, Park CW, Park I, Joung S (2020) Generative approach for data augmentation for deep learning-based bone surface segmentation from ultrasound images. *Int J Comput Assist Radiol Surg* 15:931–941
- Banerjee I, Crawley A, Bhethanabotla M, Daldrup-Link HE, Rubin DL (2018) Transfer learning on fused multiparametric MR images for classifying histopathological subtypes of rhabdomyosarcoma. *Comput Med Imaging Graph* 65:167–175
- Frangi AF, Tsafaris SA, Prince JL (2018) Simulation and synthesis in medical imaging. *IEEE Trans Med Imaging* 37:673–679
- Tricco AC, Lillie E, Zarin W et al (2018) PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med* 169:467–473
- Bandyopadhyay O, Biswas A, Bhattacharya BB (2019) Bone-cancer assessment and destruction pattern analysis in long-bone X-ray image. *J Digit Imaging* 32:300–313
- Chianca V, Albano D, Messina C et al (2021) An update in musculoskeletal tumors: from quantitative imaging to radiomics. *Radiol Med* 126:1095–1105
- Chianca V, Cuocolo R, Gitto S et al (2021) Radiomic machine learning classifiers in spine bone tumors: a multi-software, multi-scanner study. *Eur J Radiol* 137:109586
- Do NT, Jung ST, Yang HJ, Kim SH (2021) Multi-level seg-unet model with global and patch-based X-ray images for knee bone tumor detection. *Diagnostics*, 11(4):691
- Dufau J, Bouhamama A, Leporq B et al (2019) Prediction of chemotherapy response in primary osteosarcoma using the machine learning technique on radiomic data. *Bull Cancer* 106:983–999
- Eweje FR, Bao B, Wu J et al (2021) Deep learning for classification of bone lesions on routine MRI. *EBioMedicine* 68:103402
- Fields BKK, Demirjian NL, Hwang DH et al (2021) Whole-tumor 3D volumetric MRI-based radiomics approach for distinguishing between benign and malignant soft tissue tumors. *Eur Radiol* 31: 8522–8535
- Gao Y, Ghodrati V, Kalbasi A et al (2021) Prediction of soft tissue sarcoma response to radiotherapy using longitudinal diffusion MRI and a deep neural network with generative adversarial network-based data augmentation. *Med Phys* 48:3262–3372
- Gao Y, Kalbasi A, Hsu W et al (2020) Treatment effect prediction for sarcoma patients treated with preoperative radiotherapy using radiomics features from longitudinal diffusion-weighted MRIs. *Phys Med Biol* 65:175006
- García-Gómez JM, Vidal C, Martí-Bonmatí L et al (2004) Benign/malignant classifier of soft tissue tumors using MR imaging. *MAGMA* 16:194–201
- Gitto S, Cuocolo R, Albano D et al (2020) MRI radiomics-based machine-learning classification of bone chondrosarcoma. *Eur J Radiol* 128:109043
- Glass JO, Reddick WE (1998) Hybrid artificial neural network segmentation and classification of dynamic contrast-enhanced MR imaging (DEMRI) of osteosarcoma. *Magn Reson Imaging* 16:1075–1083
- Gorelik N, Chong J, Lin DJ (2020) Pattern recognition in musculoskeletal imaging using artificial intelligence. *Semin Musculoskelet Radiol* 24:38–49
- He Y, Pan I, Bao B et al (2020) Deep learning-based classification of primary bone tumors on radiographs: a preliminary study. *EBioMedicine* 62:103121
- Holbrook MD, Blocker SJ, Mowery YM et al (2020) MRI-based deep learning segmentation and radiomics of sarcoma in mice. *Tomography* 6:23–33
- Hu P, Chen L, Zhou Z (2021) Machine learning in the differentiation of soft tissue neoplasms: comparison of fat-suppressed T2WI and apparent diffusion coefficient (ADC) features-based models. *J Digit Imaging* 34:1146–1155
- Hu S, Xu C, Guan W, Tang Y, Liu Y (2014) Texture feature extraction based on wavelet transform and gray-level co-occurrence matrices applied to osteosarcoma diagnosis. *Biomed Mater Eng* 24: 129–143
- Huang B, Wang J, Sun M et al (2020) Feasibility of multiparametric magnetic resonance imaging combined with machine learning in the assessment of necrosis of osteosarcoma after neoadjuvant chemotherapy: a preliminary study. *BMC Cancer* 20:322
- Huang L, Xia W, Zhang B, Qiu B, Gao X (2017) MSFCN-multiple supervised fully convolutional networks for the osteosarcoma segmentation of CT images. *Comput Methods Programs Biomed* 143: 67–74
- Juntu J, Sijbers J, De Backer S, Rajan J, Van Dyck D (2010) Machine learning study of several classifiers trained with texture analysis features to differentiate benign from malignant soft-tissue tumors in T1-MRI images. *J Magn Reson Imaging* 31:680–689
- Leporq B, Bouhamama A, Pilleul F et al (2020) MRI-based radiomics to predict lipomatous soft tissue tumors malignancy: a pilot study. *Cancer Imaging* 20:78
- Li L, Wang K, Ma X et al (2019) Radiomic analysis of multiparametric magnetic resonance imaging for differentiating skull base chordoma and chondrosarcoma. *Eur J Radiol* 118:81–87
- Liu R, Pan D, Xu Y et al (2021) A deep learning-machine learning fusion approach for the classification of benign, malignant, and intermediate bone tumors. *Eur Radiol*. <https://doi.org/10.1007/s00330-021-08195-z>

36. Pan D, Liu R, Zheng B et al (2021) Using machine learning to unravel the value of radiographic features for the classification of bone tumors. *Biomed Res Int* 2021:8811056
37. Peeken JC, Bernhofer M, Spraker MB et al (2019) CT-based radiomic features predict tumor grading and have prognostic value in patients with soft tissue sarcomas treated with neoadjuvant radiation therapy. *Radiother Oncol* 135:187–196
38. Peeken JC, Goldberg T, Knie C et al (2018) Treatment-related features improve machine learning prediction of prognosis in soft tissue sarcoma patients. *Strahlenther Onkol* 194:824–834
39. Reinus WR, Wilson AJ, Kalman B, Kwasny S (1994) Diagnosis of focal bone lesions using neural networks. *Invest Radiol* 29:606–611
40. Shen R, Li Z, Zhang L et al (2018) Osteosarcoma patients classification using plain X-rays and metabolomic data. *Annu Int Conf IEEE Eng Med Biol Soc* 2018:690–693
41. Terunuma T, Tokui A, Sakae T (2018) Novel real-time tumor-contouring method using deep learning to prevent mistracking in X-ray fluoroscopy. *Radiol Phys Technol* 11:43–53
42. von Schacky CE, Wilhelm NJ, Schäfer VS et al (2021) Multitask deep learning for segmentation and classification of primary bone tumors on radiographs. *Radiology* 301:398–406
43. Vos M, Starmans MPA, Timbergen MJM et al (2019) Radiomics approach to distinguish between well differentiated liposarcomas and lipomas on MRI. *Br J Surg* 106:1800–1809
44. Wang B, Perronne L, Burke C, Adler RS (2021) Artificial intelligence for classification of soft-tissue masses at US. *Radiol Artif Intell* 3:e200125
45. Wang H, Zhang J, Bao S et al (2020) Preoperative MRI-based radiomic machine-learning nomogram may accurately distinguish between benign and malignant soft-tissue lesions: a two-center study. *J Magn Reson Imaging* 52:873–882
46. Yin P, Mao N, Zhao C, Wu J, Chen L, Hong N (2019) A triple-classification radiomics model for the differentiation of primary chordoma, giant cell tumor, and metastatic tumor of sacrum based on T2-weighted and contrast-enhanced T1-weighted MRI. *J Magn Reson Imaging* 49:752–759
47. Yin P, Mao N, Zhao C et al (2019) Comparison of radiomics machine-learning classifiers and feature selection for differentiation of sacral chordoma and sacral giant cell tumour based on 3D computed tomography features. *Eur Radiol* 29:1841–1847
48. Yin P, Zhi X, Sun C et al (2021) Radiomics models for the preoperative prediction of pelvic and sacral tumor types: a single-center retrospective study of 795 cases. *Front Oncol* 11:709659
49. Zhang L, Ren Z (2020) Comparison of CT and MRI images for the prediction of soft-tissue sarcoma grading and lung metastasis via a convolutional neural networks model. *Clin Radiol* 75:64–69
50. Zhang R, Huang L, Xia W, Zhang B, Qiu B, Gao X (2018) Multiple supervised residual network for osteosarcoma segmentation in CT images. *Comput Med Imaging Graph* 63:1–8
51. Zhang Y, Zhu Y, Shi X et al (2019) Soft tissue sarcomas: preoperative predictive histopathological grading based on radiomics of MRI. *Acad Radiol* 26:1262–1268
52. Rajpurkar P, Chen E, Banerjee O, Topol EJ (2022) AI in health and medicine. *Nat Med* 28:31–38
53. Picci P, Manfrini M, Donati D et al (2020) Diagnosis of Musculoskeletal Tumors and Tumor-like Conditions: Clinical, Radiological and Histological Correlations-the Rizzoli Case Archive (pp. 3–11). Cham: Springer

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

4.2. From Self-supervised Learning to Transfer Learning with Musculoskeletal Radiographs

Authors:

Florian Hinterwimmer, Sarah Consalvo, Jan Neumann, Carina Micheler, Nikolas Wilhelm, Jan Lang, Rüdiger von Eisenhart-Rothe, Rainer Burgkart, Daniel Rueckert

Conference:

Joint Annual Conference of the Austrian, German and Swiss Societies for Biomedical Engineering (BMT)

Synopsis:

In our study, we investigated the use of transfer learning with MSK radiographs to support classification tasks, specifically to distinguish Ewing's sarcoma from acute osteomyelitis in pediatric patients. We recognized the challenge of limited data in rare diseases and investigated whether transfer learning could solve this problem. Previous studies have demonstrated the effectiveness of transfer learning in various applications, including medical image interpretation. However, to our knowledge, no model has been pretrained specifically for MSK features in radiographs.

Our dataset consisted of 42,608 pseudonymized radiographs collected over a 25-year period from a MSK tumour center. The images included different MSK regions and had varying data quality, resolution, and sources. For analysis, we used a separate dataset of 63 images (22 acute osteomyelitis, 41 Ewing sarcoma) from pediatric patients. Our algorithm followed a two-step deep learning framework. In the first step, we employed a self-supervised model called DeepCluster to group the unstructured data into multiple clusters. DeepCluster used k-means to cluster the feature embeddings and used the resulting mappings as labels to update the network weights. The optimal number of clusters was determined through several test runs. In the second step, the cluster assignments served as "auxiliary" class labels for pretraining a ResNet50 model. We split the data for pretraining, validation, and hold-out testing. For the downstream classification task, we evaluated the pretrained model on a limited sample dataset and implemented cross-validation for statistical robustness.

Our results showed that the self-supervised clustering and subsequent transfer learning approach significantly improved the downstream classification accuracy. We achieved 89.6% accuracy in validation and 70.8% in testing, outperforming the accuracy of an untrained network (81.5%/54.2%) and an ImageNet pre-trained network

(81.5%/54.2%). The improvement was substantial, with a difference of 4.4 and 17.3 percentage points, respectively.

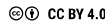
Our study has demonstrated the potential of using a large dataset of unstructured radiographs for transfer learning. In the field of orthopaedic oncology, where data is often limited, this approach can overcome the challenge of insufficient data for deep learning applications. While collecting more high-quality data remains critical for improving DL in medicine, techniques such as data augmentation and image synthesis can also support image interpretation tasks. We also found that pretraining with a larger dataset reduced classification overfitting problems. However, we acknowledge that our dataset is still relatively small compared to commonly used pretraining datasets such as ImageNet, and further validation is needed to establish general validity. Although we have achieved a significant improvement in accuracy results, our results are not yet clinically relevant. Systematic and structured data collection is essential for further development of deep learning applications. In conclusion, our study demonstrated the effectiveness of transfer learning on MSK radiographs, particularly in distinguishing between Ewing's sarcoma and acute osteomyelitis. Transfer learning proved to be a powerful technique even with limited datasets. However, it is not an overall solution, and structured data acquisition remains critical to achieving clinically relevant results in deep learning applications.

Contribution of thesis author:

Florian Hinterwimmer was the principal investigator in this study and contributed at least 50%.

Copyright:

From Self-supervised Learning to Transfer Learning with Musculoskeletal Radiographs © 2022 by Florian Hinterwimmer is licensed under Creative Commons Attribution 4.0 International. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>



ATTRIBUTION 4.0 INTERNATIONAL

Deed

Canonical URL: <https://creativecommons.org/licenses/by/4.0/>

[See the legal code](#)

You are free to:

Share — copy and redistribute the material in any medium or format for any purpose, even commercially.

Adapt — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

Attribution — You must give **appropriate credit**, provide a link to the license, and **indicate if changes were made**. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

No additional restrictions — You may not apply legal terms or **technological measures** that legally restrict others from doing anything the license permits.

Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable **exception or limitation**.

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as **publicity, privacy, or moral rights** may limit how you use the material.

Notice

This deed highlights only some of the key features and terms of the actual license. It is not a license and has no legal value. You should carefully review all of the terms and conditions of the actual license before using the licensed material.

Creative Commons is not a law firm and does not provide legal services. Distributing, displaying, or linking to this deed or the license that it summarizes does not create a lawyer-client or any other relationship.

Creative Commons is the nonprofit behind the open licenses and other legal tools that allow creators to share their work. Our legal tools are free to use.

- [Learn more about our work](#)
- [Learn more about CC Licensing](#)
- [Support our work](#)
- [Use the license for your own material](#)
- [Licenses List](#)
- [Public Domain List](#)

Footnotes

appropriate credit — If supplied, you must provide the name of the creator and attribution parties, a copyright notice, a license notice, a disclaimer notice, and a link to the material. CC licenses prior to Version 4.0 also require you to provide the title of the material if supplied, and may have other slight differences.

- [More info](#)

indicate if changes were made — In 4.0, you must indicate if you modified the material and retain an indicator of previous modifications. In 3.0 and earlier license versions, the indication of changes is only required if you create a derivative.

- [Marking guide](#)
- [More info](#)

technological measures — The license prohibits application of effective technological measures, defined with reference to Article 11 of the WPO Copyright Treaty.

- [More info](#)

exception or limitation — The rights of users under exceptions and limitations, such as fair use and fair dealing, are not affected by the CC licenses.

- [More info](#)

publicity, privacy, or moral rights — You may need to get additional permissions before using the material as you intend.

- [More info](#)

[Contact](#)

[Newsletter](#)

[Privacy](#)

[Policies](#)

[Terms](#)

CONTACT US

Creative Commons PO Box 1396, Mountain View, CA 94042

info@creativecommons.org

+1 415 433 8500

SUBSCRIBE TO OUR NEWSLETTER

Your email

[SUBSCRIBE](#)

SUPPORT OUR WORK

Our work relies on you. Help us keep the Internet free and open.

[DONATE NOW](#)

Except where otherwise noted, content on this site is licensed under [a Creative Commons Attribution 4.0 International License](#). Icons by [Font Awesome](#).

Florian Hinterwimmer*, Sarah Consalvo, Jan Neumann, Carina Micheler, Nikolas Wilhelm, Jan Lang, Rüdiger von Eisenhart-Rothe, Rainer Burgkart, and Daniel Rueckert

From Self-supervised Learning to Transfer Learning with Musculoskeletal Radiographs

<https://doi.org/10.1515/cdbme-2022-1003>

Abstract: Ewing sarcomas are malignant neoplasm entities typically found in children and adolescents. Early detection is crucial for therapy and prognosis. Due to the low incidence the general experience as well as according data is limited. Novel support tools for diagnosis, such as deep learning models for image interpretation, are required. While acquiring sufficient data is a common obstacle in medicine, several techniques to tackle small data sets have emerged. The general necessity of large data sets in addition to a rare disease lead to the question whether transfer learning can solve the issue of limited data and subsequently support tasks such as distinguishing Ewing sarcoma from its main differential diagnosis (acute osteomyelitis) in paediatric radiographs. 42,608 unstructured radiographs from our musculoskeletal tumour centre were retrieved from the PACS. The images were clustered with a DeepCluster, a self-supervised algorithm. 1000 clusters were used for the upstream task (pretraining). Following, the pretrained classification network was applied for the downstream task of differentiating Ewing sarcoma and acute osteomyelitis. An untrained network achieved an accuracy of 81.5%/54.2%, while an ImageNet-pretrained network resulted in 89.6%/70.8% for validation and testing, respec-

tively. Our transfer learning approach surpassed the best result by 4.4%/17.3% percentage points. Transfer learning demonstrated to be a powerful technique to support image interpretation tasks. Even for small data sets, the impact can be significant. However, transfer learning is not a final solution to small data sets. To achieve clinically relevant results, a structured and systematic data acquisition is of paramount importance.

Keywords: transfer learning, self-supervised learning, radiographs, sarcoma

1 Introduction

Ewing sarcomas are highly malignant tumour entities that occur predominantly in children and adolescents. Early detection and differentiation from other entities, especially acute osteomyelitis, are critical for therapy and prognosis and thus patient survival [17]. Because of the low incidence, experience especially in outpatient clinics is usually limited, so the chance of early detection is low [18]. Hence, new sophisticated diagnostic support tools are required. Deep learning (DL) has achieved great success in image interpretation in many other disciplines [11]. However, a common obstacle to the application of DL in medicine is the availability of a sufficient amount of data. Several techniques to cope with small data sets, such as data augmentation [15, 16], data synthesis, or transfer learning, have emerged. The general need of DL models for sufficient (training) data poses a challenge in the context of rare diseases. The question arises whether transfer learning can solve the problem of limited data and support specific tasks such as distinguishing Ewing sarcoma from acute osteomyelitis in pediatric radiographs. The presented study investigated if and how 42,608 unstructured radiographs can be integrated in a transfer learning approach to support minimal data sets in a classification task. In summary, we make the following contributions:

1. We demonstrate a novel transfer learning approach specifically developed with musculoskeletal radiographs by subsequent training of an already ImageNet-pretrained model.
2. We leverage a state-of-the-art self-supervised model to obtain weak auxiliary labels from 42,608 unstructured radiographs.

*Corresponding author: Florian Hinterwimmer, Technical University of Munich, Institute for AI and Informatics in Medicine & Department for Orthopaedics and Sports Orthopaedics, 81675 Munich, Germany, e-mail: florian.hinterwimmer@tum.de

Sarah Consalvo, Nikolas Wilhelm, Rüdiger von Eisenhart-Rothe, Rainer Burgkart, Technical University of Munich, Klinikum rechts der Isar, Department for Orthopaedics and Sports Orthopaedics, 81675 Munich, Germany

Jan Neumann, Technical University of Munich, Klinikum rechts der Isar, Institute for Diagnostic and Interventional Radiology and Paediatric Radiology, 81675 Munich, Germany

Daniel Rueckert, Technical University of Munich, Institute for AI and Informatics in Medicine, 81675 Munich, Germany

Carina Micheler, Technical University of Munich, Klinikum rechts der Isar, Department for Orthopaedics and Sports Orthopaedics, 81675 Munich, Germany & Institute for Machine Tools and Industrial Management, School of Engineering and Design, Technical University of Munich, Garching near Munich, Germany

Jan Lang, Technical University of Munich, Klinikum rechts der Isar, Department for Orthopaedics and Sports Orthopaedics, 81675 Munich, Germany & Chair of Non-destructive Testing, School of Engineering and Design, Technical University of Munich, Munich, Germany

3. We underline the importance of sufficient data by showing that transfer learning is a powerful technique, but not a sole solution to limited data sets.

1.1 Related work

Transfer learning was first proposed in 1976 by Bozinovski and Fulgosi [4]. Since then, it has found various applications and shown great impact [2–4, 6–8]. The most popular transfer learning models are pretrained on ImageNet [8]. While these models are trained on every day images such as landscape-, cat- and dog-images, the pretraining still shows significant improvement also in medical image interpretation. Recently, several transfer learning approaches in the context of Covid19 detection and classification tasks have been published [2, 7]. These studies, due to the nature of the disease, focus on thorax images. However, to our knowledge, no model generally pretrained for musculoskeletal features in radiographs has been demonstrated.

2 Materials and methods

2.1 Data sets

The data set consisted of 42,608 unstructured, pseudonymised radiographs from a musculoskeletal tumour centre. All images belonged to patients with sarcoma associated ICD codes. Sarcomas typically occur in extremities and joints. Additionally, the data set contained images, which were initiated to check for metastases or monitor progress after surgery or therapy. Therefore, it is to be expected that any possible musculoskeletal region is included. The DICOM images were retrieved from the local PACS (Picture Archiving and Communication System) at Klinikum rechts der Isar (Munich). The imaging data was gathered over the past 25 years and contained corrupted and false data as well as heterogeneous data quality, resolution and external images. The DICOM header information was fully blinded, so that no meta-information for statistical analysis remained. For assessment of the transfer learning approach, a second data set consisting of 63 images (22 acute osteomyelitis, 41 Ewing sarcoma) from patients under 18 years of age was used. No further restrictions regarding age, musculoskeletal features or sex were made.

2.2 Model training

Model training and inference was conducted on a DGX Station A100 with four 80GB graphical processing units

(Nvidia Corporation, Santa Clara, CA), 64 2.25 GHz cores and 512 GB DDR4 system memory running on a Linux/Ubuntu 20.04 distribution (Canonical, London, UK). Preprocessing and model implementation were performed in Python 3.9.6 (<https://www.python.org/>) using PyTorch 1.9.0 and cuda toolkit 11.1 (<https://pytorch.org/>). The pretrained model of this study will be provided upon publication.

2.3 Algorithm

We developed a two step deep learning framework to pretrain a classification network on an upstream task and subsequently evaluate it on a downstream task with different data from the same domain (musculoskeletal radiographs). In step one, the unstructured data set was clustered by a self-supervised model [5] into several clusters: DeepCluster presents a self-supervising approach to learning image representations. It iteratively groups features using k-means and uses the subsequent assignments as labels to update the weights of a network. The optimal number of clusters was determined through test runs measured by highest pretraining classification scores. In step two, the cluster assignments were used as "auxiliary" class labels for a classification task, whereby a ResNet50 [13] was pretrained. The data split for pretraining was 80%, 10%, 10% for training, validation and hold-out testing. Next, the pretrained model was applied to a two-entity classification task with limited samples for both entities with a data split of 80%, 10%, 10% for assessment of the transfer learning approach. To provide statistical robust results and avoid cross-contamination, a cross-validation was implemented. Accuracy values were calculated to evaluate the results. Figure 1 displays the workflow including the five steps from unstructured data to the final pretrained model.

2.4 Hyperparameters and runtime

For the upstream task a batch size of 512, a learning rate of 0.05 and 500 epochs were chosen. The runtime was ~ 7.5 hours. For the downstream task a batch size of 4, a learning rate of 0.0001 and 100 epochs were chosen. The runtime for the all cross-validation folds was ~ 2 hours. The inference step for all folds took ~ 7 minutes.

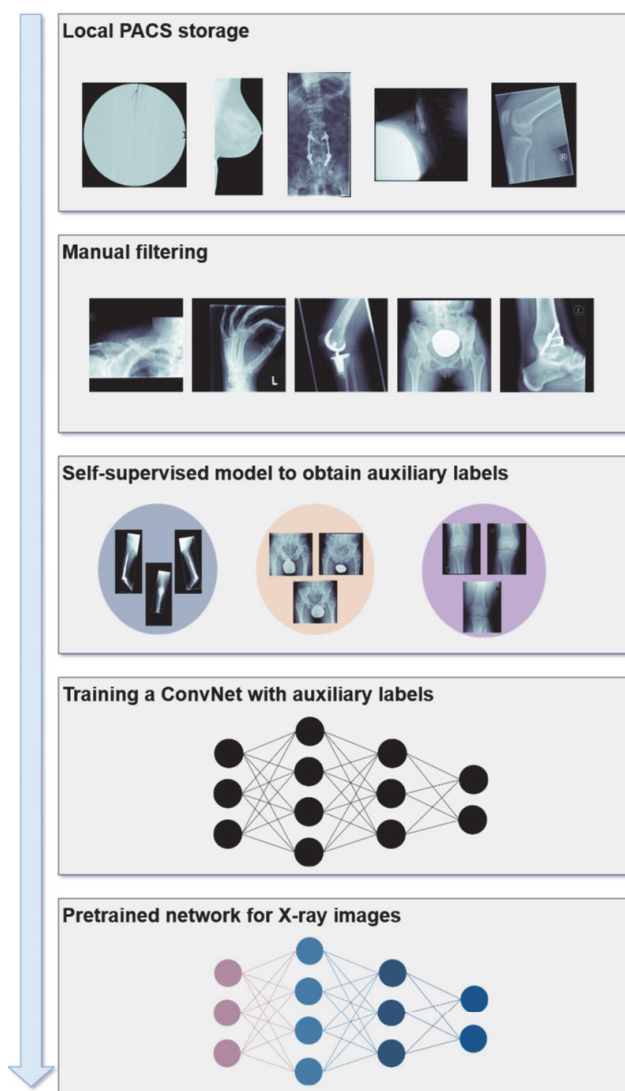


Fig. 1: Illustration of workflow with respective data samples.

Tab. 1: Classification results of Ewing sarcoma vs. acute osteomyelitis

Model	Val Acc	Test Acc
ResNet50	81.5 %	54.2 %
ResNet50 pretrained (ImageNet)	89.6 %	70.8 %
ResNet50 pretrained (our approach)	94.0 %	88.1 %

3 Results

3.1 Upstream task: from clustering to pretraining

The highest pretraining classification scores were achieved with 1000 clusters. The clustering of 42,608 images into 1000 clusters resulted in a normalised mutual information of 0.930. The smallest cluster was comprised of 2, the largest of 135 image samples with a first quartil of 20, a median of 46 and a third quartil of 55. The pretraining of the ResNet50 achieved an accuracy of 86.7%/80.0% for validation and testing respectively.

3.2 Downstream task: final classification

An untrained network achieved an accuracy of 81.5%/54.2%, while an ImageNet-pretrained network resulted in 89.6%/70.8% for validation and testing in the downstream task, respectively. Our transfer learning approach surpassed the best result by 4.4 and 17.3 percentage points (table 1).

4 Discussion

The most important finding of this study was that 42,608 unstructured radiographs can be utilised for transfer learning by leveraging a modern self-supervised model, thus significantly improving downstream classification tasks.

The obstacle of insufficient data for state-of-the-art deep learning applications is very common in medicine and especially in a field, such as orthopaedic oncology, where incidence is low and consequently data is limited. While collecting more quality data is probably the most effective way to improve deep learning applications in medicine, new techniques also need to be (further) developed. For example data augmentation [15, 16] or image synthesis [14] have shown to support various image interpretation tasks. We developed a transfer learning approach specified for radiographs with bone and soft tissue tumours. Most certainly though, our pretrained network will also improve other tasks working with radiographs of human

patients.

Another noticeable finding is that the not-pretrained network seemed to be overfitting and both pretrained networks seemed to mitigate this effect, thus, underlining the positive impact of pretraining with bigger data sets.

The major limitation of this study is that in contrast to the common data sets applied for pretraining (for example ImageNet [8], currently more than 14 million images), our data set is still comparably small. Therefore, the overall validity is still to be proven. However, for the particular task of distinguishing radiographs of Ewing and acute osteomyelitis patients, we achieved noticeable improvement. Although we were able to increase accuracy scores significantly (1), we did not reach clinically relevant results, yet. In the future, systematic and structured data collection will be of utmost importance for the improvement of DL applications.

4.1 Conclusion

Transfer learning has proven to be a powerful technique for supporting image interpretation tasks. Even for very limited data sets, the impact can be significant. However, transfer learning is not an overall solution for small data sets. To achieve clinically relevant results, structured and systematic data collection is of paramount importance.

Author Statement

Research funding: The author state no funding involved. Conflict of interest: Authors state no conflict of interest. Informed consent: Informed consent has been obtained from all individuals included in this study. Ethical approval: The research related to human use complies with all the relevant national regulations, institutional policies and was performed in accordance with the tenets of the Helsinki Declaration, and has been approved by the authors' institutional review board or equivalent committee.

References

- [1] Aiello, M., C. Cavaliere, A. D'Albore and M. Salvatore (2019). "The challenges of diagnostic imaging in the era of big data." *Journal of clinical medicine* 8(3): 316.
- [2] Al-Rakhami, M. S., M. M. Islam, M. Z. Islam, A. Asraf, A. H. Sodhro and W. Ding (2021). "Diagnosis of COVID-19 from X-rays using combined CNN-RNN architecture with transfer learning." *MedRxiv*: 2020.2008. 2024.20181339.
- [3] Banerjee, I., A. Crawley, M. Bhethanabotla, H. E. Daldrup-Link and D. L. Rubin (2018). "Transfer learning on fused multiparametric MR images for classifying histopathological subtypes of rhabdomyosarcoma." *Comput Med Imaging Graph* 65: 167-175.
- [4] Bozinovski, S. and A. Fulgosi (1976). The influence of pattern similarity and transfer of learning upon training of a base perceptron B2.(original in Croatian: Utjecaj slicnosti likova i transfera učenja na obucavanje baznog perceptrona B2). *Proc. Symp. Informatica*.
- [5] Caron, M., P. Bojanowski, A. Joulin and M. Douze (2018). Deep clustering for unsupervised learning of visual features. *Proceedings of the European conference on computer vision (ECCV)*.
- [6] Chhikara, P., P. Singh, P. Gupta and T. Bhatia (2020). Deep convolutional neural network with transfer learning for detecting pneumonia on chest X-rays. *Advances in Bioinformatics, Multimedia, and Electronics Circuits and Signals*, Springer: 155-168.
- [7] Das, N. N., N. Kumar, M. Kaur, V. Kumar and D. Singh (2020). "Automated deep transfer learning-based approach for detection of COVID-19 infection in chest X-rays." *Irbm*.
- [8] Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, IEEE.
- [9] Ghesu, F. C., B. Georgescu, A. Mansoor, Y. Yoo, D. Neumann, P. Patel, et al. (2022). "Self-supervised Learning from 100 Million Medical Images." *arXiv preprint arXiv:2201.01283*.
- [10] Nguyen, X.-B., G. S. Lee, S. H. Kim and H. J. Yang (2020). "Self-supervised learning based on spatial awareness for medical image analysis." *IEEE Access* 8: 162973-162981.
- [11] Rajpurkar, P., E. Chen, O. Banerjee and E. J. Topol (2022). "AI in health and medicine." *Nat Med* 28(1): 31-38.
- [12] Willeminck, M. J., W. A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, et al. (2020). "Preparing medical imaging data for machine learning." *Radiology* 295(1): 4-15.
- [13] He, K., X. Zhang, S. Ren and J. Sun (2016). "Deep residual learning for image recognition". *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- [14] Cao, B., H. Zhang, N. Wang, X. Gao and D. Shen (2020). Auto-GAN: self-supervised collaborative learning for medical image synthesis. *Proceedings of the AAAI Conference on Artificial Intelligence*.
- [15] Chlap, P., H. Min, N. Vandenberg, J. Dowling, L. Holloway and A. Haworth (2021). "A review of medical image data augmentation techniques for deep learning applications." *Journal of Medical Imaging and Radiation Oncology* 65(5): 545-563.
- [16] Gao, Y., V. Ghodrati, A. Kalbasi, J. Fu, D. Ruan, M. Cao, et al. (2021). "Prediction of soft tissue sarcoma response to radiotherapy using longitudinal diffusion MRI and a deep neural network with generative adversarial network-based data augmentation." *Med Phys* 48(6): 3262-3372.
- [17] Picci, P., M. Manfrini, D. Donati, M. Gambarotti, A. Righi, D. Vanel et al. (2020). Diagnosis of Musculoskeletal Tumors and Tumor-like Conditions Clinical, Radiological and Histological Correlations - The Rizzoli Case Archive: Clinical, Radiological and Histological Correlations - The Rizzoli Case Archive.
- [18] Clark, M. A. and J. M. Thomas (2005). "Delay in referral to a specialist soft-tissue sarcoma unit." *Eur J Surg Oncol* 31(4): 443-448.

4.3. SAM-X: sorting algorithm for musculoskeletal x-ray radiography

Authors:

Florian Hinterwimmer, Sarah Consalvo, Nikolas Wilhelm, Fritz Seidl, Rainer Burgkart, Rüdiger von Eisenhart-Rothe, Daniel Rueckert, Jan Neumann

Journal:

European Radiology

Synopsis:

In our study, we addressed challenges of MSK diseases and their impact on individuals, healthcare and society. These diseases cause pain, limit movement and reduce the quality of life of those affected. Medical imaging plays a critical role in the diagnosis and treatment of these conditions, and with advances in AI and neural networks, the potential exists for improved radiological examinations. However, the complexity of medical imaging and the increasing amount of data present challenges to radiologists and AI systems. To effectively analyze and understand radiological data, we propose a sorting algorithm called SAM-X for MSK radiographs. This algorithm aims to automatically classify and organize large image datasets based on anatomical features to facilitate data analysis.

We explore the use of self-supervised models and modify an existing model called DeepCluster for our classification task. This model uses a k-means clustering algorithm to group image features and uses the resulting labels to update network weights. By incorporating human interaction through weak semantic label assignment, we minimize the need for time-consuming annotation by domain experts. We collected a dataset of 42,608 pseudonymised radiographs from a MSK tumour centre. These images were collected over a 25-year period and varied in quality, resolution, and source. Our dataset included different MSK regions, anatomic variations, and medical implants. For evaluation, we used statistical measures of analysis such as normalised mutual information (NMI) and precision metric. The first clustering phase achieved an NMI of 0.930, and an experienced radiologist identified 28 main MSK classes from the clusters. Non-MSK images were discarded, giving us a subset of 29,433 images for further training. In the second phase, we trained a CNNs using the weak semantic labels assigned in the previous phase. The weights of the network were updated based on the labels provided, resulting in a cross-validated classification accuracy of 96.2% for validation data and 96.6% for hold-out test data. Accuracy increased to 99.7%

when the top two predicted class labels were considered. To gain insights into the decision-making process of the AI model, we implemented Grad-CAMs that highlight relevant information for classification. The degree CAM results confirmed that the algorithm focuses on anatomical regions that are relevant to each class.

Our study demonstrates the effectiveness of the SAM-X sorting algorithm for MSK radiographs. It achieves high accuracy and reliability in categorizing radiographs based on anatomical features. The algorithm correctly identifies and categorizes different pathologies and appearances within the same anatomical region. The integration of AI systems and neural networks in medical imaging holds great potential for improving clinical routine and research. The human-in-the-loop setup, as demonstrated in our study, enables active collaboration between humans and AI systems. Radiology is at the forefront of adopting new imaging technologies, and our proposed algorithm can improve workflow efficiency and productivity in the face of increasing data volumes. By categorizing radiographic images by anatomical features, we provide a valuable tool for managing and analysing large amounts of image data. This approach is consistent with the classification of MSK disorders based on anatomical locations. It also complements existing research on categorizing MSK disorders for occupational health care, surveillance, or research purposes.

While similar content-based image retrieval models have been proposed in other medical fields, our SAM-X algorithm fills a gap in categorizing MSK radiographs. We are aware of the limitations of our study, such as the assumption that any input image is in fact a radiographic image and related to the defined radiographic classes. In conclusion, our sorting algorithm for MSK radiographs, SAM-X, provides an effective and efficient framework for automatic classification and inference in large images.

Contribution of thesis author:

Florian Hinterwimmer was the principal investigator in this study and contributed at least 50%.

Copyright:

SAM-X: sorting algorithm for musculoskeletal x-ray radiography © 2023 by Florian Hinterwimmer is licensed under Creative Commons Attribution 4.0 International. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

SAM-X: sorting algorithm for musculoskeletal x-ray radiography



Author: Florian Hinterwimmer et al

Publication: European Radiology

Publisher: Springer Nature

Date: Oct 29, 2022

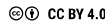
Copyright © 2022, The Author(s)

Creative Commons

This is an open access article distributed under the terms of the [Creative Commons CC BY](#) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.

To request permission for a type of use not listed, please contact [Springer Nature](#)



ATTRIBUTION 4.0 INTERNATIONAL

Deed

Canonical URL: <https://creativecommons.org/licenses/by/4.0/>

[See the legal code](#)

You are free to:

Share — copy and redistribute the material in any medium or format for any purpose, even commercially.

Adapt — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

Attribution — You must give **appropriate credit**, provide a link to the license, and **indicate if changes were made**. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

No additional restrictions — You may not apply legal terms or **technological measures** that legally restrict others from doing anything the license permits.

Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable **exception or limitation**.

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as **publicity, privacy, or moral rights** may limit how you use the material.

Notice

This deed highlights only some of the key features and terms of the actual license. It is not a license and has no legal value. You should carefully review all of the terms and conditions of the actual license before using the licensed material.

Creative Commons is not a law firm and does not provide legal services. Distributing, displaying, or linking to this deed or the license that it summarizes does not create a lawyer-client or any other relationship.

Creative Commons is the nonprofit behind the open licenses and other legal tools that allow creators to share their work. Our legal tools are free to use.

- [Learn more about our work](#)
- [Learn more about CC Licensing](#)
- [Support our work](#)
- [Use the license for your own material](#)
- [Licenses List](#)
- [Public Domain List](#)

Footnotes

appropriate credit — If supplied, you must provide the name of the creator and attribution parties, a copyright notice, a license notice, a disclaimer notice, and a link to the material. CC licenses prior to Version 4.0 also require you to provide the title of the material if supplied, and may have other slight differences.

- [More info](#)

indicate if changes were made — In 4.0, you must indicate if you modified the material and retain an indicator of previous modifications. In 3.0 and earlier license versions, the indication of changes is only required if you create a derivative.

- [Marking guide](#)
- [More info](#)

technological measures — The license prohibits application of effective technological measures, defined with reference to Article 11 of the WPO Copyright Treaty.

- [More info](#)

exception or limitation — The rights of users under exceptions and limitations, such as fair use and fair dealing, are not affected by the CC licenses.

- [More info](#)

publicity, privacy, or moral rights — You may need to get additional permissions before using the material as you intend.

- [More info](#)

[Contact](#)

[Newsletter](#)

[Privacy](#)

[Policies](#)

[Terms](#)

CONTACT US

Creative Commons PO Box 1396, Mountain View, CA 94042

info@creativecommons.org

+1 415 433 8800

SUBSCRIBE TO OUR NEWSLETTER

Your email

[SUBSCRIBE](#)

SUPPORT OUR WORK

Our work relies on you. Help us keep the Internet free and open.

[DONATE NOW](#)

Except where otherwise noted, content on this site is licensed under [Creative Commons Attribution 4.0 International License](#). Icons by [Font Awesome](#).



SAM-X: sorting algorithm for musculoskeletal x-ray radiography

Florian Hinterwimmer^{1,2} · Sarah Consalvo¹ · Nikolas Wilhelm¹ · Fritz Seidl³ · Rainer H. H. Burgkart¹ · Rüdiger von Eisenhart-Rothe¹ · Daniel Rueckert² · Jan Neumann⁴

Received: 1 July 2022 / Revised: 14 September 2022 / Accepted: 19 September 2022
© The Author(s) 2022

Abstract

Objective To develop a two-phased deep learning sorting algorithm for post-X-ray image acquisition in order to facilitate large musculoskeletal image datasets according to their anatomical entity.

Methods In total, 42,608 unstructured and pseudonymized radiographs were retrieved from the PACS of a musculoskeletal tumor center. In phase 1, imaging data were sorted into 1000 clusters by a self-supervised model. A human-in-the-loop radiologist assigned weak, semantic labels to all clusters and clusters with the same label were merged. Three hundred thirty-two non-musculoskeletal clusters were discarded. In phase 2, the initial model was modified by “injecting” the identified labels into the self-supervised model to train a classifier. To provide statistical significance, data split and cross-validation were applied. The hold-out test set consisted of 50% external data. To gain insight into the model’s predictions, Grad-CAMs were calculated.

Results The self-supervised clustering resulted in a high normalized mutual information of 0.930. The expert radiologist identified 28 musculoskeletal clusters. The modified model achieved a classification accuracy of 96.2% and 96.6% for validation and hold-out test data for predicting the top class, respectively. When considering the top two predicted class labels, an accuracy of 99.7% and 99.6% was accomplished. Grad-CAMs as well as final cluster results underlined the robustness of the proposed method by showing that it focused on similar image regions a human would have considered for categorizing images.

Conclusion For efficient dataset building, we propose an accurate deep learning sorting algorithm for classifying radiographs according to their anatomical entity in the assessment of musculoskeletal diseases.

Key Points

- Classification of large radiograph datasets according to their anatomical entity.
- Paramount importance of structuring vast amounts of retrospective data for modern deep learning applications.
- Optimization of the radiological workflow and increase in efficiency as well as decrease of time-consuming tasks for radiologists through deep learning.

Keywords Artificial intelligence · Deep learning · X-ray · Musculoskeletal diseases · Workflow

Abbreviations

AI	Artificial intelligence
CBIR	Content-based image retrieval
DL	Deep Learning
NMI	Normalized mutual information
PACS	Picture Archiving and Communication System

Introduction

Musculoskeletal diseases present a daily and also a global challenge for today’s healthcare system with far-reaching economic burdens to society and consequences to each individual who is affected by this disease. Finally, they result in pain and restriction of motion and interfering with the individuals’

✉ Florian Hinterwimmer
florian.hinterwimmer@tum.de

¹ Department of Orthopaedics and Sports Orthopaedics, Klinikum rechts der Isar, Technical University of Munich, Ismaninger Str. 25, 81675 Munich, Germany

² Institute for AI and Informatics in Medicine, Technical University of Munich, Munich, Germany

³ Department of Trauma Surgery, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany

⁴ Department of Diagnostic and Interventional Radiology, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany

quality of life [1]. Presenting a diverse group with respect to their pathophysiology, most conditions are, at least in part, classified according to the anatomical entity in which they are located. Hence, anatomy is being a crucial organizing principle for such diseases [2]. Accompanied by the ever-ongoing process of improving medical imaging with rapid changes and innovation [3], the recent dawning era of artificial intelligence (AI) and artificial neural networks will potentially lead to an increase of radiology exams. This will be especially true for medical imaging since it presents a cornerstone in the daily clinical routine and the workflow of shared care of patients with musculoskeletal diseases. Along the journey of new AI technologies, a variety of medical applications related to medical imaging has been noticed with the majority focusing on a head-to-head comparison of AI with humans [4]. However, modern imaging is more likely to involve human-in-the-loop setups, where humans actively collaborate with AI systems and provide oversight. When facing an increase of image data, the complexity of the radiology imaging workflow and corresponding amount of data promotes the need for possibilities to understand and use radiology data for gaining new knowledge and insights [5]. The complexity of medical imaging technologies implements the challenge for radiologists and neural networks to capture all details of each dataset, potentially reversing the hoped-for effect of reducing medical costs and time consumption when combining the workflow of human-AI collaboration. In order to avoid trade-offs to manage complex datasets in an active collaboration of humans with AI systems [6], pre-sorting algorithms according to the anatomical entity can be helpful to categorize large amount of image data, thus resulting in a more effective data analysis.

Multiple supervised, unsupervised, as well as self-supervised models have emerged [7–11] over the past years. These models could be utilized for sorting data. However, supervised learning requires a significant amount of annotated data and therefore demands for a substantial amount of time of a domain expert [12, 13]. In contrast, unsupervised and self-supervised models eliminate the need for time-consuming annotations, but clustering data follows mathematical rules such as similarity measures and consequently does not necessarily cluster data according to specific needs. In this study, we focus on utilization and modification of an established self-supervised model to categorize data effectively and efficiently according to specific requirements, while still keeping the demand of human interaction at a minimum. DeepCluster [8] demonstrates a self-supervising approach for learning image representation. The model iteratively groups features using a standard k-means clustering algorithm and uses the subsequent labels as supervision to update the weights of the network.

Therefore, in the present study, we propose a sorting algorithm for musculoskeletal X-ray radiography (SAM-X), a novel framework to support automatic classification and reasoning in the context of large image datasets.

Materials and methods

Dataset

The local institutional review and ethics board approved this retrospective study (N°48/20S). The study was performed in accordance with national and international guidelines. Informed consent was waived for this retrospective and anonymized study.

In total, 42,608 unstructured and pseudonymized radiographs were retrieved from the local Picture Archiving and Communication System (PACS) from a musculoskeletal tumor center (Klinikum rechts der Isar, Technical University of Munich) with sarcoma-associated ICD codes in DICOM format. The image data were collected over the past 25 years and contained heterogeneous data quality, resolution, and external images (~20%). Metadata such as DICOM header information or diagnoses are not yet validated and therefore not yet available.

All radiographs have been obtained through standard radiography techniques according to the body part imaged and in accordance with the radiographic manual procedures of our institution. Based on the aforementioned sarcoma-associated ICD code selection, data were acquired on potentially various points in time of the respective therapy status. Due to the various radiographic appearances and ubiquitous locations of sarcomas, our dataset includes a variety of all body parts, including potential anatomic variations, prosthetic devices, and medical implants. Figure 1 demonstrates examples from the dataset.

Statistical analysis

The result of the initial clustering (phase 1) was assessed through a normalized mutual information (NMI) metric. NMI is a variant of a measure commonly used in information theory, called mutual information. Mutual information indicates the “amount of information” that can be extracted from one distribution with respect to a second one. The results of sorting the images in the classification task were measured with an accuracy score calculating the amount of correctly assigned images with respect to all images from the hold-out test set. Additionally, the accuracy for considering the two most probable prediction labels was computed. Since meta-information is not yet available, no distribution analysis of sex, gender, diagnoses, etc. was conducted.

Model training

Model training and inference were conducted on a DGX Station A100 with four 80-GB graphical processing units (Nvidia Corporation, Santa Clara), 64 2.25-GHz cores, and 512-GB DDR4 system memory running on a Linux/

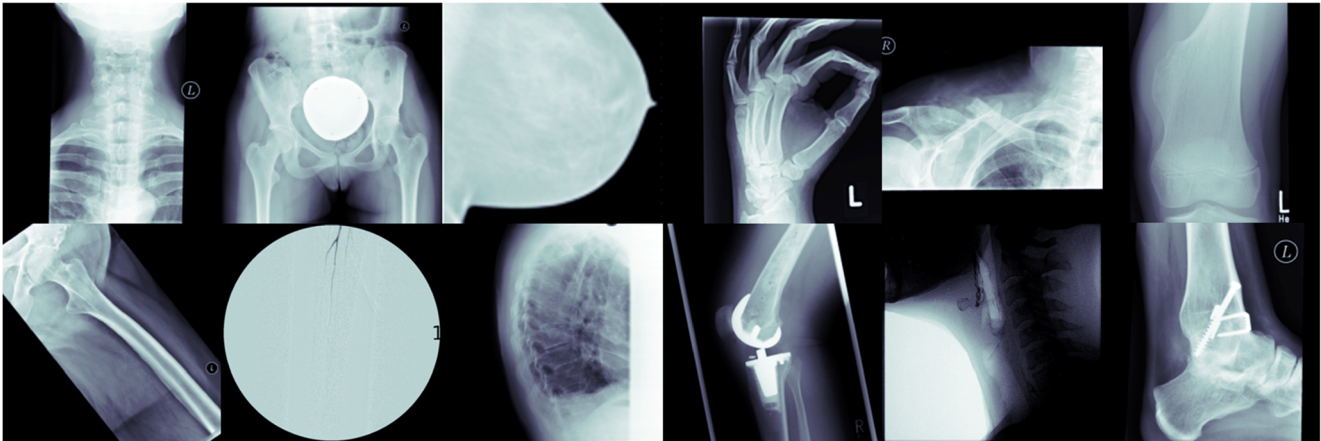


Fig. 1 Exemplary data sample showing various anatomical entities from the initial unstructured data collective

Ubuntu 20.04 distribution (Canonical). Preprocessing and model implementation were performed in Python 3.9.6 (<https://www.python.org/>) using PyTorch 1.10.2 and cuda toolkit 11.3 (<https://pytorch.org/>). The trained classification model will be available on GitHub (<https://github.com/>) upon publication.

Algorithm

A two-phase deep learning framework was developed consisting of a self-supervised model, human interaction

through weak, semantic label assignment, and implementation of a supervised learning task for final training (Fig. 2).

In phase 1, the 42,608 unstructured and pseudonymized X-ray images from a musculoskeletal research storage were clustered into 1000 clusters by application of DeepCluster [8]. Following, a senior radiologist identified several musculoskeletal labels by screening the results from phase 1. Each of the 1000 clusters was either assigned a class label or discarded, since the images were not applicable for the task at hand (e.g., not a musculoskeletal X-ray such as upper gastrointestinal series/barium swallow, mammography). Clusters with the same class label were merged, so

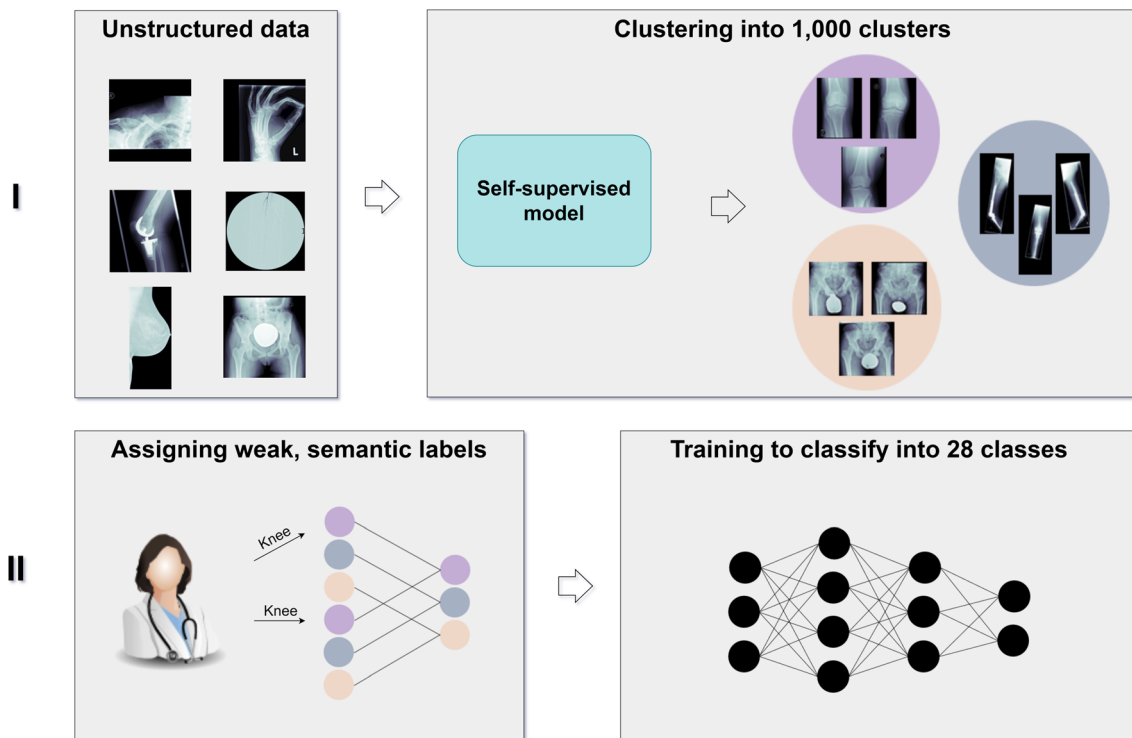


Fig. 2 Illustration of the presented framework in two phases: clustering data with a self-supervised model and training a network with human-annotated clusters

consequently “weak” (noisy) classes emerged. In phase 2, a convolutional neural network was trained on the emerged classes as weak, semantic labels: to accomplish an optimal training and update of the network’s weights, the created labels were “injected” into the same self-supervised model. The auxiliary labels for each training iteration from k-means became obsolete, since we provided the labels. It became a supervised task and the network’s weights were trained with respect to our classification requirements. In order to provide statistical significance and avoid cross-contamination, the data was split into training, validation, and hold-out test sets with a respective split ratio of 6-2-2, hence a 5-fold cross-validation was implemented. Half of the hold-out test set consisted of external imaging data to provide an independent and unbiased test set and increase significance of the results.

Plausibility

To add plausibility and additional insight into the AI model, Grad-CAMs were implemented in the final inference step [14]. Grad-CAMs utilize the gradient information from the last convolutional layer of a deep learning network to understand specific neurons and their impact for decision-making. The result is a colored heat map, which is co-registered to the original input image and indicates where the algorithm found relevant information for the task at hand. This technique was applied to acquire a better understanding where the algorithm detects relevant information. To provide a higher expressiveness, the Grad-CAM results were averaged from the 5-fold cross-validation.

Results

In phase 1, an NMI of 0.930 in clustering the entire dataset was reached. Subsequently, a senior radiologist identified the following 28 main musculoskeletal classes: abdomen, ankle, calcaneus, cervical spine, clavícula, elbow, femur, finger, foot, forearm, hand, hip, humerus, knee, lower leg, lumbar spine, paranasal sinus, patella, pelvis, ribs, sacrum, shoulder, skull, spine, thoracic spine, thorax, whole leg (standing), and wrist. In total, 13,175 non-musculoskeletal images from three hundred thirty-two clusters were discarded and a “musculoskeletal subset” of 29,433 images remained for further training. Table 1 shows the final classes with the number of images per class and percentage share with respect to the musculoskeletal subset.

In phase 2, a cross-validated classification accuracy of 96.2% for validation and 96.6% for hold-out test data was accomplished, when only considering the class with the highest prediction probability. When considering the top two predicted class labels, an accuracy of 99.7% and 99.6% for

Table 1 Final classes with the respective number of image samples and percentage share

Distribution of 28 classes		
Class	Number of images	in %
Abdomen	495	1.7%
Ankle	1033	3.5%
Calcaneus	177	0.6%
Cervical spine	3521	12.0%
Clavícula	100	0.3%
Elbow	186	0.6%
Femur	4270	14.5%
Finger	288	1.0%
Foot	904	3.1%
Forearm	192	0.7%
Hand	345	1.2%
Hip	1529	5.2%
Humerus	811	2.8%
Knee	3815	13.0%
Lower leg	1406	4.8%
Lumbar spine	564	1.9%
Paranasal sinuses	257	0.9%
Patella	269	0.9%
Pelvis	2134	7.3%
Ribs	88	0.3%
Sacrum	96	0.3%
Shoulder	1577	5.4%
Skull	79	0.3%
Spine	821	2.8%
Thoracic spine	814	2.8%
Thorax	3209	10.9%
Whole leg	212	0.7%
Wrist	241	0.8%
Total	29,433	100.0%

validation and testing was reached, respectively. Figure 3 shows examples of the final predictions of the *knee* cluster.

Furthermore, Fig. 4 displays the results of the cross-validated Grad-CAMs for the classes *pelvis* (1a, 1b) and *shoulder* (2b, 2b). Purple pixels indicate that the algorithm did not find any relevant information in these pixels in contrast to red pixels, where most relevant information was detected.

Discussion

In the present study, we demonstrate a sorting algorithm for musculoskeletal X-ray radiographs (SAM-X) to categorize images according to their anatomical entity. Based on a two-phase deep learning framework, 42,608 unstructured and

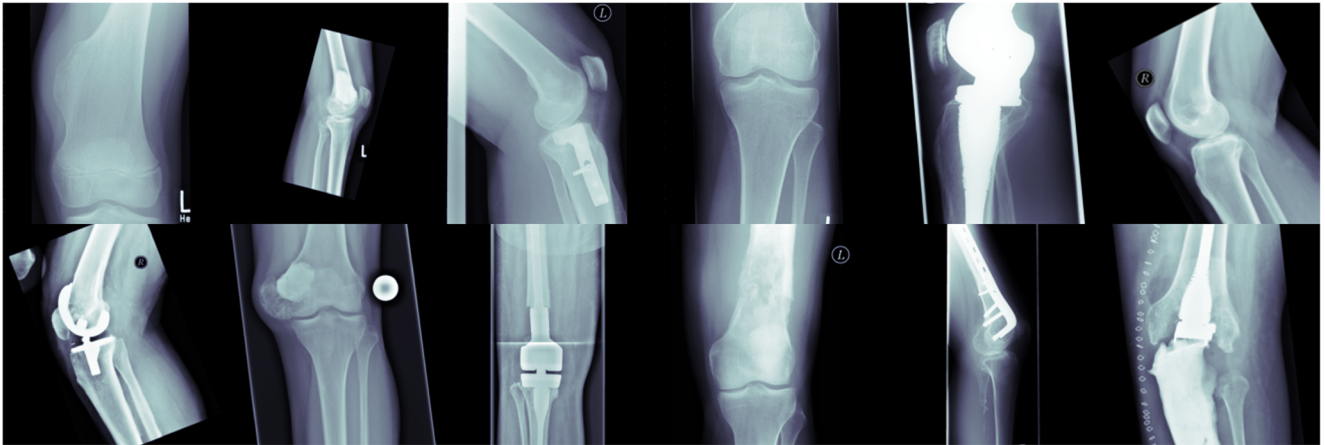
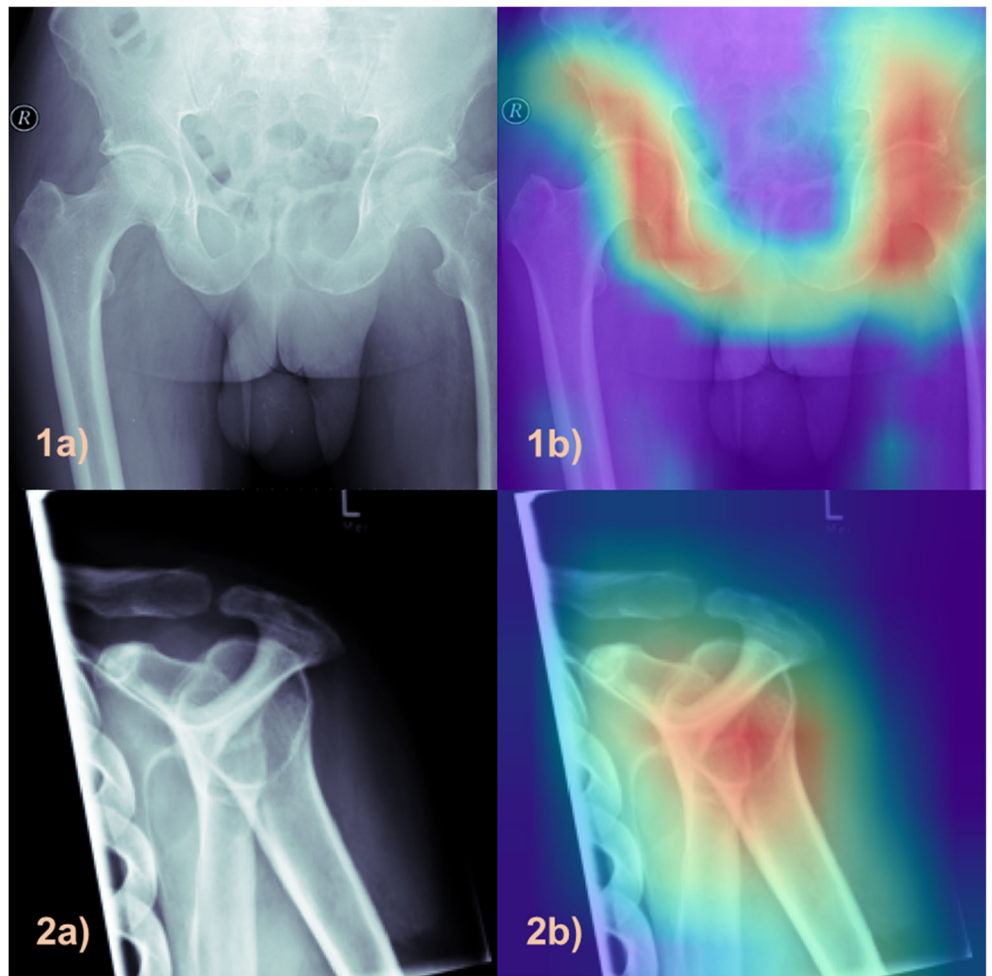


Fig. 3 Illustration of the final *knee* class showing the correct anatomical entity (knee) despite underlying heterogeneous pathologies

pseudonymized radiographs have been categorized into a total of 28 anatomical regions. Cross-validated accuracy of 96.2% for validation and 96.6% for hold-out test data indicated a high accuracy and excellent reliability. The aforementioned final predictions of the knee cluster demonstrate that even very different pathologies and appearances of knees were correctly identified and categorized. Further,

the calculated Grad-CAMs display the algorithms focus for predicting a specific class and help to unravel the black-box nature of DL methodology. These results underline the plausibility and robustness of the predictions: the algorithm primarily focused on anatomical regions, which are significant for the respective class and would also have been used by radiologists to determine the class.

Fig. 4 Grad-CAM examples from the classes pelvis (1a and 1b) and shoulder (2a and 2b) displaying the pixel areas, which were relevant for the algorithm to predict the according classes



Today, the emerging use of medical AI systems and neural networks presents an important, yet “still in the making” opportunity for our daily clinical routine and research. Radiology has always been at the front line of evolution in medical imaging since the introduction of digital imaging systems [15], teleradiology [16], computer-aided diagnosis tools [17], AI systems, and machine learning methods have emerged in the age of digitization. Ultimately, this has led to a significant increase of radiology examinations [17, 18]. Despite the broad use of previous research projects to analyze AI and humans in a head-to-head comparison [4, 9], promoting a potential level of lacking trustworthiness in these new technologies, future use is more likely practicing the human-in-the-loop setup used in our study as well, providing the possibility of humans actively collaborating with AI systems. Due to the complexity of clinical radiology examinations, accompanied by the amount of data linked to each examination, radiological daily operation may become inefficient, requiring tools to improve daily workflow and productivity [19]. Even more, with respect to image research, the most time-consuming part will become dataset building, potentially being the bottleneck [7] between data collection and creation of structured vs. unstructured data. Also, with respect to musculoskeletal disorders, radiologists, clinicians, and researchers face a diverse group of underlying pathologies which, in the setting of the aforementioned increasing number of radiology examinations and large image datasets, may benefit from pre-sorting algorithms to maintain order and effectivity. Although the aforementioned technologies have emerged, up to date, radiography still plays an essential and fundamental role for diagnosing, differentiating, and assessing the onset as well as progression of various musculoskeletal diseases [20, 24].

Mainly to harmonize occupational exposures and to flourish study comparison in meta-analyses for the use in occupational healthcare, surveillance, or research [25], the principle of categorizing musculoskeletal disease has already been widely used in the setting of identifying potential etiological or work-related factors that may lead to the onset or worsening of musculoskeletal disorders [26, 27]. However, case categorization in the setting of research or daily clinical routine needs to consider feasibility and the availability of resources. Hence, Dionne et al [28] proposed a minimal and optimal case definition for categorizing musculoskeletal diseases depending on its research purpose to promote balanced results. In contrast to the aforementioned studies, focusing on the preceding etiological aspect of musculoskeletal diseases, the sorting algorithm proposed in our study steps in to manage post-image acquisition of musculoskeletal diseases, and yet is in line with the aforementioned approach of Dionne et al since our proposed pre-sorting algorithm based on the anatomical entity provides a minimal still optimal tool for classifying radiographs in the assessment of musculoskeletal diseases. Since the expansion of radiological exams is most likely to generate large volumes of information

and in order to establish a common hub to facilitate such large image datasets with potentially underlying musculoskeletal diseases, the data in our study were categorized according to their anatomical entity since most musculoskeletal diseases are, at least in part, classified according to the anatomical entity in which they are located [2].

To the best of our knowledge, no framework for curating and categorizing medical radiographs by musculoskeletal characteristics has yet been proposed. However, related problems have been addressed. In 2005, Lehmann et al [29] proposed automatic categorization of medical images into 80 classes, e.g., by imaging modality and biological system in the context of content-based image retrieval (CBIR). Uwimana et al [30] also demonstrated a content-based image retrieval model by establishing links between low-level features of images and high-level features of text codes. Gál et al [31] proposed a CBIR model with a multidisciplinary approach to solve the classification problem by combining image features, metadata, textual, and referential information. More recently, Guo et al [32] presented an interactive algorithm for dermatological image quantification that combines computation, visualization, and expert interaction. The most comparable study was proposed by Kart et al [33]: DeepMCAT, an unsupervised clustering approach also based on DeepCluster [8]. An end-to-end training automatically categorizes large-scale cardiac MR images into 13 classes without any annotation. The main differences with the method presented by Kart et al are that we integrated weak annotation to train the CNN according to our requirements and we aim to categorize X-ray images into 28 classes. However, we hypothesize that our approach generally can achieve high results due to the implementation of powerful state-of-the-art self-supervised methodology while keeping the demand of human interaction at a limit and being adaptable to other requirements with only minor adjustments.

We acknowledge that our study has several limitations. Firstly, the classifier did assume an input of images that relate to one of the (pre)defined radiographic classes. Images that would have been derived of a different image modality, such as ultrasound or cross-sectional imaging, and falsely merged to our SAM-X model, would not be detected as such but forced into one of the classes. However, this is a common issue with AI classification models and needs to be addressed in the future. Secondly, weak supervision is an approach of machine learning that uses imprecise (noisy) data for supervised learning usually by bypassing the time-consuming task of hand-labelling the whole dataset [34] for example through obtaining weak labels with clustering methods. Assuming that some images were “incorrectly” clustered even though the region of interested is present in the image (e.g., shoulder and clavicle or femur and knee), it is reasonable to consider multiple labels for a single image. To address this issue, we calculated a second accuracy score and considered the two predictions with the highest probability (as presented in the

“Results” section). The respective scores for validation and testing reached 99.7% and 99.6% (in comparison to a single label 96.2 and 96.6). These numbers indicate that the model did not weakly label with respect to any incidental image features such as background or artefacts, but did indeed label according to similar anatomical features.

In conclusion, to facilitate the increasing amount of radiology examinations, accompanied by large image datasets, we propose a precise human-in-the-loop sorting algorithm for classifying radiographs in the assessment of musculoskeletal diseases according to the anatomical entity in which they are located. For dataset building, the algorithm proposes to be an efficient and time-saving tool in the setting of post-image acquisition.

Funding Open Access funding enabled and organized by Projekt DEAL. The authors state that this work has not received any funding.

Declarations

Guarantor The scientific guarantor of this publication is Jan Neumann.

Conflict of interest The authors of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

Statistics and biometry No complex statistical methods were necessary for this paper.

Informed consent Written informed consent was waived by the Institutional Review Board.

Ethical approval Institutional Review Board approval was obtained (N°48/20S).

Methodology

- retrospective
- diagnostic or prognostic study/experimental
- performed at one institution

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/>.

References

1. Organization, W.H. (2003) The burden of musculoskeletal conditions at the start of the new millennium: report of a WHO scientific group
2. Board, W.C.o.T.E., Soft Tissue and Bone Tumours (2020) International agency for research on cancer
3. Aiello M, Cavaliere C, D'Albore A, Salvatore M (2019) The challenges of diagnostic imaging in the era of big data. *J Clin Med* 8(3):316
4. Liu X, Faes L, Kale AU et al (2019) A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health* 1(6):e271–e297
5. Kharat AT, Singhal S (2017) A peek into the future of radiology using big data applications. *Indian J Radiol Imaging* 27(2):241–248
6. Larson DB, Magnus DC, Lungren MP, Shah NH, Langlotz CP (2020) Ethics of using and sharing clinical imaging data for artificial intelligence: a proposed framework. *Radiology* 295(3):675–682
7. Rajpurkar P, Chen E, Banerjee O, Topol EJ (2022) AI in health and medicine. *Nat Med* 28(1):31–38
8. Caron M, Bojanowski P, Joulin A, Douze M (2018) Deep clustering for unsupervised learning of visual features. In: *Proceedings of the European conference on computer vision (ECCV)*
9. von Schacky CE, Wilhelm NJ, Schäfer VS et al (2021) Multitask deep learning for segmentation and classification of primary bone tumors on radiographs. *Radiology* 301(2):398–406
10. Nguyen X-B, Lee GS, Kim SH, Yang HJ (2020) Self-supervised learning based on spatial awareness for medical image analysis. *IEEE Access* 8:162973–162981
11. Ghesu FC, Georgescu B, Mansoor A et al (2022) Self-supervised Learning from 100 Million Medical Images. *arXiv preprint arXiv: 2201.01283*
12. Montagnon E, Cerny M, Cadrin-Chênevert A et al (2020) Deep learning workflow in radiology: a primer. *Insights Imaging* 11(1):22
13. Willemink MJ, Koszek WA, Hardell C et al (2020) Preparing medical imaging data for machine learning. *Radiology* 295(1):4–15
14. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE international conference on computer vision*
15. Alhajeri M, Aldosari H, Aldosari B (2017) Evaluating latest developments in PACS and their impact on radiology practices: A systematic literature review. *Inform Med Unlocked* 9:181–190
16. Mun SK, Turner JW (1999) Telemedicine: emerging e-medicine. *Annu Rev Biomed Eng* 1:589–610
17. Doi K (2007) Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Comput Med Imaging Graph* 31(4-5):198–211
18. McDonald RJ, Schwartz KM, Eckel LJ et al (2015) The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. *Acad Radiol* 22(9):1191–1198
19. Schemmel A, Lee M, Hanley T et al (2016) Radiology workflow disruptors: a detailed analysis. *J Am Coll Radiol* 13(10):1210–1214
20. Al-Rakhami MS, Islam MM, Islam MZ, Asraf A, Sodhro AH, Ding W (2021) Diagnosis of Covid-19 from X-rays using combined CNN-RNN architecture with transfer learning. *MedRxiv: 2020.08.24.20181339*
21. Chhikara P, Singh P, Gupta P, Bhatia T (2020) Deep convolutional neural network with transfer learning for detecting pneumonia on chest X-rays. In: *Advances in bioinformatics, multimedia, and electronics circuits and signals*. Springer, pp 155–168
22. Das NN, Kumar N, Kaur M, Kumar V, Singh D (2020) Automated deep transfer learning-based approach for detection of COVID-19 infection in chest X-rays. *IRBM*

23. Ory PA (2003) Radiography in the assessment of musculoskeletal conditions. *Best Pract Res Clin Rheumatol* 17(3):495–512
24. Taljanovic MS, Hunter TB, Fitzpatrick KA, Krupinski EA, Pope TL (2003) Musculoskeletal magnetic resonance imaging: importance of radiography. *Skeletal Radiol* 32(7):403–411
25. van der Molen HF, Visser S, Alfonso JH et al (2021) Diagnostic criteria for musculoskeletal disorders for use in occupational healthcare or research: a scoping review of consensus- and synthesised-based case definitions. *BMC Musculoskelet Disord* 22(1):169
26. Ma K, Zhuang ZG, Wang L et al (2019) The Chinese Association for the Study of Pain (CASP): consensus on the assessment and management of chronic nonspecific low back pain. *Pain Res Manag* 2019:8957847
27. Sluiter JK, Rest KM, Frings-Dresen MH (2001) Criteria document for evaluating the work-relatedness of upper-extremity musculoskeletal disorders. *Scand J Work Environ Health* 27(Suppl 1):1–102
28. Dionne CE, Dunn KM, Croft PR et al (2008) A consensus approach toward the standardization of back pain definitions for use in prevalence studies. *Spine (Phila Pa 1976)* 33(1):95–103
29. Lehmann TM, Guld MO, Deselaers T et al (2005) Automatic categorization of medical images for content-based retrieval and data mining. *Comput Med Imaging Graph* 29(2-3):143–155
30. Uwimana E, Ruiz ME (2008) Automatic classification of medical images for content based image retrieval systems (CBIR). In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. SAGE Publications Sage CA, Los Angeles, CA
31. Gál V, Solt I, Gedeon T, Nachtegaele M (2011) Multi-disciplinary modality classification for medical images. *Magnetic Resonance Imaging* 17:1–7
32. Guo X, Yu Q, Li R et al (2016) An expert-in-the-loop paradigm for learning medical image grouping. In: *Pacific-Asia conference on knowledge discovery and data mining*. Springer
33. Kart T, Bai W, Glocker B, Rueckert D (2021) DeepMCAT: large-scale deep clustering for medical image categorization. Springer International Publishing, Cham
34. Zhou Z-H (2018) A brief introduction to weakly supervised learning. *Nat Sci Rev* 5(1):44–53

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

4.4. Recommender-based Bone tumour Classification – a Link to the Past

Authors:

Florian Hinterwimmer, Ricardo Smits Serena, Nikolas Wilhelm, Sebastian Breden, Sarah Consalvo, Fritz Seidl, Dominik Juestel, Rainer Burgkart, Klaus Woertler, Ruediger von Eisenhart-Rothe, Jan Neumann, Daniel Rueckert

Journal:

European Radiology

Synopsis:

Accurate classification of bone tumours is crucial for determining the appropriate course of treatment and improving patient outcomes. However, this task is challenging due to the diverse manifestations of MSK tumours and limited availability of data.

This study aimed to develop a novel algorithm for real-time classification of predefined bone tumour entities and linking undiagnosed patients with previous patient histories based on radiographic features. The algorithm combines DL, a hash-based nearest-neighbor recommender approach, and majority voting for simultaneous classification of multiple bone tumours. By doing so, it leverages dormant information in clinical systems, facilitates comparison with previous patient encounters, and ultimately impacts the diagnosis of rare and complex diseases across different medical fields. The study involved the retrospective curation of patient data from 2000-2021, encompassing 809 patients and 1792 radiographs representing ten different types of tumours. Two classification models were initially implemented to establish baseline results. Then, a novel method was proposed that involves extracting image features using DL, clustering the k-most similar images to a target image using a hash-based nearest-neighbor recommender approach, and then performing simultaneous classification by majority voting.

The results indicated that the proposed model achieved a precision-at-k of 62.58% and a mean classification accuracy of 92.86% for the optimal configuration, significantly outperforming the state-of-the-art models that only achieved 54.10% and 62.80% respectively. This underscored the potential of the proposed approach to navigate the challenges associated with MSK tumours and limited data availability, thereby enabling early and precise diagnoses.

The proposed framework offers an accurate and efficient approach to bone tumour classification by effectively dealing with limited and unstructured data and complex

classification problems. It provides real-time feedback for bone tumour assessment and leverages previously collected knowledge based on previous patient journeys. This approach not only revolutionizes the way physicians can harness dormant information but also suggests potential versatility across various medical disciplines. Furthermore, it lays the foundation for assisting general practitioners and young physicians in challenging situations, ultimately impacting the diagnosis of rare and complex diseases across different medical fields.

Contribution of thesis author:

Florian Hinterwimmer was the principal investigator in this study and contributed at least 50%.

Copyright:

Recommender-based bone tumour classification with radiographs—a link to the past © 2024 by Florian Hinterwimmer is licensed under Creative Commons Attribution 4.0 International. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Recommender-based bone tumour classification with radiographs—a link to the past

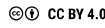
SPRINGER NATURE**Author:** Florian Hinterwimmer et al**Publication:** European Radiology**Publisher:** Springer Nature**Date:** Mar 15, 2024*Copyright © 2024, The Author(s)*

Creative Commons

This is an open access article distributed under the terms of the [Creative Commons CC BY](#) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.

To request permission for a type of use not listed, please contact [Springer Nature](#)



ATTRIBUTION 4.0 INTERNATIONAL

Deed

Canonical URL: <https://creativecommons.org/licenses/by/4.0/>

[See the legal code](#)

You are free to:

Share — copy and redistribute the material in any medium or format for any purpose, even commercially.

Adapt — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

Attribution — You must give **appropriate credit**, provide a link to the license, and **indicate if changes were made**. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

No additional restrictions — You may not apply legal terms or **technological measures** that legally restrict others from doing anything the license permits.

Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable **exception or limitation**.

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as **publicity, privacy, or moral rights** may limit how you use the material.

Notice

This deed highlights only some of the key features and terms of the actual license. It is not a license and has no legal value. You should carefully review all of the terms and conditions of the actual license before using the licensed material.

Creative Commons is not a law firm and does not provide legal services. Distributing, displaying, or linking to this deed or the license that it summarizes does not create a lawyer-client or any other relationship.

Creative Commons is the nonprofit behind the open licenses and other legal tools that allow creators to share their work. Our legal tools are free to use.

- [Learn more about our work](#)
- [Learn more about CC Licensing](#)
- [Support our work](#)
- [Use the license for your own material](#)
- [Licenses List](#)
- [Public Domain List](#)

Footnotes

appropriate credit — If supplied, you must provide the name of the creator and attribution parties, a copyright notice, a license notice, a disclaimer notice, and a link to the material. CC licenses prior to Version 4.0 also require you to provide the title of the material if supplied, and may have other slight differences.

- [More info](#)

indicate if changes were made — In 4.0, you must indicate if you modified the material and retain an indicator of previous modifications. In 3.0 and earlier license versions, the indication of changes is only required if you create a derivative.

- [Marking guide](#)
- [More info](#)

technological measures — The license prohibits application of effective technological measures, defined with reference to Article 11 of the WPO Copyright Treaty.

- [More info](#)

exception or limitation — The rights of users under exceptions and limitations, such as fair use and fair dealing, are not affected by the CC licenses.

- [More info](#)

publicity, privacy, or moral rights — You may need to get additional permissions before using the material as you intend.

- [More info](#)

[Contact](#)

[Newsletter](#)

[Privacy](#)

[Policies](#)

[Terms](#)

CONTACT US

Creative Commons PO Box 1396, Mountain View, CA 94042

info@creativecommons.org

+1 415 439 4200

SUBSCRIBE TO OUR NEWSLETTER

Your email

[SUBSCRIBE](#)

SUPPORT OUR WORK

Our work relies on you. Help us keep the Internet free and open.


[DONATE NOW](#)

Except where otherwise noted, content on this site is licensed under [a Creative Commons Attribution 4.0 International License](#). Icons by [Font Awesome](#).

IMAGING INFORMATICS AND ARTIFICIAL INTELLIGENCE



Recommender-based bone tumour classification with radiographs—a link to the past

Florian Hinterwimmer^{1,2*} , Ricardo Smits Serena^{1,2}, Nikolas Wilhelm¹, Sebastian Breden¹, Sarah Consalvo¹, Fritz Seidl³, Dominik Juestel^{3,4,5}, Rainer H. H. Burgkart¹, Klaus Woertler⁶, Ruediger von Eisenhart-Rothe¹, Jan Neumann⁶ and Daniel Rueckert²

Abstract

Objectives To develop an algorithm to link undiagnosed patients to previous patient histories based on radiographs, and simultaneous classification of multiple bone tumours to enable early and specific diagnosis.

Materials and methods For this retrospective study, data from 2000 to 2021 were curated from our database by two orthopaedic surgeons, a radiologist and a data scientist. Patients with complete clinical and pre-therapy radiographic data were eligible. To ensure feasibility, the ten most frequent primary tumour entities, confirmed histologically or by tumour board decision, were included. We implemented a ResNet and transformer model to establish baseline results. Our method extracts image features using deep learning and then clusters the k most similar images to the target image using a hash-based nearest-neighbour recommender approach that performs simultaneous classification by majority voting. The results were evaluated with precision-at- k , accuracy, precision and recall. Discrete parameters were described by incidence and percentage ratios. For continuous parameters, based on a normality test, respective statistical measures were calculated.

Results Included were data from 809 patients (1792 radiographs; mean age 33.73 ± 18.65 , range 3–89 years; 443 men), with Osteochondroma (28.31%) and Ewing sarcoma (1.11%) as the most and least common entities, respectively. The dataset was split into training (80%) and test subsets (20%). For $k = 3$, our model achieved the highest mean accuracy, precision and recall (92.86%, 92.86% and 34.08%), significantly outperforming state-of-the-art models (54.10%, 55.57%, 19.85% and 62.80%, 61.33%, 23.05%).

Conclusion Our novel approach surpasses current models in tumour classification and links to past patient data, leveraging expert insights.

Clinical relevance statement The proposed algorithm could serve as a vital support tool for clinicians and general practitioners with limited experience in bone tumour classification by identifying similar cases and classifying bone tumour entities.

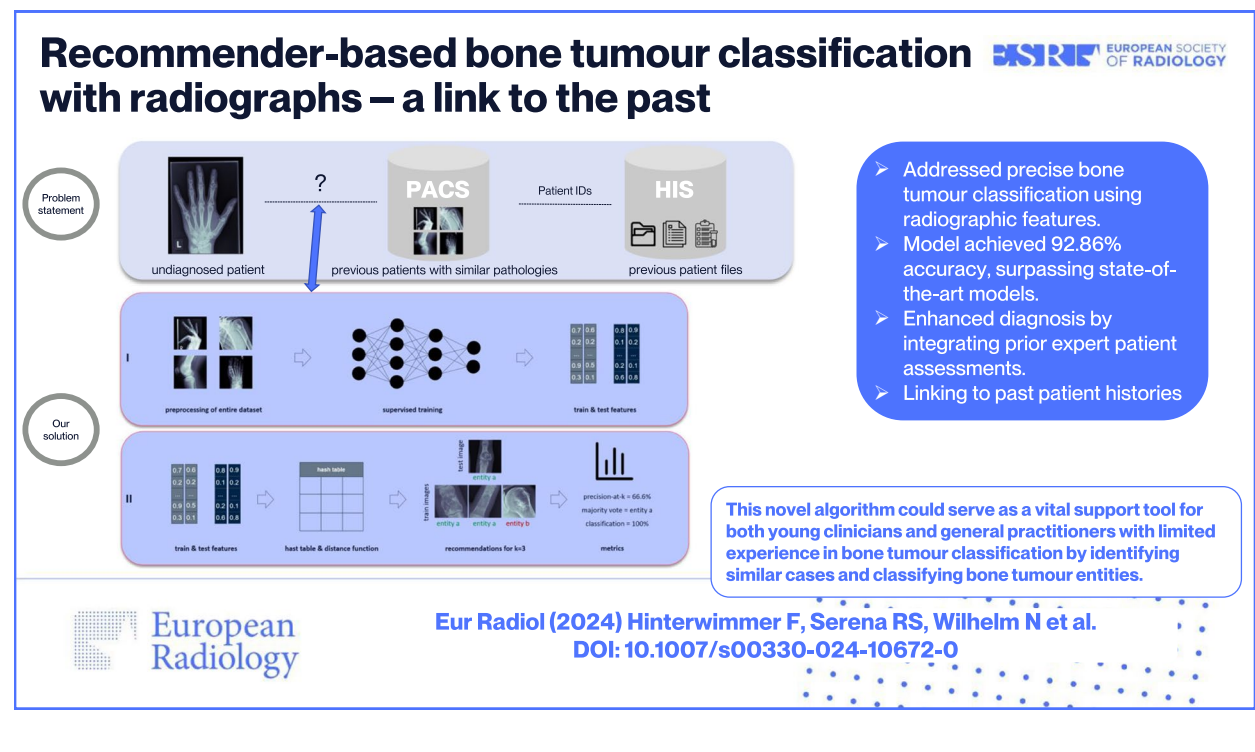
Key Points

- Addressed accurate bone tumour classification using radiographic features.
- Model achieved 92.86%, 92.86% and 34.08% mean accuracy, precision and recall, respectively, significantly surpassing state-of-the-art models.
- Enhanced diagnosis by integrating prior expert patient assessments.

*Correspondence:
Florian Hinterwimmer
florian.hinterwimmer@tum.de
Full list of author information is available at the end of the article

Keywords Bone neoplasms, Deep learning, Classification, Machine learning, Radiography

Graphical abstract



Introduction

Bone tumours are a group of rare and diverse types of neoplasms [1–4]. The vast majority of primary bone tumours are benign, whereas malignant primary bone tumours account for 0.2% of all malignancies in adults [3, 5]. It is crucial to diagnose bone tumours early, as this directly affects the patient's prognosis and curability [1]. Hence, prompt referral to a specialised tumour centre to determine tumour malignancy, establish a specific diagnosis and initiate early treatment is essential [6]. Unfortunately, delays of more than 1 year often occur in clinical practice, partly due to the lack of specific symptoms in the early stages and the fact that non-oncologically trained orthopaedic surgeons [4, 7, 8], primary care physicians or paediatricians only encounter about three malignant musculoskeletal (MSK) tumours in their professional career and therefore lack the experience in unequivocally identifying these complex tumour entities [7].

Imaging is crucial in diagnosing bone tumours [5]. The Musculoskeletal Tumor Society and American Academy of Orthopedic Surgeons recommend radiographs as the initial screening tool [5, 8]. While CT and MRI provide additional diagnostic information, they should not delay initial medical care [5]. Definitive diagnosis typically requires a

combination of imaging, histopathologic findings and clinical presentation, with further detailed imaging assessments recommended at specialised MSK tumour centres [9].

Diagnostic imaging is rapidly advancing with significant technological and market growth, leading to an increase in imaging data [10–12]. In MSK radiology and orthopaedic oncology, precision medicine and image interpretation are increasingly critical. Despite the growing use of artificial intelligence (AI) and deep learning (DL) in cancer research, their application in MSK tumour research remains limited [2, 13]. However, these advanced data analysis techniques hold promise for revolutionising MSK tumour diagnostics and enhancing healthcare delivery [14].

As AI technologies evolve, various medical imaging applications are being developed, often focusing on comparing AI's performance with that of human experts in tasks like pathology classification [15–17]. Among these, recommender systems (RS) offer a novel approach, primarily suggesting options based on user preferences, bypassing extensive algorithm training [18]. While traditionally used in commercial settings, RS are increasingly recognised for their potential in

medical decision-making, such as recommending drug therapies or identifying similar patient cases based on medical history and imaging data [19, 20].

MSK tumour centres have extensive knowledge and experience lying dormant in their hospital information system (HIS) and picture archiving and communication system (PACS) based on patients treated for MSK tumours in the past. In this study, we present a DL-based algorithm that recommends similar patients based on clustering of radiographic features, draws on the extensive experience dormant in clinical systems based on previous patient histories and simultaneously classifies multiple bone tumour pathologies to enable early and specific diagnosis.

Materials and methods

The local institutional review and ethics board approved this retrospective study (no. 48/20S). The study was performed in accordance with national and international guidelines. Informed consent was waived for this retrospective and anonymised study. The general structure of the manuscript follows the *Checklist for artificial intelligence in medical imaging* (CLAIM [21]).

Eligibility criteria

For this single-centre study, we conducted a search through the database of our MSK tumour centre. All patients treated for primary bone neoplasms (based on the according ICD

codes) between 2000 and 2021 were screened. Patients with the following primary tumours were selected, as these are the most frequent ones in our database: aneurysmal bone cyst (ABC), chondroblastoma, chondrosarcoma, enchondroma, Ewing sarcoma, fibrous dysplasia, giant cell tumour, non-ossifying fibroma (NOF), osteochondroma and osteosarcoma. The diagnosis of malignant lesions was verified by histopathology as standard of reference. Benign and intermediate lesions were either verified by histopathology, if available, or discussed in the local tumour board and classified according to radiological features known from the literature [22]. The clinical and imaging data were retrieved from our HIS and PACS, respectively. To ensure the feasibility of the proposed model, the ten most frequent entities were considered. Any tumour representation in the radiographs was eligible. Forty-four patients with inadequate imaging (no pre-operative/pre-therapy radiographs), two patients with incomplete clinical data and 31 patients lost to follow-up were excluded. Subsequently, 809 patients with 1792 respective radiographs were found (Fig. 1). The curation and validation of the data were conducted by two orthopaedic residents (S.C., S.B.) and a senior MSK radiologist (J.N.), respectively, with support of a data scientist (E.H.).

Demographics and statistical evaluation

Descriptive data is presented according to the *Strengthening the Reporting of Observational studies in Epidemiology*

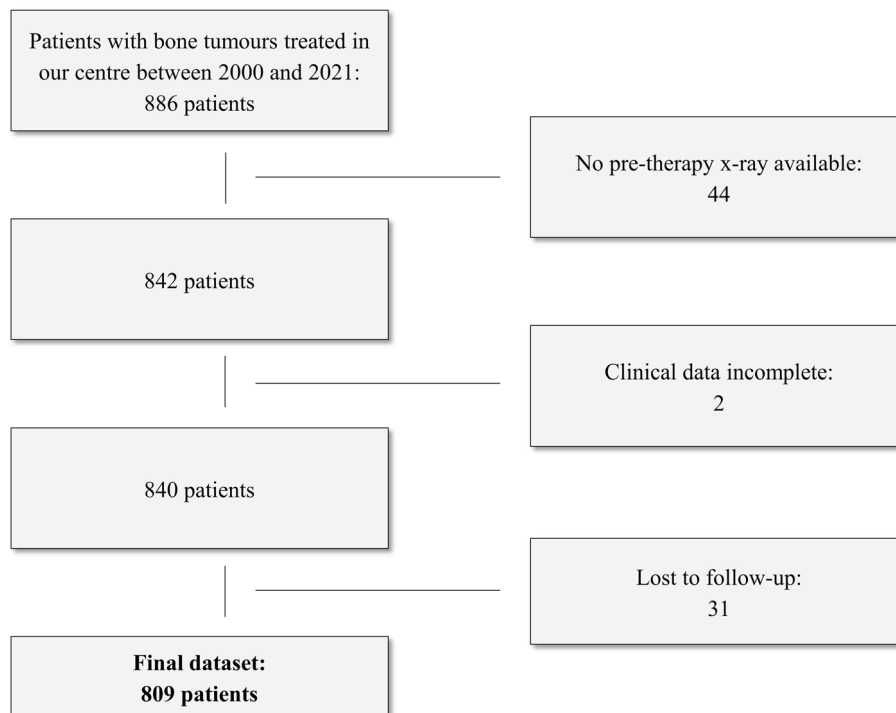


Fig. 1 Flow diagram showing the application of eligibility criteria to create a final dataset

(STROBE [23]) guidelines. Discrete parameters were described by incidence and percentage ratios. For continuous parameters, based on a Shapiro-Wilk normality test, respective statistical measures were computed.

The mean classification accuracy, precision and recall of the baseline models are calculated based on their performance on the test data. In our multiclass setting, classification accuracy is calculated as the ratio of correctly predicted instances to the total number of test data instances. Precision and recall are measured for each class individually and then averaged: precision is the ratio of true positive predictions of each class to all predictions made for that class, and recall is the ratio of true positive predictions of each class to all actual instances of that class in the test data. The RS clustering results are assessed using a precision-at- k metric, which calculates the proportion of relevant items within the top- k recommendations. To compute the final classification accuracy, precision and recall of the proposed model, we compared the correct predictions obtained through a majority vote from the k -closest images in the RS against the labels of the respective target images in the test data. About 10% of the total dataset represents external imaging data obtained from other institutions and integrated into our Health Information System (HIS) and Picture Archiving and Communication System (PACS). The dataset is divided into training (80%) and test data (20%), with the metrics being calculated solely on the test data. This test subset exclusively contains patients with a single image to avoid any overlap with the training dataset. The dataset was stratified based on the types of bone tumours, ensuring that each tumour type was proportionally represented in both the training and test

subsets. The final metrics, including classification accuracy, precision and recall, were determined three times using randomly shuffled data, and the corresponding mean values were calculated. In addition, the normality of the distribution of performance results was assessed. Based on the outcome of normality tests, suitable statistical methods were chosen to evaluate the significance of model performance metrics.

Model training

Model training and inference were conducted on a DGX Station A100 with four 80 GB graphical processing units (Nvidia Corporation), 64 2.25 GHz cores and 512 GB DDR4 system memory running on a Linux/Ubuntu 20.04 distribution (Canonical). Preprocessing and model implementation were performed in Python 3.11.1 (<https://www.python.org/>) using PyTorch 1.13.1 and cuda toolkit 12.0 (<https://pytorch.org/>).

Algorithm

The general concept of the proposed framework is shown in Fig. 2: identification of the most similar cases from previous patients based on radiographs with respect to an undiagnosed image. First, to create baselines for bone tumour entity classification, we calculated classification metrics by straightforward application of a standard [24] (baseline 1) and a state-of-the-art [25] (baseline 2) DL model to a multi-entity classification task. For the implementation of our proposed approach, we performed two main steps: (I) to emphasise on tumourous tissue rather than background or non-relevant tissue, we created bounding boxes around the region of interest, which can be accomplished algorithmically [26] or through

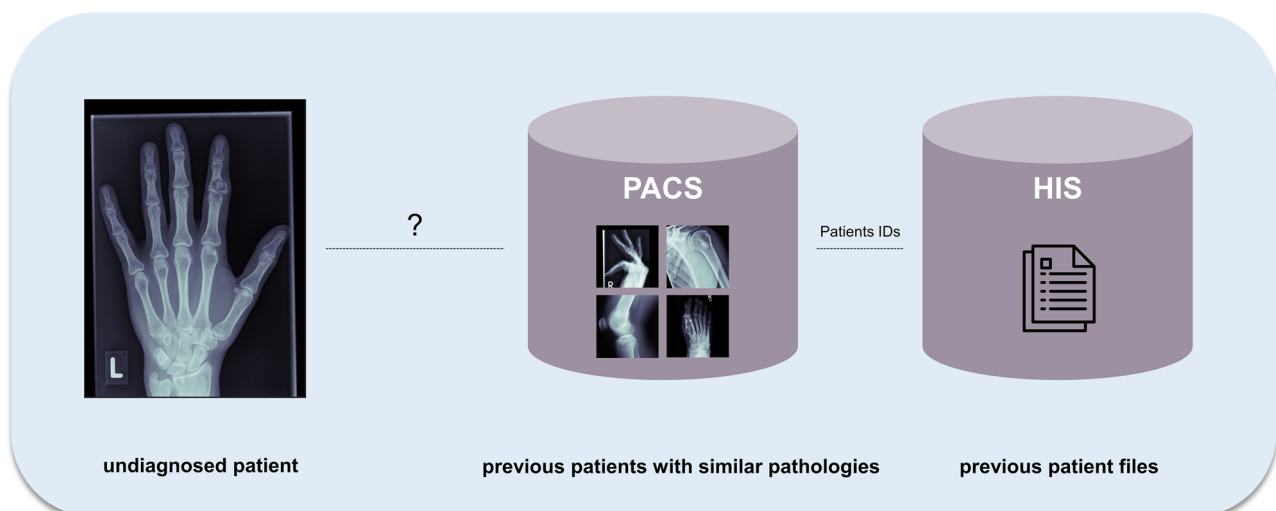


Fig. 2 General concept of the proposed method—clustering new patients with previous patients based on radiographs to identify similar cases and classify tumour entity (PACS, picture archiving and communications systems; HIS, hospital information system)

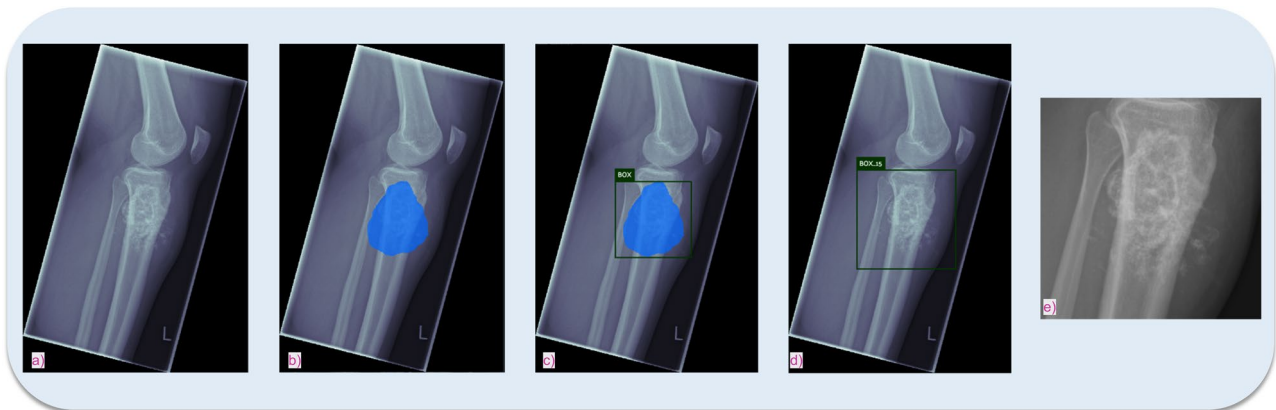


Fig. 3 Exemplary creation of bounding boxes focusing the tumorous tissue by the segmentation algorithm of Bloier et al [26]: (a) initial image, (b) segmented tumour, (c) calculated bounding box, (d) bounding box with 15% margin to assure all tumour tissue is captured, (e) cropped image

manual cropping by a domain expert (Fig. 3). We employ the model from baseline 1. The trained model as well as the extracted features from the training data was saved. After training was completed, we calculated the image features of the test data by running the data through the trained convolutional neural network model. (II) We created a hash table. Instead of comparing each set of new image features to the training data features, we used locality-sensitive hashing (LSH), an approximate nearest neighbour algorithm that reduces the computational complexity from $O(N^2)$ to $O(\log N)$. LSH generates a hash value for image features by taking the spatiality of the data into account. Data elements that are similar in

high dimensional space have a higher chance of obtaining the same hash value [27]. Based on a hamming distance function, we computed the k -nearest neighbours with respect to each target image. By assigning the k -nearest neighbours (from training images) to one cluster along with the target image (test image), we established a link between the undiagnosed patient and past patient cases stored in our database. Since local patient identifiers from the training data patients are known, this allowed us to potentially link to experiences from previous patients in our clinical systems, e.g. radiology reports, laboratory results and therapy results. Furthermore, we obtained a classification of tumour entities by applying a majority

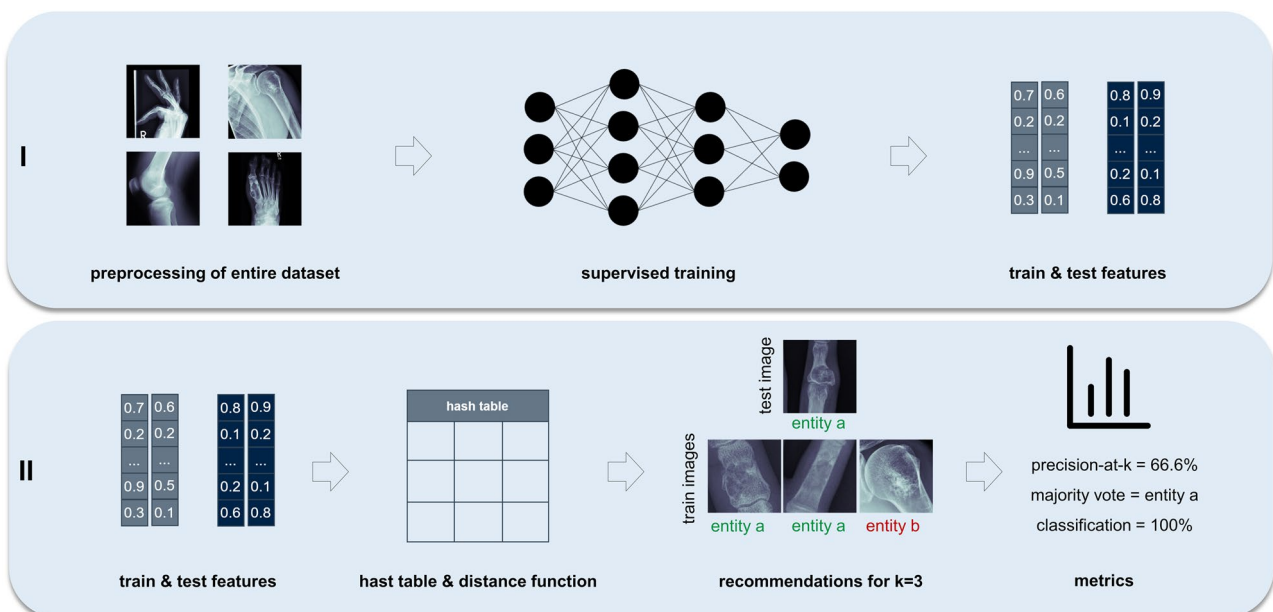


Fig. 4 Flow chart of the proposed model—(I) preparing the images, training of the convolutional neural network, saving the model and features; (II) calculating the high dimensional distances with a distance function, adding a hash tables, clustering of the most similar x-rays and calculating a precision-at-k and a tumour entity classification with a majority vote of the k -clustered images

Table 1 Distribution of continuous and discrete characteristics (*IQR* interquartile range)

Characteristic	Shapiro Wilk test	Median	IQR	#	%
Patients					
Age	$W(809) = 0.94, p < .001$	30.00	30.00	–	–
Entity					
Aneurysmal bone cyst (ABC)	–	–	–	49	6.06%
Chondroblastoma	–	–	–	18	2.22%
Chondrosarcoma	–	–	–	124	15.33%
Enchondroma	–	–	–	181	22.37%
Ewing sarcoma	–	–	–	9	1.11%
Fibrous dysplasia	–	–	–	31	3.83%
Giant cell tumour	–	–	–	51	6.30%
Non ossifying fibroma (NOF)	–	–	–	33	4.08%
Osteochondroma	–	–	–	229	28.31%
Osteosarcoma	–	–	–	84	10.38%
Gender					
Female	–	–	–	366	45.24%
Male	–	–	–	443	54.76%
Location					
Clavicula	–	–	–	7	0.87%
Columna vertebralis	–	–	–	4	0.49%
Femur	–	–	–	297	36.71%
Fibula	–	–	–	42	5.19%
Humerus	–	–	–	124	15.33%
Manus	–	–	–	62	7.66%
Os ilium	–	–	–	24	2.97%
Os ischii	–	–	–	8	0.99%
Os pubis	–	–	–	11	1.36%
Os sacrum	–	–	–	1	0.12%
Patella	–	–	–	7	0.87%
Pes	–	–	–	42	5.19%
Radius	–	–	–	12	1.48%
Scapula	–	–	–	17	2.10%
Tibia	–	–	–	146	18.05%
Ulna	–	–	–	5	0.62%

Table 2 Tumour entity classification results—mean of accuracy, precision and recall with standard deviation

	Baseline 1: ResNet50	Baseline 2: Transformer	Our approach
Accuracy			
Training	77.01 ± 1.11	82.37 ± 2.20	–
Validation	58.69 ± 3.04	69.54 ± 2.87	–
Test	54.10 ± 2.91	62.80 ± 1.90	92.86 ± 0.59
Precision			
Training	79.10 ± 0.76	80.44 ± 2.29	–
Validation	60.23 ± 2.09	67.91 ± 1.39	–
Test	55.57 ± 2.00	61.33 ± 2.10	92.86 ± 0.59
Recall			
Training	28.26 ± 0.41	30.23 ± 0.99	–
Validation	21.53 ± 1.12	25.52 ± 0.81	–
Test	19.85 ± 1.07	23.05 ± 0.70	34.08 ± 2.76

Table 3 Statistical significance of model performance metrics—ANOVA results demonstrate the overall significance of differences in accuracy, precision and recall among all tested models. Tukey’s HSD (Honestly Significant Difference) post hoc analysis further identifies the specific pairwise comparisons that are statistically significant. The *p* values indicate that the performance of ‘Our Approach’ is significantly different from both baseline models, and there is a significant difference in performance between the two baseline models

	ANOVA	Tukey’s HSD post hoc test
<i>p</i> values of test metrics		
Accuracy	< 0.0001	–
Precision	< 0.0001	–
Recall	< 0.0001	–
<i>p</i> values of model comparison		
Baseline 1: ResNet50 vs. Baseline 2: Transformer	–	0.0035
Our approach vs. Baseline 1: ResNet50	–	0.001
Our approach vs. Baseline 2: Transformer	–	0.001

vote to the entities of the images clustered to the target image. Figure 4 illustrates the proposed approach.

Results

Dataset

The mean age of patients was 33.73 with a standard deviation of 18.65 and a range of 3 to 89. Osteochondroma was

the most common entity, accounting for 28.31% of the total dataset, while Ewing sarcoma was the least frequent entity representing only 1.11% of the dataset. The gender was close to similarly distributed (males 54.76%, females 45.24%) with a slight tendency towards males. The most frequent location of tumour occurrence was the femur with 36.71%, while a tumour only occurred once at the os sacrum representing 0.12% of the whole dataset. Further details of the continuous and discrete characteristics are displayed in Table 1.

Model performances

For both baseline models, we conducted extensive hyperparameter tuning to optimise their performance as well as a fivefold cross-validation. Key hyperparameters adjusted included learning rate, batch size and number of training epochs. Additionally, we employed several data augmentation techniques (rotations, horizontal and vertical flipping) to enhance the dataset and prevent overfitting. The optimised values for each hyperparameter were as follows: learning rate 0.003/0.0025, batch size 8/8 number of training epochs 85/77 and probability for applying data augmentation 0.3/0.3 respectively for the ResNet and the transformer model. We accomplished a mean test accuracy/precision/recall of 54.10%, 55.57% and 19.85% with a pretrained ResNet32 [24] model and 62.80%, 61.33% and 23.05% with a state-of-the-art Vision Transformer model [25, 28] for classifying the tumour entities on a test split. For our proposed method, the respective precision-at-*k* for *k* = 1/3/5/7 was 65.46%/62.58%/62.06%/61.48%. The

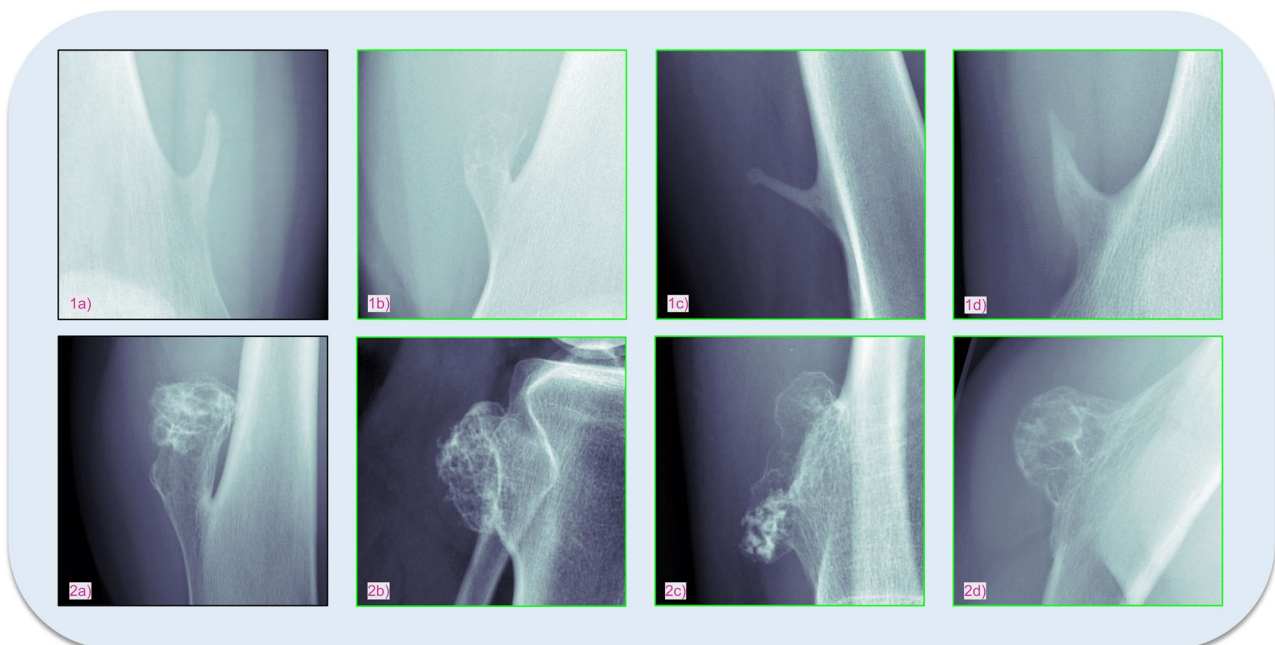


Fig. 5 Examples of osteochondroma X-rays showcasing the model’s ability to accurately cluster different appearances of the same tumour entity. The target image is marked with a black frame, while correctly matched images are highlighted with a green frame

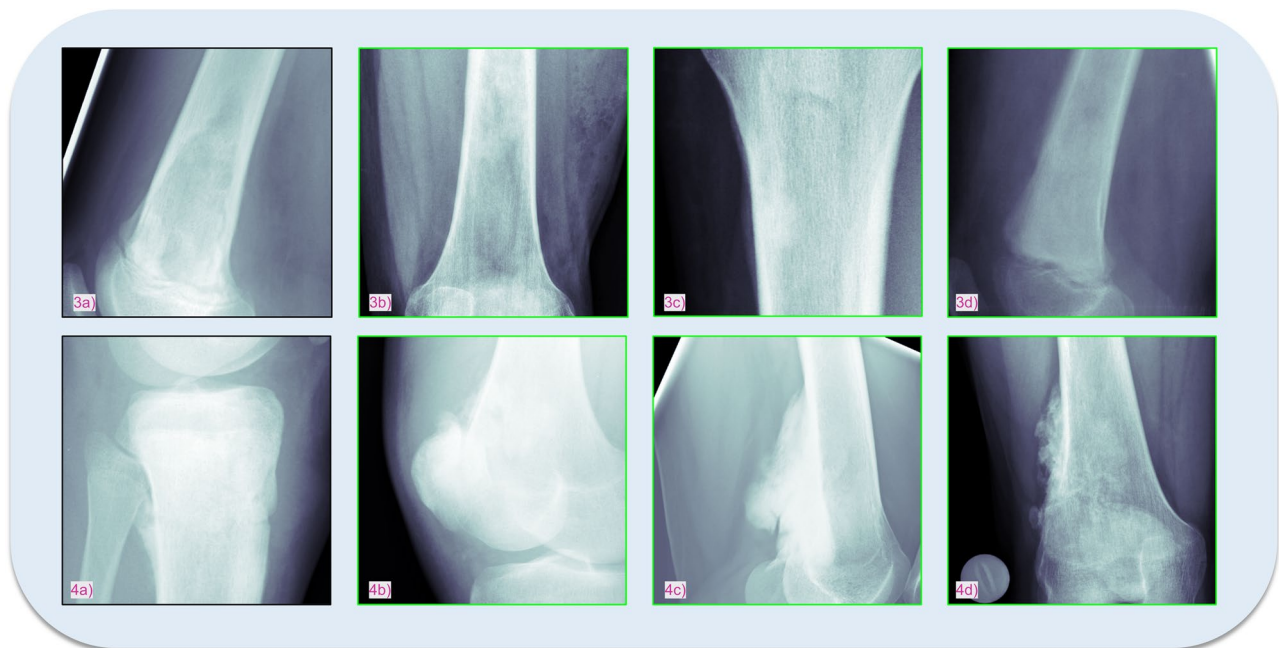


Fig. 6 Examples of osteosarcoma X-rays illustrating the model's effectiveness in clustering diverse manifestations of the same tumour entity. The target image is enclosed in a black frame, and correctly clustered images are indicated with a green frame

classification metrics based on the described majority vote on the clustered images was 65.46%/92.86%/92.13%/92.01%. For k , only odd values were used to facilitate meaningful calculation of the majority vote. No higher value than seven was chosen because the lowest number of entity samples was only nine (Ewing sarcoma) and, therefore, considering only odd values, a maximum of seven samples could be assigned. Table 2 displays the results for the two baseline models as well as the result of the best configuration.

Initial Shapiro-Wilk tests were performed to assess the normality of the distribution of model performance metrics. The results suggested a normal distribution for most metrics, providing a basis for the use of parametric tests. Consequently, ANOVA was utilised to analyse the significance of differences in model performance, revealing significant disparities across the models ($p < 0.0001$ for all metrics, threshold at $p = 0.05$). Despite the limited sample sizes, ANOVA was considered appropriate due to the normality of the data and the robustness of this test under certain conditions. Following the ANOVA, Tukey's HSD (Honestly Significant Difference) post hoc tests were conducted for pairwise model comparisons, which identified statistically significant differences, indicating that our approach significantly outperformed the baseline models (Table 3). Figures 5 and 6 show examples of correctly mapped osteochondromas and osteosarcomas from different patients and visually different appearances. The first images (1a, 2a, 3a, 4a - black) show the target images and the second to fourth images (1b-1d,

2b-2d, 3b-3d, 4b-4d - green) in each row show the correspondingly clustered images.

Discussion

The main result of this study was that we were able to develop an algorithm for real-time classification of ten preselected primary bone tumour entities that significantly outperformed a widely used [24] and a state-of-the-art model [25] and those shown in similar studies [17, 19] by circumventing the problem of confounding factors through clustering of the k most similar radiographs to a target image rather than classifying all different appearances and different anatomical structures of the same tumour pathology into one class. Further, identifying the most similar cases also allows large amounts of knowledge and experience lying dormant in clinical systems, such as previous diagnoses, treatments, etc., to be attributed to new and undiagnosed patients, potentially supporting an early and specific diagnosis. We hypothesise that the poor performance of the baseline models as well as the poor scores for recall across all approaches originate in overfitting due to limited available data and even more so because of significant class imbalances.

Similar studies were published [17, 19]. For example, von Schacky et al [17] presented a multitask DL model for simultaneous detection, segmentation and classification of bone lesions and compared the results with those of radiologists with different levels of experience. The general task of classifying bone lesions as well as the investigated entities are similar to our study. Their model achieved a classification accuracy

of 43.2%, whereas a radiology resident achieved 44.1% and an MSK fellow radiologist 58.6% in classifying bone lesions by entity. While our metric scores are significantly higher, von Schacky et al had to cope with a lower ratio of samples per class. A major problem for the DL model probably was that bone tumour entities can occur in different anatomical regions and demonstrate different appearances. Therefore, a DL model has to classify the same pathology with different anatomical and visual features into the same class to predict correctly. As illustrated in Figs. 5 and 6, our model was able to bypass this issue by clustering only the k most similar cases and calculating the final prediction of the entity based on a majority vote. Their study underlines the complexity of precise identification of bone neoplasms for DL models as well as for human experts. Despite the widespread use of previous research projects analysing AI and humans in a direct comparison [16], the future use of AI to support instead of replacing medical experts is more likely. Another similar study was recently published by Kuanr et al [19]. The main concept behind their study was to identify similar COVID-19 patients based on comparably homogenous chest radiography by applying feature extraction accomplished by a DL model. The approach of comparing similar patients based on x-ray images is similar to that of our study. However, by implementing a majority vote on top of the clustered images for final metric calculation, we additionally demonstrated a classification for multiple entities and heterogenous pathologies. To the best of our knowledge, no study has yet shown a RS approach with majority vote to conclude in a classification of several bone tumour entities or link to previous sarcoma patient data.

The general approach of utilising retrospective datasets, training a DL model to extract meaningful image features and clustering similar cases based on imaging data with a nearest neighbour model is adaptable to other pathologies and scenarios as well. However, we hypothesise that the heterogeneity and multiple manifestations of bone tumours are one of the main reasons why we have achieved such a significant improvement with our algorithm compared to conventional classification approaches. It has been shown before that ensemble methods tend to give better results when the models and datasets have a large variety [29]. For tumour entities that occur more frequently in the same anatomic region, a classical approach would yield better results to begin with. Nevertheless, the concept of finding similar cases to compare with previous treatments of patients may be relevant to any other pathology.

The major limitation of this study is that we did not consider clinical data in the assessment of the tumour entity. Although plain radiographs are crucial for the initial screening for a possible bone tumour [5, 8, 30, 31], further classification requires the inclusion of clinical data (and possibly additional imaging) [9]. However, we hypothesise that some clinical information such as the

patient's age, anatomical region, or tumour location is partially represented in the x-ray images and therefore indirectly integrated into our prediction model. Inclusion of clinical data and other bone tumour entities will be explored in future studies. Another limitation arises from the limited data set. While 1792 radiographs are a considerable number for the rare entities of MSK tumours, a mean of 179 samples per class is rather low in view of the heterogeneity of MSK lesions and additionally in the context of DL applications. Although approximately 10% of the data set consists of external radiographs from general practitioners, external radiologists, etc. uploaded to our clinical systems, another limitation is that the model needs to be tested on external data to further assess generalisability [32] before suitability for clinical use can be evaluated [33]. Although we managed to circumvent problems with confounding factors, the fact that most of the data were collected in a single centre could still affect the robustness of the model: different image characteristics associated with different radiographic devices or different patient characteristics could cause this.

In conclusion, we have demonstrated a way to deal with limited data and complex classification problems, providing a real-time feedback for bone tumour assessment. The proposed framework can link undiagnosed patients with previous experience and knowledge lying dormant in our clinical systems. Additionally, we have used AI methodology to leverage previously collected knowledge based on previous patient journeys, allowing us to draw on human experts to potentially assist general practitioners and young physicians in difficult situations and enable early and specific diagnosis.

Abbreviations

AI	Artificial intelligence
DL	Deep learning
LSH	Locality-sensitive hashing
RS	Recommender system

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1007/s00330-024-10672-0>.

Below is the link to the electronic supplementary material. Supplementary file1 (PDF 307 KB)

Funding

Open Access funding enabled and organized by Projekt DEAL. The authors state that this work has not received any funding.

Declarations

Guarantor

The scientific guarantor of this publication is Daniel Rueckert.

Conflict of interest

The authors of this manuscript declare no relationships with any companies, whose products or services may be related to the subject matter of the article.

Statistics and biometry

One of the authors has significant statistical expertise. No complex statistical methods were necessary for this paper.

Informed consent

Written informed consent was waived by the Institutional Review Board.

Ethical approval

Institutional Review Board approval was obtained (no. 48/205).

Study subjects or cohorts overlap

No study subjects or cohort overlap has been reported.

Methodology

- retrospective
- diagnostic or prognostic study
- performed at one institution

Author details

¹Department of Orthopaedics and Sports Orthopaedics, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany. ²Institute for AI and Informatics in Medicine, Technical University of Munich, Munich, Germany. ³Institute of Biological and Medical Imaging, Helmholtz Zentrum München, Neuherberg, Germany. ⁴Institute at Helmholtz: Institute of Computational Biology, Oberschleißheim, Germany. ⁵Chair of Biological Imaging at the Central Institute for Translational Cancer Research (TranslaTUM), School of Medicine, Technical University of Munich, Munich, Germany. ⁶Musculoskeletal Radiology Section, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany.

Received: 20 October 2023 Revised: 16 January 2024 Accepted: 5 February 2024

Published online: 15 March 2024

References

- Grimer RJ, Carter SR, Pynsent PB (1997) The cost-effectiveness of limb salvage for bone tumours. *J Bone Joint Surg Br* 79:558–561
- Hinterwimmer F, Consalvo S, Neumann J, Rueckert D, von Eisenhart-Rothe R, Burgkart R (2022) Applications of machine learning for imaging-driven diagnosis of musculoskeletal malignancies—a scoping review. *Eur Radio* 32:7173–7184
- Picci P, Manfrini M, Donati DM et al (2020) Diagnosis of musculoskeletal tumors and tumor-like conditions: clinical, radiological and histological correlations—the Rizzoli Case Archive. Springer
- Rechl H, Kirchhoff C, Wortler K, Lenze U, Topfer A, von Eisenhart-Rothe R (2011) [Diagnosis of malignant bone and soft tissue tumors]. *Orthopade* 40:931–941; quiz 942–933
- Gaume M, Chevret S, Campagna R, Larousserie F, Biau D (2022) The appropriate and sequential value of standard radiograph, computed tomography and magnetic resonance imaging to characterize a bone tumor. *Scientific Rep* 12:1–9
- Grimer RJ, Briggs TW (2010) Earlier diagnosis of bone and soft-tissue tumours. *J Bone Joint Surg Br* 92:1489–1492
- Clark MA, Thomas JM (2005) Delay in referral to a specialist soft-tissue sarcoma unit. *Eur J Surg Oncol* 31:443–448
- Miller BJ (2019) Use of imaging prior to referral to a musculoskeletal oncologist. *J Am Acad Orthop Surg* 27:e1001–e1008
- Kindblom LG (2009) Bone tumors: epidemiology, classification, pathology. In: Davies A, Sundaram M, James S (eds) *Imaging of bone tumors and tumor-like lesions*. Medical Radiology. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-77984-1_1
- Aiello M, Cavaliere C, D'Albore A, Salvatore M (2019) The challenges of diagnostic imaging in the era of big data. *J Clin Med* 8:316
- Hinterwimmer F, Consalvo S, Wilhelm N et al (2022) SAM-X: sorting algorithm for musculoskeletal x-ray radiography. *Eur Radiol* 33:1537–1544
- Kharat AT, Singhal S (2017) A peek into the future of radiology using big data applications. *Indian J Radiol Imaging* 27:241–248
- Lacroix M, Aouad T, Feydy J et al (2023) Artificial intelligence in musculoskeletal oncology imaging: a critical review of current applications. *Diag and Interv Imaging* 104:18–23
- Rajpurkar P, Chen E, Banerjee O, Topol EJ (2022) AI in health and medicine. *Nat Med* 28:31–38
- Consalvo S, Hinterwimmer F, Neumann J et al (2022) Two-phase deep learning algorithm for detection and differentiation of ewing sarcoma and acute osteomyelitis in paediatric radiographs. *Anticancer Res* 42:4371–4380
- Liu X, Faes L, Kale AU et al (2019) A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health* 1:e271–e297
- von Schacky CE, Wilhelm NJ, Schäfer VS et al (2021) Multitask deep learning for segmentation and classification of primary bone tumors on radiographs. *Radiology* 301:398–406
- Singh PK, Pramanik PKD, Dey AK, Choudhury P (2021) Recommender systems: an overview, research trends, and future directions. *Int J Business Syst Res* 15:14–52
- Kuanr M, Mohapatra P, Mittal S, Maindarkar M, Fouda MM, Saba L, Saxena S, Suri JS (2022) Recommender system for the efficient treatment of COVID-19 using a convolutional neural network model and image similarity. *Diagnostics* 12(11):2700. <https://www.mdpi.com/2075-4418/12/11/2700>
- Saadat H, Shah B, Halim Z, Anwar S (2022) Knowledge graph-based convolutional network coupled with sentiment analysis towards enhanced drug recommendation. *IEEE/ACM Trans Comput Biol Bioinform*. <https://pubmed.ncbi.nlm.nih.gov/36441898/>
- Mongan J, Moy L, Kahn Jr CE (2020) Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. *Radiol Artif Intell* 2(2):e200029. <https://pubs.rsna.org/doi/full/10.1148/ryai.2020200029>
- Board WCOTE (2020) Soft tissue and bone tumours. International Agency for Research on Cancer. <https://publications.iarc.fr/Book-And-Report-Series/Who-Classification-Of-Tumours/Soft-Tissue-And-Bone-Tumours-2020>
- Vandenbroucke JP, von Elm E, Altman DG et al (2007) Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *Epidemiol* 18:805–835
- He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
- Zhang Z, Zhang H, Zhao L, Chen T, Arik SÖ, Pfister T (2022) Nested hierarchical transformer: towards accurate, data-efficient and interpretable visual understanding. Proceedings of the AAAI Conference on Artificial Intelligence, 3417–3425
- Bloier M, Hinterwimmer F, Breden S et al (2022) Detection and segmentation of heterogeneous bone tumours in limited radiographs. *current directions in biomedical engineering*. De Gruyter, 69–72
- Jafari O, Maurya P, Nagarkar P, Islam KM, Crushev C (2021) A survey on locality sensitive hashing algorithms and their applications. arXiv preprint arXiv:210208942
- Murphy ZR, Venkatesh K, Sulam J, Yi PH (2022) Visual transformers and convolutional neural networks for disease classification on radiographs: a comparison of performance, sample efficiency, and hidden stratification. *Radiol Artif Intell* 4:e220012
- Kuncheva LI, Whitaker CJ (2003) Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learn* 51:181
- Taljanovic MS, Hunter TB, Fitzpatrick KA, Krupinski EA, Pope TL (2003) Musculoskeletal magnetic resonance imaging: importance of radiography. *Skeletal Radiol* 32:403–411
- Ory PA (2003) Radiography in the assessment of musculoskeletal conditions. *Best Pract Res Clin Rheumatol* 17:495–512
- Yu AC, Mohajer B, Eng J (2022) External validation of deep learning algorithms for radiologic diagnosis: a systematic review. *Radiol Artif Intell* 4:e210064
- Park SH, Han K (2018) Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology* 286:800–809

Publisher's Note

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

Abbreviations

AI artificial intelligence

ANNs artificial neural networks

AUC area under the curve

CNNs convolutional neural networks

CT computed tomography

DICOM Digital Imaging and Communications in Medicine

DL deep learning

GANs generative adversarial networks

GPUs graphical processing units

GPT generative pre-trained transformer

HIS hospital information system

LLMs large language models

LSTM long short-term memory

LSH locality-sensitive hashing

ML machine learning

Abbreviations

MRI magnetic resonance imaging

MSK musculoskeletal

NLP natural language processing

NMI normalised mutual information

PACS picture archiving and communication system

PET positron emission tomography

ResNets residual neural networks

RNNs recurrent neural network

VAEs variational autoencoders

List of Figures

1.1. Exemplary radiological (a-c) and pathological (d) imaging data of a musculoskeletal tumour patient: a) preoperative x-ray, b) preoperative MRI, c) CT-guided needle biopsy, d) histology.	2
1.2. Example protocol from weekly MSK tumour board meeting.	4
1.3. proposed approach of deep learning methods for musculoskeletal tumour diagnostics - the overall problem of musculoskeletal tumour diagnostic is that early diagnostics is complex and delays occur. Deep learning models can help based on retrospective data to handle limited data, optimise workflows and finally build diagnostic support tools. (This figure has been designed using images from Flaticon.com)	14
2.1. Selection process for final references [1].	22
2.2. Transfer learning approach through self-supervised pre-structuring of the data to obtain auxiliary labels [2].	25
2.3. Illustration of the presented framework in two phases: clustering data with a self-supervised model and training a network with human-annotated clusters [3].	27
2.4. Flow chart of the proposed model – (I) preparing the images, training of the convolutional neural network, saving the model and features, (II) calculating the high dimensional distances with a distance function, adding a hash tables, clustering of the most similar x-rays and calculating a precision-at-k and a tumour entity classification with a majority vote of the k-clustered images [4].	29

List of Tables

2.1. Distribution of discrete parameters with incidence and percentage ratio.	18
2.2. Distribution of continuous parameters with median and interquartile range.	19

A. Related publications

Two-Phase Deep Learning Algorithm for Detection and Differentiation of Ewing Sarcoma and Acute Osteomyelitis in Paediatric Radiographs

SARAH CONSALVO¹, FLORIAN HINTERWIMMER^{1,2}, JAN NEUMANN³,
MARC STEINBORN⁴, MAYA SALZMANN¹, FRITZ SEIDL⁵, ULRICH LENZE¹,
CAROLIN KNEBEL¹, DANIEL RUECKERT² and RAINER H.H. BURBKART¹

¹Department of Orthopaedics and Sports Orthopaedic, Klinikum Rechts der Isar, Technical University of Munich, Munich, Germany;

²Institute for AI and Informatics in Medicine, Technical University of Munich, Munich, Germany;

³Department of Diagnostic and Interventional Radiology, Klinikum Rechts der Isar, Technical University of Munich, Munich, Germany;

⁴Institute for Diagnostic and Interventional Radiology and Paediatric Radiology, Klinikum Schwabing, Munich, Germany;

⁵Department of Trauma Surgery, Klinikum Rechts der Isar, Technical University of Munich, Munich, Germany

Abstract. *Background/Aim:* Ewing sarcoma is a highly malignant tumour predominantly found in children. The radiological signs of this malignancy can be mistaken for acute osteomyelitis. These entities require profoundly different treatments and result in completely different prognoses. The purpose of this study was to develop an artificial intelligence algorithm, which can determine imaging features in a common radiograph to distinguish osteomyelitis from Ewing sarcoma. *Materials and Methods:* A total of 182 radiographs from our Sarcoma Centre (118 healthy, 44 Ewing, 20 osteomyelitis) from 58 different paediatric (≤ 18 years) patients were collected. All localisations were taken into consideration. Cases of acute, acute on chronic osteomyelitis and intraosseous Ewing sarcoma were included. Chronic osteomyelitis, extra-skeletal Ewing sarcoma, malignant small cell tumour and soft tissue-based primitive neuroectodermal tumours were excluded. The algorithm development was split into two phases and two different classifiers were built and

combined with a Transfer Learning approach to cope with the very limited amount of data. In phase 1, pathological findings were differentiated from healthy findings. In phase 2, osteomyelitis was distinguished from Ewing sarcoma. Data augmentation and median frequency balancing were implemented. A data split of 70%, 15%, 15% for training, validation and hold-out testing was applied, respectively. *Results:* The algorithm achieved an accuracy of 94.4% on validation and 90.6% on test data in phase 1. In phase 2, an accuracy of 90.3% on validation and 86.7% on test data was achieved. Grad-CAM results revealed regions, which were significant for the algorithms decision making. *Conclusion:* Our AI algorithm can become a valuable support for any physician involved in treating musculoskeletal lesions to support the diagnostic process of detection and differentiation of osteomyelitis from Ewing sarcoma. Through a Transfer Learning approach, the algorithm was able to cope with very limited data. However, a systematic and structured data acquisition is necessary to further develop the algorithm and increase results to clinical relevance.

Correspondence to: Sarah Consalvo, Ismaninger Straße 22, 81675 Munich, Germany. Tel: +49 8941408012, e-mail: sarah.consalvo@mri.tum.de

Key Words: Artificial intelligence, Ewing sarcoma, osteomyelitis, tumor detection, early diagnosis, deep learning, transfer learning.



This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-NC-ND) 4.0 international license (<https://creativecommons.org/licenses/by-nc-nd/4.0>).

Ewing sarcomas (ES) represent 7-10% of all bone malignancies and have the second highest incidence after osteosarcomas (1). The main differential diagnoses of Ewing sarcoma are acute osteomyelitis (OM) and Langerhans Histiocytosis. Acute osteomyelitis is a severe bone infection which most often has a haematogenous origin (2). Other causes can be trauma, surgery, or contiguously infected soft tissue. It occurs in 8 out of 100,000 children per year in high-income countries, yet it is extremely common in developing countries as well. Male

children are affected twice as often as female children (3). Clinical and laboratory exams might be normal. Blood cultures and biopsy samples are positive for bacteria in only 32-62% and 40-60%, respectively. *Staphylococcus aureus*, β -haemolytic *Streptococcus*, *Streptococcus pneumoniae*, *Escherichia coli* and *Pseudomonas aeruginosa* are the most common bacteria involved in this acute bone infection (4). The symptoms include pain, ROM (Range of Motion) limitations and fever (5). After all, with proper treatment, the outcome for OM is usually good. Conservative treatment with antibiotics is effective in 90% of the early diagnosed paediatric cases (5, 6).

However, Ewing sarcoma is a highly malignant blue round cell tumour, 90% of whose cases occur in patients between age 5 to 25. Worldwide, 2.9 out of 1,000,000 children per year are affected by this malignancy, with a slightly higher incidence in male patients (1.5 male: 1 female) (7). Children usually present with load-independent local pain and ROM limitation without a history of trauma, lasting for at least four to six weeks. Ewing sarcoma treatment begins and ends with chemotherapy. Surgery to remove the cancer is normally performed after neoadjuvant chemotherapy.

Taking into consideration the completely contrasting course of these two diseases, early diagnosis and referral to a specialised centre is crucial for a successful treatment. However differential diagnosis is extremely difficult.

Radiographs and MR images have a relatively low diagnostic value in this crucial differential diagnosis (8, 9), if not interpreted by a trained and experienced musculoskeletal radiologist.

In brief summary, the symptoms, blood screening, as well as the localisation (10) are extremely similar in both diseases. The first radiological exam to conduct a differential diagnosis apart from an ultrasound will be an X-ray. Even with this imaging modality, the diagnosis will not be clear. Although methods of nuclear medicine such as PET and SPECT are currently the most accurate techniques, they are too elaborate to be used in the phase of differential diagnosis and they are usually not available for outpatient clinics (11, 12).

In radiographs, both entities can present bone destruction and periosteal reaction. The typical periosteal reaction associated with Ewing sarcoma – lamellated, “onion skin” – or “Codman’s triangle” can also be present in acute osteomyelitis due to a subperiosteal abscess (4). Instead, MR T1-weighted images in comparison with short tau inversion recovery (STIR) showing sharp margins are one of the most significant signs of Ewing sarcoma for the differentiation from osteomyelitis (13). Hence, MRI, PET and SPECT are complex techniques that are indicated when a solid suspicion is provided or when the diagnosis is to be validated. The resemblance of the radiological features as well as the clinical course makes it demanding to distinguish these two entities.

According to Bacci *et al.* (14), the overall delay between initial symptoms and biopsy for Ewing sarcoma is

approximately four months. If we consider that the estimated five-year survival for Ewing sarcoma patients shifts from 50-70% in early diagnosed localised cases to 18-30% in metastatic cases (15) and that unfortunately, 25% of all Ewing sarcoma patients have a metastatic disease at the time of diagnosis (16), four months “until” or “since” the first diagnosis make a huge difference in the prognosis of these young patients. To shorten the delay of referral to a specialised centre, it is crucial to improve the ability of outpatient clinics to address a suspicious case. In this process, radiographs represent the first obligatory step. In order to prevent delays and limitation of the prognosis, it is decisive to develop a new form of assistance which can support precision and accuracy of the diagnostic process.

Image interpretation as a part of precision medicine will play an increasingly important role in the future of orthopaedic oncology and novel, more comprehensive and specific analysis tools are urgently needed, especially for outpatient clinics with limited experience and resources for detection and interpretation of rare bone malignancies. Deep learning (DL) represents a subset of Machine Learning and a distinct application of artificial intelligence (AI), which evolved from pattern recognition and learning theory. While complex data analysis of cancerous tissue by AI models and imaging data is already widely applied in some medical specialties (*e.g.* lung and breast cancer), the application of these methods in orthopaedic oncology is still very limited (17). The fact that globally no far-reaching structures for systematic data acquisition have yet been established and that sarcomas are very rare and heterogeneous entities makes modern AI applications, for which a sufficient and qualitative amount of data is crucial, considerably more difficult. While this is a common obstacle – particularly in medicine – several techniques to cope with limited data have emerged. One popular technique is called data augmentation (18), in which new data is created artificially by applying minor transformations to initial data. Another even more powerful method is Transfer Learning (19), where a model is developed for a source task and then reused as a starting point for the target task (Figure 1).

The focus of this study was to develop a real-time support tool for the detection and distinction between Ewing sarcoma and acute osteomyelitis using a two-phase DL algorithm.

Materials and Methods

Data and ethics approval. The local institutional review and ethics board (Klinikum rechts der Isar, Technical University of Munich) approved this retrospective study (N°48/20S). The study was performed in accordance with national and international guidelines. The study is a purely retrospective study in which all data are already available and are collected in pseudonymised form with the help of the musculoskeletal tumour database or by studying files. To increase the quality of the presented observational study and its prediction model, reporting was derived from the Transparent

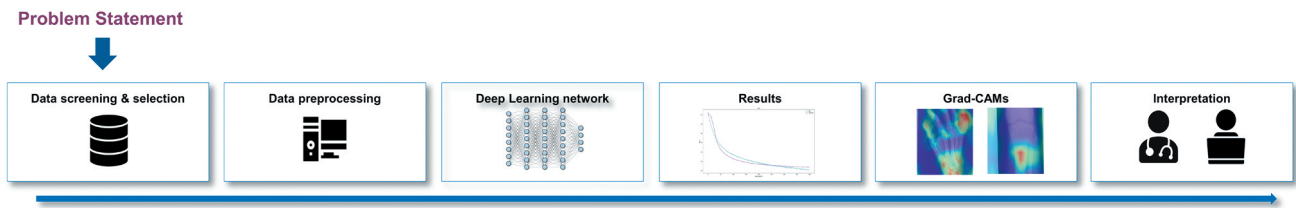


Figure 1. Study work flow.

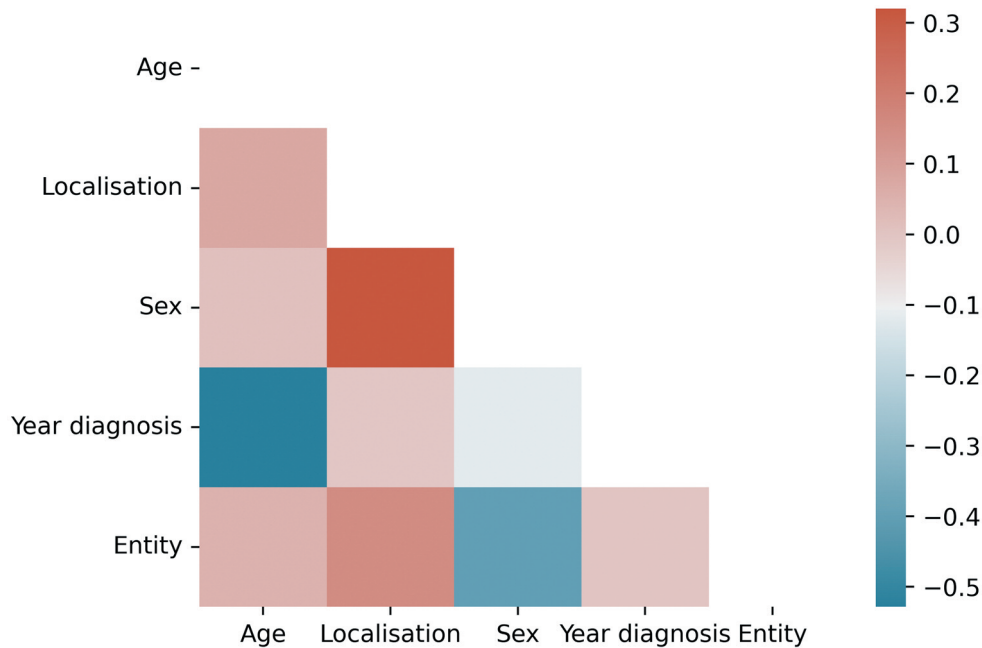


Figure 2. Correlation matrix: a matrix describing the correlation between age, localization, sex, year of diagnosis and entity.

Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines (20) and the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement (21).

Eligibility criteria. All patients from our database with the according ICD-10 code for OM and ES were selected. For all patients, the diagnoses were validated through a histopathological examination as reference standard. The data was retrieved from our hospital information system (HIS) and the picture archiving and communications system (PACS).

The following inclusion criteria were applied:

- patients younger than or equal to 18 years;
- intraosseous Ewing sarcoma;
- histopathologically confirmed cases of acute osteomyelitis or acute on chronic osteomyelitis;
- images prior to treatment.

Patients older than 18 years, chronic osteomyelitis, extraosseous Ewing sarcoma, malignant small cell tumour, soft tissue-based primitive neuroectodermal tumours (PNET) cases were excluded.

Statistical analysis. For statistical analysis and evaluation, accuracy, sensitivity, and specificity were computed for each phase, cross-validated and interpreted by an orthopaedic surgeon (S.C.) and a computer scientist (F.H.). The metrics were implemented using the scikit-learn library (https://scikit-learn.org/stable/modules/model_evaluation.html).

Considering that the control group was selected, only the patients with acute osteomyelitis and Ewing sarcoma were included in the statistical analysis. Nevertheless, a control group was needed to develop an algorithm for detection of pathological cases in the first place.

Except for the 'Localisation', none of the patient meta data is normally distributed according to normality test by D'Agostino-Pearson. Figure 2 shows a correlation matrix according to values of Spearman's rank-order correlation coefficient, which is a measure for linear correlation between two datasets and does not assume that both datasets are normally distributed. Only 'Age'/'Year' of diagnosis' and 'Sex'/'Entity' show a slight indirect correlation ($|r| > 0.4$). It is to be expected with small datasets that no high and stable correlations can be found.

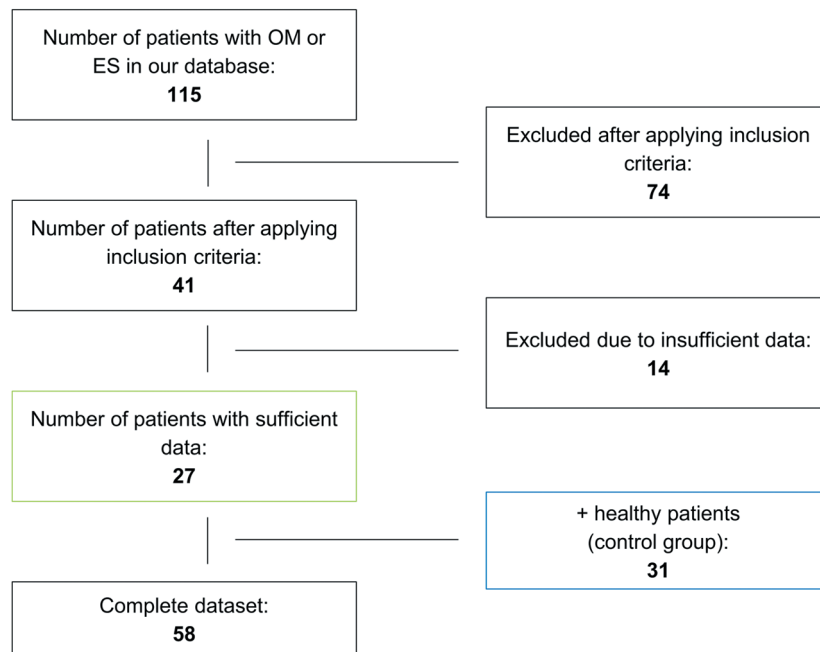


Figure 3. Flow chart: description of patient selection according to eligibility criteria.

Model training. Model training and inference was conducted on a DGX Station A100 with four 80GB graphical processing units (Nvidia Corporation, Santa Clara, CA, USA), 64 2.25 GHz cores and 512 GB DDR4 system memory running on a Linux/Ubuntu 20.04 distribution (Canonical, London, England). Preprocessing and model implementation were performed in Python 3.9.6 (<https://www.python.org/>) using PyTorch 1.9.0 and cuda toolkit 11.1 (<https://pytorch.org/>).

The source code for this study is provided on GitHub (<https://github.com/FlorianH3000/ewing>).

A supervised DL algorithm for image classification in two phases was developed: phase 1 for detection of pathological cases and phase 2 for differentiation of ES and OM cases. For both phases, a ResNet (22) was selected. Beforehand, the model was pretrained on 42,608 sarcoma related X-ray images for Transfer learning. For phase 1 and 2 a ResNet18 architecture was chosen. To tackle the overall limited amount of data and integrate regularization, extensive data augmentation was implemented to artificially create more input data during training. In order to manage the class imbalances in both phases, median frequency balancing was utilized to weight the loss of classes accordingly and support the robustness of the algorithm (23). A data split of 70%, 15%, 15% was applied for training, validation, and testing, respectively. Since up to four images from single patients were collected, the data split was applied to patients in order to avoid cross-contamination and therefore provide a higher statistical significance. An additional 6-fold cross validation supported this task, while random chosen hold-out test data for final evaluation remained untouched.

Plausibility. To provide plausibility and more insight into the AI model, Grad-CAMs were implemented in the final inference step (24). Grad-

CAMs utilize the gradient information from the last convolutional layer of a deep learning network to understand specific neurons and their impact for decision-making. The result is a coloured heat map, which is co-registered to the original input image and indicates where the algorithm found relevant information for the task at hand. This technique was applied to get a better understanding where the algorithm detects relevant information. To provide a higher statistical significance, the Grad-CAM results were averaged from the 6-fold cross validation.

Results

Dataset. A total of 115 patients treated in our institution for OM or ES between 2000 and 2021 were retrospectively reviewed. After applying the inclusion criteria, 74 cases were excluded, and 41 cases remained. After screening the data, another 14 cases were excluded due to insufficient or invalid data. Ultimately, 27 cases, 9 with acute osteomyelitis and 18 with Ewing sarcoma were collected.

Additionally, 31 healthy cases were included in order to balance the dataset and create a “control group”. These patients were treated in our emergency room with a history of acute trauma of a joint. The performed X-ray could exclude any kind of fracture or bone anomalies so that these cases were diagnosed as bruises or contusions. Consequently, a “healthy group” without exposing children to X-ray radiation for our study was obtained. The control group was chosen with similar localisation to our “pathological group”. Overall, 182 radiographs (healthy 118, 44 Ewing, 20

Table I. *Distribution of Ewing sarcoma (ES) and osteomyelitis (OM) dataset according to patient characteristics (sex and localisation).*

	Entity		Sex		Localisation			
	#	%	#	%		#		
Pathological cases (ES & OM)	Acute osteomyelitis	9	33.3%	Female	5	18.5%	Upper extr.	2
				Male	4	14.8%	Lower extr.	6
	Ewing sarcoma	18	66.7%	Female	3	11.1%	Other	1
				Male	15	55.6%	Upper extr.	6
							Lower extr.	9
	Total	27	100.0%	Female	8	29.6%	Other	3
Male				19	70.4%	Upper extr.	8	
Control group (healthy)	Total	31	100.0%	Female	15	48.4%	Lower extr.	15
				Male	16	51.6%	Other	4
							Upper extr.	11
Complete dataset	Pathological cases (relative to whole dataset)	27	46.6%	Female	8	13.8%	Upper extr.	11
				Male	19	32.8%	Lower extr.	20
	Control group (relative to whole dataset)	31	53.4%	Female	15	25.9%	Other	4
				Male	16	27.6%	Upper extr.	11
							Lower extr.	20
	Total	58	100.0%	Female	23	39.7%	Other	0
Male				35	60.3%	Upper extr.	19	
						Lower extr.	35	
						Other	4	

Table II. *Age distribution of involved patients classified in Ewing sarcoma (ES) group and osteomyelitis (OM) group: pathological cases, control group and complete dataset.*

	Entity		Age			
	#	%	Average	Variance	Standard deviation	
Pathological cases (ES & OM)	Acute osteomyelitis	9	33.3%	13.6	5.6	2.4
	Ewing sarcoma	18	66.7%	12.8	20.5	4.5
	Total	27	100.0%	13.0	15.7	4.0
Control group (healthy)	Total	31	100.0%	6.4	16.6	4.1
Complete dataset	Pathological cases	27	46.6%	13.0	15.7	4.0
	Control group	31	53.4%	6.4	16.6	4.1
	Total	58	100.0%	9.5	27.1	5.2

osteomyelitis) from 58 patients were collected including data from external imaging data (Figure 3).

Patient characteristics. The dataset including the healthy control group, Ewing sarcoma and acute osteomyelitis consists of 23 females (39.7%) and 35 males (60.3%). While 19 (32.8%) of the patients were affected at their upper extremities, 35 (60.3%) were affected at their lower extremities and 4 (6.9%)

at other localisations. The average age of patients at the time of the initial diagnosis resulted in 9.5 years with a variance of 27.6 and a standard deviation of 5.2 (Table I and Table II).

Model performance in phase I. All results were cross-validated. The first two-entity classification of the healthy control group and the pathological group resulted in an accuracy of 94.4%/90.6%, sensitivity of 90.0%/89.4% and

Performance phase 1

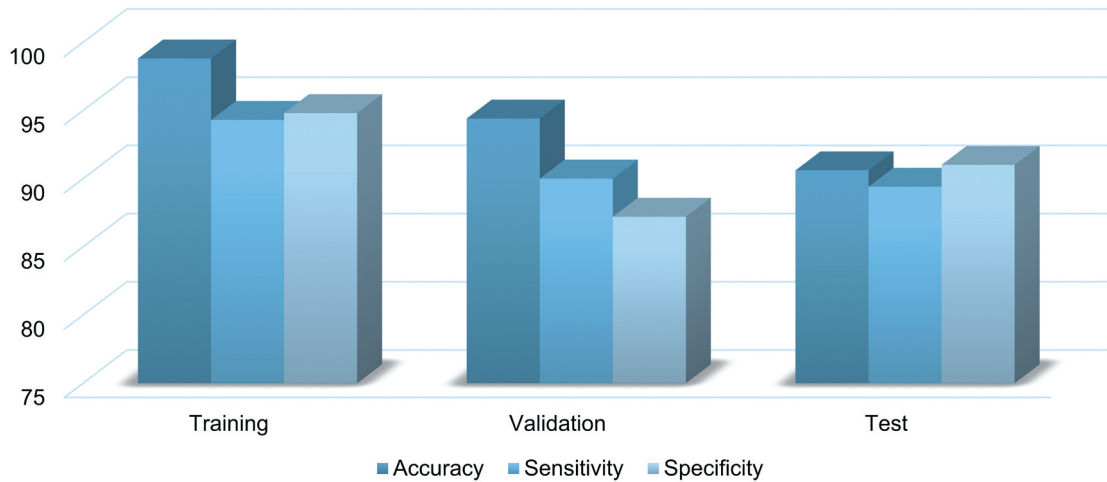


Figure 4. Prediction of performance in Phase 1.

Performance phase 2

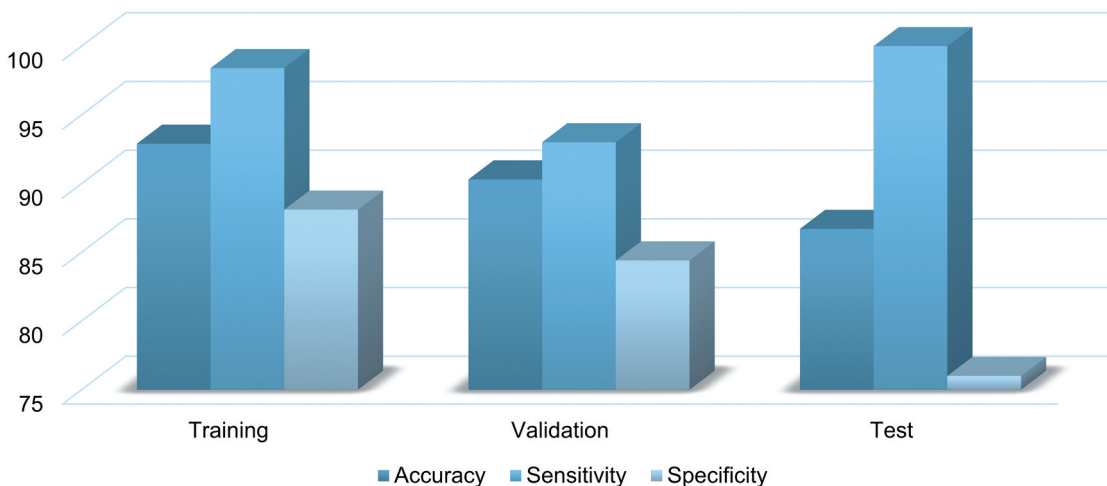


Figure 5. Prediction of performance in Phase 2.

specificity of 87.2%/91.0% for the validation and test split, respectively (Figure 4).

Model performance in phase 2. All results were cross-validated. The second two-entity classification of OM and ES cases resulted in an accuracy of 90.3%/86.7%, sensitivity of 93.0%/100.0% and specificity of 84.4%/76.0% for the validation and test dataset, respectively (Figure 5).

Grad-CAM results. Figure 6 and Figure 7 display the results of Grad-CAM visualizations from the test dataset of each

entity. The displayed Figures show that the algorithm did in fact find relevant information in very similar areas where a trained radiologist or an orthopaedic surgeon would look at when diagnosing a patient based on a radiograph.

Discussion

The most important finding of this study is that even with a very limited amount of data, good results in detecting and distinguishing Ewing sarcoma from acute osteomyelitis can be achieved through data augmentation and particularly

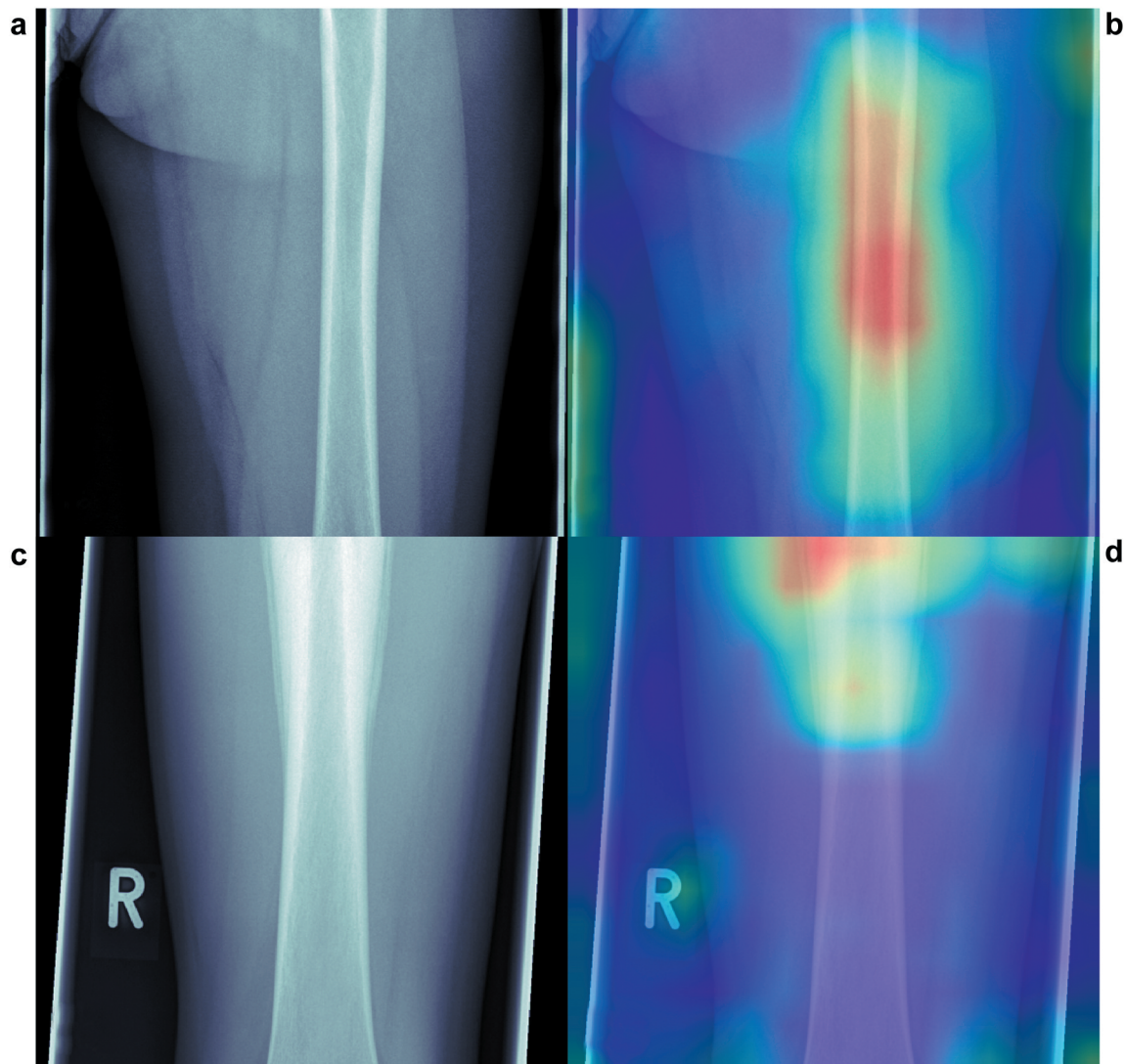


Figure 6. Grad-CAM of healthy (a/b) and pathological cases (c/d) in phase 1: Grad-CAM results displaying that the algorithm focused on pixels similar to the areas a radiologist or orthopaedic surgeon would look at.

Transfer Learning. Nevertheless, to further increase the results, a systematic and structured data acquisition is necessary to gather sufficient data and improve the overall accuracy.

Limitations. The main limitation of studying these entities is the extreme rarity of Ewing sarcoma. This makes it very challenging to acquire sufficient imaging data that could enhance the accuracy and stability of the algorithm. Additionally, in most centres data infrastructures are not yet fully adapted to the needs of modern AI applications. Current HIS and PACS systems were often initially set up years ago and were not designed to retrieve data for AI research. Thus,

a considerable amount of data was lost over the years (14 patients excluded due to insufficient data).

While several precautions to provide statistical significance were applied – such as cross validation, loss weighting or incorporating Transfer Learning via pretrained networks - limited amount of data for final validation and testing might still bias the accuracy of the algorithm compared to real-world scenarios. However, this issue can most likely be addressed with further establishment of collaboration of specialised centres, the according data infrastructure, and therefore more sufficient datasets.

Another limitation of this study is that the DL model did not use demographics or other important patient characteristics as

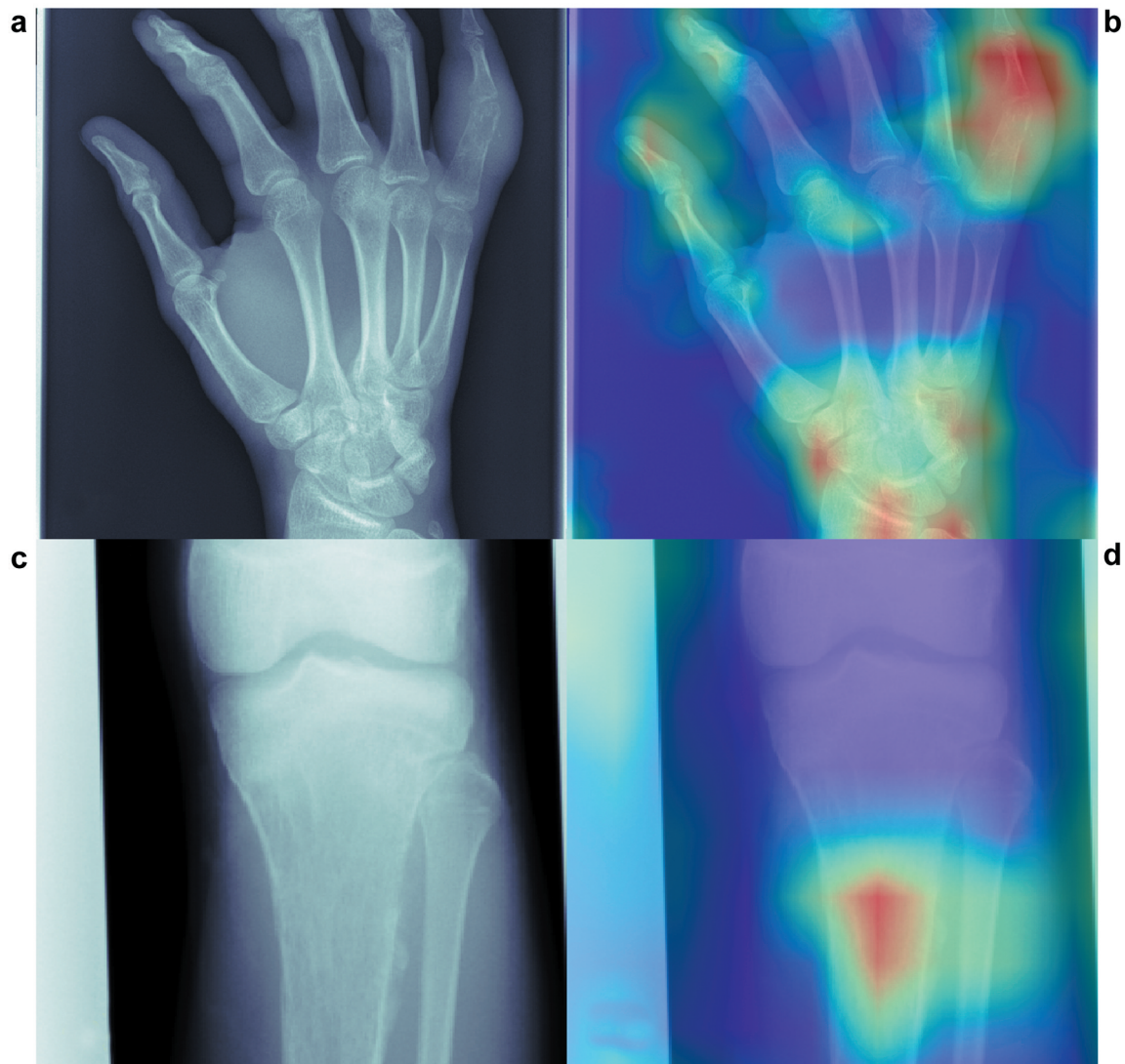


Figure 7. Grad-CAMs of an OM patient (a/b) and an ES patient (c/d) in phase 2: Grad-CAM results showing that for ES the algorithm did also focus on similar areas a radiologist or orthopaedic surgeon would look at, but in the acute osteomyelitis case several areas were in focus of the algorithm.

input. This study is supposed to be a feasibility study for radiographs, ES and OM. Nevertheless, integrating meta information into the algorithm is one of the next steps.

Interpretation of results. From a medical as well as a computer science point of view, the performances are very promising considering the complexity of the radiological manifestation of the diseases and the very limited amount of available data. Not only the overall accuracy, but the sensitivity and specificity (also incorporating true positive rate and the true negative rate), concluded in considerably high results.

The model accuracy obtained in the study of von Schacky *et al.* (25) involving all primary bone tumours was comparable

with a musculoskeletal fellowship-trained radiologist (71.2% and 64.9%, respectively) and even higher than the one obtained by radiologic residents (83.8% and 82.9%; respectively). Therefore, we can hypothesize that deep learning algorithms, such as the one presented in this study, can potentially become a significant support - particularly for outpatient clinic doctors who do not have access to expert orthopaedic tumour radiologists. The algorithm could help to reduce the delay of referral to a specialised centre and improve the overall survival of young patients.

While this study demonstrates the feasibility of interpreting X-ray images with ES and OM through DL and most likely also surpasses the accuracy of outpatient clinics (no literature

was found to underline this statement), however, statistical robustness must be further investigated before a decision support tool can be integrated into a clinical environment.

Interpretation of Grad-CAMs. While the significance and validity of Grad-CAMs is for some tasks also controversially discussed in the field of computer science, we still believe that it is worth analysing and interpreting specific Grad-CAM results. For example, Figure 7 (Grad-CAM c) shows that the suspect region around the middle phalange of the 4th finger was detected by the algorithm, but additionally several other spots in the wrist area affected the algorithm's decision. Such findings can help to unravel the "black box" behind state-of-the-art DL algorithms, might indicate new ways to evaluate radiographs (and also other imaging modalities) and on the long run assist the process of making precise and fast diagnosis.

Future application. The primary application of the developed algorithm is focused on outpatient clinics. While specialised centres usually have several sarcoma experts as well as more sophisticated imaging modalities, an outpatient clinic doctor has to rely on his/her expertise and radiographic diagnostics to conclude a first diagnosis and potentially refer a patient to a specialised centre, while having seen only about three musculoskeletal malignancies in his/her professional life (26). In such a case, a support tool to highlight suspect cases and even identify ES or OM could have a significant impact.

Conclusion

Radiography is a common and largely available imaging technique that is often used for first clinical assessment. Although radiographs only consist of two-dimensional greyscale information, the high resolution and the considerably standardised technique still make it a very suitable input for modern algorithms. We believe that AI algorithms can become a valuable real-time support for any outpatient clinic involved in the crucial processes of detecting and differentiating a case of acute osteomyelitis from a possible case of an Ewing sarcoma. This allows for a minimal loss of time between diagnosis and specific treatment, which is crucial for patients with Ewing sarcoma. While our algorithm was developed for a specific dataset, it can function as a template for other entities with minor adjustments, where a radiograph can be utilised for early and precise detection for various diseases.

Conflicts of Interest

The Authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors' Contributions

FH - Project administration, Software, Writing - Original Draft, Visualization. SC - Project administration, Conceptualization, Writing - Original Draft, Data Curation. JN - Writing - Review & Editing, Formal analysis. MS - Investigation, Data Curation. AS - Investigation, Data Curation. MS - Investigation, Data Curation. FS - Writing - Review & Editing. UL - Writing - Review & Editing. CK - Writing - Review & Editing. DR - Methodology, Supervision, Validation. RE - Supervision, Validation. RB - Supervision, Validation

Acknowledgements

The Authors would like to thank Bernhard Renger from the Department of Diagnostic and Interventional Radiology (Klinikum Rechts der Isar, Technical University of Munich, Munich, Germany) for his support and his valuable help in retrieving the data from the database backend of the clinical systems (27).

References

- Weber MA, Papakonstantinou O, Nikodinovska VV and Vanhoenacker FM: Ewing's sarcoma and primary osseous lymphoma: Spectrum of imaging appearances. *Semin Musculoskelet Radiol* 23(1): 36-57, 2019. PMID: 30699452. DOI: 10.1055/s-0038-1676125
- Peltola H and Pääkkönen M: Acute osteomyelitis in children. *N Engl J Med* 370(4): 352-360, 2014. PMID: 24450893. DOI: 10.1056/NEJMra1213956
- Chiappini E, Mastrangelo G and Lazzeri S: A case of acute osteomyelitis: an update on diagnosis and treatment. *Int J Environ Res Public Health* 13(6): 539, 2016. PMID: 27240392. DOI: 10.3390/ijerph13060539
- Pineda C, Vargas A and Rodríguez AV: Imaging of osteomyelitis: current concepts. *Infect Dis Clin North Am* 20(4): 789-825, 2006. PMID: 17118291. DOI: 10.1016/j.idc.2006.09.009
- McNeil JC: Acute hematogenous osteomyelitis in children: Clinical presentation and management. *Infect Drug Resist* 13: 4459-4473, 2020. PMID: 33364793. DOI: 10.2147/IDR.S257517
- Peltola H, Pääkkönen M, Kallio P, Kallio MJ and Osteomyelitis-Septic Arthritis Study Group: Short- versus long-term antimicrobial treatment for acute hematogenous osteomyelitis of childhood: prospective, randomized trial on 131 culture-positive cases. *Pediatr Infect Dis J* 29(12): 1123-1128, 2010. PMID: 20842069. DOI: 10.1097/INF.0b013e3181f55a89
- Picci P, Manfrini M, Fabbri N, Gambarotti M and Vanel D: Atlas of musculoskeletal tumors and tumorlike lesions: The rizzoli case archive. Springer International Publishing, 2016.
- McCarville MB, Chen JY, Coleman JL, Li Y, Li X, Adderson EE, Neel MD, Gold RE and Kaufman RA: Distinguishing osteomyelitis from Ewing sarcoma on radiography and MRI. *AJR Am J Roentgenol* 205(3): 640-50; quiz 651, 2015. PMID: 26295653. DOI: 10.2214/AJR.15.14341
- Kuleta-Bosak E, Kluczevska E, Machnik-Broncel J, Madziara W, Ciupińska-Kajor M, Sojka D, Rogala W, Juszczyk J and Wilk R: Suitability of imaging methods (X-ray, CT, MRI) in the diagnostics of Ewing's sarcoma in children - analysis of own material. *Pol J Radiol* 75(1): 18-28, 2010. PMID: 22802757.

- 10 Mar WA, Taljanovic MS, Bagatell R, Graham AR, Speer DP, Hunter TB and Rogers LF: Update on imaging and treatment of Ewing sarcoma family tumors: what the radiologist needs to know. *J Comput Assist Tomogr* 32(1): 108-118, 2008. PMID: 18303298. DOI: 10.1097/RCT.0b013e31805c030f
- 11 Aggarwal H, D'souza M, Panwar P, Jyotsna N, Alvi T, Solanki Y, Kumar T and Sharma R: Role of fluoroethyl tyrosine positron emission tomography-computed tomography scan in differentiating ewing's sarcoma from osteomyelitis. *World J Nucl Med* 18(1): 77-80, 2019. PMID: 30774555. DOI: 10.4103/wjnm.WJNM_23_18
- 12 Santiago Restrepo C, Giménez CR and McCarthy K: Imaging of osteomyelitis and musculoskeletal soft tissue infections: current concepts. *Rheum Dis Clin North Am* 29(1): 89-109, 2003. PMID: 12635502. DOI: 10.1016/s0889-857x(02)00078-9
- 13 Henninger B, Glodny B, Rudisch A, Trieb T, Loizides A, Putzer D, Judmaier W and Schocke MF: Ewing sarcoma versus osteomyelitis: differential diagnosis with magnetic resonance imaging. *Skeletal Radiol* 42(8): 1097-1104, 2013. PMID: 23685708. DOI: 10.1007/s00256-013-1632-5
- 14 Bacci G, Di Fiore M, Rimondini S and Baldini N: Delayed diagnosis and tumor stage in Ewing's sarcoma. *Oncol Rep* 6(2): 465-466, 1999. PMID: 10023023.
- 15 Bellan DG, Filho RJ, Garcia JG, de Toledo Petrilli M, Maia Viola DC, Schoedl MF and Petrilli AS: Ewing's sarcoma: Epidemiology and prognosis for patients treated at the pediatric oncology institute, IOP-GRAACC-UNIFESP. *Rev Bras Ortop* 47(4): 446-450, 2015. PMID: 27047848. DOI: 10.1016/S2255-4971(15)30126-9
- 16 Esiashvili N, Goodman M and Marcus RB Jr: Changes in incidence and survival of Ewing sarcoma patients over the past 3 decades: Surveillance Epidemiology and End Results data. *J Pediatr Hematol Oncol* 30(6): 425-430, 2008. PMID: 18525458. DOI: 10.1097/MPH.0b013e31816e22f3
- 17 Vogrin M, Trojner T and Kelc R: Artificial intelligence in musculoskeletal oncological radiology. *Radiol Oncol* 55(1): 1-6, 2020. PMID: 33885240. DOI: 10.2478/raon-2020-0068
- 18 Moradi M, Madani A, Karargyris A and Syeda-mahmood T: Chest x-ray generation and data augmentation for cardiovascular abnormality classification. *Medical Imaging 2018: Image Processing*, 2018. DOI: 10.1117/12.2293971
- 19 Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, Xiong H and He Q: A comprehensive survey on transfer learning. *Proceedings of the IEEE* 109(1): 43-76, 2022. DOI: 10.1109/JPROC.2020.3004555
- 20 Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF and Collins GS: Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 162(1): W1-73, 2015. PMID: 25560730. DOI: 10.7326/M14-0698
- 21 von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP and STROBE Initiative: The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies. *Int J Surg* 12(12): 1495-1499, 2014. PMID: 25046131. DOI: 10.1016/j.ijsu.2014.07.013
- 22 He K, Zhang X, Ren S and Sun J: Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019. DOI: 10.1109/CVPR.2016.90
- 23 Eigen D and Fergus R: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. 2015 IEEE International Conference on Computer Vision (ICCV), 2017. DOI: 10.1109/ICCV.2015.304
- 24 Selvaraju R, Cogswell M, Das A, Vedantam R, Parikh D and Batra D: Grad-CAM: Visual explanations from deep networks via gradient-based localization. 2017 IEEE International Conference on Computer Vision (ICCV), 2022. DOI: 10.1109/ICCV.2017.74
- 25 von Schacky CE, Wilhelm NJ, Schäfer VS, Leonhardt Y, Gassert FG, Foreman SC, Gassert FT, Jung M, Jungmann PM, Russe MF, Mogler C, Knebel C, von Eisenhart-Rothe R, Makowski MR, Woertler K, Burgkart R and Gersing AS: Multitask deep learning for segmentation and classification of primary bone tumors on radiographs. *Radiology* 301(2): 398-406, 2021. PMID: 34491126. DOI: 10.1148/radiol.2021204531
- 26 Clark MA and Thomas JM: Delay in referral to a specialist soft-tissue sarcoma unit. *Eur J Surg Oncol* 31(4): 443-448, 2005. PMID: 15837054. DOI: 10.1016/j.ejso.2004.11.016
- 27 Biswas B, Rastogi S, Khan SA, Mohanti BK, Sharma DN, Sharma MC, Mridha AR and Bakhshi S: Outcomes and prognostic factors for Ewing-family tumors of the extremities. *J Bone Joint Surg Am* 96(10): 841-849, 2014. PMID: 24875025. DOI: 10.2106/JBJS.M.00411

Received April 28, 2022

Revised May 19, 2022

Accepted July 1, 2022



Magdalena Bloier*, Florian Hinterwimmer, Sebastian Breden, Sarah Consalvo, Jan Neumann, Nikolas Wilhelm, Rüdiger von Eisenhart-Rothe, Daniel Rueckert, and Rainer Burgkart

Detection and Segmentation of Heterogeneous Bone Tumours in Limited Radiographs

<https://doi.org/10.1515/cdbme-2022-1019>

Abstract: Bone tumours are a rare and often highly malignant entity. Early clinical diagnosis is the most important step, but the difficulty of detecting and assessing bone malignancies is in its radiological peculiarity and limited experience of non-experts. Since X-ray imaging is the first imaging method of bone tumour diagnostics, the purpose of this study is to develop an artificial intelligence (AI) model to detect and segment the tumorous tissue in a radiograph. We investigated which methods are necessary to cope with limited and heterogeneous data. We collected 531 anonymised radiographs from our musculoskeletal tumour centre. In order to adapt to the complexity of recognizing the malignant tissue and cope with limited data, transfer learning, data augmentation as well as several architectures, some of which were initially designed for medical images, were implemented. Furthermore, dataset size was varied by adding another bone tumour entity. We applied a data split of 72%, 18%, 10% for training, validation and testing, respectively. To provide statistical significance and robustness, we applied a cross-validation and image stratification with respect to tumour pixels present. We achieved an accuracy of 99.72% and an intersection over union of 87.43% for hold-out test data by applying several methods to tackle limited data. Transfer learning and additional data brought the greatest performance increase. In conclusion, our model

was able to detect and segment tumorous tissue in radiographs with good performance, although it was trained on a very limited amount of data. Transfer Learning and data augmentation proved to significantly mitigate the issue of limited data samples. However, to accomplish clinical significance, more data has to be acquired in the future. Through minor adjustments, the model could be adapted to other musculoskeletal tumour entities and become a general support tool for orthopaedic surgeons and radiologists.

Keywords: deep learning, sarcoma, bone tumour, detection, segmentation

1 Introduction

Bone tumours are a rare disease overall [8], but are among the most common cancers in children and adolescents (>10% of all paediatric cancers). The complex and time-consuming diagnosis in a specialised centre includes clinical, radiological and histopathological steps as well as the subsequent interdisciplinary assessment in a specialised tumour board. A general practitioner, on the other hand, usually has only X-ray diagnostics and, because of the incidence, statistically encounters bone tumours less than three times in his/her professional life. As a result, sarcomas are often misdiagnosed [8] and prognostically essential time is lost, and patients are delayed in being referred to specialised sarcoma centres. Hence, new and more sophisticated techniques for early and reliable detection and evaluation of bone tumours are urgently needed. Deep learning (DL) is poised to reshape medicine and potentially improve the experience of physicians as well as patients [9]. DL has already had ample success in many medical disciplines [12]. In comparison, the impact and number of publications of DL in orthopaedics are very limited [4, 11]. Most certainly this can be explained by the low incidence of bone tumours (and soft tissue tumours as well) and the lack of sufficient data infrastructures. While limited data is a common obstacle to DL applications in medicine, even more so for image interpretation of bone malignancies. Thus, we developed a segmentation framework for heterogeneous and limited radiographs of

*Corresponding author: **Magdalena Bloier**, Department for Orthopaedics and Sports Orthopaedics, Klinikum rechts der Isar, Technical University of Munich, 81675 Munich, Germany, e-mail: magdalena.bloier@tum.de

Florian Hinterwimmer, Institute for AI and Informatics in Medicine, Technical University of Munich & Department for Orthopaedics and Sports Orthopaedics, Klinikum rechts der Isar, Technical University of Munich, 81675 Munich, Germany
Sebastian Breden, Sarah Consalvo, Nikolas Wilhelm, Rüdiger von Eisenhart-Rothe, Rainer Burgkart, Department for Orthopaedics and Sports Orthopaedics, Klinikum rechts der Isar, Technical University of Munich, 81675 Munich, Germany
Jan Neumann, Institute for Diagnostic and Interventional Radiology and Paediatric Radiology, Klinikum rechts der Isar, Technical University of Munich 81675 Munich, Germany
Daniel Rueckert, Institute for AI and Informatics in Medicine, Technical University of Munich, 81675 Munich, Germany

bone tumours, focusing on Transfer Learning [2], data augmentation [1, 3, 7] and leveraging various segmentation models [5, 10, 14] and configurations. In summary, we make the following contributions:

1. We demonstrate a novel approach for bone tumour assessment by detecting and segmenting malignancies with DL and limited and heterogeneous radiographs.
2. We illustrate the impact of transfer learning, data augmentation, different architectures and dataset size.
3. We provide a potential support tool to identify bone tumours, not only for expert centres, but also potentially targeting outpatients clinics and young physicians.

1.1 Related work

With the rise of DL, especially the task of segmentation [6] of medical images became more and more popular over the past decade [9]. Nonetheless, segmentation of bone tumours has only been presented a few times. Zhang et al. [16] proposed a multiple supervised residual network to segment osteosarcomas in CT images with good results (Dice score 0.89). In contrast, Schacky et al. [15] demonstrated a multi-task DL approach to classify, detect and segment bone tumours in radiographs with fair segmentation performance (Dice score 0.6). CT imaging is usually the imaging modality of choice for bone pathologies, because it provides more detailed information about the potential destruction of cortical bone. However, similar to Schacky et al., this study focused on detecting and segmenting bone tumours in standardised radiographs and identifying the most impactful methods for an imperfect dataset.

2 Materials and Methods

2.1 Dataset

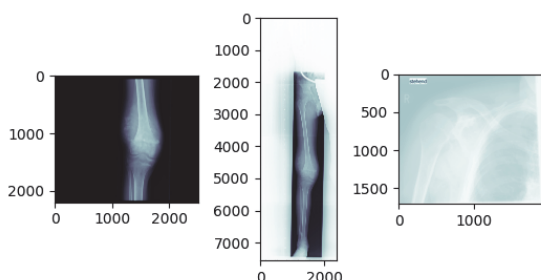


Fig. 1: Sample images before preprocessing.

We collected 531 preoperative radiographs from our musculoskeletal tumour centre from paediatric sarcoma patients. The dataset includes two subsets. The first subset with 44% of the entire dataset contains osteosarcoma and the second subset with 56% of the entire dataset contains chondrosarcoma. Typically, sarcoma occur in joints and long tubular bones and no restrictions regarding anatomical regions were applied. Also, the images available are heterogeneous in character, as can be seen in figure 1. They vary in dimension, resolution and data quality containing black or white background and marks. External images were also included in the dataset. No meta-information for statistical analysis was available and no further restrictions regarding age, musculoskeletal features or sex were made. Additionally, masks of the X-ray images including the location of the tumour were created. The masks are a binarised representation of tumour tissue and healthy tissue which were manually segmented by orthopaedic surgeons.

2.2 Model training

Model training and inference was conducted on a DGX Station A100 with four 80GB graphical processing units (Nvidia Corporation, Santa Clara, CA), 64 2.25 GHz cores and 512 GB DDR4 system memory running on a Linux/Ubuntu 20.04 distribution (Canonical, London, England). Preprocessing and model implementation were performed in Python 3.8.5 (<https://www.python.org/>) using PyTorch 1.10.2 and cuda toolkit 11.3 (<https://pytorch.org/>). The pretrained ConvNet model of this study will be provided upon publication.

2.3 Algorithm and experimental setup



Fig. 2: Illustration of workflow.

We developed a DL framework to train a segmentation network and evaluate it through the metrics accuracy and intersection over union (IoU). The dataset was preprocessed by removing the background areas, padding to create a square image, scaling to 256x256 and normalizing. The data split for pretraining was 72%, 18%, 10% for training, validation and hold-out testing, respectively. To provide statistical significance and robustness, images from all subsets were equally distributed with respect to tumour pixels present. An additional 5-fold cross-validation supported the task.

The initial dataset contained only the images with osteosar-

coma. As a baseline, we used U-Net-architecture as common performance baseline with ResNet34 to train our neural network and successively adapted several extensions to boost the performance of the model as shown in figure 2. In the first setup, instead of training from scratch, a pretrained model with Imagenet weights was implemented. Afterwards, the impact of data augmentation techniques was investigated in setup 2. Therefore, the current best model of setup 1 was extended by 12 different data augmentations including geometric transformations, cropping, filtering and intensity operations. Then, different data augmentation methods were combined by successively extending the amount of data augmentations using the operations with the highest IoU of the 12 data augmentation methods first. Among other things, so-called unrealistic data augmentations were used, which describe in particular operations that alter the images to such an extent that they no longer correspond to a medically realistic X-ray image as shown in figure 3 [3]. After determining the data augmentations leading to the highest IoU, we varied the model architecture. Model architectures selected were UNet++, DeepLabv3, DeepLabv3+ and MA-Net. In the last setup the data were duplicated by using the chondrosarcoma dataset to determine the influence of the amount of data.

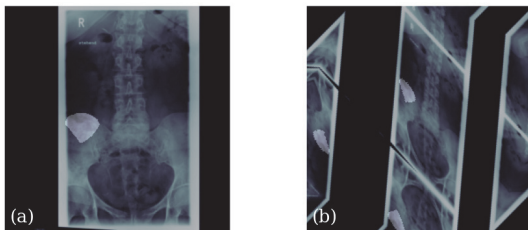


Fig. 3: Preprocessed sample image (a) and sample image with excessive data augmentation (b).

3 Results

Tab. 1: Overall results of the test dataset.

Setup number	Description	Accuracy	IoU
Baseline	U-Net with ResNet34	99.51 %	73.78 %
Setup 1	Transfer learning	99.59 %	79.58 %
Setup 2	Data augmentation	99.62 %	82.70 %
Setup 3	Model architecture	99.64 %	83.39 %
Setup 4	Amount of data	99.72 %	87.43 %

Transfer learning improved the baseline results to an IoU of 5.80% in setup 1 as shown in table 1. The best combination of additional data augmentations from configurations selected was the combined application of several affine transformations as shown in figure 1. In setup 3, we choose UNet++ architecture as model selected with the highest IoU compared to the model architectures U-Net, MA-Net, DeepLabv3 and DeepLabv3+. While MA-Net also had a relatively high IoU and accuracy of 83.19% and 99.62%, DeepLabv3 only reached an IoU 10.78% lower than UNet++. Through adding the chondrosarcoma images, we reached a final accuracy of 99.72% and an IoU of 87.43%. A prediction of a sample image of the final framework is shown in figure 4.

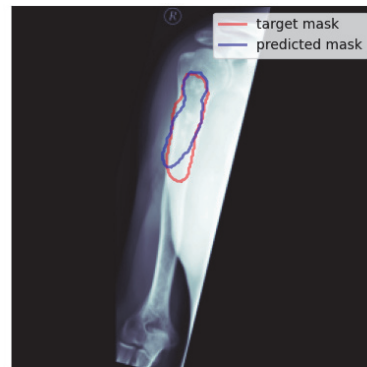


Fig. 4: Sample image with predicted and target mask.

4 Discussion

The main finding of this study was that significant results in detecting and segmenting heterogeneous bone tumour appearances in limited radiographs can be reached by implementing several methods to fine-tune the algorithm and tackle the issue of small datasets.

An improvement in accuracy and IoU could be achieved in each setup through an extending framework. For this task, using a pretrained neural network and adding more data leads to the biggest improvement of the segmentation task. It should be emphasized that the added data is that of another entity. Nevertheless, the datasets contain similar image features and therefore lead to improved performance. Additionally, data augmentation methods have shown to support image segmentation tasks. Further, while in literature realism is a goal for

many authors [1, 3], in our approach unrealistic data augmentations through for example affine transformations lead to better segmentation results than the realistic ones. However, the use of more data augmentations is not guaranteed to be beneficial as they might lead to poorer results. Choosing an appropriate model is crucial to get good segmentation results. In our task, UNet++ [14] and MA-Net [5, 13] lead to the best performance.

The major limitation of this study is the low amount of data available, hence, robustness of the model has to be further validated. In addition, the interpretation of the results must take into account that there is no gold standard for segmentation labels. Therefore, the segmentation labels need to be evaluated by an interdisciplinary team.

4.1 Conclusion

Transfer learning and an increased quantity of data even from another entity lead to the largest improvement of segmentation results, while varying the model architecture leads to the biggest differences in IoU and accuracy. In addition, unrealistic data augmentation through affine transformations supported the task. To achieve clinically relevant results, a systematic and structured collection of data to increase dataset size is of paramount importance.

Author Statement

Research funding: The author state no funding involved. **Conflict of interest:** Authors state no conflict of interest. **Informed consent:** Informed consent has been obtained from all individuals included in this study. **Ethical approval:** The research related to human use complies with all the relevant national regulations, institutional policies and was performed in accordance with the tenets of the Helsinki Declaration, and has been approved by the authors' institutional review board or equivalent committee.

Equal Contribution

Magdalena Bloier and Florian Hinterwimmer equally contributed to this article (shared first).

References

- [1] Abdollahi, B., N. Tomita and S. Hassanpour (2020). Data augmentation in training deep learning models for medical image analysis. *Deep learners and deep learner descriptors for medical applications*, Springer: 167-180.
- [2] Al-Rakhami, M. S., M. M. Islam, M. Z. Islam, A. Asraf, A. H. Sodhro and W. Ding (2021). "Diagnosis of COVID-19 from X-rays using combined CNN-RNN architecture with transfer learning." *MedRxiv: 2020.2008. 2024.20181339*.
- [3] Chlap, P., H. Min, N. Vandenberg, J. Dowling, L. Holloway and A. Haworth (2021). A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology* 65(5): 545-563.
- [4] Do, N.-T., S.-T. Jung, H.-J. Yang and S.-H. Kim (2021). Multi-level seg-unet model with global and patch-based X-ray images for knee bone tumor detection. *Diagnostics* 11(4): 691.
- [5] Fan, T., Wang, G., Li, Y. and Wang, H. (2020). MA-Net: A Multi-Scale Attention Network for Liver and Tumor Segmentation. *IEEE Access*, 8, 179656–179665.
- [6] Hesamian, M. H., Jia, W., He, X. and Kennedy, P. (2019). Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges. *Journal of Digital Imaging*, 32(4), 582–596.
- [7] Hussain, Z., F. Gimenez, D. Yi and D. Rubin (2017). Differential data augmentation techniques for medical imaging classification tasks. *AMIA annual symposium proceedings*, American Medical Informatics Association.
- [8] Picci, P., M. Manfrini, D. M. Donati, M. Gambarotti, A. Righi, D. Vanel et al. (2020). *Diagnosis of Musculoskeletal Tumors and Tumor-like Conditions: Clinical, Radiological and Histological Correlations-the Rizzoli Case Archive*, Springer.
- [9] Rajpurkar, P., E. Chen, O. Banerjee and E. J. Topol (2022). AI in health and medicine. *Nat Med* 28(1): 31-38.
- [10] Ronneberger, O., P. Fischer and T. Brox (2015). U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention*, Springer.
- [11] Hinterwimmer, F., Lasic, I., Suren, C., Hirschmann, M., Pohlig, F., Rueckert, D., et al. (2022). "Machine learning in knee arthroplasty: specific data are key—a systematic review." *Knee Surgery, Sports Traumatology, Arthroscopy*: 1-13.
- [12] Singh, P., N. Singh, K. K. Singh and A. Singh (2021). Diagnosing of disease using machine learning. *Machine Learning and the Internet of Medical Things in Healthcare*, Elsevier: 89-111.
- [13] Shang, R., J. Zhang, L. Jiao, Y. Li, N. Marturi and R. Stolkin (2020). Multi-scale adaptive feature fusion network for semantic segmentation in remote sensing images. *Remote Sensing* 12(5): 872.
- [14] Zhou, Z., M. M. Rahman Siddiquee, N. Tajbakhsh and J. Liang (2018). Unet++: A nested u-net architecture for medical image segmentation. *Deep learning in medical image analysis and multimodal learning for clinical decision support*, Springer: 3-11.
- [15] von Schacky, C. E., N. J. Wilhelm, V. S. Schäfer, Y. Leonhardt, F. G. Gassert, S. C. Foreman, et al. (2021). Multitask deep learning for segmentation and classification of primary bone tumors on radiographs. *Radiology* 301(2): 398-406.
- [16] Zhang, R., L. Huang, W. Xia, B. Zhang, B. Qiu and X. Gao (2018). Multiple supervised residual network for osteosarcoma segmentation in CT images. *Computerized Medical Imaging and Graphics* 63: 1-8.

1 **Title**

2 Impact of Metadata in Multimodal Classification of Bone
3 Tumours

4

5 **Authors:**

6 ^{1,2} Florian Hinterwimmer, M.Sc. & ¹ Michael Guenther, M.Sc. ([shared first](#))

7 ¹ Sarah Consalvo, MD

8 ³ Jan Neumann, MD

9 ³ Alexandra Gersing, MD

10 ³ Klaus Woertler, Prof. MD

11 ¹ Ruediger von Eisenhart-Rothe, Prof. MD, Dipl. Kfm.

12 ¹ Rainer Burgkart, Prof. MD & ^{2,4} Daniel Rueckert, Prof. PhD ([shared last](#))

13

14 **Affiliations:**

15 ¹ Department of Orthopaedics and Sports Orthopaedics, Klinikum rechts der Isar, Technical
16 University of Munich, Munich, Germany

17 ² Institute for AI and Informatics in Medicine, Technical University of Munich, Munich,
18 Germany

19 ³ Musculoskeletal Radiology Section, Klinikum rechts der Isar, Technical University of
20 Munich, Munich, Germany

21 ⁴ Department of Computing, Imperial College London, London, UK

22

23

24 **Corresponding Author:**

25 Florian Hinterwimmer

26 Trogerstraße 4

27 81675 Munich, Germany

28 florian.hinterwimmer@tum.de

29

30 **Funding:**

31 None

32

33 **Manuscript type:**

34 Original research

35

36 **Word Count:**

37 Abstract: 247 (+42)

38 Manuscript: 2977

39

40

41 **Abstract**

42 Objective: To propose a multimodal deep learning model that integrates clinical metadata and
43 X-ray imaging to enhance the classification of primary bone tumours, while providing
44 explainability through Shapley additive explanations.

45 Methods: For this retrospective single centre study a dataset of 1,785 radiographs of 804
46 patients from 2000 to 2020 was used. Additionally, the respective metadata was collected (age
47 = patient's age, site = affected bone, position = position of the tumour at the bone, gender =
48 gender of the patient). The dataset encompassed ten selected tumour types and employed
49 histopathology or tumour board decision as the reference standard. The proposed approach was
50 based on the NesT image classification model and a multilayer perceptron with a joint fusion
51 architecture. The study followed STROBE guidelines to present descriptive data, reporting
52 discrete parameters using incidence and percentage ratio and continuous parameters using
53 mean, standard deviation, median, and interquartile range, while the reporting and validation
54 of the prediction model were based on TRIPOD statement.

55 Results: The mean age was 33.62 ± 18.60 [SD] and 54.73% of patients were male. The
56 multimodal deep learning model outperformed a state-of-the-art model (Vision Transformer)
57 and similar studies in classifying primary bone tumours with 69.7% accuracy. SHAP values
58 elucidated that age exerted the most substantial influence among the considered metadata.

59 Conclusion: We developed a joint fusion approach outperforming state-of-the-art models and
60 comparable studies by incorporating clinical metadata and imaging data into one model.
61 Furthermore, we provide a magnitude of the impact of metadata through SHAP values.

62 Clinical relevance statement: The developed algorithm for bone tumour classification,
63 combining imaging data and clinical metadata, improves accuracy and aids timely referrals. It
64 offers comprehensive assessment, assisting e.g. non-tumour experts in diagnosing bone
65 tumours more effectively, with potential for precision medicine applications.

66

67 **Keywords:**

68 radiography, bone neoplasm, classification, deep learning, metadata,

69

70 **Key points:**

- 71 • The study introduces a model based on transformers and multilayer perceptrons (MLPs)
72 for classifying ten primary bone tumour entities.
- 73 • We found that including clinical metadata along with radiography significantly
74 improved the classification accuracy of the bone tumour entities. This highlights the
75 importance of incorporating patient-specific information alongside imaging data for
76 more accurate classification.
- 77 • While the results of the study showed promise in terms of increased accuracy, it is
78 important to note that the model's results are not yet suitable for direct clinical
79 application. Further refinement and evaluation are necessary before considering its use
80 in a clinical setting. This highlights the ongoing nature of research and the need for
81 extensive testing before deploying AI models in healthcare.

82

83 **Abbreviations:**

84 AI – artificial intelligence

85 DL – deep learning

86 NesT - Nested Hierarchical Transformer

87 MLP – multilayer perceptron

88 SHAP – Shapley additive explanations

89 ViT – Vision transformer

90 NLP – natural language processing

91

92 **Summary statement:**

93 We developed a novel joint fusion model that integrates imaging data and clinical metadata for
94 bone tumour classification, improving performance and bringing us closer to gold standards of
95 clinical diagnostic workflows. Further improvements and data collection can enhance the
96 model's precision and make it a valuable tool for a range of medical professionals in bone
97 tumour assessment.

98

100 Bone tumors encompass rare lesions comprising various tumour entities [1-3], with the vast
101 majority being benign [4]. Malignant primary bone tumours, though accounting for a mere
102 0.2% of adult malignancies [4; 5], rank as the sixth most common cancer in children and the
103 third most common in adolescents [1; 6]. Diagnosis in the early stages is challenging due to
104 the absence of specific symptoms, resulting in significant treatment delays [3]. Timely referral
105 to specialized tumour centres is crucial for comprehensive evaluation and differentiation
106 between benign, intermediate, and malignant tumours [1]. Unfortunately, non-oncology-
107 trained medical professionals encounter only a few malignant primary bone tumours
108 throughout their careers [7; 8], leading to potential delays of over a year and a lack of
109 experience in identifying these complex tumour entities unequivocally.

110 In 2018, the Musculoskeletal Tumor Society and American Academy of Orthopedic Surgeons
111 Working Group recommended plain radiographs as the initial screening for bone tumours [5],
112 including for children [9]. Even if only a radiograph is available, patients with suspected
113 malignant lesions should be referred to musculoskeletal tumour centres to prevent treatment
114 delays. The need for further imaging studies should be assessed at referral centres [5]. The final
115 diagnosis relies on synthesizing clinical presentation, imaging features, and histopathologic
116 findings if a specific radiologic diagnosis of a benign entity proves inconclusive [10].

117 The field of diagnostic imaging is rapidly advancing, with technology, innovation, and market
118 expansion leading to increased production of imaging and clinical data [2; 11]. Precision
119 medicine plays an increasingly important role in musculoskeletal radiology and orthopaedic
120 oncology, necessitating advanced analysis tools [12; 13]. While artificial intelligence (AI),
121 specifically Deep Learning (DL), is widely employed in lung, breast, and CNS cancer research
122 [14], its application in musculoskeletal tumour research remains limited [2]. Nevertheless,
123 these advanced data analysis techniques hold the potential to revolutionize the medical field,
124 benefiting both physicians and patients [15].

125 In this study, we propose a DL model that emulates expert radiologists by incorporating clinical
126 metadata alongside imaging data for diagnostic assessment and dataset enrichment. By doing
127 so, our research question aligns closer with clinical reality: Can a state-of-the-art DL model
128 accurately classify ten different bone tumour entities, leveraging the inclusion of clinical
129 metadata from patients and providing insights into the significance of each clinical parameter?

130

131 **Methods**

132 This retrospective study (N°48/20S) was approved by the local institutional review and ethics
133 board, following national and international guidelines. Informed consent was waived for this
134 retrospective and anonymized study.

135

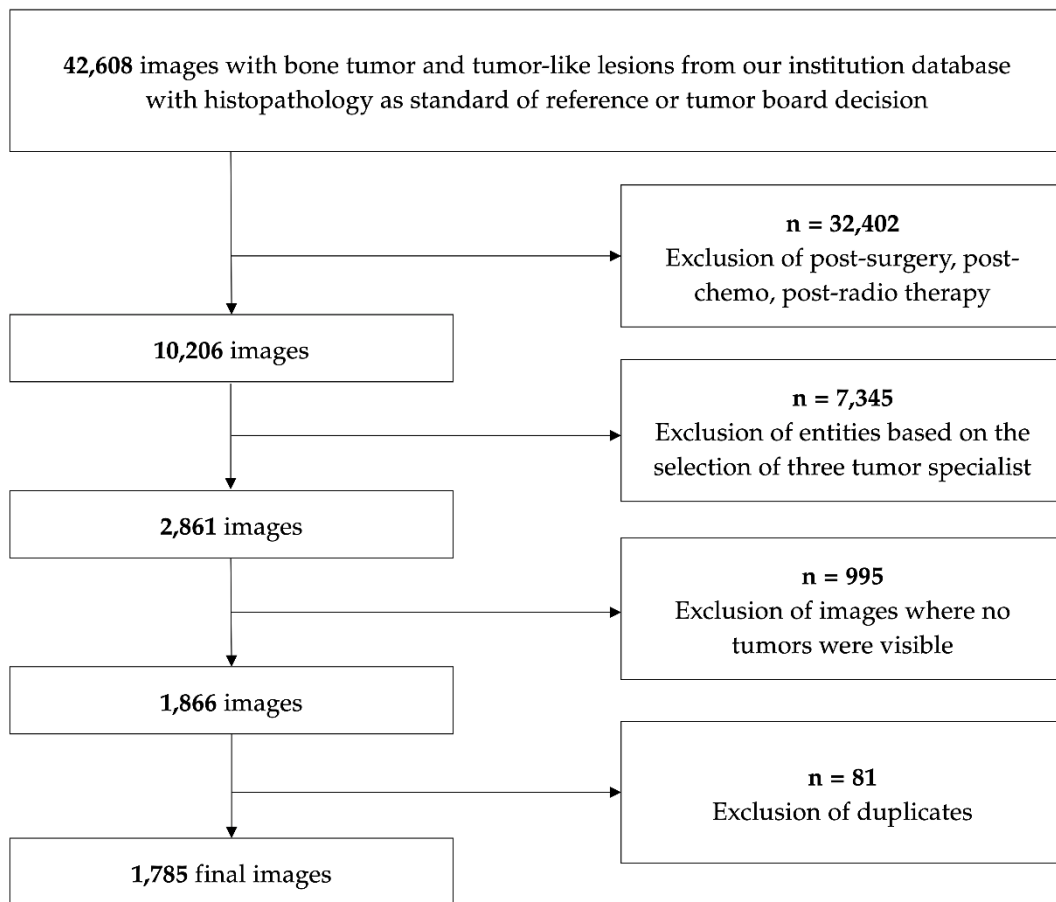
136 **Eligibility criteria**

137 In this single centre study, we screened our musculoskeletal tumour centre's database from
138 2000 to 2020 for patients treated for primary bone neoplasms based on ICD codes. The selected
139 tumours were the most frequent in our database: Aneurysmal bone cyst (ABC),
140 chondroblastoma, chondrosarcoma, enchondroma, Ewing sarcoma, fibrous dysplasia, giant
141 cell tumour, non-ossifying fibroma (NOF), osteochondroma, and osteosarcoma. Malignant
142 lesions were verified by histopathology, while benign and intermediate lesions were confirmed
143 by histopathology or discussed in the local tumour board and classified based on radiological
144 features [16]. Clinical and imaging data were retrieved from our HIS and PACS, respectively.
145 The data curation and validation were performed by an orthopaedic resident, (SC) a senior
146 musculoskeletal radiologist (JN), and a data scientist (MG).

147

148 **Patients**

149 We had access to a total of 42,608 radiographic images of bone tumours, tumour-like lesions,
150 and their differential diagnoses, including e.g. osteomyelitis. However, approximately three-
151 fourths of the images were excluded as they were acquired post-surgery, after systemic therapy,
152 or post-radiotherapy. The dataset ultimately included 1,785 images representing ten entities of
153 benign, intermediate, and malignant bone tumours, after excluding 7,345 images that did not
154 meet the criteria (e.g., exostosis, tumour-like lesions). Additionally, 995 images were discarded
155 due to the absence of visible tumours (e.g., wrong angle, artefacts). These images were
156 accompanied by patient metadata from 922 patients. Figure 1 displays the according flow
157 diagram.



158 Figure 1 - Flow diagram showing the application of eligibility criteria to create a final dataset.

159

160 **Statistical analysis**

161 Descriptive data follows STROBE guidelines [17], presenting discrete parameters as incidence
 162 and percentage ratio, and continuous parameters as mean, standard deviation, median, and
 163 interquartile range. The reporting and validation of the prediction model adhere to TRIPOD
 164 guidelines [18]. Statistical analysis was conducted by two data scientists (MG, FH).

165

166 **Image processing**

167 To accommodate the large size difference between radiographs and the standard DL model
 168 input size of 224 x 224 pixels [19; 20], we performed a ROI crop to remove non-relevant
 169 information. Segmentation masks, created by medical experts (SC, JN), were used for this
 170 dataset. An automated segmentation model by Bloier et al. [21] achieved a 99.72% accuracy

171 in predicting these masks. The images were then cropped using a square bounding box around
172 the segmentation mask, with a 15% padding for uncertainty. In the final preprocessing step,
173 the cropped images were converted to standard grey scale and rescaled to 224 x 224 pixels.

174

175 **Model development**

176 Our approach combines imaging data and clinical metadata in a single deep learning model.
177 Image features are extracted using a state-of-the-art classification network. Clinical data is
178 processed through fully connected layers with ReLU activation functions and batch
179 normalization, and the resulting features are concatenated. The final softmax layer provides a
180 probability prediction for the entity. During training, the loss is propagated through the entire
181 model, including the image and metadata networks. According to Huang et al. [22] our model
182 configuration is classified as a joint fusion model that is independent of any specific image
183 classification network, chosen because it weighs image and metadata equally.

184 To compute baseline results, we implemented a XGboost [23] model for classification only
185 with metadata. Further, to compute baseline results for solely imaging data, we compared a
186 ResNet [19] model with the NestT [24] model. ResNet utilizes residual connections to build
187 deeper and more powerful CNNs. NestT, on the other hand, is based on the ViT architecture
188 from the NLP domain, which employs self-attention instead of convolutions [25]. ViT
189 generally outperforms CNNs when datasets are large [26], while NestT performs well even
190 with small datasets [24]. The model was trained using the Adam optimizer with specific
191 parameters according to Kingma et al.'s paper [27]. The weights of the model were pretrained
192 with the ImageNet dataset, and early stopping was implemented with a patience of 20 to prevent
193 unnecessary training.

194

195 The model training and inference were conducted on a DGX Station A100 equipped with four
196 graphical processing units, 64 cores, and 512 GB system memory. The system ran on a
197 Linux/Ubuntu 20.04 distribution. Pre-processing and model implementation were performed
198 using Python 3.9.12, PyTorch Lightning 1.7.1, PyTorch 1.12.1, and the CUDA toolkit 11.3.1.
199 The trained classification model will be made available on GitHub upon publication.

200

201 **Outcome and model evaluation**

202 For hyperparameter optimization, we compared the mean validation accuracy using 5-fold
203 cross-validation. The standard deviation was assessed to evaluate model robustness. In our
204 multi-class classification setting, we used macro averaging, which calculates metrics for each
205 class separately and then averages them with equal weight. For evaluating individual class
206 performance, we calculated a confusion matrix. To compare different model architectures and
207 approaches, we assessed the mean test accuracy with 5-fold cross-validation. Additionally, we
208 analysed the ensemble test accuracy, which combines all training data from the five cross-
209 validation folds.

210

211 **Model interpretation**

212 Understanding the reasoning behind a model's prediction is often crucial, but complex models
213 can be difficult to interpret, posing challenges for humans. Deep learning models, in particular,
214 are considered black boxes due to their nested nonlinear structure [28]. To address this issue,
215 we implemented SHAP (SHapley Additive exPlanations) introduced by Lundberg et al. [29].
216 This method calculates the impact of features on individual model predictions. To assess
217 overall performance, we computed average SHAP values across all test samples. As the
218 DeepLiftShap approach does not support Vision Transformers (ViTs), we used the
219 GradientShap algorithm to calculate SHAP values. For instance, the binary-encoded metadata
220 *gender* is represented by two input features, whose sum yields the SHAP value. Similarly,
221 SHAP values for *tumour site* and *position at bone* were computed using the same procedure.

222

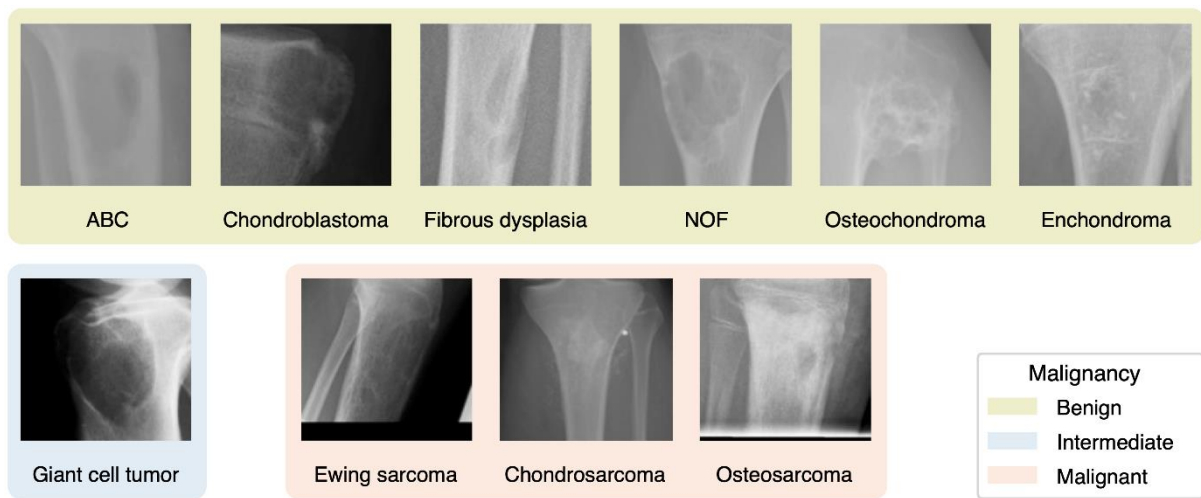
223 **Results**

224 **Dataset**

225 The patients' mean age was 33.62 ± 18.60 . For more information, refer to Table 1. Among the
226 entities in the dataset, osteochondroma was the most common (28.48%), while Ewing sarcoma

227 was the least frequent (0.37%). This indicates an imbalanced dataset with significant variations
 228 in sample distribution. The gender distribution was relatively balanced, with females
 229 accounting for 45.27% and males for 54.73% of cases, slightly favouring males. The femur
 230 was the most common tumour site (36.82%), whereas the os sacrum had only one occurrence
 231 (0.12%) in the entire dataset. Additional discrete patient characteristics can be found in Table
 232 2. Figure 2 illustrates one example image per entity after pre-processing.

233



234

235

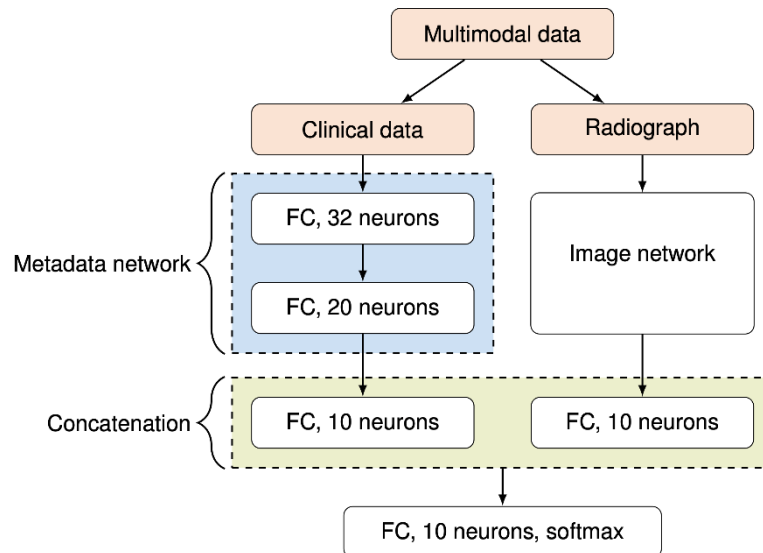
Figure 2 - Examples of radiographs of each entity: one example for each entity after pre-processing (cropping, rescaling).

236

237 Model performance

238 Table 3 presents the different tested models, including three “baseline” models (only meta- or
 239 only imaging data) and three models based on our proposed architecture. The XGBoost [23]
 240 model trained solely with metadata achieved an accuracy of 0.185. For image-only models, we
 241 compared ResNet18 [19] and NesT [24], with NesT achieving a higher accuracy of $0.628 \pm$
 242 0.019 compared to ResNet18's 0.541 ± 0.029 . The NesT architecture with a small model size,
 243 a learning rate of $\alpha = 0.0005$, and a batch size of 16 performed best. Concatenating the features
 244 from the NesT model and a separate MLP for metadata, we constructed our multimodal
 245 approach. The architecture is depicted in Figure 3.

246



247 Figure 3 - Proposed multimodal model: metadata (clinical data) being processed through a multilayer perceptron, imaging data through
 248 common image classification networks and concatenation (fusion) of both in feature space before a final fully connected and softmax layer.
 249

250 With the multimodal model, we achieved an accuracy of 0.641. When optimizing the
 251 hyperparameters, we found that the encoding of the metadata had an impact on performance.
 252 For our final model, the features were coded as follows:

- 253 • gender: *binary* [male, female]
- 254 • age: *ordinal* [1,2,3 etc.]
- 255 • site: *one-hot* [clavicle, femur, etc.]
- 256 • position: *one-hot* [epiphysis, epi-metaphysis, metaphysis, meta-diaphysis, diaphysis]

257 All shown results were obtained after hyperparameter tuning.

258

259 **Improvement through ensemble model**

260 The average accuracy results from cross-validation were presented. Combining all folds into
 261 an ensemble model improved the accuracy by summing up the logits of the five models and
 262 applying a softmax layer. The multimodal model showed the highest improvement, achieving
 263 a final accuracy of 0.697.

264

265 **Comparison of the performance of each entity**

266 Figure 4 shows a confusion matrix with the individual accuracy of each entity. Each test sample
 267 of Chondroblastomas and Ewing sarcomas were classified incorrectly. The best performance
 268 was observed for osteochondromas with an accuracy of 88% and Enchondromas with 75%.

Ground truth	Chondroblastoma	0 0%	3 50%	0 0%	0 0%	0 0%	1 17%	0 0%	1 17%	1 16%	0 0%
	Chondrosarcoma	0 0%	36 63%	0 0%	12 21%	0 0%	2 4%	0 0%	4 7%	3 5%	0 0%
	Fibrous dysplasia	0 0%	3 19%	6 38%	1 6%	0 0%	4 25%	2 12%	0 0%	0 0%	0 0%
	Enchondroma	0 0%	4 5%	0 0%	55 75%	0 0%	2 3%	0 0%	7 10%	3 4%	2 3%
	Ewing's sarcoma	0 0%	0 0%	0 0%	0 0%	0 0%	1 100%	0 0%	0 0%	0 0%	0 0%
	Aneurysmal bone cyst	0 0%	0 0%	0 0%	0 0%	0 0%	9 45%	3 15%	2 10%	1 5%	5 25%
	Non-ossifying fibroma	0 0%	1 8%	0 0%	0 0%	0 0%	3 23%	8 61%	0 0%	1 8%	0 0%
	Osteochondroma	1 1%	5 5%	2 2%	1 1%	0 0%	0 0%	0 0%	88 88%	3 3%	0 0%
	Osteosarcoma	0 0%	3 7%	2 5%	6 14%	0 0%	0 0%	1 2%	3 7%	28 63%	1 2%
	Giant cell tumor	0 0%	3 11%	0 0%	2 7%	0 0%	1 4%	0 0%	1 4%	1 4%	19 70%
			Chondroblastoma	Chondrosarcoma	Fibrous dysplasia	Enchondroma	Ewing's sarcoma	Aneurysmal bone cyst	Non-ossifying fibroma	Osteochondroma	Osteosarcoma
		Prediction									

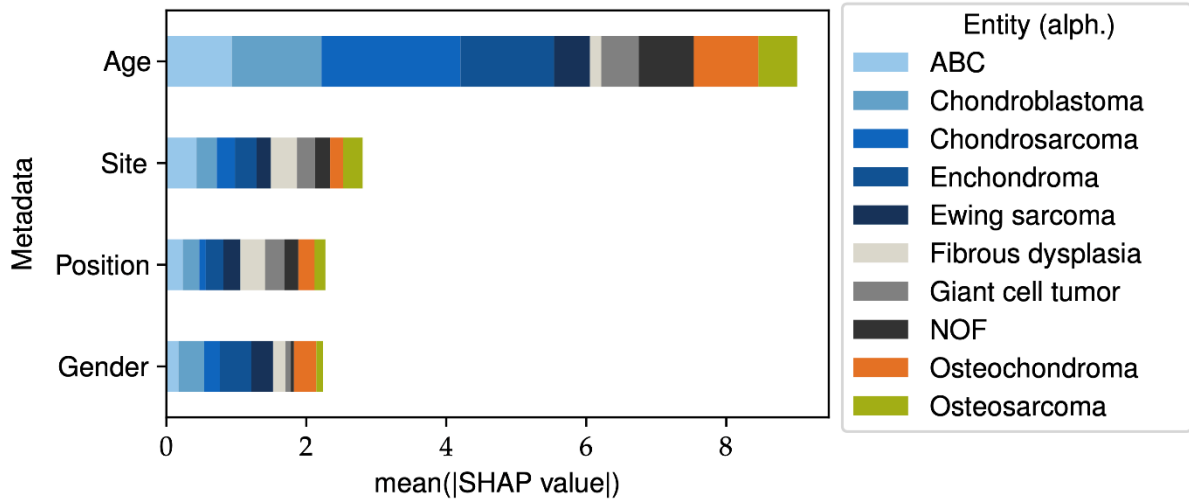
269 Figure 4 - Accuracy scores per entity and respective predictions: No Chondroblastoma or Ewing was classified correctly, therefore all metrics
 270 for these two entities were 0.00. Osteochondromas and enchondromas achieved the best performance with 88% and 75%, respectively.

271

272 **Explainability through SHAP**

273 We used the SHAP framework to evaluate the impact of metadata on the best-performing NesT
 274 model, as introduced by Lundberg et al. [29]. In Figure 5, the averaged SHAP results are
 275 visualized.

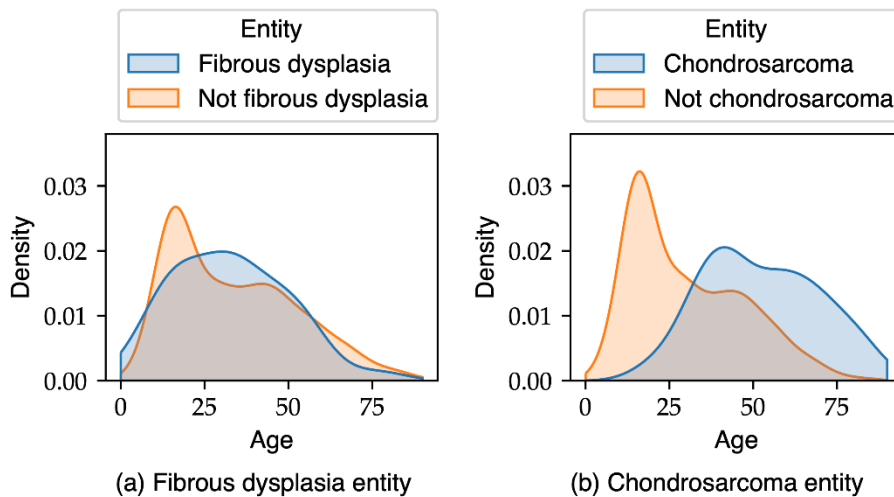
276



277 Figure 5 - SHAP results for the best-performing NesT model: the contribution of the features of the tumor at the bone, gender = gender of the
 278 patient).

279

280 Age was found to have the highest impact on the prediction, while the contribution of other
 281 metadata features was similar. Additionally, the influence of metadata on output predictions
 282 varied for each entity. Chondrosarcoma showed the highest SHAP value for age, while fibrous
 283 dysplasia had the lowest. To gain further insights, we examined the age distributions for these
 284 entities, comparing them with the rest of the dataset. Figure 6 illustrates the age distribution for
 285 fibrous dysplasia, and a similar comparison is made for chondrosarcoma.



286 Figure 6 - Age distribution for two entities compared to the overall distribution of the other entities. Normalization is calculated separately for
 287 both groups (entity/not-entity), since the not-entity group contains many more samples.

288

289 **Discussion**

290 The main finding of this study was the successful development of a transformer and MLP-
291 based model for classifying ten primary bone tumour entities. Including clinical metadata along
292 with radiography significantly improved the classification accuracy. While the results are not
293 yet ready for clinical application, they highlight the potential of addressing the complex
294 diagnostic classification task by increasing data quality and quantity. Integration of more
295 comprehensive data, as suggested in literature [2; 5; 6], could further enhance the model's
296 performance. Additionally, the implementation of SHAP helped provide insights into the
297 impact of metadata and interpret the black-box nature of DL models. It is important to note that
298 the age plot for fibrous dysplasia and other entities showed a significant intersection, making
299 it challenging to extract information based on age alone for this entity. Conversely, the age plot
300 for chondrosarcoma displayed a larger impact according to SHAP values, with less intersection
301 with other entities (Figure 6b). This suggests that age is a useful feature for distinguishing the
302 presence of chondrosarcoma.

303 A major limitation of this study is the size of the dataset. While having 1,785 radiographs is
304 significant for rare primary bone tumour entities, the average of 178 samples per class is
305 relatively low considering the heterogeneity of bone lesions and the requirements of deep
306 learning. This limited dataset results in class imbalance, which is common in medical datasets.
307 The most common entity (osteochondroma) has 501 occurrences, while the least common
308 entity (Ewing sarcoma) has only 6 occurrences. The underrepresentation of less frequent
309 classes can lead to overfitting and poor generalization [26]. Another limitation is the lack of
310 evaluation using external data. Although around 10% of the dataset consists of radiographs
311 from external sources, such as general practitioners and external radiologists, it is necessary to
312 test the model on additional external data to assess its generalizability [30]. This assessment is
313 crucial before considering the model's suitability for clinical use [31].

314 Several prior studies have explored the classification of bone tumours using imaging data [12;
315 13] or have demonstrated multimodal approaches for integrating imaging and tabular data in
316 medical classification [26; 32]. For instance, von Schacky et al. [13] developed a multitask
317 deep learning (DL) model capable of simultaneously detecting, segmenting, and classifying
318 bone lesions, comparing its performance against radiologists of varying experience levels. The
319 overall task of classifying bone lesions and the specific entities examined in their study were
320 similar to those in our research. While their model achieved a classification accuracy of 43.2%,

321 a musculoskeletal radiologist achieved 58.6% accuracy in classifying bone lesions on an entity
322 level. Although our metric values were significantly higher, von Schacky et al. had to contend
323 with a lower sample ratio per class, fewer patients, and thus a smaller overall dataset. Their
324 study primarily focused on a multitasking model and comparison with human experts, while
325 our emphasis centered on integrating clinical metadata in conjunction with imaging data using
326 state-of-the-art techniques. Nonetheless, their study underscores the intricate nature of
327 accurately identifying bone neoplasms for both DL models and clinical professionals. In a
328 similar study, Liu et al. [33] proposed a deep learning-machine learning model for classifying
329 bone tumours using patient clinical metadata and radiographs. They collected 982 radiographs
330 from 643 patients, incorporating clinical metadata such as age, gender, and location. Their
331 approach involved using an Inception V3 model to process imaging data and fusing its output
332 with clinical features to train an XGBoost model. Their fusion model achieved a top macro
333 area under the curve of 0.872, outperforming five radiologists by 0.819. The main difference
334 between their study and ours is their focus on predicting tumour malignancy, while we aimed
335 to classify ten tumour entities. The classification task differs due to the number of classes and
336 sample sizes. Our fusion approach captures both image and metadata information, while Liu et
337 al. combined DL model probabilities with metadata before using a secondary model. We
338 hypothesize that our approach better aligns with the clinical algorithm used by radiologists and
339 surgeons, as it simultaneously evaluates metadata and imaging data for comprehensive and
340 accurate bone tumour assessment, leading to improved performance. Xu et al. [34] presented a
341 notable study that employed multimodal data and a fusion approach for accurate differential
342 diagnosis of skin tumours. They introduced a transformer model capable of leveraging
343 multimodality imaging and non-imaging data to enhance diagnostic performance. Their
344 approach involved integrating a cross-modality fusion module with a transformer-based
345 multimodal classification system, enabling the fusion of data from multiple sources. The
346 dataset used in their study encompassed dermoscopy, clinical imaging, and patient metadata.
347 To evaluate the effectiveness of their proposed model, Xu et al. conducted experiments on both
348 a public dataset (Derm7pt, 1,011 cases) and an in-house dataset (5,601 cases). The results were
349 highly promising, surpassing the state-of-the-art performance with a 2.8% increase and
350 achieving an impressive accuracy of 88.5%, respectively. In comparison to our model, the
351 approach described by Xu et al. demonstrated the capability to incorporate multimodal imaging
352 in addition to metadata. While "remixing" metadata within disease classes yielded positive
353 results in their specific domain, we posit that in our case, metadata and image features are
354 closely intertwined and should not be interchangeably treated. Nevertheless, it is important to

355 note that no existing model, to the best of our knowledge, has proposed a multimodal approach
356 that integrates both imaging and patient-specific metadata for bone tumour classification.

357 The framework with a transformer model and MLP, combined through feature join for image
358 and metadata processing, is transferable to other scenarios where integration of different types
359 of information is crucial for decision-making. The retrospective dataset, spanning 20 years and
360 including diverse patient populations and imaging devices, ensures a lack of strong bias and
361 good generalizability. However, although the dataset is considerable for rare bone tumours, it
362 is not extremely large in terms of DL. To ensure broader generalizability, a larger dataset
363 should be collected in the future.

364 The proposed model's results do not yet have direct clinical relevance, but the increased
365 accuracy achieved through state-of-the-art methodology shows promise. Enriching the imaging
366 dataset with clinical metadata brings AI models closer to the approach of human experts. These
367 promising results, along with other applications of AI models in medicine, could raise
368 awareness among domain experts. Optimal AI model performance relies on domain experts
369 supporting the collection of complete, accurate, and comprehensive medical data, as data
370 quality and quantity are vital factors.

371 In conclusion, we developed a novel fusion model, combining NesT and MLP, to integrate
372 imaging data and clinical metadata for bone tumour classification. By enriching the imaging
373 dataset with patient-specific clinical metadata, such as age, gender, tumour position, and site,
374 we improved performance and surpassed similar studies. This approach aligns with current
375 clinical diagnostic workflows, where imaging data and patient characteristics are evaluated
376 together for tumour assessment. While the results are not yet suitable for clinical application,
377 we believe that structured data collection can further enhance our model's performance, making
378 it a valuable tool for radiologists, surgeons, and general practitioners in bone tumour
379 assessment.

380 **References**

- 381 1 Grimer RJ, Carter SR, Pynsent PB (1997) The cost-effectiveness of limb salvage for
382 bone tumours. *J Bone Joint Surg Br* 79:558-561
- 383 2 Hinterwimmer F, Consalvo S, Neumann J, Rueckert D, von Eisenhart-Rothe R,
384 Burgkart R (2022) Applications of machine learning for imaging-driven diagnosis of
385 musculoskeletal malignancies—a scoping review. *European Radiology* 32:7173-7184
- 386 3 Rechl H, Kirchhoff C, Wortler K, Lenze U, Topfer A, von Eisenhart-Rothe R (2011)
387 [Diagnosis of malignant bone and soft tissue tumors]. *Orthopade* 40:931-941; quiz
388 942-933
- 389 4 Picci P, Manfrini M, Donati DM et al (2020) Diagnosis of Musculoskeletal Tumors
390 and Tumor-like Conditions: Clinical, Radiological and Histological Correlations-the
391 Rizzoli Case Archive. Springer
- 392 5 Gaume M, Chevret S, Campagna R, Larousserie F, Biau D (2022) The appropriate
393 and sequential value of standard radiograph, computed tomography and magnetic
394 resonance imaging to characterize a bone tumor. *Scientific reports* 12:1-9
- 395 6 Grimer RJ, Briggs TW (2010) Earlier diagnosis of bone and soft-tissue tumours. *J*
396 *Bone Joint Surg Br* 92:1489-1492
- 397 7 Clark MA, Thomas JM (2005) Delay in referral to a specialist soft-tissue sarcoma
398 unit. *Eur J Surg Oncol* 31:443-448
- 399 8 Miller BJ (2019) Use of imaging prior to referral to a musculoskeletal oncologist.
400 *JAAOS-Journal of the American Academy of Orthopaedic Surgeons* 27:e1001-e1008
- 401 9 Salom M, Chiari C, Alessandri JMG, Willegger M, Windhager R, Sanpera I (2021)
402 Diagnosis and staging of malignant bone tumours in children: what is due and what is
403 new? *Journal of Children's Orthopaedics* 15:312-321
- 404 10 Kindblom LG (2009) Bone tumors: epidemiology, classification, pathology. *Imaging*
405 *of bone tumors and tumor-like lesions: techniques and applications:1-15*
- 406 11 Kharat AT, Singhal S (2017) A peek into the future of radiology using big data
407 applications. *Indian J Radiol Imaging* 27:241-248
- 408 12 Consalvo S, Hinterwimmer F, Neumann J et al (2022) Two-Phase Deep Learning
409 Algorithm for Detection and Differentiation of Ewing Sarcoma and Acute
410 Osteomyelitis in Paediatric Radiographs. *Anticancer Research* 42:4371-4380

411 13 von Schacky CE, Wilhelm NJ, Schäfer VS et al (2021) Multitask deep learning for
412 segmentation and classification of primary bone tumors on radiographs. *Radiology*
413 301:398-406

414 14 Savage N (2020) How AI is improving cancer diagnostics. *Nature* 579:S14+

415 15 Rajpurkar P, Chen E, Banerjee O, Topol EJ (2022) AI in health and medicine. *Nat*
416 *Med* 28:31-38

417 16 Moch H (2020) Soft Tissue and Bone Tumours WHO Classification of
418 Tumours/Volume 3. WHO Classification of Tumours 3

419 17 Vandembroucke JP, von Elm E, Altman DG et al (2007) Strengthening the Reporting
420 of Observational Studies in Epidemiology (STROBE): explanation and elaboration.
421 *Epidemiology* 18:805-835

422 18 Collins GS, Reitsma JB, Altman DG, Moons KG (2015) Transparent reporting of a
423 multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the
424 TRIPOD statement. *Annals of internal medicine* 162:55-63

425 19 He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image
426 recognition. *Proceedings of the IEEE conference on computer vision and pattern*
427 *recognition*, pp 770-778

428 20 Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale
429 image recognition. *arXiv preprint arXiv:1409.1556*

430 21 Bloier M, Hinterwimmer F, Breden S et al (2022) Detection and Segmentation of
431 Heterogeneous Bone Tumours in Limited Radiographs. *Current Directions in*
432 *Biomedical Engineering*. De Gruyter, pp 69-72

433 22 Huang S-C, Pareek A, Seyyedi S, Banerjee I, Lungren MP (2020) Fusion of medical
434 imaging and electronic health records using deep learning: a systematic review and
435 implementation guidelines. *NPJ digital medicine* 3:1-9

436 23 Chen T, Guestrin C (2016) XGBoost: A Scalable Tree Boosting System. *Proceedings of*
437 *the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data*
438 *Mining*. Association for Computing Machinery, San Francisco, California, USA, pp
439 785–794

440 24 Zhang Z, Zhang H, Zhao L, Chen T, Arik SÖ, Pfister T (2022) Nested hierarchical
441 transformer: Towards accurate, data-efficient and interpretable visual
442 understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, pp
443 3417-3425

444 25 Raghu M, Unterthiner T, Kornblith S, Zhang C, Dosovitskiy A (2021) Do vision
445 transformers see like convolutional neural networks? *Advances in Neural Information*
446 *Processing Systems* 34:12116-12128

447 26 Li Z, Kamnitsas K, Glocker B (2020) Analyzing overfitting under class imbalance in
448 neural networks for image segmentation. *IEEE Transactions on Medical Imaging*
449 40:1065-1077

450 27 Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint
451 arXiv:14126980

452 28 Samek W, Wiegand T, Müller K-R (2017) Explainable artificial intelligence:
453 Understanding, visualizing and interpreting deep learning models. arXiv preprint
454 arXiv:170808296

455 29 Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions.
456 *Advances in Neural Information Processing Systems* 30

457 30 Yu AC, Mohajer B, Eng J (2022) External validation of deep learning algorithms for
458 radiologic diagnosis: a systematic review. *Radiology: Artificial Intelligence*
459 4:e210064

460 31 Park SH, Han K (2018) Methodologic guide for evaluating clinical performance and
461 effect of artificial intelligence technology for medical diagnosis and prediction.
462 *Radiology* 286:800-809

463 32 Cai G, Zhu Y, Wu Y, Jiang X, Ye J, Yang D (2022) A multimodal transformer to fuse
464 images and metadata for skin disease classification. *The Visual Computer*:1-13

465 33 Liu R, Pan D, Xu Y et al (2021) A deep learning-machine learning fusion approach
466 for the classification of benign, malignant, and intermediate bone tumors. *Eur Radiol.*
467 10.1007/s00330-021-08195-z

468 34 Xu J, Gao Y, Liu W et al (2022) RemixFormer: A Transformer Model for Precision
469 Skin Tumor Differential Diagnosis via Multi-modal Imaging and Non-imaging
470 DataMedical Image Computing and Computer Assisted Intervention–MICCAI 2022:
471 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part
472 III. Springer, pp 624-633
473
474

475 **Figures and Tables**

- 476 • Figure 1 - Flow diagram showing the application of eligibility criteria to create a final
477 dataset.
- 478 • Figure 2 - Examples of radiographs of each entity: one example for each entity after
479 pre-processing (cropping, rescaling).
- 480 • Figure 3 - Proposed multimodal model: metadata (clinical data) being processed
481 through a multilayer perceptron, imaging data through common image classification
482 networks and concatenation (fusion) of both in feature space before a final fully
483 connected and softmax layer.
- 484 • Figure 4 - Accuracy scores per entity and respective predictions: No Chondroblastoma
485 or Ewing was classified correctly, therefore all metrics for these two entities were 0.00.
486 Osteochondromas and enchondromas achieved the best performance with 88% and
487 75%, respectively.
- 488 • Figure 5 - SHAP results for the best-performing NesT model: the contribution of the
489 features based on the entity measured in SHAP values. The values are calculated based
490 on the test dataset, taking the mean of the absolute values (age = patient's age, site =
491 affected bone, position = position of the tumor at the bone, gender = gender of the
492 patient).
- 493 • Figure 6 - Age distribution for two entities compared to the overall distribution of the
494 other entities. Normalization is calculated separately for both groups (entity/not-entity),
495 since the not-entity group contains many more samples.
- 496
- 497 • Table 1: Distribution of continuous characteristics (std = standard deviation, IQR =
498 interquartile range).
- 499 • Table 2: Distribution of discrete characteristics with incidence and percentage ratio.
- 500 • Table 3: Experiment results reporting the test accuracy with standard deviation.

501

502 **Acknowledgements**

503 Thank you very much to Fritz Seidl, MA Interpreting and Translating, for the support in
504 language editing and revising the manuscript.

505

506 **Supplemental Materials**

507 None

508

Bibliography

- [1] F. Hinterwimmer, S. Consalvo, J. Neumann, D. Rueckert, R. von Eisenhart-Rothe, and R. Burgkart. "Applications of machine learning for imaging-driven diagnosis of musculoskeletal malignancies - a scoping review." In: *European Radiology* 32.10 (2022), pp. 7173–7184.
- [2] F. Hinterwimmer, S. Consalvo, J. Neumann, C. Micheler, N. Wilhelm, J. Lang, R. v. Eisenhart-Rothe, R. Burgkart, and D. Rueckert. "From Self-supervised Learning to Transfer Learning with Musculoskeletal Radiographs." In: *Current Directions in Biomedical Engineering*. Vol. 8. 2. De Gruyter. 2022, pp. 9–12.
- [3] F. Hinterwimmer, S. Consalvo, N. Wilhelm, F. Seidl, R. H. Burgkart, R. von Eisenhart-Rothe, D. Rueckert, and J. Neumann. "SAM-X: sorting algorithm for musculoskeletal x-ray radiography." In: *European Radiology* 33.3 (2023), pp. 1537–1544.
- [4] F. Hinterwimmer, R. S. Serena, N. Wilhelm, S. Breden, S. Consalvo, F. Seidl, D. Juestel, R. H. Burgkart, K. Woertler, R. von Eisenhart-Rothe, et al. "Recommender-based bone tumour classification with radiographs—a link to the past." In: *European Radiology* (2024), pp. 1–10.
- [5] S. Consalvo, F. Hinterwimmer, J. Neumann, M. Steinborn, M. Salzmann, F. Seidl, U. Lenze, C. Knebel, D. Rueckert, and R. H. Burgkart. "Two-Phase Deep Learning Algorithm for Detection and Differentiation of Ewing Sarcoma and Acute Osteomyelitis in Paediatric Radiographs." In: *Anticancer Research* 42.9 (2022), pp. 4371–4380.
- [6] M. Bloier, F. Hinterwimmer, S. Breden, S. Consalvo, J. Neumann, N. Wilhelm, R. v. Eisenhart-Rothe, D. Rueckert, and R. Burgkart. "Detection and Segmentation of Heterogeneous Bone Tumours in Limited Radiographs." In: *Current Directions in Biomedical Engineering*. Vol. 8. 2. De Gruyter. 2022, pp. 69–72.
- [7] F. Hinterwimmer, M. Günther, S. Consalvo, A. Gersing, K. Woertler, R. von Eisenhart-Rothe, J. Neumann, R. Burgkart, and D. Rueckert. "Impact of meta-data in multimodal bone tumour classification." In: *PLOS Digital Health* (2024). submitted 01/2024.

- [8] F. Lenze, F. Hinterwimmer, L. Fleckenstein, I. Lazic, D. Dammerer, R. VON Eisenhart-Rothe, N. Harrasser, and F. Pohlig. "Minimally invasive total hip arthroplasty: a comparison of restoring hip biomechanics with and without a traction table." In: *in vivo* 36.1 (2022), pp. 424–429.
- [9] F. Hinterwimmer, I. Lazic, C. Suren, M. T. Hirschmann, F. Pohlig, D. Rueckert, R. Burgkart, and R. von Eisenhart-Rothe. "Machine learning in knee arthroplasty: specific data are key - a systematic review." In: *Knee Surgery, Sports Traumatology, Arthroscopy* 30.2 (2022), pp. 376–388.
- [10] C. Zanzinger, N. Harrasser, O. Gottschalk, P. Dolp, F. Hinterwimmer, H. Hoerterer, and M. Walther. "One-year Follow-Up Results with Hydrogel Implant in Therapy of Hallux Rigidus: Case Series with 44 Patients." In: *Zeitschrift für Orthopädie und Unfallchirurgie* 160.04 (2022), pp. 414–421.
- [11] D. M. Hedderich, M. Keicher, B. Wiestler, M. J. Gruber, H. Burwinkel, F. Hinterwimmer, T. Czempel, J. E. Spiro, D. Pinto dos Santos, D. Heim, et al. "AI for Doctors - a course to educate medical professionals in artificial intelligence for medical imaging." In: *Healthcare*. Vol. 9. 10. MDPI. 2021, p. 1278.
- [12] F. Hinterwimmer, I. Lazic, S. Langer, C. Suren, F. Charitou, M. T. Hirschmann, G. Matziolis, F. Seidl, F. Pohlig, D. Rueckert, et al. "Prediction of complications and surgery duration in primary TKA with high accuracy using machine learning with arthroplasty-specific data." In: *Knee Surgery, Sports Traumatology, Arthroscopy* 31.4 (2023), pp. 1323–1333.
- [13] I. Lazic, F. Hinterwimmer, S. Langer, F. Pohlig, C. Suren, F. Seidl, D. Rückert, R. Burgkart, and R. von Eisenhart-Rothe. "Prediction of Complications and Surgery Duration in Primary Total Hip Arthroplasty Using Machine Learning: The Necessity of Modified Algorithms and Specific Data." In: *Journal of Clinical Medicine* 11.8 (2022), p. 2147.
- [14] R. von Eisenhart-Rothe, F. Hinterwimmer, H. Graichen, and M. T. Hirschmann. "Artificial intelligence and robotics in TKA surgery: promising options for improved outcomes?" In: *Knee Surgery, Sports Traumatology, Arthroscopy* 30.8 (2022), pp. 2535–2537.
- [15] N. J. Wilhelm, S. Haddadin, J. J. Lang, C. Micheler, F. Hinterwimmer, A. Reiners, R. Burgkart, and C. Glowalla. "Development of an Exoskeleton Platform of the Finger for Objective Patient Monitoring in Rehabilitation." In: *Sensors* 22.13 (2022), p. 4804.

- [16] I. Lazic, F. Hinterwimmer, and R. von Eisenhart-Rothe. "Vorhersage von irregulären Operationsdauern bei Knie totalendoprothesen mit Daten aus dem Endoprothesenregister Deutschland und EndoCert." In: *Knie Journal* 4.4 (2022), pp. 224–229.
- [17] C. M. Micheler, J. J. Lang, N. J. Wilhelm, I. Lazic, F. Hinterwimmer, C. Fritz, R. v. Eisenhart-Rothe, M. F. Zäh, and R. H. Burgkart. "Scaling Methods of the Pelvis without Distortion for the Analysis of Bone Defects." In: *Current Directions in Biomedical Engineering*. Vol. 8. 2. De Gruyter. 2022, pp. 797–800.
- [18] J. J. Lang, V. Baylacher, C. M. Micheler, N. J. Wilhelm, F. Hinterwimmer, B. Schwaiger, D. Barnewitz, R. v. Eisenhart-Rothe, C. U. Grosse, and R. Burgkart. "Improving Equine Intramedullary Nail Osteosynthesis via Fracture Adjacent Polymer Reinforcement." In: *Current Directions in Biomedical Engineering*. Vol. 8. 2. De Gruyter. 2022, pp. 129–132.
- [19] S. Consalvo, F. Hinterwimmer, N. Harrasser, U. Lenze, G. Matziolis, R. von Eisenhart-Rothe, and C. Knebel. "C-Reactive Protein Pretreatment-Level Evaluation for Ewing's Sarcoma Prognosis Assessment - A 15-Year Retrospective Single-Centre Study." In: *Cancers* 14.23 (2022), p. 5898.
- [20] N. Harrasser, F. Hinterwimmer, S. Baumbach, K. Pfahl, C. Glowalla, M. Walther, and H. Hörterer. "The distal metatarsal screw is not always necessary in third-generation MICA: a case-control study." In: *Archives of Orthopaedic and Trauma Surgery* (2022), pp. 1–7.
- [21] C. Glowalla, S. Langer, U. Lenze, I. Lazic, M. T. Hirschmann, F. Hinterwimmer, R. von Eisenhart-Rothe, and F. Pohlig. "Postoperative full leg radiographs exhibit less residual coronal varus deformity compared to intraoperative measurements in robotic arm-assisted total knee arthroplasty with the MAKO™ system." In: *Knee Surgery, Sports Traumatology, Arthroscopy* (2023), pp. 1–7.
- [22] D. Jüstel, H. Irl, F. Hinterwimmer, C. Dehner, W. Simson, N. Navab, G. Schneider, and V. Ntziachristos. "Spotlight on nerves: Portable multispectral optoacoustic imaging of peripheral nerve vascularization and morphology." In: *arXiv preprint arXiv:2207.13978* (2022).
- [23] S. Breden, F. Hinterwimmer, S. Consalvo, J. Neumann, C. Knebel, R. von Eisenhart-Rothe, R. H. Burgkart, and U. Lenze. "Deep learning-based detection of bone tumors around the knee in X-rays of children." In: *Journal of Clinical Medicine* 12.18 (2023), p. 5960.

- [24] S. Breden, F. Hinterwimmer, S. Beischl, S. Consalvo, A. S. Gersing, U. Lenze, R. von Eisenhart-Rothe, and C. Knebel. "A New Method for Assessing Patients' Obesity-Associated Infection Risk Using X-rays in Hip Arthroplasties." In: *Journal of Clinical Medicine* 12.23 (2023), p. 7277.
- [25] N. Wilhelm, C. M. Micheler, J. J. Lang, F. Hinterwimmer, V. Schaack, R. Smits, S. Haddadin, and R. Burgkart. "Development and Evaluation of a Cost-effective IMU System for Gait Analysis: Comparison with Vicon and VideoPose3D Algorithms." In: *Current Directions in Biomedical Engineering*. Vol. 9. 1. De Gruyter. 2023, pp. 254–257.
- [26] P. Picci, M. Manfrini, D. M. Donati, M. Gambarotti, A. Righi, D. Vanel, and A. P. Dei Tos. "Diagnosis of musculoskeletal tumors and tumor-like conditions: clinical, radiological and histological correlations-the Rizzoli case archive." In: (2020).
- [27] E. Hipp, W. Plötz, R. Burgkart, and R. Schelter. "Limb salvage, 6." In: *Aufl. Hipp E, Munich* (1998).
- [28] M. Gaume, S. Chevret, R. Campagna, F. Larousserie, and D. Biau. "The appropriate and sequential value of standard radiograph, computed tomography and magnetic resonance imaging to characterize a bone tumor." In: *Scientific Reports* 12.1 (2022), pp. 1–9.
- [29] B. J. Miller. "Use of imaging prior to referral to a musculoskeletal oncologist." In: *JAAOS-Journal of the American Academy of Orthopaedic Surgeons* 27.22 (2019), e1001–e1008.
- [30] C. F. Gould, J. Q. Ly, G. E. Lattin Jr, D. P. Beall, and J. B. Sutcliffe III. "Bone tumor mimics: avoiding misdiagnosis." In: *Current problems in diagnostic radiology* 36.3 (2007), pp. 124–141.
- [31] S. De Salvo, V. Pavone, S. Coco, E. Dell'Agli, C. Blatti, and G. Testa. "Benign bone tumors: an overview of what we know today." In: *Journal of Clinical Medicine* 11.3 (2022), p. 699.
- [32] R. Grimer and T. Briggs. "Earlier diagnosis of bone and soft-tissue tumours." In: *The Journal of bone and joint surgery. British volume* 92.11 (2010), pp. 1489–1492.
- [33] H. Moch. "Soft Tissue and Bone Tumours WHO Classification of Tumours/Volume 3." In: *WHO Classification of Tumours 3* (2020).
- [34] R. Lalam, J. L. Bloem, I. M. Noebauer-Huhmann, K. Wörtler, A. Tagliafico, F. Vanhoenacker, V. V. Nikodinovska, H. T. Sanal, H.-J. van der Woude, O. Papanikolaou, et al. "ESSR consensus document for detection, characterization, and referral pathway for tumors and tumorlike lesions of bone." In: *Seminars in musculoskeletal radiology*. Vol. 21. 05. Thieme Medical Publishers. 2017, pp. 630–647.

- [35] M. Clark and J. Thomas. "Delay in referral to a specialist soft-tissue sarcoma unit." In: *European Journal of Surgical Oncology (EJSO)* 31.4 (2005), pp. 443–448.
- [36] H. Rechl, C. Kirchhoff, K. Wörtler, U. Lenze, A. Toepfer, and R. von Eisenhart-Rothe. "Diagnosis of malignant bone and soft tissue tumors." In: *Der Orthopäde* 40 (2011), pp. 931–944.
- [37] L. M. Goedhart, J. G. Gerbers, J. J. Ploegmakers, and P. C. Jutte. "Delay in diagnosis and its effect on clinical outcome in high-grade sarcoma of bone: a referral oncological centre study." In: *Orthopaedic Surgery* 8.2 (2016), pp. 122–128.
- [38] Y. LeCun, Y. Bengio, and G. Hinton. "Deep learning." In: *nature* 521.7553 (2015), pp. 436–444.
- [39] A. Krogh. "What are artificial neural networks?" In: *Nature biotechnology* 26.2 (2008), pp. 195–197.
- [40] W. S. McCulloch and W. Pitts. "A logical calculus of the ideas immanent in nervous activity." In: *The bulletin of mathematical biophysics* 5 (1943), pp. 115–133.
- [41] F. Rosenblatt. "The perceptron: a probabilistic model for information storage and organization in the brain." In: *Psychological review* 65.6 (1958), p. 386.
- [42] M. Minsky and S. Papert. "An introduction to computational geometry." In: *Cambridge tiass., HIT* 479 (1969), p. 480.
- [43] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. "Learning representations by back-propagating errors." In: *nature* 323.6088 (1986), pp. 533–536.
- [44] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition." In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.
- [45] A. Sherstinsky. "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network." In: *Physica D: Nonlinear Phenomena* 404 (2020), p. 132306.
- [46] S. Hochreiter and J. Schmidhuber. "Long short-term memory." In: *Neural computation* 9.8 (1997), pp. 1735–1780.
- [47] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "Imagenet classification with deep convolutional neural networks." In: *Communications of the ACM* 60.6 (2017), pp. 84–90.
- [48] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

- [49] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative adversarial networks." In: *Communications of the ACM* 63.11 (2020), pp. 139–144.
- [50] Z. Zhang, H. Zhang, L. Zhao, T. Chen, S. Ö. Arik, and T. Pfister. "Nested hierarchical transformer: Towards accurate, data-efficient and interpretable visual understanding." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 3. 2022, pp. 3417–3425.
- [51] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. "Going deeper with convolutions." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.
- [52] M. Tan and Q. Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." In: *International conference on machine learning*. PMLR. 2019, pp. 6105–6114.
- [53] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. "Imagenet: A large-scale hierarchical image database." In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.
- [54] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition." In: *arXiv preprint arXiv:1409.1556* (2014).
- [55] R. Girshick, J. Donahue, T. Darrell, and J. Malik. "Rich feature hierarchies for accurate object detection and semantic segmentation." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 580–587.
- [56] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. "You only look once: Unified, real-time object detection." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.
- [57] O. Ronneberger, P. Fischer, and T. Brox. "U-net: Convolutional networks for biomedical image segmentation." In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III* 18. Springer. 2015, pp. 234–241.
- [58] V. Badrinarayanan, A. Kendall, and R. Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." In: *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017), pp. 2481–2495.
- [59] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." In: *IEEE transactions on pattern analysis and machine intelligence* 40.4 (2017), pp. 834–848.

- [60] A. Radford, L. Metz, and S. Chintala. “Unsupervised representation learning with deep convolutional generative adversarial networks.” In: *arXiv preprint arXiv:1511.06434* (2015).
- [61] T. Karras, S. Laine, and T. Aila. “A style-based generator architecture for generative adversarial networks.” In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 4401–4410.
- [62] J. Devlin, M. Chang, K. Lee, and K. Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” In: *CoRR abs/1810.04805* (2018). arXiv: 1810.04805.
- [63] M. J. Willeminck, W. A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, L. R. Folio, R. M. Summers, D. L. Rubin, and M. P. Lungren. “Preparing medical imaging data for machine learning.” In: *Radiology* 295.1 (2020), pp. 4–15.
- [64] N. Savage. “Breaking into the black box of artificial intelligence.” In: *Nature* (2022).
- [65] S. Reddy. “Explainability and artificial intelligence in medicine.” In: *The Lancet Digital Health* 4.4 (2022), e214–e215.
- [66] S. Kundu. “AI in medicine must be explainable.” In: *Nature medicine* 27.8 (2021), pp. 1328–1328.
- [67] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King. “Key challenges for delivering clinical impact with artificial intelligence.” In: *BMC medicine* 17 (2019), pp. 1–9.
- [68] J. Feng, R. V. Phillips, I. Malenica, A. Bishara, A. E. Hubbard, L. A. Celi, and R. Pirracchio. “Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare.” In: *npj Digital Medicine* 5.1 (2022), p. 66.
- [69] D. J. Blezek, L. Olson-Williams, A. Missert, and P. Korfiatis. “AI integration in the clinical workflow.” In: *Journal of Digital Imaging* 34 (2021), pp. 1435–1446.
- [70] G. Marcus. “Deep learning: A critical appraisal.” In: *arXiv preprint arXiv:1801.00631* (2018).
- [71] K. N. Vokinger, S. Feuerriegel, and A. S. Kesselheim. “Mitigating bias in machine learning for medicine.” In: *Communications medicine* 1.1 (2021), p. 25.
- [72] M. Hort, Z. Chen, J. M. Zhang, F. Sarro, and M. Harman. “Bias mitigation for machine learning classifiers: A comprehensive survey.” In: *arXiv preprint arXiv:2207.07068* (2022).
- [73] T. L. D. Health. *There is no such thing as race in health-care algorithms*. 2019.

- [74] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren. "Secure, privacy-preserving and federated machine learning in medical imaging." In: *Nature Machine Intelligence* 2.6 (2020), pp. 305–311.
- [75] X. Ferrer, T. van Nuinen, J. M. Such, M. Coté, and N. Criado. "Bias and discrimination in AI: a cross-disciplinary perspective." In: *IEEE Technology and Society Magazine* 40.2 (2021), pp. 72–80.
- [76] A. C. Tricco, E. Lillie, W. Zarin, K. K. O'Brien, H. Colquhoun, D. Levac, D. Moher, M. D. Peters, T. Horsley, and L. Weeks. "PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation." In: *Annals of internal medicine* 169.7 (2018), pp. 467–473. issn: 0003-4819.
- [77] M. Caron, P. Bojanowski, A. Joulin, and M. Douze. "Deep clustering for unsupervised learning of visual features." In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 132–149.
- [78] E. Von Elm, D. G. Altman, M. Egger, S. J. Pocock, P. C. Gøtzsche, J. P. Vandenbergue, S. Initiative, et al. "The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies." In: *International journal of surgery* 12.12 (2014), pp. 1495–1499.
- [79] O. Jafari, P. Maurya, P. Nagarkar, K. M. Islam, and C. Crushev. "A survey on locality sensitive hashing algorithms and their applications." In: *arXiv preprint arXiv:2102.08942* (2021).
- [80] J. Mongan, L. Moy, and C. E. Kahn Jr. *Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers*. 2020.
- [81] C. E. von Schacky, N. J. Wilhelm, V. S. Schäfer, Y. Leonhardt, F. G. Gassert, S. C. Foreman, F. T. Gassert, M. Jung, P. M. Jungmann, and M. F. Russe. "Multitask deep learning for segmentation and classification of primary bone tumors on radiographs." In: *Radiology* 301.2 (2021), pp. 398–406. issn: 0033-8419.
- [82] M. S. Taljanovic, T. B. Hunter, K. A. Fitzpatrick, E. A. Krupinski, and T. L. Pope. "Musculoskeletal magnetic resonance imaging: importance of radiography." In: *Skeletal radiology* 32 (2003), pp. 403–411.
- [83] P. A. Ory. "Radiography in the assessment of musculoskeletal conditions." In: *Best Practice & Research Clinical Rheumatology* 17.3 (2003), pp. 495–512.
- [84] L. Kindblom. *Bone tumors: epidemiology, classification, pathology. Imaging of bone tumors and tumor-like lesions techniques and applications*. 2009.

- [85] R. Liu, D. Pan, Y. Xu, H. Zeng, Z. He, J. Lin, W. Zeng, Z. Wu, Z. Luo, G. Qin, et al. "A deep learning-machine learning fusion approach for the classification of benign, malignant, and intermediate bone tumors." In: *European Radiology* 32.2 (2022), pp. 1371–1383.
- [86] J. Xu, Y. Gao, W. Liu, K. Huang, S. Zhao, L. Lu, X. Wang, X.-S. Hua, Y. Wang, and X. Chen. "RemixFormer: A Transformer Model for Precision Skin Tumor Differential Diagnosis via Multi-modal Imaging and Non-imaging Data." In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part III*. Springer. 2022, pp. 624–633.