Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Comprehensive review of deep learning in orthopaedics: Applications, challenges, trustworthiness, and fusion[☆]

Laith Alzubaidi[a,b,c,*,1], Khamael AL-Dulaimi[d,e,1], Asma Salhi[b,c,1], Zaenab Alammar[f,1], Mohammed A. Fadhel[c,1], A.S. Albahri[g,1], A.H. Alamoodi[h,1], O.S. Albahri[i,1], Amjad F. Hasan[j,1], Jinshuai Bai[a,b,1], Luke Gilliland[b,c,1], Jing Peng[c,1], Marco Branni[b,c,1], Tristan Shuker[b,k,1], Kenneth Cutbush[b,k,1], Jose Santamaría[l,1], Catarina Moreira[m,1], Chun Ouyang[n,1], Ye Duan[o,1], Mohamed Manoufali[p,q,1], Mohammad Jomaa[b,k,1], Ashish Gupta[a,b,c,1], Amin Abbosh[q,1], Yuantong Gu[a,b,1]

[a] *School of Mechanical, Medical, and Process Engineering, Queensland University of Technology, Brisbane, QLD 4000, Australia*
[b] *QUASR/ARC Industrial Transformation Training Centre—Joint Biomechanics, Queensland University of Technology, Brisbane, QLD 4000, Australia*
[c] *Research and Development department, Akunah Med Technology Pty Ltd Co, Brisbane, QLD 4120, Australia*
[d] *Computer Science Department, College of Science, Al-Nahrain University, Baghdad, Baghdad 10011, Iraq*
[e] *School of Electrical Engineering and Robotics, Queensland University of Technology, Brisbane, QLD 4000, Australia*
[f] *School of Computer Science, Queensland University of Technology, Brisbane, QLD 4000, Australia*
[g] *Technical College, Imam Ja'afar Al-Sadiq University, Baghdad, Iraq*
[h] *Institute of Informatics and Computing in Energy, Universiti Tenaga Nasional, Kajang 43000, Malaysia*
[i] *Australian Technical and Management College, Melbourne, Australia*
[j] *Faculty of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65211, USA*
[k] *St Andrew's War Memorial Hospital, Brisbane, QLD 4000, Australia*
[l] *Department of Computer Science, University of Jaén, Jaén 23071, Spain*
[m] *Data Science Institute, University of Technology Sydney, Australia*
[n] *School of Information Systems, Queensland University of Technology, Brisbane, QLD 4000, Australia*
[o] *School of Computing, Clemson University, Clemson, 29631, SC, USA*
[p] *CSIRO, Kensington, WA 6151, Australia*
[q] *School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, QLD 4067, Australia*

ARTICLE INFO

Keywords:
Deep learning
Orthopaedics
Fracture detection
Trustworthy AI
Osteoarthritis
Fusion
Orthopaedic technologies

ABSTRACT

Deep learning (DL) in orthopaedics has gained significant attention in recent years. Previous studies have shown that DL can be applied to a wide variety of orthopaedic tasks, including fracture detection, bone tumour diagnosis, implant recognition, and evaluation of osteoarthritis severity. The utilisation of DL is expected to increase, owing to its ability to present accurate diagnoses more efficiently than traditional methods in many scenarios. This reduces the time and cost of diagnosis for patients and orthopaedic surgeons. To our knowledge, no exclusive study has comprehensively reviewed all aspects of DL currently used in orthopaedic practice. This review addresses this knowledge gap using articles from Science Direct, Scopus, IEEE Xplore, and Web of Science between 2017 and 2023. The authors begin with the motivation for using DL in orthopaedics, including its ability to enhance diagnosis and treatment planning. The review then covers various applications of DL in orthopaedics, including fracture detection, detection of supraspinatus tears using MRI, osteoarthritis, prediction of types of arthroplasty implants, bone age assessment, and detection of joint-specific soft tissue disease. We also examine the challenges for implementing DL in orthopaedics, including the scarcity of data to train DL and the lack of interpretability, as well as possible solutions to these common pitfalls. Our work highlights the requirements to achieve trustworthiness in the outcomes generated by DL, including the need for accuracy, explainability, and fairness in the DL models. We pay particular attention to fusion techniques as one of the ways to increase trustworthiness, which have also been used to address the common multimodality in orthopaedics. Finally, we have reviewed the approval requirements set forth by the US Food and Drug

[☆] Review of Deep Learning in Orthopaedics.

[*] Corresponding author at: School of Mechanical, Medical, and Process Engineering, Queensland University of Technology, Brisbane, QLD 4000, Australia.
E-mail address: l.alzubaidi@qut.edu.au (L. Alzubaidi).

[1] The authors contributed equally to this work.

https://doi.org/10.1016/j.artmed.2024.102935
Received 1 June 2023; Received in revised form 18 March 2024; Accepted 22 July 2024
Available online 25 July 2024

Administration to enable the use of DL applications. As such, we aim to have this review function as a guide for researchers to develop a reliable DL application for orthopaedic tasks from scratch for use in the market.

Contents

1. Introduction

The branch of surgery related to the human musculoskeletal system, including the spine, extremities, and corresponding structures, is known as orthopaedics [1]. This discipline emphasises the prevention, treatment, and rehabilitation of the different structural and/or functional diseases affecting bones, articulations, ligaments, musculotendinous units, and the surrounding neuro-vasculature. A subcategory of this discipline deals with injuries sustained during traumatic events, including accidents at work and in sports. Common injuries include fractures, joint dislocations, tendon tears, ligament ruptures, traumatic disc herniations, and acute nerve compressions. Other orthopaedic diagnoses relate to degenerative pathologies, mainly due to age and/or overuse-related wear and tear of the involved structure(s). Osteoarthritis and chronic disc herniations are examples of the latter.

Orthopaedics also covers some paediatric congenital conditions that appear at birth and occur during childhood, including extremity deformities. Furthermore, there are conditions that are described as neoplastic. These include benign and cancerous bone tumours. In light of this broad scope of practice, we can infer that musculoskeletal disorders not only significantly impact the health of individual patients but also impose a significant burden on the healthcare system, negatively impacting any country's overall health and economy.

For instance, in the United Kingdom (UK) alone, 25% of all surgical operations are related to musculoskeletal problems, which represents £4.7 billion of the expenditure of the National Health Service (NHS) annually [2]. Meanwhile, it is believed that one in two adults in the United States of America (USA) suffers from musculoskeletal problems, inflating the cost of treatment and economic productivity to an estimated 213 billion dollars, equivalent to approximately 1. 4% of its Gross Domestic Product (GDP) [3]. Due to the significant economic pressures, there is a growing demand for effective and reproducible orthopaedic diagnostic and treatment approaches. There is great potential to adopt innovative technological applications, including deep learning (DL) based technologies in this space [4]. Successful treatment, or even more efficient diagnosis of a particular condition, can substantially reduce the economic burdens in today's industrial and ageing populations.

The diversity of characteristics in orthopaedic diagnostic and therapeutic schemes makes it a suitable environment to implement creative and innovative DL methods, addressing the scalability and budget issues that plague traditional healthcare models. Currently, technology is utilised mainly in orthopaedic "software" and "hardware", ranging from a simple fracture–fixation screw to highly advanced robotics that can navigate and possibly implant a prosthetic joint. Due to the wide range of possible applications, orthopaedic experts must acquire comprehensive technical knowledge to appropriately evaluate, plan, and implement these technologies when treating patients. Thus, orthopaedic surgeons must assess and adopt new technologies once these technologies' effectiveness, precision, and accuracy have been proven and verified. A prime example of this process is the growing implementation of mixed reality [5,6].

In addition, orthopaedic experts already have deep professional relationships with technologically innovative industries. This professional, patient-focused partnership provides mechanisms for collecting real-time feedback and improving improvements and improvements. The diagnostic and therapeutic methodologies of the orthopaedic practice are generally well established, facilitating the large-scale optimisation of processes. This is aided by the fact that the majority of orthopaedic operations, such as hip replacements, are commonly reproducible with highly effective outcomes [7]. After decades of refinement, such procedures' progressive and consistent success provides reasonable stability for developing up-to-date technologies and advanced optimisation systems. More importantly, as a result of the integral relationships with technologically innovative partners, the orthopaedic community is considered a pioneer in applications from the field of Big Data. Historically, orthopaedics has been the leading medical branch that established national and international databases (e.g., joint replacement registries) with a huge collection of procedure-specific data repositories [8]. Considering that a significant obstacle in the rapid advancement of DL is the requirement for adequate data to optimise performance and accuracy, the maturity of orthopaedic databases, with tens to hundreds of thousands of patients in individual national registries, is a significant boon for DL development. These factors have influenced the rapid research and development (R&D) of numerous DL-based techniques and applications in orthopaedics.

Roughly, Artificial Intelligence (AI) is defined as the ability of machines to process information similar to human intelligence [9]. Over the years, AI has achieved significant advances through Machine Learning (ML), which has demonstrated the capability of machines to conduct diagnostic tasks for medical imaging with an outcome similar to that of human specialists [4]. These performances have been achieved using novel contributions based on DL, e.g. Deep Convolutional Neural Networks (DCNN), which consist of multiple algorithmic layers that approximate the structure of a human visual cortex [10,11].

Specifically, DL algorithms have shown the ability to perform a wide range of radiographic tasks that involve reading musculoskeletal images [12]. The accurate and precise identification of the types of prosthetic implants has been reported for hip [13], knee [14], and shoulder [15,16] arthroplasty using DL models. The accuracy of these results was equivalent to or superior to that of expert human readers, while also being much faster than traditional methods. This opens obvious avenues to alleviate the workload of physicians in identifying orthopaedic implants from radiographs. Despite such outstanding findings, there is a lack of a systematic assessment of the extent and operation of AI algorithms to classify orthopaedic implants in patients. There are some reviews on the application of machine learning/deep learning in orthopaedics [4,9,17–19], but none are as comprehensive as our review. Existing reviews have not focused on challenges and possible solutions. Furthermore, there has been little discussion on how to create trustworthy DL systems for this field. Moreover, there are no details on technologies, software, or fusion techniques. This review aims to assist researchers and practitioners by highlighting challenges and providing information on trustworthy DL systems that can support the orthopaedic sector. We have formulated and will address the following five questions that are of great importance and are attracting increasing attention from the medical field of orthopaedics.

- *What are the main* ***applications*** *of DL in orthopaedics*?
- **Contribution#1:** DL applications in orthopaedics have been described, including fracture classification, prediction of arthroplasty implants, bone age regression, MRI-based tear segmentation, and soft tissue disease segmentation.
- *What are the major* ***challenges*** *encountered in applying DL techniques for orthopaedic tasks, and what are the potential* ***solutions*** *to overcome them?*
- **Contribution#2:** Data scarcity and lack of interpretability are challenges in applying DL techniques for orthopaedic tasks. Potential solutions include data enhancement, transfer learning, and the development of interpretable DL models.
- *What are the main* ***technologies*** *associated with DL for tasks in orthopaedics?*
- **Contribution#3:** Technologies associated with DL for ortho paedics encompass integration with robotic surgery, mixed reality (MR), wearable sensors, and 3D printing, enhancing preoperative software for diagnosis, planning, and outcome prediction.
- *What are the requirements needed for developing* ***reliable*** *DL applications in the field of orthopaedics?*
- **Contribution#4:** Requirements for developing reliable DL applications in orthopaedics entails ensuring data quality, model reliability, and a transparent process, which includes data verification, model validation, and compliance with regulatory standards such as FDA approval.
- *What techniques can be employed to address the* ***integration*** *of multimodal data in Orthopaedics?*
- **Contribution#5:** Techniques employed to integrate multimodal data in orthopaedics include feature fusion, image fusion, decision fusion, and multi-modal fusion, enhancing diagnostic and predictive capabilities.

The structure of this review is as follows:

2. Review methodology

Four digital databases were utilised to search for the target publications: Science Direct (SD), Scopus, IEEE Xplore (IEEE), and Web of Science (WoS). SD provides reliable technological, scientific, and engineering references. Scopus contains reliable resources in several fields, including medicine, health, technology, science, and engineering. The IEEE includes a comprehensive technical and scientific literature on electrical engineering, electronics, and computer science. WoS contains all cross-disciplinary research papers in science, technology, art, and social science. These databases provide comprehensive insights for researchers by covering most research disciplines from a scientific and technological perspective. We have focused on scientific publications between Jan 2017 and March 2023. The articles of this literature review were selected based on the following queries: ("Deep learning" AND "artificial intelligence" AND "orthopaedics" OR "orthopedics" OR "orthopaedic surgery" OR "orthopedics surgery"), ("Deep learning" AND "orthopaedics" OR "orthopedics"), ("Deep learning" AND "fracture Detection"), ("Deep learning" AND "Supraspinatus Tears Detection"), ("Deep learning" AND "Osteoarthritis"), ("Deep learning" AND "Bone Age Assessment"), ("Deep learning" AND "Transfer learning"), ("Deep learning" AND "MURA"), ("Deep learning" AND "Interpretability" OR "orthopaedics" OR "orthopedics"), ("Deep learning" AND "Adversarial Attacks" OR "orthopaedics" OR "orthopedics"), ("Deep learning" AND "Robotic surgery" AND "orthopaedics" OR "orthopedics"), ("Deep learning" AND "Mixed Reality" AND "orthopaedics" OR "orthopedics"), ("Deep learning" AND "Wearable sensors" AND "orthopaedics" OR "orthopedics"), ("Deep learning" OR "Preoperative

Software” AND “orthopaedics” OR “orthopedics”), (“Deep learning” AND “Trustworthy” AND “orthopaedics” OR “orthopedics”), (“Deep learning” AND “Fusion Techniques” OR “orthopaedics” OR “orthopedics”). The keywords were selected based on the recommendations of AI experts and orthopaedic surgeons.

3. Motivation: Deep learning in diagnosis of orthopaedics

Orthopaedic practice relies heavily on radiological imaging, with the most frequent modalities including ultrasonography, radiography, computed tomography scanning, and magnetic resonance imaging. Most injuries are initially addressed in an emergency department or by a primary care practitioner, who requests some form of imaging. In some cases, the radiographic images may be inconclusive or insufficiently good to properly diagnose an injury that requires a more advanced imaging modality. An orthopaedic specialist may request supporting opinions or a more sensitive scanning analysis for further treatment. The delay in initiating treatment or misdiagnosis of a traumatic lesion may result in poor recovery of the patients, increasing the need for complex medical management, resulting in a higher cost of treatment [20,21].

Additionally, missed or occult fracture detections contribute to a substantial portion of medicolegal claims. For example, a scaphoid fracture often leads to an elevated incidence of avascular necrosis [20]. As many as 20% of scaphoid fractures are missed radiologically, and detection of these fractures can be enhanced through serial imaging, reasonable prospects, or an urgent protocol for MRI or CT scan [20,22]. Another example of missed fracture cases is Lisfranc fracture-dislocations, of which roughly 20% of cases are missed [23], increasing to 50% if the X-ray images are misinterpreted [24]. The risk of missed fractures can be minimised by early determination of the mechanism that caused the injury, a well-planned image protocol, and a low threshold to use a more advanced imaging modality, which could reduce the rates of morbidity and cost of care. Using DL in orthopaedic diagnostics has the potential to reduce the risk of missed injuries and achieve more timely treatment. Possible applications include enhanced upstream applications, such as rapid image acquisition and more effective protocols, as well as downstream functions, such as computerised image analysis and interpretation [11]. Specific applications include:

- DL utilisation to improve the precision of surgical planning. 3D models of bones and joints can be utilised to simulate surgical procedures and make outcome predictions. This can lead to more effective and less invasive surgical procedures, resulting in faster recovery times for patients.
- DL in orthopaedics can also assist in reducing the workload of healthcare professionals, leading to more efficient and cost-effective healthcare delivery.
- DL can offer more accurate and efficient diagnosis and treatment planning. Most notably, DL can be utilised to analyse medical images such as X-rays and MRIs to detect and diagnose traumatic or degenerative lesions and other orthopaedic tasks. This can lead to more accurate and timely diagnoses, resulting in better patient outcomes.
- DL can also help improve the overall quality of life of patients by reducing the time spent in hospitals and facilitating the provision of more personalised treatment plans.

4. Deep learning applications in orthopaedics

Recently, there has been a growing interest in the adoption of DL techniques in various areas of orthopaedics. The emergence of AI and DL technologies in medical imaging is supported by Convolutional Neural Networks (CNNs), which serve as the backbone for this class of algorithms, and leverage enhanced computational processing capacity to perform various tasks, including segmentation, injury recognition, and image reconstruction [25]. Despite the fact that DL methods have received significant attention in diagnostic imaging in recent years, the implementation of DL in musculoskeletal imaging has been largely overlooked compared to imaging in other fields [26], and DL is used primarily in musculoskeletal imaging for anatomical segmentation and injury classification [27]. Given the ever-expanding imaging volumes in current practice, DL has been introduced as a promising workflow aid. Current applications include acting as a ’second observer’ to minimise inaccurate diagnoses and improve efficiency, as well as in non-interpretative usages [28] such as image acquisition and protocolling [29]. Segmentation models are introduced to integrate with applications related to downstream quantitative modelling, with these models being developed out of early developmental stages [30].

This section will explore some of the current and potential applications of DL in orthopaedics.

4.1. *Fracture classification*

Fractures are one of the most common ailments assessed by orthopaedists [31,32]. As such, it is logical that DL methods were first employed in fracture detection. Since then, multiple studies have demonstrated the capabilities of DL algorithms for automated fracture detection in radiographs. These models call can also be extended to include the classification of fracture type (see Fig. 1). Fig. 2 illustrates the typical DL workflow to predict the type of fracture. It is worth mentioning that this workflow is almost the same for all orthopaedic tasks, with small changes based on the target task.

Previously, Kalmet et al. [34] presented a brief overview of DL technology, describing how DL has been used to detect fractures on radiographs and CT examinations. Chung et al. [35] proposed a CNN model to diagnose and classify proximal humerus fractures. The study employed three specialists to identify 1891 anteroposterior shoulder radiographs that had normal proximal humerus (n = 515) or one of four types of proximal humerus fractures (greater tuberosity: 346; surgical neck: 514; 3-part: 269; and 4-part: 247). Subsequently, a dataset-trained ResNet-152 CNN model was generated using the augmented data. Compared to a trained CNN model, the accuracy of 96% was achieved for normal shoulders and proximal humerus fractures, which was higher compared to a general orthopaedist who had an accuracy of 92 8%. The fracture type classification by the CNN model recorded a top-1 accuracy of 65%–86% with an Area Under the Curve (AUC) of 0.90–0.98.

Demir et al. [36] presented a DL model with enhanced classification accuracy to diagnose and classify humerus fractures via a novel stable feature extraction approach. Referred to as the exemplar pyramid method, the model returned an exceptional 99.12% classification accuracy.

Similarly to shoulder fracture classification, several studies have attempted to classify hip fractures by training CNN-based models. Yamada et al. [37] trained an Xception architectural CNN model using 3123 plain hip and lateral radiography images. The trained model was able to classify the fractures with up to 98% accuracy compared to 92.2% by orthopaedists. Urakawa et al. [38] employed hip plain radiographs (1573 normal hip and 1773 intertrochanteric hip fracture images, respectively) to train a VGG-16 CNN model and recorded 95.5% accuracy.

In another study, Lee et al. [39] trained a GoogLeNet-InceptionV3 CNN model using 786 anteroposterior pelvic plain radiographs. The trained model recorded a reasonable accuracy of 86.8% and successfully classified the proximal femur fracture into three types, type A (trochanteric region), type B (femur neck), and type C (femoral head), following the AO/OTA classification system [40]. Lind et al. [40] also trained a ResNet-based CNN model using 6768 images of anteroposterior and lateral knee radiographs. The trained CNN model classified knee radiographic images following the AO/OTA classification system and classified the patellar fractures, proximal tibia fractures, and distal

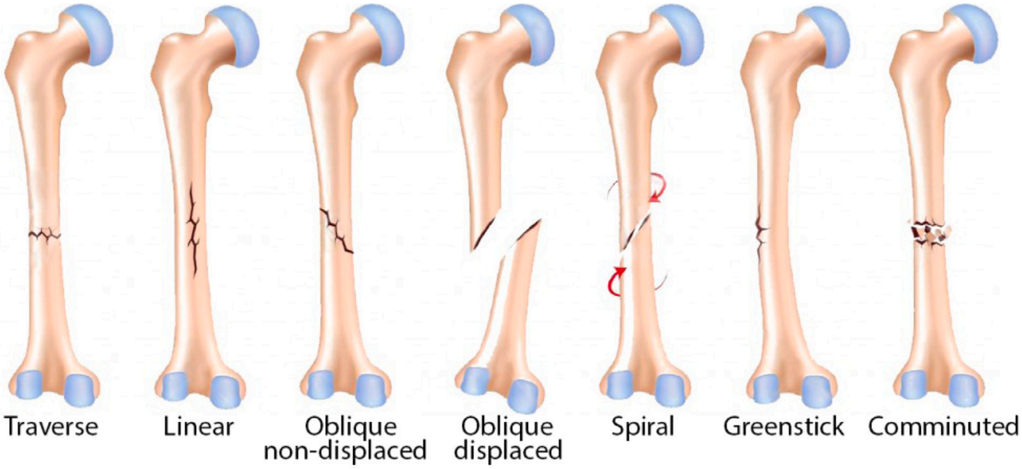


Fig. 1. Femur bone fractures types [33].

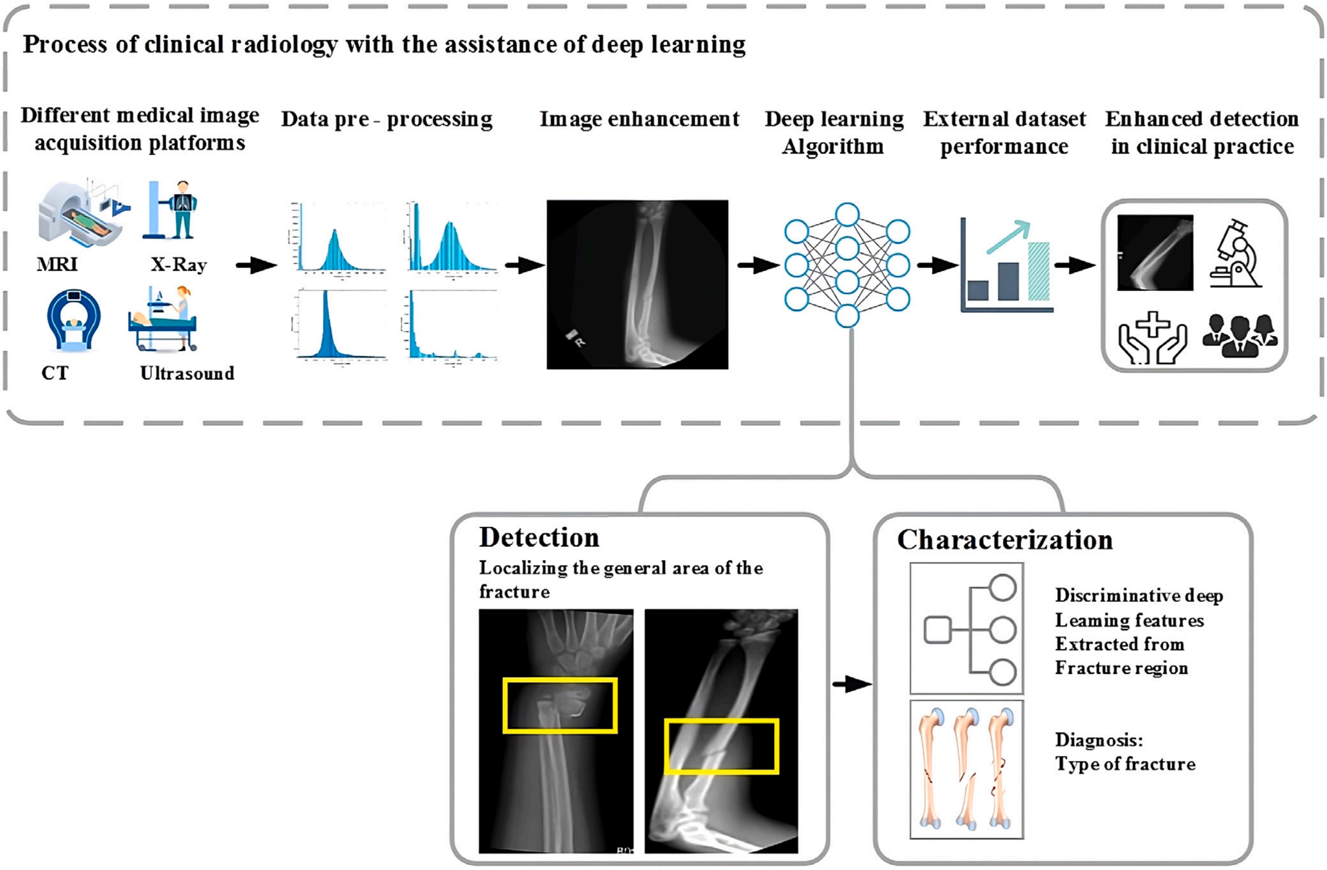

Process of clinical radiology with the assistance of deep learning
Different medical image acquisition platforms
MRI
X-Ray
CT
Ultrasound
Data pre - processing
Image enhancement
Deep learning Algorithm
External dataset performance
Enhanced detection in clinical practice
Detection
Localizing the general area of the fracture
Characterization
Discriminative deep Leaming features Extracted from Fracture region
Diagnosis: Type of fracture

Fig. 2. Fracture classification process with DL.

femur fractures with AUCs of 0.89, 0.87, and 0.89, respectively. In addition, the trained CNN model diagnosed and classified fractures in the large appendices of the shoulder, hip, and knee at a relatively high AUC and accuracy level. Conversely, the trained CNN model only achieved a fairly poor AUC and accuracy when diagnosing and classifying fractures in the small or axial joints.

Farda et al. [41] trained a PCANet-based CNN model to classify calcaneal fractures according to Sanders classification using a dataset containing 5534 CT scans and achieved 72% accuracy. Apart from that, Ozkaya et al. [42] trained a ResNet50 CNN-based model using 390 anteroposterior wrist radiographic images. The trained CNN model recorded an AUC of 0.84, indicating a moderately satisfactory outcome, although the value was lower compared to orthopaedic specialists. The outcome was similar to the report by Langerhuizen et al. [43], where the accuracy of a trained VGG16 CNN-based model was only 72% compared to the higher accuracy by an orthopaedic surgeon (84%). The scaphoid fracture diagnostic was performed using 150 radiographic scaphoid fracture images and 150 radiographic normal wrist images without fracture. A total of 23 out of the 150 scaphoid fracture images could not be assessed using the radiographic images and were only verified via MRI imaging. It should be noted that all orthopaedic surgeons missed five out of six occult scaphoid fractures.

A report was also published by Chen et al. [44], who diagnosed compression fractures in the spine using a trained ResNet-based CNN model. Spinal results seem to be more heavily influenced by the imaging modality used for training. Previous results showed a significant difference with an accuracy of 73.59% when the CNN model was trained using plain spine X-rays compared to Yabu et al. [45] who introduced a trained CNN model using MRI images with a higher accuracy of 88% compared to the surgeons.

In 2023, Wang et al. [46] proposed a diagnostic tool that uses DL to identify and classify fatigue fractures in X-ray images of the tibiofibular and foot regions. They obtained a sensitivity of 95.4%/85.5%, a specificity of 80.1%/77.0%, and an AUC of 0.965/0.877 of internal testing/external validation set, respectively.

In 2022, Meena and Roy [31] evaluated multiple methods for identifying and categorising fractures across multiple regions. They determined that utilising CNN-based models, specifically the InceptionNet and XceptionNet, yields better results than other CNN-based models. Liao et al. [47] presented a new method using CNN attention guidance to assist CNN classification networks in making decisions based on more meaningful visual patterns. This was achieved by incorporating self-attention and human-guided regularisation into state-of-the-art CNN models. The study used a ResNet-50 backbone and showed great improvements in prediction accuracy on evaluated fracture datasets, which are representative of typical data sizes for medical image analysis issues.

Overall, a higher accuracy level is achieved for fracture diagnosis (binary classification) using DL-trained CNN models than fracture classification (multiclass classification), with an expected narrowing of the gap as more advanced CNN models are introduced. It was noted that the classification of fractures of small and axial joints results in poorer accuracy when compared to that of large joints. The poor outcome is viewed as a drawback of CNN-based approaches, which adjudicate the outcome by recognising the contrast information (such as the average margin of the cortical bone and the fracture line or normal joint line) and the spatial information of the images. The authors assume that more powerful CNN models can be utilised to overcome these limitations. So far, DL methods have been employed to diagnose and classify osteoporotic fractures, while the investigation of low-frequency osteoporotic fracture joints has reported fairly poor outcomes. This could be due to the high percentage of osteoporotic fractures observed amongst all types of fractures and the relatively standardised fracture pattern that makes it suitable to be applied in fracture classification.

Generally, the two primary factors governing the performance of the previous DL models for fracture detection are the need for a large number of high-quality images and the interpretability of the results. Table 1 lists the latest state-of-the-art DL methods in fracture classification. Most of the previous fracture classification methods are listed in [48].

4.2. Osteoarthritis and prediction of arthroplasty implants classification

Osteoarthritis is the degenerative wear and tear of the articular cartilage, progressively leading to joint destruction. To date, several investigations have used DL algorithms to diagnose and classify osteoarthritis [29]. Lee et al. [57] summarised the use of AI to diagnose knee osteoarthritis and predict the outcomes of subsequent total knee arthroplasty. The study found that ML-based models show promising results in grading knee radiographs and predicting the need for total knee arthroplasty. This study also showed the ability of AI algorithms to predict patient-reported outcomes and satisfaction after surgery. However, the study also highlighted weaknesses in the model, such as the lack of validation, biases in clinical data, and the need for large datasets for training.

One of the pioneering studies applying DL methods in orthopaedics was performed by Xue et al. [58], who trained a VGG-16 CNN-based model using 420 plain hip X-rays. The diagnosis of hip osteoarthritis by the trained model achieved 92.8% accuracy. Another pioneering study was carried out by Üreten et al. [59], who introduced a similar design model for diagnosing hip osteoarthritis and recorded 90. 2% precision. Meanwhile, Tiulpin et al. [60] reported a multiclass accuracy of 66.7% from a trained Siamese classification CNN model using plain knee X-rays to classify knee osteoarthritis based on the Kellgren–Lawrence grading scale.

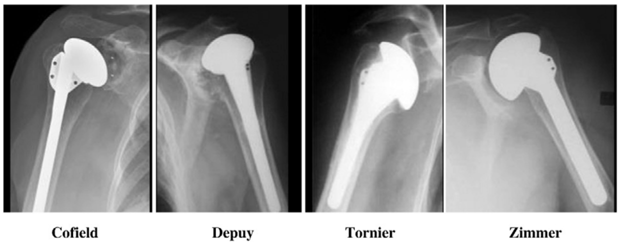


Fig. 3. Types of shoulder implants in X-ray Classification [69].

Furthermore, Swiecicki et al. [61] trained a Faster Region-based CNN (R-CNN) using knee X-rays from the Multicentre Osteoarthritis Study (MOST) dataset. The model recorded a multiclass accuracy of 71.9%, indicating enhanced performance compared to Tiulpin et al. [60]. Pedoia et al. [62] trained a DenseNet-based CNN model using MRI images instead of X-ray data (normally used in previous studies) and achieved a high AUC of 0.83. Furthermore, Kim et al. [63] used 4366 knee anteroposterior X-rays as the data set together with various demographic information (body mass index (BMI), sex, and age), alignment, and pertinent metabolic information that can train a CNN model based on SE-ResNet. The study achieved a significantly higher AUC by coupling image data with additional patient information.

Zhuang et al. [64] presented a new approach to using MRI to diagnose knee osteoarthritis using multiple views and integrating them to improve accuracy, unlike traditional methods which used a single-view MRI. They proposed a Local Graph Fusion Network (LGF-Net), which models multiple MRI views as a unified graph and uses graph-based representation and fusion for osteoarthritis diagnosis.

Given that arthroplasty is frequently needed in the advanced hip or knee osteoarthritis, several researchers have proposed DL-based models to identify arthroplasty implants. Previously, Karnuta et al. [65] trained an InceptionV3 network-based CNN model using anteroposterior knee X-rays with nine varying implant models. The trained model classified implant models at a nearly perfect level with accuracy and AUC of 99% and 0.99, respectively. Borjali et al. [66] performed a similar study on hip joints by training a CNN model using 252 plain hip X-rays with three implant types. The model successfully classified the implants with 100% accuracy. In addition, Kang et al. [67] trained a CNN model using 170 plain hip X-rays with 29 implant types. The trained model achieved an outstanding performance value of AUC of 0.99. However, the small training dataset can raise the issue of overfitting and lack of generalisation. Urban et al. [68] also trained an implant identification CNN model using 597 plain shoulder X-rays with four implant types (see Fig. 3) and 16 different DL models, recording up to 80% accuracy.

Sultan et al. [70] proposed a modified ResNet and DenseNet model to classify different implant designs from four manufacturers, which obtained an accuracy of 85.9%. In 2021, Yılmaz [15] utilised a DL model to classify four types of implants based on CNNs with a new layer through a channel selection formula to improve filter properties. This model achieved an accuracy rate of 97.2%. In 2022, Sivari et al. [71] proposed a combination of 10 different hybrid DL models and ML algorithms for the same task. The best performing model was the integrated DenseNet201-logistic regression model, which had an accuracy of 95.07%. Table 2 lists the latest state-of-the-art DL methods in osteoarthritis and prediction of arthroplasty implants. More details on the previous methods can be found in [72–74].

4.3. Bone age regression

The development of the human skeletal structure is a continuous differentiation process with different maturity markers that can be

Table 1
The state-of-the-art methods of DL applied to fracture classification.

Reference and year	Body part	CNN model	Number of samples	Best results
Yamada et al. [37]	Hip	Xception	3123	Accuracy = 0.98
Lee et al. [39]	Hip	Inceptionv3	686	Accuracy = 0.86
Langerhuizen et al. [43]	Wrist	VGG-16	300	AUC = 0.77
Lind et al. [40]	Knee	ResNet	6768	AUC = 0.89
Farda et al. [41]	Ankle	PCANet	5534	Accuracy = 0.72
Yabu et al. [45]	Vertebra	VGG-16,19, Inception V3,ResNet50	1624	AUC = 0.95
Chen et al. [44]	Vertebra	ResNeXt	1306	Accuracy = 0.73
Oakden-Rayner et al. [49]	Proximal femur	DenseNet	4577	AUC = 0.99
Wang et al. [50]	Mandible	U-Net and ResNet	22 256	AUC = 0.95
Dupuis et al. [51]	Several body parts	DL algorithm (Rayvolve)	5865	AUC = 0.92
Guan et al. [52]	femur	ResNeXt	3842	Precision = 0.88
Ashkani-Esfahani et al. [53]	Ankle	Inception V3 and Renet-50	2100	Sensitivity = 0.98
Wang et al. [46]	Foot	ResNet-50	3993	AUC = 0.96
Huang et al. [54]	Rib	AlexNet , GoogLeNet, EfficientNet, DenseNet201, and MobileNet	2000	Accuracies: 92.6%, 92.2%, 92.3%, 92.4%, 91.2%.
Cheng et al. [55]	Vertebral	YOLOv4 and ResUNet	3634	Precision = 0.99, 0.74, and 0.94
Schilcher et al. [56]	Femur	Fusion techniques	1124+4014	AUC = 0.98

Table 2
The state-of-the-art methods of the DL in osteoarthritis and prediction of arthroplasty implants.

Reference and year	Body part	CNN model	Best results
Xue et al. [58]	Hip	VGG-16	Accuracy = 0.92
Tiulpin et al. [60]	Knee	Siamese CNN	Accuracy = 0.66
Pedoia et al. [62]	Knee	DenseNet	AUC = 0.83
Kim et al. [63]	Knee	SE-ResNet	AUC = 0.75
Üreten et al. [59]	Hip	VGG-16	Accuracy = 0.90
Leung et al. [75]	Knee	ResNet34	AUC = 0.87
Swiecicki et al. [61]	Knee	Faster R-CNN	Accuracy = 0.71
Yılmaz [15]	Shoulder	Multichannel model	Accuracy = 0.97.2
Karaci [76]	Shoulder	YOLOV3 & DenseNet201	Accuracy = 0.84
Sivari et al. [71]	Shoulder	DenseNet201 & Logistic Regression	Accuracy = 0.95

recognised and analysed by paediatricians and radiologists to assess bone age. With this in mind, bone age is a quantitative measure of skeletal maturity [77,78]. The difference between bone age and chronological age is strongly associated with physical growth, such as body size, changes in sex characteristics, the significant appearance of the pubertal growth spurt (fast growth) and the level of endocrine hormones [79–81]. In medical practise, bone age assessment is carried out by analysing specific patterns of skeletal maturity markers on hand-wrist X-ray images of a patient.

Standard clinical bone age assessment techniques include atlas and scoring methods. The Greulich and Pyle (G&P) method is an example of an atlas method [82]. Generally, radiologists compare target X-ray images with the atlas as a reference and use the closest match as the evaluation outcome. However, it is challenging to accurately assess bone age when the target X-ray image is between two proximate atlas references, and the description of bone development is less detailed. In contrast, the Tanner–Whitehouse (TW) approach is a type of scoring method in which radiologists initially observe 20 specific regions of interest (ROI) before evaluating bone age based on the analysis of each ROI [83]. The method has been revised and updated over time, with the latest version being the TW3 method. In the TW3 method, a higher number of parameters and indicators are used compared to the 20 ROIs in the original TW method to build a more detailed description of bone development and improve the accuracy of the assessment. Two common features of the various procedures for bone age assessment is that they are largely subjective and time-consuming. Consequently, a modern radiology department has difficulty obtaining consistent evaluation results within an acceptable error margin. Therefore, several available computer-aided techniques were evaluated to determine factors that aid or hinder the performance of clinical procedures. Evaluation of bone age using computer-aided methods can be grouped into two classes: non-DL and DL. The former basically utilises classic ML techniques and image processing technology [84]. Non-DL methods also utilise hand-crafted visual features from entire images or local informative regions, and the classifiers are constructed using a small-scale private dataset. The results range from 10 to 28 months of Mean Absolute Error (MAE) and are easily affected by hand-wrist X-ray images with unexpected image quality. The generalisation ability of the models is also disputable.

For example, Pietka et al. [85] distinguished bone tissue from other regions by applying numerous window sizes with adaptive thresholds. Additionally, ROIs were employed to produce feature descriptors in terms of geometrical description and pixel property values. Ultimately, the generated feature descriptors were passed into the decision-making approaches to estimate bone age. A recent study illustrated the extraction of epiphyseal and metaphyseal tissues [86]. The feature descriptors were acquired by measuring the critical bone area diameter and the ratios of the crucial distance.

A commercial automated method called the BoneXpert adopts a generative model to generate images while retaining realistic shapes and densities, which collectively resemble bone structure [87]. The features include information on the bone's shape, texture, and intensities. In short, the method implements an automated assessment by mapping functions to generate a relative score depending on the selected TW or G&P methods. Nevertheless, the process would sometimes have to be performed manually since poor-quality images or abnormal bone structures are rejected. Compared to non-DL methods, BoNet [88] is a type of DL method that used a purpose-designed CNN model to extract low- and middle-level feature descriptors and applied an extra layer of deformation to represent non-rigid deformed objects. The evaluation of bone age was then achieved by implementing fully connected layers with an approximate MAE of 9.6 months. Previously, the Radiological Society of North America (RSNA) developed a large-scale bone age assessment data set containing 12,611 images of various resolutions to support the formation and establishment of ML models for medical image analysis [89].

Data processing that comprises multiple subtasks is a significant element in identifying informative regions. In one study, both DL and classic ML were utilised to generate a credible prediction [90]. Pre-trained CNNs were implemented for the DL-based method to extract image features and construct a regressor model automatically. On the other hand, the canny edge detection was adopted for the classic ML method to extract the image features and develop five traditional ML regressors: Random Forest, Linear Regression, XGBoost, Support Vector Regressor (SVR), and Multilayer Perceptron (MLP). Based on the results, the pre-trained CNNs achieved the best performance with an MAE of 14.78 months. A U-Net model was first trained in another study to acquire vital point regions with manually labelled hand masks [91]. A vital point detection model was then utilised to align the hand radiographs into a common coordinate space. Overall, the study achieved an MAE of 6.30 months and 6.49 months for males and females, respectively.

A novel experimental design with manually labelled bounding boxes and key point annotations was proposed during training in [92]. Local information was exploited to perform pose estimation and region detection for bone age assessment. The findings recorded the best RSNA with an MAE of 4.14 months. Despite the high accuracy and efficiency of the large number of models that perform well with accurate manual annotations, additional annotations were considered time-consuming. They restricted the algorithm conversion into useful clinical practice.

In [93], a hybrid model was proposed that combines learning of DL characteristics and a fast ELM method to monitor the skeletal development of children using bone age prediction from hand-wrist radiographs. ROIs were used to assess bone maturity. The previously mentioned Tanner–Whitehouse (TW) method, a common Bone Age Assessment (BAA) alternative, was used to estimate bone age. The proposed method obtained high performance with the RSNA dataset when using a hybrid model of MAE 6.0737.

DL was also used by Li et al. [94] to assess the bone age of a child diagnosed with growth disorders. Unsupervised learning methods with CNNs were used to extract high-level features using a batch normalisation layer and an argmax layer for feature clustering. This work is based on the MLP technique for the prediction head and MobileNetV3 for the backbone, and the MAE was 6.2 and 5.1, respectively.

The deep neural network (DNN) model has been proposed in [95] to assess bone age using a database of paediatric left-hand radiographs. The result has shown that the MAE values for male and female models were between 0.33 and 0.25 years, respectively.

Research by Zulkifley et al. [96] proposed an Attention-Xception Network (AXNet) method to predict bone growth in the paediatric population. This method used ROIs to calculate the rotational alignment module of the hand region based on a key-point detector. The results showed that the proposed framework achieved the lowest MAE and MSE values of 7.699 and 108.869 months, respectively.

In [97], the deep convolutional neural network technique (DCCN) has been proposed to estimate age-based sex information. They used hand-crafted key point detection-based affine transformation to register the hand pose. The MEA obtained was 5.31 months higher than other existing BAA methods.

The hyperparameter optimisation-based DL-based model for the automated assessment and classification of bone age (HPTDL-BAAC) method has been proposed by Palaniswamy [98] to assess bone age and classify it into stages using the Digital Hand Atlas (DHA) database. A SqueezeNet-based regional convolutional neural network (RCNN) mask was used for feature extraction. The result showed that it produced higher average accuracy of 98.30% and an average F-score of 98.31% than other techniques.

Estimating bone maturity using hand-bone radiographs and X-ray images has been proposed in [99]. This research employed Multi-scale Multi-reception Attention Net (MMANet) and Multi-scale Multi-reception Complement Attention (MMCA) network to improve the representation of features of key regions and eliminate background regions' influence. The method achieved a higher performance of MAE 3.88 months using the RSNA Paediatric Bone Age Challenge dataset than other compared methods. The research of Wang et al. [100] proposed multi-instance learning-based attention networks using patch features. These features have been ranked by an attention backbone and aggregated to predict bone age. The proposed method achieved MAE 4.17 months using RSNA 2017 dataset.

Recently, DL techniques have been used to assess age based on the bone structure of the teeth. For example, the study of Upalananda et al. [101] developed the GoogLeNet method to assess the developmental stage of the mandibular third molars using the method of Demirjian et al. [102] classification stages D to H, generating sound results. Table 3 lists the latest state-of-the-art DL methods in bone age assessment tasks.

4.4. Classification and segmentation of supraspinatus tears using MRI

DL methods can help analyse data sets for the diagnosis, management of risk and treatment of patients with musculoskeletal injury, which benefits both patients and their providers [106]. One of the promising applications of DL methods relates to rotator cuff tears. Rotator cuff tears are a broad entity of diagnoses encompassing different combinations of tendon involvement, varying depths and dimensions of tendon tears, diverse tear configurations, and different retraction patterns, all requiring advanced medical imaging for evaluation. In general, tears occur in all demographics [107], and the supraspinatus tendon is the most frequently affected tendon. Cuff tears can be treated with conservative or surgical methods, and decision making is influenced by tear severity and other patient factors [108]. As such, the clinical presentation and imaging features usually dictate the need for surgery. Moosmayer et al. [109], Longo et al. [110]. MRI is the gold standard imaging to diagnose and classify rotator cuff tears, offering a sensitivity of 90% (95% CI: 84%–96%) and specificity of 90% (95% CI: 84%–95%) for binary classification [111]. Fig. 4 shows an example where DL can be applied for this task.

Several DL methods have been employed in shoulder imaging, particularly for the rotator cuff anatomy and segmentation of the glenohumeral joint [27,27,112]. Although three individual studies have reported on the use of both DL [113] and non-DL [114] methods for MR classification of supraspinatus tears, single-scanner data in these studies may not represent the heterogeneity in standard clinical practices, including the variations in magnetic field strength. Thus, one of the feasible applications of DL this paper is reviewing is the detection of supraspinatus tears using MRI. The study also examined the variable accuracy under different MR protocols and injury subtypes to gain further insights into the generalisability of the DL models for future clinical applications. Fig. 4 shows the Goutallier classification system, where the four rotator cuff muscles are classified into five escalating grades of fatty infiltration into the muscle belly. By classifying the quality of muscle–tendon tissue,this system allows clinicians to assess the chronicity and repairability of rotator cuff tears. DL can be used to classify rotator cuff muscles using this system, as explained in the following paragraphs [115].

Shim et al. [113] trained a Voxception-ResNet (VRN)-based CNN model using 2124 shoulder MRIs to assess the presence and size of rotator cuff tears. The identification and classification of these tears were recorded with an accuracy of 92.5% and 76.5%, respectively. Lee et al. [116] constructed an innovative DL architecture based on an integrated positive loss function and a pre-trained encoder, which accurately determined the position of the rotator cuff tear even with imbalanced and noisy ultrasound images.

Kim et al. [117] trained a CNN model using shoulder MRI of 240 patients as the dataset. The trained CNN model achieved an accuracy of 99.9% and determined the muscle area of the rotator cuff with fatty infiltration of high grade. Similarly, Taghizadeh et al. [118] trained a CNN model using shoulder CT of 103 patients as the dataset and measured the fatty infiltration with 91% accuracy. In another study, Medina

Table 3
The state-of-the-art methods of the DL in bone age assessment.

Reference and year	Role	Feature extraction	Methods	MAE
Spampinato et al. [88]	Bone age assessment	low-middle-level feature descriptors	BoNet	9.6
Iglovikov et al. [91]	Aligning hand radiographs	Vital Point region	U-Net	6.3
Wibisono et al. [90]	Identifying informative regions in bone	RF, LR, XGBoost, SVR, and MLP	CNNs	14.78
Escobar et al. [92]	Conducting the pose estimation and region detection	–	BoNet	4.14
Li et al. [94]	assessment of bone age	using CNNs for high level features	MLP and MobileNetV3	6.2–5.1
Cheng et al. [95]	Assess bone age based on pediatric left-hand radiographs	–	DNNA	0.311–0.25
Zulkifley et al. [96]	To detect any anomaly in bone growth among kids and babies	RIOs to get hands key points	AXNet	7.699
Guo et al. [93]	Monitoring the skeletal development of children' hand-wrist	using RIOs to assess the maturity of bone	CNNs and ELM	6.0737
Nguyen et al. [97]	Bone age prediction based sex	using information hand's key-points features	DCNNs	5.31
Palaniswamy [98]	Bone age assessment	using R-CNN based mask with SqueezeNet for features	HPTDL-BAAC	–
Jabbar and Abdulmunem [103]	bone evaluation growth stage of younger	–	DCNN	–
Yang et al. [99]	Estimating bone maturity from hand bone	improving key and background regions features	MMCA and MMANet	3.88
Upalananda et al. [101]	Tooth development of mandibular third molars	–	GoogLeNet	–
Wang et al. [100]	Evaluating children's endocrine, genetic, and growth disorders	Patch features using feature extraction network	DL based on multiple-instance learning	4.17
Rassmann et al. [104]	Bone age assessment validated on skeletal dysplasias	Patch features using feature extraction network	Convolutional neural network models	3.87 & 5.84
Wu et al. [105]	Bone age prediction in children	Patch features using feature extraction network	Vision transformers	0.5 & 0.4

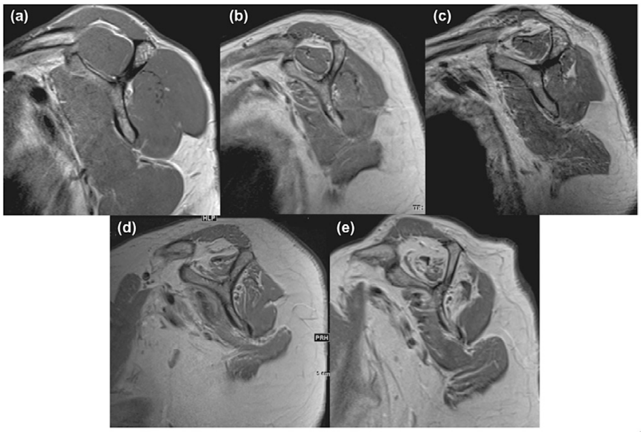


Fig. 4. The Goutallier classification: oblique-sagittal proton density-weighted images show different degrees of fatty degeneration of the supraspinatus muscle: normal = grade 0 (a), some fat streaks = grade 1 (b), less fat than muscle = grade 2 (c), as much fat as muscle = grade 3 (d), and fatter than muscle = grade 4 (e) [115].

et al. [119] segmented rotator cuff muscles on 258 shoulder MRIs with mean Dice scores > 0.93.

Ro et al. [120] developed a DL model to detect and evaluate rotator cuff tears on shoulder MRI scans. They used a DL algorithm to detect the supraspinatus muscle and its fossa and calculated the occupation ratio of the muscle in the fossa. The authors also evaluated the fatty infiltration of the muscle using an automated region-based Otsu thresholding technique. They found that the proposed CNN was able to detect and evaluate the occupation ratio and fatty infiltration accurately, and these had a moderate negative correlation.

In 2022, Yao et al. [112] used DL to detect tears in the supraspinatus tendon using MRI images automatically. The study used 200 shoulder MRI scans and divided them into three categories: full-thickness tears, partial-thickness tears, or intact tendons. They created a 3-stage process using different types of computer networks to analyse the images and then compared the results to the findings of radiologists. They found that the DL was able to detect tears with a high level of accuracy, with a test sensitivity and specificity of 85%. They also found that the DL was able to detect full-thickness tears with 100% sensitivity.

In 2023, Lin et al. [27] aimed to develop a DL model to detect and classify rotator cuff tears on shoulder MRI images. They used 11,925 MRI scans from 2 institutions, with 11,405 for training the program and 520 for testing. They used an algorithm that analysed four different series of images from each scan and used the radiologist's report as the "ground truth" for what the DL should be looking for. They found that the DL was able to detect and classify tears with high accuracy, with an overall AUC of 0.93. The DL performed especially well detecting full-thickness tears, with an AUC of 0.98. They also found that the DL's accuracy was similar to that of radiologists.

Based on the aforementioned studies, it can be stated that DL has become a feasible approach for detecting supraspinatus tears on MRI. Extra training data is required to further characterise the tears.

4.5. Joint-specific soft tissue disease segmentation

Specialised DL algorithms for diagnosis based on learned images have key structural differences from algorithms used for segmentation

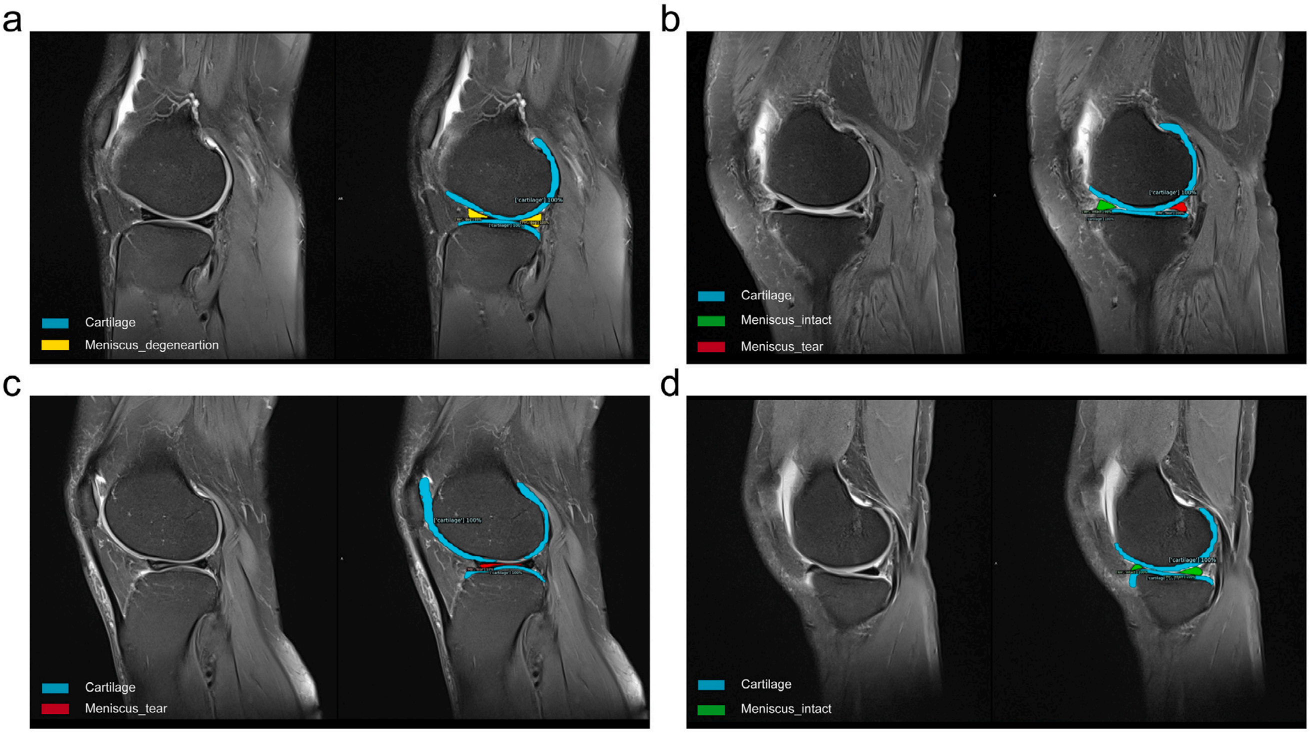


Fig. 5. Example of results on meniscus MR images from [123]. (a) Degenerations in the anterior and posterior meniscal horns, (b) Tears in the meniscal body, (c) Tears in the posterior horn, (d) Healthy meniscus.

using analysed features. As such, both algorithms have been developed for specific applications in different areas [121]. Segmentation algorithms suffer from technical complexity, which is easily lost in the outer layer process of synthesising the results of the CNN model during training. It is necessary to preserve spastic information [122]. Several techniques, such as the Fully Convolutional Network (FCN)-based semantic segmentation, have been assessed to overcome these limitations. The presence of numerous DL algorithms also influences the level of utilisation of DL in the field of orthopaedics.

The aforementioned DL-based studies represent study cases that used X-ray images for diagnosis and classification, which often does not require specialised CNN models for segmentation [122]. Conversely, a satisfactory accuracy level can only be achieved using specialised CNN models to segment diseases diagnosed and classified based on MRI or ultrasound images. For example, a CNN model is more suitable for diagnosing rotator cuff tears based on the normal outline of the rotator cuff (segmentation) than a diagnostic approach specifying the point of the tear occurrence (regional detection). Thus, the use of CNN models to diagnose soft tissue disorders in orthopaedics only appeared after 2018, when the segmentation technology was fully developed [112].

Recent publications have proposed the use of a CNN model to diagnose cartilage lesions, meniscal tears, and ruptures of the anterior cruciate ligament (ACL) in the knee joint [124]. According to Couteaux et al. [125], a Mask-RCNN model was trained using 1828 knee MRI images to identify the torn segment of the meniscus from the normal segment and classify the tear according to its position. The model diagnosed and classified meniscal tears with a high AUC of 0.91. A comparable model was also reported by Roblot et al. [126], which also recorded an exceptional AUC of 0.94. Additionally, Chang et al. [127] proposed a trained CNN model based on U-Net to diagnose complete ACL tears using 320 MRI images, which achieved an excellent AUC of 0.97. Flannery et al. [128] assessed the segmentation of a modified trained CNN model based on U-Net, which showed a statistically insignificant difference with respect to ground truth, represented by the actual value suggested by an expert.

Li et al. [129] investigated the value of using magnetic resonance imaging to diagnose anterior cruciate ligament (ACL) injuries using a multimodal fusion model based on DL. 30 patients with knee injuries were diagnosed using both MRI and arthroscopy. Empirical results showed that DL-based magnetic resonance imaging achieved a high precision of 96. 28% in predicting ACL tears.

Li et al. [123] aimed to improve the diagnostic accuracy and efficiency of meniscal tears. A DL model was proposed to be trained using standard knee MRI images from 924 patients. The Mask Regional Convolutional Neural Network (R-CNN) and ResNet50 were used to build the structure of the DL network. The results showed that the DL model accurately recognised healthy and injured menisci with an average precision ranging from 68% to 80% and diagnostic accuracy for healthy, torn and degenerated menisci of 87. 50%, 86. 96%, and 84. 78%, respectively (see Fig. 5). Validation in an external data set showed that the precision of diagnosing intact and torn meniscus tears through 3.0T MRI images was greater 80% and greater 50% when verified by arthroscopic surgery.

Recently, Key et al. [130] proposed a method for detecting meniscal tear and anterior cruciate ligament (ACL) injuries using MR imaging. MR images were collected in three different slices (sagittal, coronal, and axial) and grouped into three data sets (sagittal database (sDB), coronal database (cDB), and axial database (aDB)). The proposed method employs deep feature extraction using the AlexNet architecture, significant feature selection using the iterative RelifF algorithm, and classification using the k-nearest-neighbour (kNN) method. The proposed method was applied to three data sets and achieved high precision with 98. 42% for sDB, 100% for cDB and 100% for aDB.

This DL model demonstrated excellent potential in diagnosing specific soft tissue diseases in joints. However, its accuracy can be further enhanced by increasing the training sample size, which will be discussed in the Challenges section, along with strategies to overcome the issue of limited data availability.

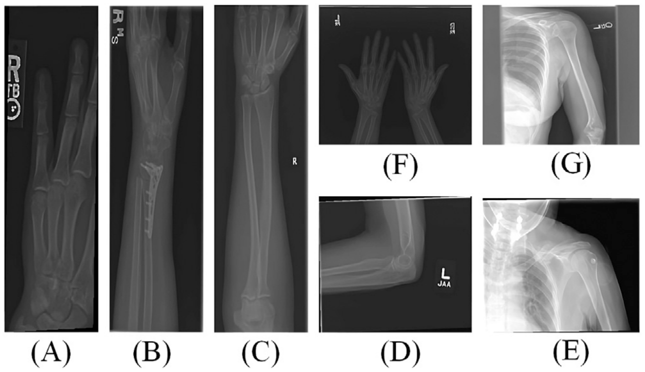


Fig. 6. The MURA dataset classes where (A) Fingers; (B) Wrist; (C) Forearm; (D) Elbow; (E) Shoulder; (F) hands; (G) Humerus.

4.6. Other applications

In addition to the groups of applications mentioned above, there are other miscellaneous applications for DL in orthopaedics. For example, DL can be widely adopted in the detection of bone density for the identification of osteoporosis from radiographs, with high accuracy [131–134]. DL can also be used to check whether a bone tumour is malignant or not [135–137]. von Schacky et al. [136] developed a DL model to categorise bone tumours, evaluated 934 patient radiographs and achieved 80. 2% accuracy of the model in the classification of malignant or benign tumours.

DL can also be implemented in the diagnosis of paediatric development dysplasia of the hip (DDH) [138–140]. Zhang et al. [138] studied 1130 patients with an average age of 1.5 years, trained a DL model using 9081 radiographs; 1138 test radiographs, and showed that the effectiveness of the DL system for hip development dysplasia diagnosis is greater than that of the clinician-led diagnosis.

Another application of DL in musculoskeletal imaging is the automatic determination of clinically important geometric angles and lengths. Automatic measurement of the hip-knee-ankle angle (HKA) is another application of DL in orthopaedics [141]. The authors used DL to automatically segment the hip, knee, and ankle in the X-ray images. They determined the HKA angle accordingly, showing that the difference between the automatic HKA angle measurement and the average manual measurement of 3 orthopaedic surgeons was 0.49° on average.

DL can also be applied in automated Cobb angle measurement [142–144]. Horng et al. [142] obtained the vertebrae segmentation via the DL-based neural network, calculated the Cobb angle result, compared it with the manual results of 2 physicians, and stated that these results are similar, demonstrating the ability of the DL model to help doctors in the clinical diagnosis and treatment of scoliosis.

DL models, such as the developed convolutional neural network (CNN), can be applied for automated angle measurement in flatfoot diagnosis. The model exhibited superior accuracy and reliability in angle measurements compared to human observers, highlighting the potential of DL to improve diagnostic precision. The guidance provided by the CNN resulted in reduced errors and a more efficient measurement process for human observers, showcasing the practical benefits of employing DL in medical diagnostics [145].

A modified U-Net architecture was developed to diagnose vertebral compression fractures (VCF) and detect vertebral levels in lumbar spine lateral radiographs (LSLRs) simultaneously. The multi-task model showed better performance in sensitivity and area under the receiver operating characteristic curve compared to the single-task model. During internal and external validation, the model demonstrated high accuracy, sensitivity, and specificity for fracture detection per patient or vertebral body and successful vertebral-level detection. This suggests that the model has the potential to assist radiologists in real-life medical examinations [146].

Cascade Convolutional Neural Network algorithms, such as the newly developed Flatfoot Landmarks Annotating Network (FlatNet), have proven to be valuable for accurate and efficient landmark detection in flatfoot radiographs. This is crucial for the analysis of foot deformities. The DL model has outperformed human observers in the identification of X- and Y-coordinates, with the average difference in absolute distance being reduced under FlatNet guidance. The overall accuracy and reliability of landmark identification have improved for both experienced orthopaedic surgeons and a general physician. This highlights the potential of FlatNet to improve diagnostic precision in the analysis of foot deformities [147].

The use of U-Net in semantic segmentation for weight-bearing lateral radiographs, especially for flatfoot-related deformities, is an effective and precise method. The active learning strategy has shown better values of the Dice similarity coefficient (DSC) and Hausdorff distance (HD) compared to learning with a pooled dataset. Active learning has also reduced angle measurement errors based on segmentation results, with a shorter labelling time of 0.82 min, compared to learning with a pooled dataset which took 0.88 min. These results indicate that active learning is a promising strategy for developing accurate and efficient flatfoot classifiers through semantic segmentation in radiographic analysis [148].

A novel approach for detecting and diagnosing adolescent idiopathic scoliosis in chest X-rays (CXRs) involves utilising the discriminative ability of the latent space in a generative adversarial network (GAN) and a simple multi-layer perceptron (MLP). Trained in a two-step process, the GAN served as a feature extractor for various scoliosis severities, and an optimised 2-layer MLP achieved the best classification results. The model exhibited an area under the receiver operating characteristic (AUROC) of 0.850 in internal and 0.847 in external datasets, with a specificity of 0.697 and 0.646, respectively, at a fixed sensitivity of 0.9. This innovative classifier, based on learning of generative representations, demonstrates promising diagnostic capabilities for adolescent idiopathic scoliosis in chest radiograph screening [149].

In conclusion, DL has enormous potential to revolutionise the field of orthopaedics by working with various imaging applications. With the use of DL, orthopaedic diagnosis can become more efficient and accurate, improving surgical outcomes and reducing the risk of complications. DL is likely to play an increasingly important role in orthopaedics by providing new and innovative solutions to complex medical problems.

5. Datasets

5.1. Public datasets

There are several public datasets in the area of orthopaedics with possible applications in DL. We have summarised the top ten datasets for different orthopaedics tasks as listed in Tables 4 and 5.

5.2. MURA dataset

The MURA (Musculoskeletal Radiographs) dataset is a large collection of X-ray images of various bones in the human body. The data set includes images of seven different skeletal bones: elbow, finger, forearm, hand, humerus, shoulder, and wrist (see Fig. 6). Each of these bones is divided into two subclasses: positive (abnormal) and negative (normal). The total number of images in the data set is 40,561. The data set is divided into training and test sets, as detailed in Table 6. This data set is useful for training ML models to detect and diagnose abnormalities in these bones, which can help diagnose and treat musculoskeletal conditions. The data set was introduced in 2017 by Rajpurkar et al. [157]. The dataset can be downloaded from the following link: https://stanfordmlgroup.github.io/competitions/mura/. Table 7 lists the latest DL methods using the MURA data set.

Table 4
Public dataset in the area of orthopaedics.

Dataset	Number of samples	Task	Download link	Ref.
RSNA benchmarking dataset of American Society of Neuroradiology (ASNR) and the American Society of Spine Radiology (ASSR).	Number of Patients: 2019 Patients with fracture: 961	CT scans of the cervical spine (neck)	https://www.kaggle.com/code/andir16/rsna-2022-cervical-spine-fracture-detection-eda	–
Knee Osteoarthritis Dataset with Severity Grading: Grade 0: Healthy knee image. Grade 1 (Doubtful): Doubtful joint narrowing with possible osteophytic lipping Grade 2 (Minimal): Definite presence of osteophytes and possible joint space narrowing. Grade 3 (Moderate): Multiple osteophytes, definite joint space narrowing, with mild sclerosis. Grade 4 (Severe): Large osteophytes, significant joint narrowing, and severe sclerosis.	Val = 826 Test = 1656 Train = 5778	This dataset contains knee X-ray data for both knee joint detection and knee KL grading.	https://data.mendeley.com/datasets/56rmx5bjcr/1	Chen [150]
The MRNet dataset consists of 1370 knee MRI exams performed at Stanford University Medical Center labels were obtained through manual extraction from clinical reports.	The dataset contains 1104 (80.6%) abnormal exams, with 319 (23.3%) ACL tears and 508 (37.1%) meniscal tears	Abnormal knee, meniscal tears and ACL tears	https://stanfordmlgroup.github.io/competitions/mrnet/	Azcona et al. [151]
RSNA public dataset containing radiological images are taken for the BAA study, such as its statistical features, training the proposed DL models and evaluating the performance on predicting bone ages. Here, the dataset is taken from the Pediatric Bone Age Challenge (RSNA, 2017) organised by the Radiological Society of North America (RSNA).	12,611 images with labels, which consists of 54.2% male and 45.8% female infants' hand images	Bone age assessment	https://www.rsna.org/education/ai-resources-and-training/ai-image-challenge/rsna-pediatric-bone-age-challenge-2017	Halabi et al. [89]
IRMA public data set compiled anonymous radiographs, which have been arbitrarily selected from the routine at the Department of Diagnostic Radiology, Aachen University of Technology (RWTH), Aachen, Germany. The images consisted of different ages, genders body bone parts.	15,363 images of 193 categories of bone Training data: 12,677 radiographs with known categories Class distribution: Distribution of classes in the training data. Test data: 1733 radiographs without classification.	Bone age assessment	https://www.kaggle.com/datasets/raddar/irma-xray-dataset	Karthik and Kamath [152]

Table 5
More of a public dataset in the area of orthopaedics.

Dataset	Number of samples	Task	Download link	Ref.
The Osteoarthritis Initiative (OAI) database contains the permanent archive of the clinical data, patient-reported outcomes, biospecimen analyses, quantitative image analyses, radiographs (X-rays) and magnetic resonance images (MRIs) acquired during this study.	There are bone assessments and measurements from 4796 subjects, with data from over 431,000 clinical and imaging, and almost 26,626,000 images Men and women ages 45–79 With, or at risk for, symptomatic femoral–tibial knee OA All ethnic minorities (focus on African–Americans).	Bone age assessment, knee, hand, foot	https://nda.nih.gov/oai	Soh et al. [153]
This dataset of X-rays of wrist fracture for male and female collected from Al-huda Digital X-ray Laboratory, Pakistan.	193 x-ray images of wrist fracture and normal	Wrist fracture	https://data.mendeley.com/datasets/xbdsnzr8ct	Malik et al. [154]
RibFrac dataset is a benchmark for the development of algorithms for rib fracture detection, segmentation, and classification. This is a large-scale dataset that could facilitate clinical research for automatic rib fracture detection and diagnoses.	RibFrac Training Set has divided into two parts: Training Set Part 1 of RibFrac dataset, including 300 CTs and their corresponding annotation images.	Rib fracture	https://zenodo.org/record/3893508#.Y851dMlBw2w	Jin et al. [155]
Osthersit dataset: is formed from thermal knee images which are composed by radiologists from Trichy and Chennai. Images were collected from the standard scan centres in Tamilnadu after ethical clearance.	This dataset consists of 100 OA thermal images which are collected from 30 cases. Among the 30 subjects, OSTHERSIT consists of 9 males subjects and 21 females subject cases.	Osteoarthritis Knee	https://sethu.ac.in/osthersit/	Lohchab et al. [156]

Table 6
Number of images of the MURA dataset.

Type of	Training		Testing	
bone	Negative	Positive	Negative	Positive
Elbow	2925	2006	234	230
Finger	3138	1968	214	247
Hand	4059	1484	271	189
Humerus	673	599	148	140
Forearm	1164	661	150	151
Shoulder	4211	4168	285	278
Wrist	5765	3987	364	295

5.3. MedShapeNet dataset

MedShapeNet stands as a pivotal resource in the field of medical shape analysis, providing an extensive collection of anatomical shapes and surgical instrument models [170]. Unlike traditional shape descriptors, which were prevalent prior to the deep learning era, MedShapeNet leverages contemporary algorithms, predominantly diverging from computer vision approaches. Currently, MedShapeNet encompasses 23 datasets comprising over 100,000 shapes, each paired with annotations (ground truth). These datasets are freely accessible through a user-friendly web interface and a Python application programming interface (API). They can be utilised for a wide range of purposes, including discriminative, reconstructive, and variational benchmarks, as well as applications in virtual, augmented, or mixed reality and 3D printing. Several exemplary use cases of MedShapeNet include brain tumour classification, skull reconstructions, multi-class anatomy completion, educational purposes, and 3D printing applications. The dataset page (https://medshapenet.ikim.nrw/). The dataset is a new one with only a few publications available on it Jayakumar et al. [171], Li et al. [172], Krieger et al. [173] and Luijten et al. [174]. MedShapeNet serves as a valuable resource for various applications in the medical domain, including orthopaedics, as shown in Fig. 7.

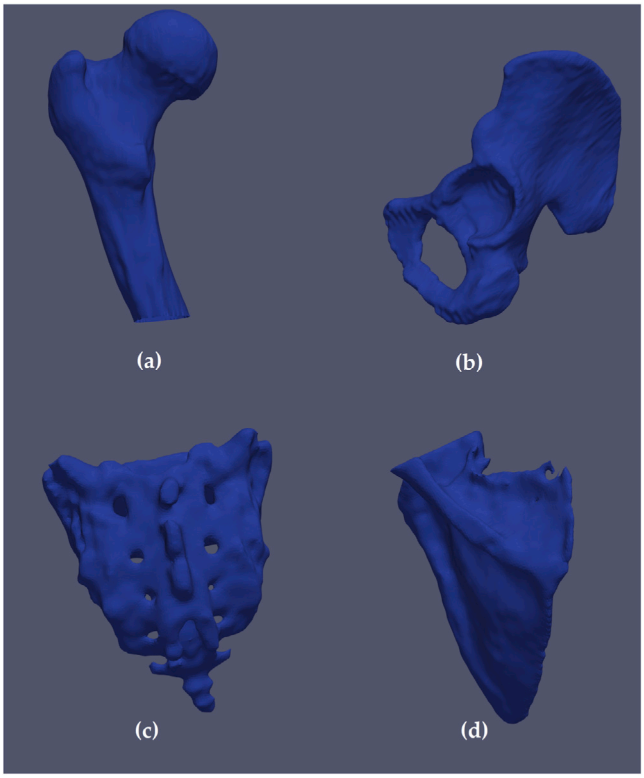


Fig. 7. Some of the orthopaedics tasks in MedShapeNet dataset where (a) Left femur; (b) Left hip; (c) Sacrum; (d) Scapula.

6. Deep learning challenges & solutions in orthopaedics

DL has been gaining momentum in the field of orthopaedics as a powerful tool for image analysis, surgical planning, and rehabilitation. However, as with any new technology, there are challenges that need to be addressed. These challenges include the need for a large number of high-quality images and the interpretability of the results. In this section, we will explore some of the main challenges facing the application of DL in orthopaedics and discuss potential solutions to overcome them.

6.1. Data scarcity

Advancements in the development of DL and increased access to larger datasets have led to the emergence of medical AI. Algorithm-based medical AI has been designed to perform specific medical tasks in terms of diagnosis, prediction, and recommendation of appropriate treatments on a wide range of medical modalities and data types [112,175]. A significant challenge during the construction of algorithms for medical AI is the heavy dependence on the availability of annotated input data, frequently at a scale of hundreds of thousands (if not millions) of data points. Overcoming this drawback would extend the advancement of accurate and precise AI algorithms to cover a wider variety of tasks in healthcare and combating disease, ranging from rapid diagnostics to effective monitoring and reliable treatment decisions.

Despite DL's significance in the medical AI space, its adoption depends largely on the availability of large-scale annotated datasets. Typically, DL models are trained using a supervised learning model in which the proposed model learns to map input data (for example, a shoulder MRI image or a hip X-ray image) to output data (such as the detection of supraspinatus tears on MRI or the detection of fractures on X-ray). The training process of these models through supervised learning requires large datasets. Each input is annotated with its corresponding output so that the model can learn appropriate patterns in the data. Instead of focussing on building annotated datasets, a greater focus has been on constructing and evaluating the models. This emphasis is in part due to the extremely high cost of building the required datasets for most medical tasks [176].

As such, there is a relatively weak commitment to invest the resources required to establish annotated data sets compared to the resources invested in model design. Although the currently available datasets have been used repetitively for common image types, such as skin lesion images, chest radiographs, and brain CT scans, task-specific orthopaedic datasets are lacking [11,158].

As previously mentioned, creating large-scale medical datasets that experts precisely annotate is difficult when compared to annotating nonmedical datasets. For example, remarkably, non-medical DL models have been successfully trained on ImageNet, where 49,000 Amazon Mechanical Turk workers (a crowd-sourcing marketplace for outsourcing tasks that require human intelligence) and hundreds of citizen scientists and academics labelled up to 15 million images from 21,000 classes (such as 'hummingbird' and 'broccoli') [177]. On the contrary, considerable time and expert input from the medical field are required to label medical datasets. Compared to other studies, more time is needed to interpret and label each medical image, such as a shoulder magnetic resonance image or tissue slide image, than is required to label diagnostic data or other natural objects in clinical applications. For example, sample images can be labelled in ImageNet at an average rate of 50 images per minute, while shoulder magnetic resonance imaging could take an average of 2–5 min per case to label. Krizhevsky et al. [177].

In addition, developmental time and domain expertise are essential for automated labelling methods. This approach enabled poorly supervised learning, which is a technique that leverages imprecise or

Table 7
The state-of-the-art DL methods using the MURA dataset.

Reference	Task	CNN Model	Best results	Limitations
Rajpurkar et al. [157]	All MURA tasks	DenseNet	AU-ROC of 0.929, 0.815 sensitivity and 0.887 specificity	TL from different domain
Kandel et al. [158]	All MURA tasks	VGG, Xception, ResNet, GoogLeNet, Inception ResNet, DenseNet	Accuracy of 84.88% for the elbow dataset.	TL from different domains and lack of interpretability
Kandel et al. [159]	All MURA tasks	VGG19, InceptionV3, ResNet50, Xception, and DenseNet	Xception using TL without FC achieved 83.58% accuracy for the elbow images.	TL from different domains and no performance explainability.
He et al. [160]	All MURA tasks	ConvNet, ResNet, and DenseNet	AUC: 0.97, Accuracy: 0.93, Precision: 0.90, Recall: 0.97, Cohen's kappa: 0.85.	Not tested against different changes.
Uysal et al. [11]	Shoulder	Thirteen DL-based models: ResNet (34,50,101,152), ResNeXt (50,101), DenseNet (169,201), VGG (13,16,19), InceptionV3, and MobileNet	Ensemble model EL2: Accuracy: 0.8472, Precision: 0.85, Recall: 0.845, F1-score: 0.845, Cohen's kappa: 0.6942.	TL from different domains and no performance explainability.
Liang and Gu [161]	All MURA tasks	Multi-scale convolution neural network (MSCNN-GCN)	Average Cohen's kappa score = 0.83	Imbalanced classes for some tasks.
Saif et al. [162]	All MURA tasks	Capsule Network	Average Cohen's kappa score = 0.80	Small dataset.
Fang et al. [163]	All MURA tasks	Iterative fusion convolutional neural network (IFCNN)	Average accuracy = 0.73	Imbalanced classes for some tasks.
Harini et al. [164]	Finger, Wrist, Shoulder	Xception, Inception v3, VGG-19, DenseNet, and MobileNet	Accuracy = 0.56 in wrist	TL from different domain.
Malik et al. [165]	Elbow	Xception and DarkNetwork-53	Accuracy of 97.1% and a kappa score of 94.3%	No performance explainability.
Alammar et al. [166]	Humerus, wrist	Feature fusion	Accuracy of 87.8% for humerus%	Requires high computational resources.
Alzubaidi et al. [167]	Forearm	Feature fusion	Accuracy of 90.7%	Requires high computational resources.
Kumar et al. [168]	All MURA tasks	Multistage feature map	Average accuracy 85% for all tasks.	Requires high computational resources.
Alzubaidi et al. [169]	Shoulder	Feature fusion	Accuracy of 99.2% and a kappa score of 98.5%	Requires high computational resources.

noisy sources to reduce the burden of obtaining hand-labelled datasets. One of the related challenges when studying complex topics like those in orthopaedics is that data sets must be comprehensive and fully reflect the diversity of the data (particularly the relevant patients and pathologies). In view of the difficulty of labelling medical domains, one of the methods to develop more effective models is to train the model using a general and massive dataset, similar to ImageNet, followed by retraining the model using a smaller and specific medical task. Various applications, such as the Transfer Learning (TL) process (from a general to a specific domain), offer models that perform better than models trained from scratch [178]. However, training models using TL for medical AI is an essential issue. The primary training is commonly unrelated to medical tasks. Hence, the properties that the model learns may not be relevant to carry out medical tasks [179].

Another solution to address the issue of data scarcity in orthopaedics is self-supervised learning. This technique trains a model on a dataset without explicit labels or supervision [180,181]. In orthopaedics, it can be used for tasks such as image segmentation, bone age prediction, and diagnosis of musculoskeletal disorders. For instance, self-supervised learning can be employed to predict the bone age of patients using X-ray images, eliminating the need for manual annotation by experts [182]. Meta-learning techniques are designed to learn from a diverse range of tasks or datasets, which helps them identify common patterns or meta-knowledge that can be applied to new tasks. This approach enables the model to quickly adapt to new tasks by utilising its acquired meta-knowledge, instead of starting from scratch every time. Meta-learning is particularly advantageous when dealing with situations where there are limited data available for each new task, as it facilitates efficient adaptation and generalisation [183,184].

In-domain TL is another option to address the issue of data scarcity. TL is a method that involves utilising the knowledge gained by a model through training on a task, to improve the performance of a different yet related task [185]. In orthopaedics, this technique can be applied by using a model that has been trained on a large dataset of orthopaedic medical images, as the starting point to train a model for a specific orthopaedic application. For instance, a model that was trained on X-ray, MRI, and CT images of different body parts can be used for a dataset of MRI images of knee joints, which will help the model learn the relevant characteristics, thus increasing its accuracy in identifying knee abnormalities. TL can also be used to adapt a model that was trained on a large dataset to a smaller dataset, which can be helpful in situations where the training dataset is small [10].

The solutions mentioned previously have already demonstrated success in various medical applications. However, the following two solutions have been less explored in the medical field, particularly in orthopaedics, and it may be worth investigating them further:

Federated learning can aid to overcome the challenge of data scarcity in orthopaedics by combining data from multiple institutions to train the model. This improves the performance of the model and increases its generalisability [186]. Federated learning is a distributed DL-based approach that allows institutions or hospitals to train a DL model on their data without sharing it. This is particularly useful in orthopaedics, where privacy and regulatory concerns often restrict data sharing. The approach allows each institution or hospital to train a model locally and share the learned model parameters with a central server. The central server then aggregates the model parameters from all institutions to create a global model. This process is repeated until the global model converges [187,188].

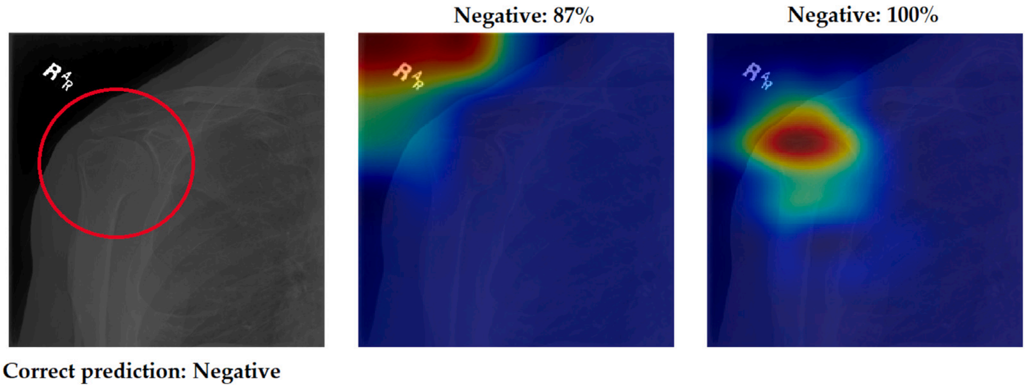


Fig. 8. Grad-CAM visualisation of two DL models on a shoulder X-ray image [169]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

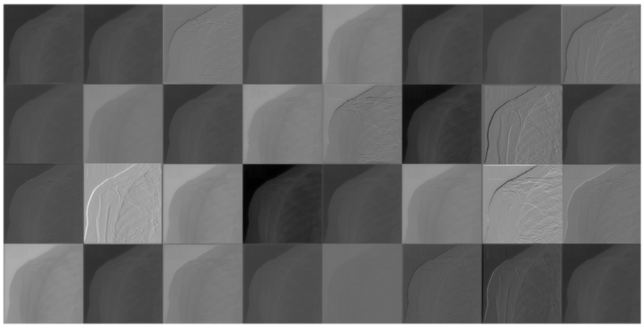


Fig. 9. Activation visualisation of the first convolutional layer of DL model from ongoing work.

Lastly, Physics-guided deep learning (PGDL) is a technique that combines the power of DL with the physical laws of the problem at hand [189]. PGDL can be used in various fields, including mechanics, finance, and medicine [190,191]. It has been shown to be highly accurate and effective in situations where data is limited. One example is using physics laws combined with limited medical imaging, such as MRI or CTI, to study the condition of organs [192,193].

In orthopaedics, PGDL can be used to model complex biomechanical systems and enhance the performance of DL models in tasks such as image segmentation, bone age prediction, and diagnosis of musculoskeletal conditions. For example, utilising PGDL for accurate bone age prediction by incorporating the physical laws of bone growth and remodelling can lead to more precise predictions of bone age and improve the performance of DL models. PGDL is still a relatively new field, and more research is needed to fully explore its potential in orthopaedics. For a more in-depth understanding of the aforementioned solutions in this article, detailed information can be found in [194]

6.2. Lack of interpretability

There is a persistent threat of creating DL applications that make unjustifiable or illegitimate decisions or are prohibited from providing in-depth explanations of their behaviour. Therefore, key stakeholders have voiced their growing concern for greater transparency in the increasing application of black-box DL models to make predictions in crucial contexts [195–198]. It is vital for a model to provide supporting explanations for a given output. This is particularly true in medicine, where experts need extensive information from the model to support the proposed diagnosis rather than a simple binary prediction. Other examples of applications that stress the significant need for supporting explanations include security, autonomous vehicles in transportation, healthcare, and finance [199].

In response to the growing demand for ethical AI, humans are generally less inclined to implement interpretable, trustworthy, and tractable methods indirectly [200]. Traditionally, it is believed that a system would become more transparent when it is solely focussing on performance. Such an assumption is valid under a trade-off between the model's performance and transparency. However, the limitations of a system can be rectified by improving understanding of the system. Thus, interpretability techniques can be used to translate the behaviour of the network into easily interpretable output [201]. Subsequently, the output can respond to questions related to the network's predictions. Three factors should be considered during the development of an ML model to enhance the implementation of interpretability as an additional design driver:

- Interpretability warrants impartiality when making decisions, such as detecting and correcting biases in the training dataset.
- Interpretability helps to provide robustness by highlighting potential adverse perturbations that could affect the prediction.
- Interpretability serves as a preventative measure so that the output is inferred only by meaningful variables to guarantee the presence of an underlying truthful causality in the model reasoning.

There is an apparent paucity of interpretability of DL in orthopaedics, as the data can be complex and the decision-making process may involve multiple factors. However, efforts are being made to improve the interpretability of DL models through the use of techniques such as attention mechanisms and visualisation techniques.

Interpretability techniques include model selection, learning, bias assessment, verification, and debugging. Interpretability techniques can be adopted after network training or integrated into network training. The unnecessary construction of an interpretable DL network is the main advantage for post-training methods, which saves time and generally makes post-training methods the preferable option. Therefore, this review highlights post-training methods that utilise test images to describe the predictions of a trained network using image data. One of the interpretability techniques is the visualisation method, which exploits visual representations of the network observation to describe its predictions. Numerous techniques have been proposed to visualise network behaviour, such as low-dimensional projections, heat maps, feature importance maps, and saliency maps.

The comparison of two models trained for shoulder abnormality detection from our ongoing work is shown in Fig. 8. Both models correctly predicted the images, according to their confidence values. However, the Grad-CAM heatmap revealed that the first model is biased and less accurate, failing to identify the region of interest outlined by the red circle. In contrast, the second model correctly detected the region of interest with high confidence. This illustrates the significance of visualisation, as even models with high confidence values may not

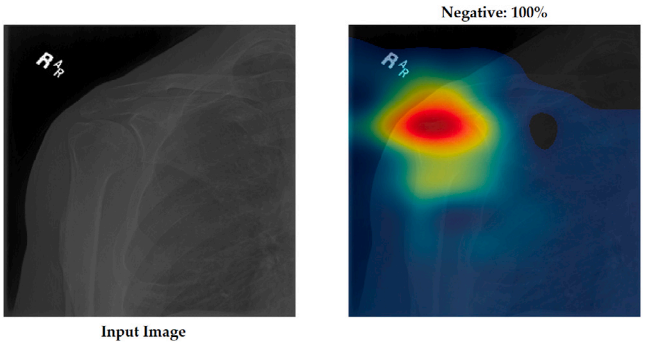


Fig. 10. Example of CAM [169].

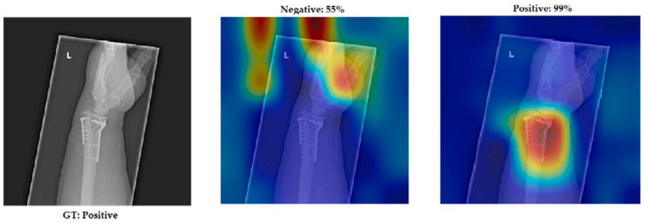


Fig. 11. Grad-CAM visualisation of two DL models on a Wrist X-ray image [166].

be accurate. By incorporating visualisation techniques, we can enhance both the results and confidence in those results before implementing models in real-world scenarios. The following section explains several visualisation methods that will help to understand DL decisions and detect bias in orthopaedics:

- **Activation visualisation:** Activation visualisation refers to a simple approach to understanding the behaviour of the network. Most CNN models learn to identify characteristics in the first convolutional layer, such as colour and edges. The network continues to learn more complex characteristics in deeper convolutional layers [202]. By using the input image in Fig. 8, Fig. 9 shows an example of Activation visualisation of the first convolutional layer of the DL model.
- **Class Activation Mapping (CAM):** CAM is also a simple technique that generates visual descriptions of CNN predictions [203]. This method utilises the global average pooling layer in a CNN model to produce a map that emphasises the distinct area of an image the network uses with respect to a specific class label. Fig. 10 shows an example of CAM.
- **Gradient-weighted Class Activation Mapping (Grad-CAM):** Derived from the CAM method, the Grad-CAM method utilises the classification score gradient in terms of the convolutional features determined by the network to understand the most essential parts of an image for the classification process [204]. Locations with a larger gradient are also where the final score relies mostly on the data. Additionally, the Grad-CAM method generates similar outcomes to the general CAM without the design limitations of CAM. Fig. 11 shows an example of Grad-CAM.
- **Occlusion Sensitivity (OS):** OS determines the sensitivity of the network towards small perturbations in the input data. The method involves perturbing small areas of the input by replacing sections with an occluding mask, typically a grey square. The change in probability score is measured for a given class as the mask moves across the image. The occlusion sensitivity can be used to highlight the vital parts of the image for classification. The appropriate values must be selected for the MaskSize and Stride options to obtain the best results from occlusion sensitivity. Thus, this tuning offers more flexibility in assessing the input features under various length scales. Fig. 12 shows an example of OS.

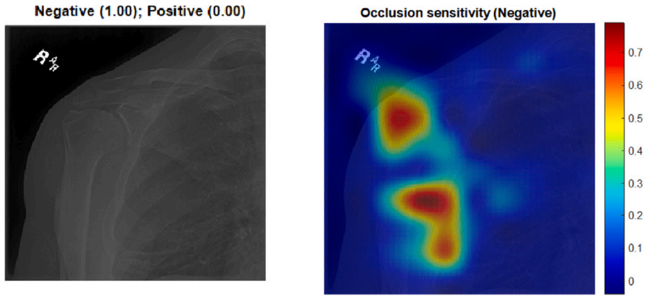

Negative (1.00); Positive (0.00)
Occlusion sensitivity (Negative)
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

Fig. 12. Example of occlusion sensitivity from ongoing work.

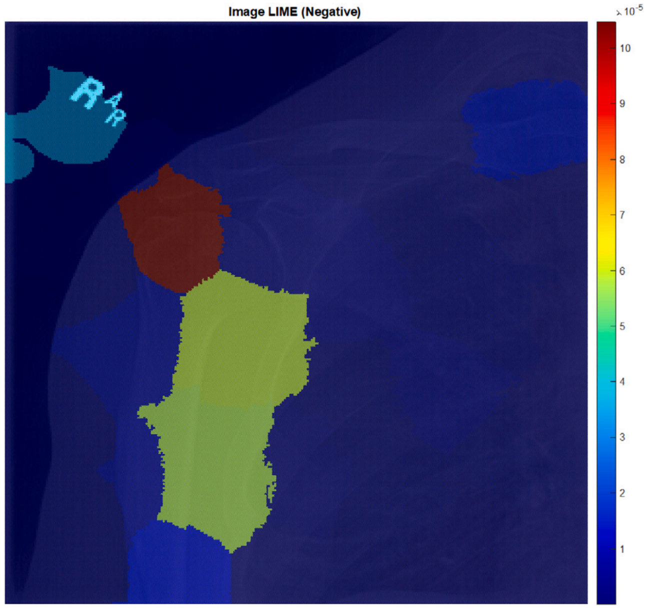


Fig. 13. Example of LIME [169].

- **Locally Interpretable Model-agnostic Explanations (LIME):** The LIME method estimates the classification behaviour of a DL network using a more simple and interpretable model, such as the regression tree or linear model [205]. This simple model evaluates the important features of the input data as a proxy for the importance of the features to the DL network. Fig. 13 shows an example of LIME.
- **Gradient Attribution (GA):** GA methods offer pixel-resolution maps that show the most essential pixels to the network classification decisions [206]. These methods calculate the class score gradient with respect to the input pixels. The maps show the pixels with the most affected class score when altered. The gradient attribution methods produce maps with the same size as the input image. Despite the high resolution of the gradient attribution maps, they tend to be much noisier, as well-trained deep networks are poorly dependent on the precise value of specific pixels.
- **DeepDream (DD):** DD is a feature visualisation technique synthesising images that strongly activate network layers [207]. The image features learned by a network are highlighted by visualising these images, which are valuable to understanding and diagnosing the behaviour of the network.
- **t-Distributed Stochastic Neighbor Embedding (t-SNE):** The t-SNE is a dimension-reduction technique that maintains the distance of both points close to each other in the high-dimension and low-dimensional representation [208]. This method is used

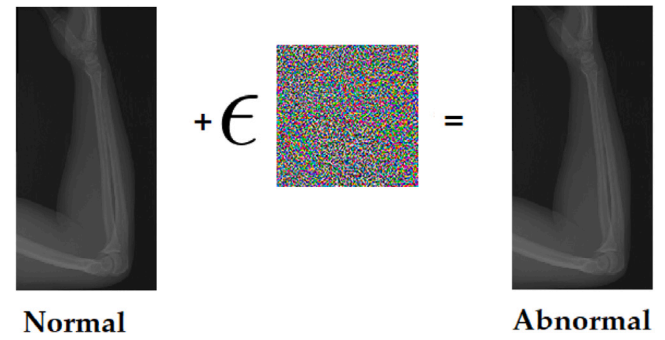


Fig. 14. Adversarial attacks.

to visualise the changing input data representation by the DL networks as it penetrates the network layers.

- **Maximal and Minimal Activating Images:** A simple technique to understand a network is to visualise images that weakly or strongly activate the network for each class. Strongly activated images indicate how the network thinks of the appearance of a "typical" image from that class. In contrast, weakly activated images facilitate identifying the image element that led the network to give inaccurate classification predictions.

6.3. Other challenges

Data scarcity and interpretability are not the only challenges facing the use of DL in orthopaedics. There are several other significant issues that must be addressed. These include:

- Generalisation: Generalisation is a significant problem in the use of DL in orthopaedics, as models tend to be trained and evaluated on a specific set of data and may not perform well on new or unseen data [209]. This specificity can be caused by a lack of diversity in the training dataset, which may not adequately reflect the diversity of patient populations [210]. Furthermore, models may be too complex and tailored to training data, failing to generalise well to new patients. This can pose a significant challenge in clinical settings, where the ability of the model to perform well on unseen patients is critical for accurate diagnosis and treatment.
- Adversarial attacks: DL models in orthopaedics can be vulnerable to adversarial attacks, where input data is intentionally manipulated to produce incorrect predictions. For instance, adding a small perturbation to an input image results in an incorrect decision [211–213] (see Fig. 14). This is a serious concern, as it can lead to inaccurate diagnoses and worse patient outcomes. These attacks can be challenging to identify and prevent, making them a significant obstacle to DL's secure and dependable utilisation in orthopaedics. It is critical to consider the possibility of an adversarial attack when building a DL application for orthopaedics or for clinical applications in general [214].
- Scalability: The use of DL in the field of orthopaedics often requires a large amount of computational power. This can pose a challenge when trying to implement these models on devices with limited resources, such as mobile phones or low-resource environments. To overcome this limitation, researchers have been investigating ways to decrease the computational demands of these models, such as model compression, quantisation, and adopting more shallow DL models, such as MobileNet and ShuffleNet. Some researchers have also utilised FPGAs to accelerate the processing of DL models [215–217].

7. Exploring the limitations of data labelling in orthopaedics: A professional analysis

Orthopaedic surgery is likely one of the most challenging medical disciplines in terms of data labelling [218,219]. Practicing orthopaedics often requires highly sophisticated analytical reasoning that is prone to human subjectivity and errors at multiple levels. It is difficult for an orthopaedic surgeon to verbally explain his clinical choices, which are often highly case-specific and surgeon-personalised. Often it is prior training and previous experience that inform the choices made by clinicians. Although evidence-based guidelines try to sharpen the profession and standardise its practice, these guidelines are often broad-reaching and general in nature, intended as broad guiding references rather than strict rules. In addition, it is very common to have a lack of consensus in orthopaedic societies and academies, and disagreements in opinion are frequent in orthopaedic literature. Several unconfirmed dogmas exist in the field, and the postulates of 'orthopaedic masters' are often regarded by members of the orthopaedic community as a ground truth. Orthopaedic surgery is a relatively new branch of surgery, with almost all the body of knowledge condensed in the second half of the 20th century. Rapid advances have been made in the last 20 years, including the evolution of pin-hole, or arthroscopic, procedures. The role of precision and accuracy in surgery has only recently been addressed and has been exponentially increasing, thanks to the introduction of a variety of imaging and device technologies in the field.

Like all branches of medicine, approaching an orthopaedic case starts with a dedicated patient interview. The diagnosis is a long process that is fundamentally based on the patient's medical history, including a detailed description of symptoms or a description of the traumatic injury. Some patients are more expressive than others and can give a description that is highly suggestive of a specific diagnosis or category of diagnoses, especially if they used classic descriptions of the concerned pathology and if their demographics (gender, age, ethnicity, etc…) have been shown to be statistically linked with this pathology. For example, a teenager who visits a knee clinic with a frequent knee "giving way" sensation after a pivoting injury to this knee caused by a kick during a soccer match and who reports hearing a loud 'pop' during this kick has an anterior cruciate ligament (ACL) tear until proven otherwise. The pathognomonic keywords, in this case, are giving-way, pivoting injury, and a loud "pop" sound. Having a magnetic resonance imaging (MRI) of the knee of this patient and finding that the ACL is not torn would be particularly surprising. In fact, if the physical examination suggests, along with the patient's story, an incompetent or deficient ACL, then an ACL tear can be confidently diagnosed, even if the MRI is negative. Although magnetic resonance imaging is the gold standard imaging modality for diagnosing ACL ruptures, it is not accurate 100%, and arthroscopic confirmation (during surgery) provides the ground truth for diagnosis. In a study by Zhao et al. [220], it was found that 4 out of 66 ACL tears confirmed during arthroscopy were mislabelled by preoperative MRI: 2 partial ACL tears that were labelled as complete tears, one complete ACL tear that was labelled as a partial tear, and one complete tear that fully escaped MRI detection. Thus, the patient's interview constitutes the essence of any diagnosis. It should be highlighted that clinical examination is only one piece of the puzzle. Special orthopaedic tests are meant to stress specific anatomic structures, but none are 100% specific or sensitive to the disease tested. A large meta-analysis revealed that none of the ten commonly used specific tests for rotator cuff pathology were consistent in a diagnostic setting [221]. Thus, it's always wise to consider a constellation of tests rather than a single test.

The human process for analysing musculoskeletal images is highly sophisticated. It's difficult for human language to describe this process in words, and secondarily, it's difficult for computer language to describe it in code. The radiologic diagnosis of an orthopaedic injury (e.g., ACL rupturing) is intrinsically asynchronous because of the fact that the mechanism of injury has already taken place and ceased prior to imaging, and what we are seeing is a static screenshot of the "aftermath" of the injury (e.g., interruption of ACL continuity, fibre edema, bone contusion, etc.). Zhao et al. [220] studied four direct and eight indirect MRI signs of ACL injury and found that these 12 signs differ by their sensitivities (a biostatistical term for the ability of a test to rule-in a disease) and specificities (a biostatistical term for the ability of a test to rule out a disease), yet none is perfect. Even a very logical and straightforward sign, such as "interruption of the continuity of ACL fibres", was absent in 15 out of 66 ACL tears and presented in 2 out of 12 normal ACLs in this study. In simpler words, 23% of torn ACLs showed continuous fibres, and 17% of normal ACLs showed disruption of fibres on MRI. We can teach a machine to identify the ACL in a knee MRI. We can then teach the machine to look for disruptions in the ligament's continuity (which is a pixilated challenge for artificial intelligence). Even with 100% performance in performing this task, the machine might miss 23% of ACL tears and would misidentify 17% of normal ACLs as tears! Thus, the "interruption of the continuity of ACL fibres" criteria is not enough to diagnose a tear. Li et al. [94] taught a multimodal feature fusion deep learning model based on deep learning algorithms to diagnose ACL tears based on a continuum of "fibre discontinuity features" rather than binary criteria (continuous/discontinuous) and added other signs related to ligament boundaries/edges, ligament thickness, ligament signal, and percentage of damage from the whole ligament. They identified three grades of ACL tear: Grade I: the ligament continuity was still good, the contour was still intact, the ligament was not thickened or slightly thickened and expanded, small patches or streaks of a signal can be seen, and damage area was less than 50%. Grade II: ligamentous continuity was poor, but some continuous fibres were still visible; locally thickened or diffused ligaments were visible; incomplete or well-defined edges were at the site of ligament injury, or there were locally notched areas; the abnormally high signal can be seen, with damage area greater than or equal to 50%. Grade III: there was an intact rupture of the ligament, characterised by broken continuity of the ligament, displacement of the bent or broken end, clumpy ligament, increased signal, and unclear boundary. This model's sensitivity, specificity, and accuracy in diagnosing ACL injury were 96.78%, 90.62%, and 92.17%, respectively. This high success rate can be attributed to better criteria definition, adding more criteria to the labelling process, generous and diverse data output, and not committing the neural network to difficult binary decisions. In fact, it makes more sense for clinicians to have a probability continuum than a 0–1 dichotomy. This would allow DL outputs to be added to, rather than replacing, the other pieces of the diagnosis puzzle.

Like in the human experience, incorporating clinical data into radiologic training improves the computer model's accuracy. Liu et al. [222] taught a deep learning model of how to classify bone tumours on X-ray into three categories: benign, intermediate, and malignant. It achieved an accuracy of 73.3%, an AUC of 0.813, a specificity of 84.4%, and a sensitivity of 62.7%. Combining clinical characteristics in the fusion model (such as the value of erythrocyte sedimentation rate, the presence of pain, etc...) improved accuracy by 4.9%, the AUC by 0.059, the specificity by 3.3%, and the sensitivity by 7.2%. Including clinical characteristics, the fusion model's performance was comparable with that of senior radiologists. This inclusion of relevant additional data is a potential additive to improve the performance of AI.

In any field of therapeutic medicine, a classification system is a categorising system that aims to grade the severity of the diagnosis, suggest treatment options, predict prognosis, standardise the reporting of clinical and epidemiologic data, unify the communication between clinicians, and homogenise data collection for research. As such, classification systems show promise in data labelling for ML. An effective classification system must be valid, reliable, and reproducible. Orthopaedic surgery is probably the specialty in medicine where we find the greatest abundance of classification systems. The vast majority are introduced by single surgeons based on small case series and have never been validated. Despite this, some are still in widespread use despite independent evidence of low interobserver and/or intraobserver reliability. In a review of 185 published orthopaedic classifications [223], only four (2.1%) had a validation process described in the initial paper that introduced that classification to the literature. Over 70% of these systems have never been independently validated and assessed for intra-observer and inter-observer error. Of those that have (54/185), only 10 (18.5%) demonstrated either an intra-observer or inter-observer error that is described as excellent (kappa score ≥ 0.8). Only two classification systems of the 54 (3.7%) were shown to have both intra-observer and inter-observer errors as excellent, meaning only 2 of the 185 classification systems reviewed (1.1%) have been shown to be highly reproducible [224].

With the so-many subjective, nonvalidated, or unreliable diagnostic and classification systems, it is hard to formulate a ground truth for ML. Building a unified reference for data labelling in orthopaedics is particularly challenging. There is no 1 + 1 = 2 in medicine; however, in orthopaedics specifically, 1 + 1 = 3 to some, and 1 + 1 = 4 to others, with evidence pointing in different directions simultaneously. Teaching a model all of these possibilities might lead to the model's replication of human confusion. Thus, it is particularly important to have any orthopaedic classification system objectively validated before calling it a ground truth, as DL outputs will only be as reliable as the frameworks on which they are based.

8. Deep learning associated with technologies in orthopaedics

DL has increasingly been adopted for various tasks in Orthopaedics, including (i) Image analysis such as X-rays, CT scans, and MRI scans which can aid in diagnosing conditions and determining the best treatment plan. (ii) DL can assist in surgical planning, which helps reduce the risk of complications and improve outcomes. DL can offer more by associating with other technologies to improve the outcomes for patients.

8.1. DL and robotic surgery in orthopaedics

With the aim of improving the precision of bone resection, soft-tissue balancing of the gap, and reducing intraoperative hand tremors in order to achieve better clinical outcomes (e.g., less pain, better restoration of joint kinematics, and better implant survival) for patients, orthopaedic surgical robotic systems have been extensively adopted in various types of arthroplasty in recent decades. In Total Knee Arthroplasty (TKA), robot-assisted surgeries demonstrated higher reproducibility and accuracy for restoring mechanical alignment compared to traditional surgeries [225]. Meanwhile, the radiological outcomes of Total Hip Arthroplasty (THA) with the help of surgical robots were reported to be superior to those outcomes of conventional arthroplasty [226]. Clinical outcomes of joint function in robotic arthroplasty were studied, and the results showed that outcomes of robotic surgeries were comparable to those of traditional arthroplasty [227]. As such, robotic surgeons have emerged as competitors in carrying out orthopaedic surgeries, with advancements in robotic technologies overcoming early issues such as blood loss and infection during long surgeries [228,229]. While comparative studies with humans have shown that robots are better in terms of limb lengthening, patient satisfaction, and cost [230], the replacement of human expertise with technology is unlikely to occur in the near future as the long-term results of traditional methods are still being observed [231].

DL has been integrated into robotic surgery in Orthopaedics to improve its accuracy and outcomes [57,232]. This combination of DL and robotics in Orthopaedics can result in:

- Safety: DL models can be adapted to monitor the surgical procedure in real-time to ensure that the procedure is performed safely and that any potential complications are identified and addressed immediately.

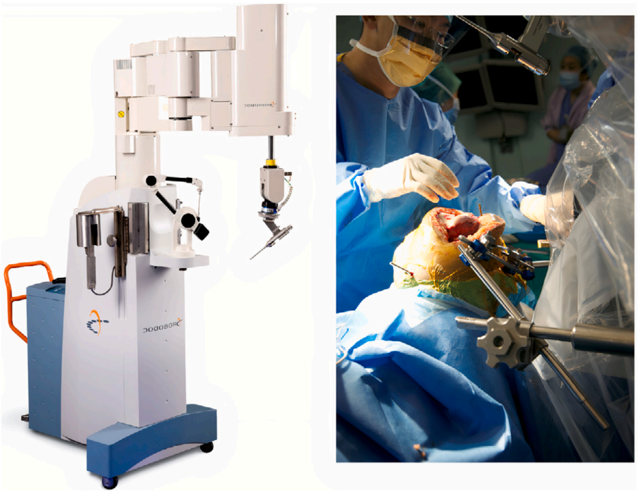


Fig. 15. Example of robotic surgery in orthopaedics [233].

- Enhanced Accuracy: DL models can be applied to process and analyse large amounts of data, including images and patient data, to support surgical decision-making and enhance surgical accuracy.
- Enhanced instrument control: DL algorithms can be utilised to optimise robotic arm movements and control, resulting in smoother and more precise surgical procedures.
- Reduced procedure time: The use of robotics in combination with DL models can lead to faster and more efficient surgical procedures, reducing the overall time required for surgery.

Many orthopaedic surgical robot systems are currently in clinical use, including in knee replacement where DL can be used (see Fig. 15). Stryker's Mako Robot and Smith & Nephew's NAVIO Surgical System were used in Unicompartmental Knee Arthroplasty (UKA), Patellofemoral Knee Arthroplasty (PFA), as well as TKA. At the same time, most other surgical robotic systems [233–235] focused on TKA. Regarding preoperative planning, ROSA [236], iBlock [237], NAVIO [238], and CORI [239] systems are examples of preoperative planning robots.

To realise a wider implementation of robot systems in orthopaedic surgeries, the cost of the robot systems and the length of operation time induced by the adoption of surgical robots should be further reduced. The mitigation of risks in robot-assisted orthopaedic surgeries, such as infections, human error due to inadequate training and malfunction of electronic components, should be considered with more care to ensure that the robotic technologies adopted in orthopaedic surgeries are safe for patients and surgeons. Although DL can play a role in reducing operation time and human error, more effort is required to integrate DL with surgical robots.

8.2. DL and Mixed Reality (MR) in orthopaedics

Mixed Reality (MR) technology is designed to bridge the gap between the preoperative plan and surgical execution, especially in revision surgeries (replacement of old or failed implants) where anatomical guidance is needed. MR technology allows 3D visualisation of deformities and the ability to manipulate the preoperative plan in real time with holograms, leading to better anatomical understanding and more confident surgical decisions [240]. In revision surgeries, surgeons often have multiple plans and options to address potential challenges during the operation. MR technology provides the advantage of generating multiple holograms based on the preoperative plan and backup plans established for various surgical scenarios (see Fig. 16) [5].

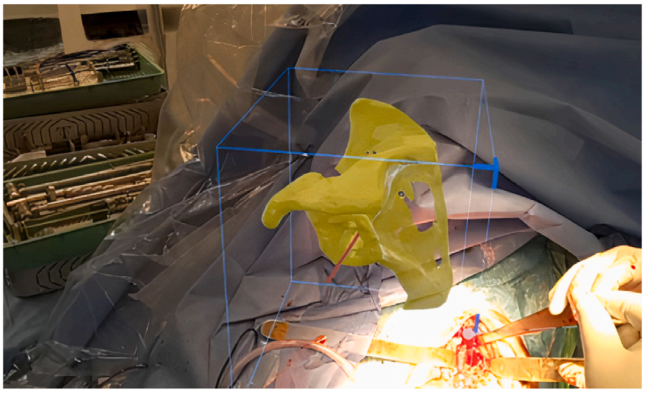


Fig. 16. Example of mixed reality [5].

MR techniques face several challenges in which DL can help to solve some of these issues, such as improving interactivity, improving realism, improving tracking and adapting the experience for each user. DL models can be used to provide real-time natural interaction through computer vision and gesture recognition, add realistic lighting and shading, improve textures and materials, and accurately represent objects and scenes for a more immersive experience. It can also improve the tracking of user and object movements and personalise the experience for each user by adapting to their preferences and current state [241].

8.3. DL and wearable sensors in orthopaedics

Use of wearable sensor technology has expanded greatly in orthopaedic diagnosis, rehabilitation, and data collection. Typically, a portable wireless body area network (WBAN) is constructed to aid in physical rehabilitation. The WBAN can be made up of different sensors, such as vital signal sensors and motion sensors. These sensors collect data and send it to a central hub to help supervise and monitor post-operative rehabilitation. The wireless network is commonly based on different modalities such as Bluetooth, Zigbee, UWB, or WLAN. The network architecture is chosen based on trade-offs between power consumption, interference level, overall system configuration, and integration [242,243].

Data are usually transmitted to a server connected to the Internet and can be accessed by clinicians and the medical registry. Continuous data acquisition then serves as a window to customise rehabilitation and investigate patient progress trends after surgery. Data collected in this fashion can be used to train DL models and improve monitoring programmes delivered to patients by clinicians [244].

In addition to wearable sensors, continuous efforts have been made to develop microwave-based imaging systems for bones. These imaging systems are classified into two categories, wearable vs. free-space systems. Free-space systems can be converted into wearable systems considering antenna conformity, impedance mismatch, and tissue loading. In [245], a feasibility study was conducted to investigate the development of a microwave imaging system for bone operating at 0.5–4 GHz. This feasibility study was carried out with a realistic phantom and 3D image reconstruction was performed to show the tibia and fibula as shown in Fig. 17. More research is needed to confirm the ability to differentiate between the tibia and fibula with different cross sections with background tissues. Here, DL models can help in the classification, as an extension of its previously established imaging applications.

The imaging of bone fractures was investigated in [246]. The system was developed to detect fractures in the tibia at a higher frequency of 8.3–11.1 GHz. The system itself is not considered wearable, but it can be changed into a conformal array once all the challenges of

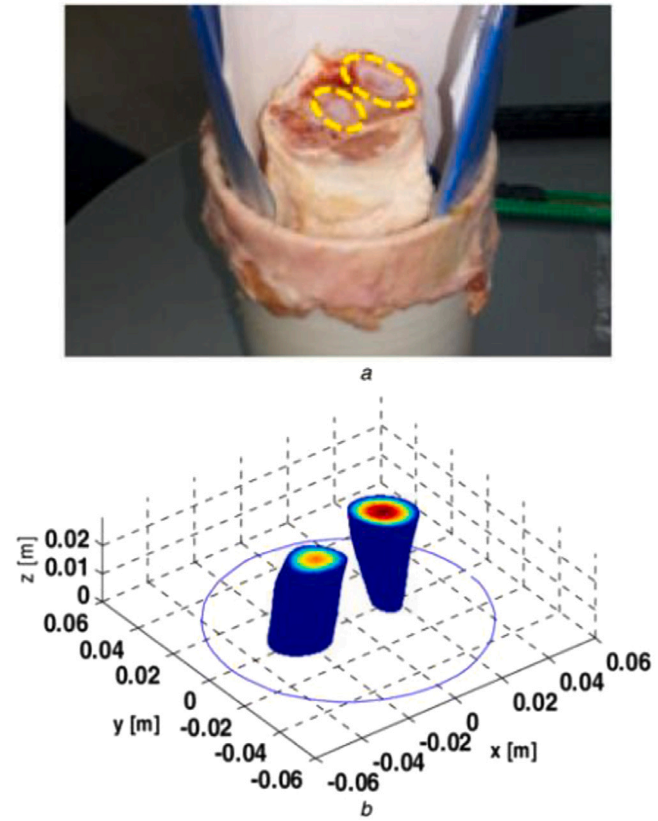


Fig. 17. 3D reconstruction of realistic phantom with tibia and fibula [245].

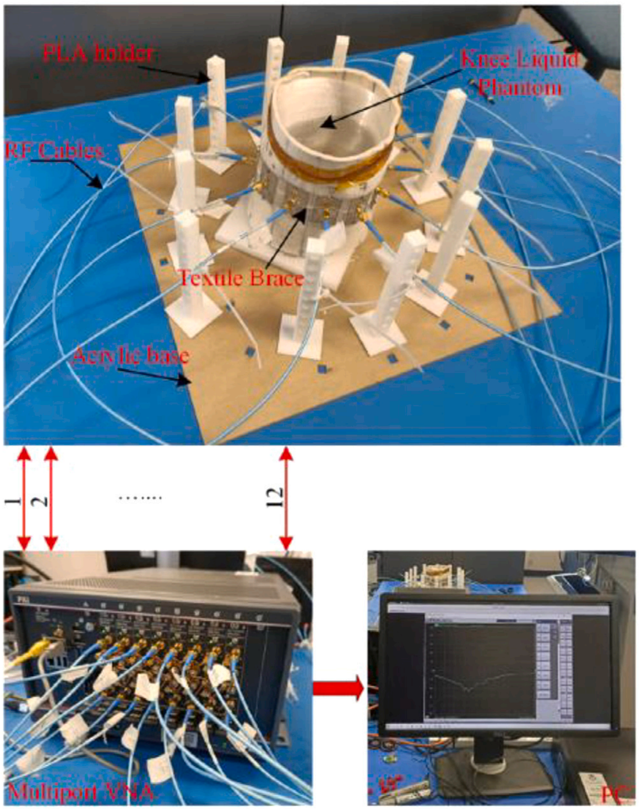


Fig. 18. Wearable knee microwave imaging system [247].

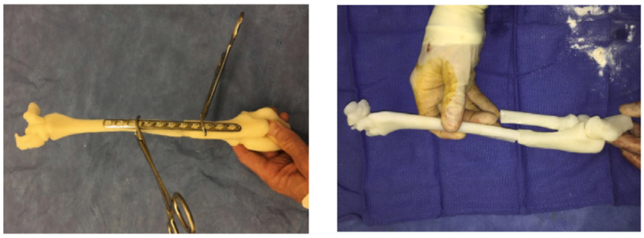


Fig. 19. Example of 3D printing in orthopaedics where both images are forearm [254].

tissue loading, antenna sensitivity to bending, and others are addressed. To remove the effect of background and skin artefacts, SVD was used instead. DL models would provide a promising image enhancement strategy to improve the quality of microwave imaging systems once a large dataset is collected. A wearable brace microwave imaging system was developed in [247] (see Fig. 18). Although the system is made with textile antennas, it is still in the research stages, where a lot of images need to be taken with volunteers to build a dataset for the improvement of image detection with DL. This system can also be combined with WBAN for rehabilitation after ACL/PCL tears surgery. However, continuous effort needs to be made to understand the interplay between the role of sensor data and images from that system. Advancements in the field of microwave imaging systems can be fused with results collected from other sensors to guide rehabilitation after orthopaedic surgery.

8.4. DL and 3D printing in orthopaedics

The deployment of 3D printing in orthopaedic surgery has significantly grown in recent years (see Fig. 19). This is due to advances in technology and favourable outcomes, as proven in the existing literature [248]. 3D printing has become more accessible and flexible due to lower costs and new developments such as bioprinting and metal 3D printing [249]. The application of this technology is likely to transform the future of healthcare delivery. Studies have shown that 3D printing can improve understanding of patient-specific anatomy and enhance outcomes, particularly in complex cases of hip and knee reconstruction [250]. Studies have also shown that 3D printing can be used to manufacture orthopaedic implants with improved mechanical strength and tribological and corrosion behaviour [251].

DL has been employed in 3D printing to improve the quality and productivity of the final product through in situ monitoring, as well as to optimise the design and process parameters. Furthermore, it has facilitated the prediction of microstructure evolution by developing surrogate models driven by physics-based data [252,253].

The US Food and Drug Administration (FDA) has established clearer regulatory pathways for 3D printed medical devices in orthopaedic surgery and other fields. The cost and time savings associated with the use of 3D printing are significant, and there is great potential for even more advances in the future [254].

DL models can be used to design and optimise custom orthopaedic implants. This can be achieved by using data from medical images such as CT and magnetic resonance imaging and patient-specific anatomical information. DL models can analyse this information to create a 3D implant model that fits the patient’s anatomy as closely as possible [255].

In summary, the use of DL in combination with 3D printing has the potential to revolutionise the way orthopaedic implants are designed and manufactured, which could lead to better patient outcomes and more efficient surgical processes [256].

9. Orthopaedic preoperative software

Orthopaedic solution-based software has emerged as a valuable tool that has the potential to assist surgeons in improving their surgical procedures and enhancing patient outcomes. These tools support the clinical decision-making process from diagnosis to treatment. According to the US Food and Drug Administration (FDA), software intended to diagnose, cure, mitigate, treat, or prevent disease in humans is considered a medical device [257]. An increasing number of medical devices and algorithms using AI and its different approaches, from ML to DL, have been approved by the FDA [258]. Consequently, the number of commercially available orthopaedic software packages has increased in recent years, especially with recent advances in digital technologies and computer vision created primarily by DL. However, compared to other medical fields, such as breast cancer and cardiology, orthopaedics still lags behind the adoption of DL. Many orthopaedic solution-based companies claim to use DL to automate their workflow and help surgeons make informed decisions for better patient outcomes. However, companies have either not formally stated how they are using DL, or it is at its very early stage of development, and regulatory approval has not been received to commercialise their products.

9.1. Computer-aided diagnosis (CAD)

Computer-aided diagnosis (CAD) is an important research topic in medical imaging and diagnostic radiology [259]. CAD is the use of the computer output as a ‘second opinion’ to assist radiologists, and has become an essential component of medical image analysis [260]. CAD is complementary to radiologists’ precision and helps in early detection of abnormalities, especially breast cancers on mammograms [261]. CAD schemes could be assembled as packages and implemented as part of a picture archiving and communication system (PACS), including computerised detection of lung nodules, vertebral fractures, and interval changes in chest radiographs, as well as the classification of benign and malignant nodules and the differential diagnosis of interstitial lung diseases [262].

We identified three CAD devices that both formally claimed the use of DL in diagnostic tasks and have been cleared by the FDA to the best of our knowledge. These devices are KOALA by Image Biopsy Lab, OsteoDetect by Imagen Technologies, and FractureDetect by Imagen Technologies. KOALA (Knee Osteoarthritis Labelling Assistant) is a knee osteoarthritis assistive diagnostic tool that uses deep learning to detect signs of knee osteoarthritis from X-ray images. The software helps provide an automated scoring of the osteoarthritis stage according to the Kellgren and Lawrence grading system, providing precise and automated measurements of the minimum joint space width and evaluation of the severity of joint space narrowing, osteophytosis, and sclerosis. The device provides a reliable measurement with 87% sensitivity and 83% specificity. It also helps standardise radiographic reading by increasing the physician’s agreement rate to the gold standard by 23% and saving workflow time. The device obtained 510(k) FDA approval in 2019.

OsteoDetect is an AI software for detecting and diagnosing wrist fractures in the adult population. This software is based on AI algorithms that are capable of analysing two-dimensional X-ray images and detecting distal radius fractures, which is a very common wrist fracture. The region of the radius fracture would be highlighted on posteroanterior and lateral radiographs of the adult’s wrists. The Osteodetect tool is intended for use by clinicians, emergency physicians, urgent care, and orthopaedic surgeons. The software was cleared by the FDA in 2018. Later, in 2020, the same company obtained 510(k) FDA approval for FractureDetect software. This computer-assisted diagnosis and detection software helps detect fractures of twelve different musculoskeletal structures (ankle, clavicle, elbow, femur, forearm, hip, humerus, knee, pelvis, shoulder, tibia/fibula and wrist) using two-dimensional radiographs of adults only. It has been reported that the use of this software has improved fracture detection by 45%.

9.2. Preoperative planning software

For preoperative planning software in the orthopaedic setting, DL-technology can be used to process medical images of different modalities (CT, MRI, radiographs, etc.), 3D reconstruct bones and soft tissue structures to provide the surgeon with additional visual insight into the disease and its severity, provide the surgeons with the critical measured parameters to guide implant positioning, and simulate a postoperative range of motion. The following are the leading companies that use DL-based algorithms in their preoperative software.

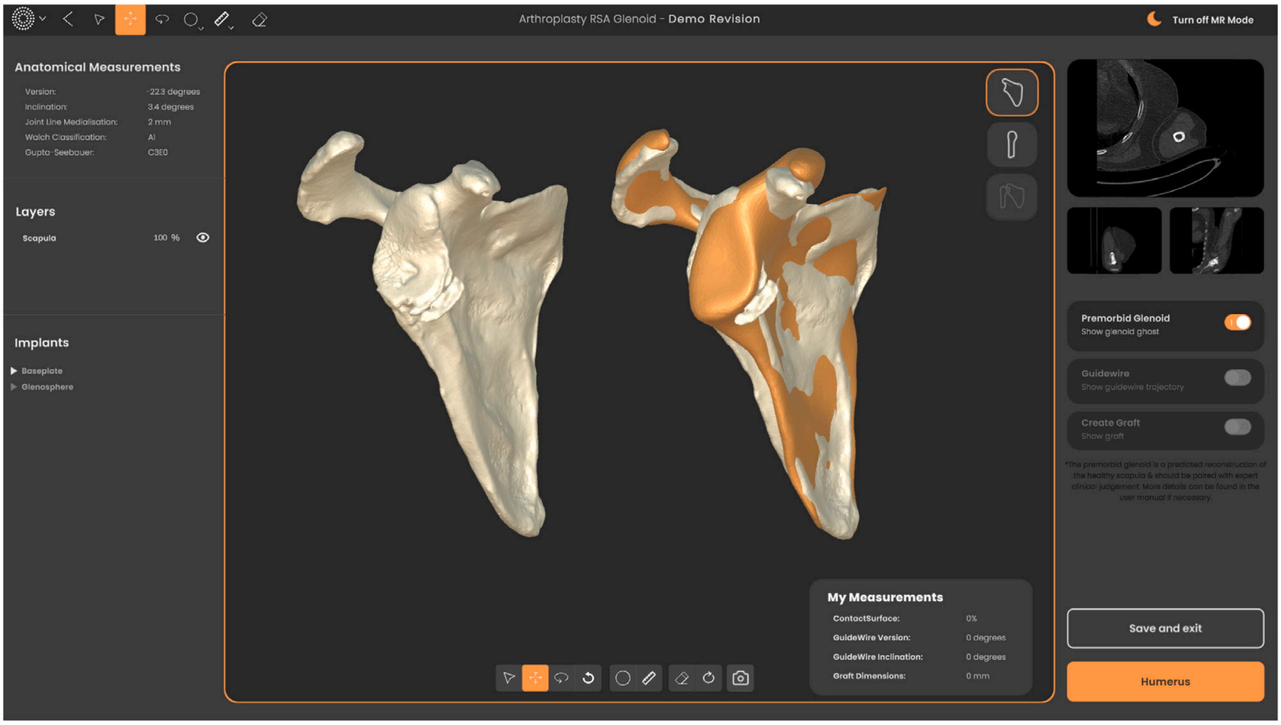
Formus Labs, based in Auckland, New Zealand, offers an automated 3D preoperative planning solution for primary total hip arthroplasty using DL and population-based computational modelling. The software provides an automated image preprocessing module to segment the bony anatomy of the joint and find the optimal implant selection and positioning. According to the company, their platform can alleviate the manual tasks involved in this process, reduce the cost of joint replacement by 25%, and minimise the risk of revision surgery. The company has submitted a 510(k) premarket approval for FDA clearance as a primordial step to access the US market. More than 450,000 total hip replacements are performed annually

Akunah, an Australian-based medical technologies company, has recently announced that it has FDA approval for its preoperative planning software (Akunah Reflect). This software aims mainly to empower surgeons to plan any primary procedures, revisions, fractures, and instability of the shoulder, and the software is agnostic to the implant of the surgeon’s choice. Akunah Reflect uses advanced ML technologies to segment bone geometries from CT scans of patients. Reconstructed 3D bones models are checked and fine-tuned using the gold standard manual technique to ensure product safety and compliance. The software features, such as anatomical measurements of the scapula and humerus, medialisation of the joint line, subluxation of the humerus, and quantification of glenoid bone loss, help surgeons make an informed decision for an optimised patient outcome. The software also has a module to reconstruct the premorbid anatomy of the scapula to provide visual information on glenoid bone loss and its severity (see Fig. 20). This module uses statistical shape model technology.

Precision AI provides AI-driven preoperative planning solutions that help improve the accuracy of surgical procedures. This company is also based in Australia and provides software that empowers shoulder surgeons to create preoperatively patient-specific guides for shoulder replacement. As of the writing of this paper, no FDA has been cleared or approved for this software. However, it is approved for use in Australia, New Zealand, and the UK.

The Signature One system by Zimmer Biomt is a planning software providing patient-specific shoulder replacement guides. The software uses an ML algorithm to automatically segment and generate 3D models of the scapula that is then fine-tuned manually by specialists to make sure that the output meets some accuracy standards. This preoperative planner uses a semi-automated approach to provide the required preoperative information to help surgeons plan the glenoid component. The FDA has cleared this software as a class II device. Overall, we found minimal preoperative planning software that claims to have implemented DL to automate the planning workflow. As such, the key features of the available software that do not use DL to support automated preoperative planning orthopaedic workflows are analysed below to understand better how we can use DL to support this automated workflow. We focused on shoulder orthopaedic surgical planners that have been FDA-approved and highlighted key features that can help surgeons efficiently prepare the case preoperatively.

Blueprint 3D planning software by Stryker is one of the surgical planners intended to be used for planning shoulder joint replacement. This software does not claim any use of DL technology. It requires CT scan images in DICOM format. The scapula and humerus bones are then automatically segmented. Then, the software allows the surgeons to select and position glenoid and humeral implants and simulate


Arthroplasty RSA Glenoid - Demo Revision

Anatomical Measurements

Version: -22.3 degrees
Inclination: 3.4 degrees
Joint Line Medialisation: 2 mm
Walch Classification: AI
Gupta-Seebauer: C3E0

Layers

Scapula 100 %

Implants

Baseplate
Glenosphere

Premorbid Glenoid — Show glenoid ghost
Guidewire — Show guidewire trajectory
Create Graft — Show graft

*The premorbid glenoid is a predicted reconstruction of the healthy scapula & should be paired with expert clinical judgement. More details can be found in the user manual if necessary.

My Measurements

ContactSurface: 0%
GuideWire Version: 0 degrees
GuideWire Inclination: 0 degrees
Graft Dimensions: 0 mm

Fig. 20. Akunahh Reflect software user interface. The scapula on the left (gold colour) represents the 3D segmented bone geometry, and on the right is the same bone geometry overlayed with the premorbid scapula (Orange colour). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the prosthetic range of motion accordingly. The software has been cleared as class II in the USA (510(k) pathway). The ExactechGPS Total Shoulder Application is a software component of the ExactechGPS system developed by Exactech, a medical device company specialising in orthopaedic solutions. This application is specifically designed to assist surgeons during total shoulder arthroplasty procedures. It is integrated with the ExactechGPS navigation system, which combines preoperative planning, intraoperative guidance, and real-time feedback to enhance the precision and accuracy of shoulder surgeries. FDA has cleared this device.

Surgicase is a shoulder pre-surgical planner by Materialise that can be used to simulate surgical shoulder interventions. The software allows surgeons to visualise the 2D and the 3D anatomy of the glenohumeral joint, reconstruct the premorbid shape of the scapula, provide humeral head diameter measurements, quantify glenoid defect, and provide glenoid and humeral component positioning based on critical measurements (such as glenoid version, lateralisation, inclination and humeral version, inclination, and resection level). The software also features a range of motion simulation, similar to Blueprint. The software is built on state-of-the-art manual segmentation and surgical planning techniques and has been cleared as a class II device in the USA (510(k) FDA).

Considering the above software packages, it is highly likely that DL could be used at each stage of the preoperative planning workflow to provide surgeons with a more informative and accurate preoperative plan for a better patient outcome. It has been proven that DL can be used for the following modules of the planning process:

- DL to segment the shoulder bones and soft tissues.
- Predicting the best-fit implant size and position based on the size and orientation of articular surfaces, bone density, and the patient's clinical history.
- SSM modelling technology to help quantify glenoid bone defects and virtually reconstruct the premorbid shape of the glenoid to better inform surgeons about the required amount of joint lateralisation and the size of the required bone graft or metallic augment.
- DL to predict impingement-Free ROM, thus preventing potential complications such as notching and instability.
- Patient-specific guides customised based on patient-specific anatomy for a more efficient patient outcome.

Table 8 list the aforementioned software's indication for use and the key features.

9.3. Prediction of clinical outcomes in orthopaedics

Several studies explored the potential of ML techniques to predict clinical outcomes and stratify risk among patients [263,264]. Kumar et al. [263], Franceschetti et al. [265] used preoperative data from 6210 primary patients undergoing shoulder arthroplasty of one prosthesis design to create predictive models for multiple clinical outcome measures using three supervised ML techniques. The results showed that each ML technique accurately predicted each outcome measure at each postoperative point for anatomic total shoulder arthroplasty (aTSA) or reversed total shoulder arthroplasty (rTSA). However, small differences in prediction accuracy were observed between techniques. The models accurately identified patients who achieved and did not achieve clinical improvement that exceeded the minimal clinically important difference (MCID) and substantial clinical benefit thresholds for each outcome measure. These findings suggest that ML techniques can accurately predict clinical outcomes at multiple postoperative points after shoulder arthroplasty and stratify risk by identifying those who may or may not achieve MCID and substantial clinical benefit improvement thresholds for each outcome measure.

Kumar et al. [266] built predictive models for clinical outcomes after shoulder arthroplasty using ML analysis on a dataset of 5774 patients. The full-feature set model and the minimal-feature set model were compared to assess the efficacy of using a minimal feature set as a decision support tool. The XGBoost ML technique created and tested

Table 8
Orthopaedic software using machine learning for different applications.

Software product	Indication for use	Use of ML or DL	Key features	Link
KOALA by Image Biopsy Lab	Knee osteoarthritis	Yes	• Automated knee osteoarthritis grading system • Minimum joint space width • Severity of the joint space narrowing • osteophytosis and sclerosis	https://www.imagebiopsy.com/product/koala-ce
OsteoDetect by Imagen Technologies	Wrist fracture	Yes	• Detection of distal radius fracture, a common type of wrist fracture, using two-dimensional radiographs of adults only.	https://www.fda.gov/news-
FractureDetect by Imagen Technologies	Musculosk-eletal fracture	Yes	• Detection of fracture of twelve different musculoskeletal structures (ankle, clavicle, elbow, femur, forearm, hip, humerus, knee, pelvis, shoulder, tibia/Fibula, and wrist) using two-dimensional radiographs of adults only.	https://imagen.ai/ai-software/
Formus labs	Presurgical planner for Primary hip arthroplasty	Yes	• Automatic segmentation of 3D models of hip joint • Optimal implant selection and positioning.	https://www.formuslabs.com/
Precision AI	Presurgical planner for Shoulder joint replacement	Yes	• Patient-specific surgical plan for shoulder replacement.	https://www.precisionai.com.au/
Akunah Reflect by Akunah	Presurgical planner for shoulder primaries, revisions, fractures and shoulder Instability.	Yes	• Segmentation of the 3D models of the scapula and the Humerus • Anatomic measurements of the Scapula and Humerus • Native joint Line Medialisation • Humerus subluxation • Glenoid bone Loss • Premorbid anatomy of the scapula • Generating planning reports	https://akunah.com/reflect-complex
Signature one by Zimmer Biomet	Presurgical planner for Shoulder joint replacement	Yes	• Preoperative planning of the glenoid component for total shoulder arthroplasty.	https://www.zimmerbiomet.com/content/dam/zbcorporate/en/products/specialties/shoulder/signature-one-planner/2619.1-GLBL-en-Signature-ONE-Features-and-Benefits-of-New-System-and-Guides-Brochure-digital.pdf
The ExactechGPS Total Shoulder Application by Exactech	Presurgical planner for total shoulder arthroplasty	No	• Segmentation of the scapula and the humerus. • Select different implants and sizes.	https://www.exac.com/extremities/exactechgps-shoulder-application/#equinoxe-planning-app
BluePrint 3D planning by Stryker	Presurgical planner for shoulder replacement surgery	No	• Automatic segmentation of scapula and humerus bones from CT scans of adults only. • Visualise, measure, reconstruct and annotate anatomic data • Position and select glenoid and humeral implants. • Simulate the range of motion of the prosthetic. • Generate planning reports.	https://www.shoulderblueprint.com/
Surgicase by Materialise	Presurgical planner for shoulder replacement surgery	No	• 3D and 3D visualisation • Reconstruction of the premorbid scapula • Automated humeral Head diameters • Glenoid and humeral components positioning • Glenoid defect quantification • Generate planning reports	https://www.materialise.com/en/healthcare/mimics-innovation-suite/surgicase
VIP by Arthrex	Presurgical planner for shoulder	No	• Segmentation of the scapula • Position of the glenoid implant	https://www.arthrexvip.com/

predictive models for multiple outcome measures. The study found that the full and abbreviated models had similar precision in predicting clinical outcomes at multiple postoperative time points. The findings suggest that the tool can be easily used during a surgical consultation to improve decision-making related to shoulder arthroplasty.

Roche et al. [267] proposed a new clinical assessment tool, the Smart Shoulder Arthroplasty Score (SAS), constructed using ML, to quantify outcomes after total shoulder arthroplasty (TSA). The SAS score was compared with five other historical assessment tools using data from 3667 TSA patients. The results demonstrated that the SAS score has equivalent or better validity, responsiveness, and clinical interpretability than the other measures analysed. Additionally, the SAS score has an appropriate response range without floor or ceiling effects and without bias in any target patient characteristic, unlike the

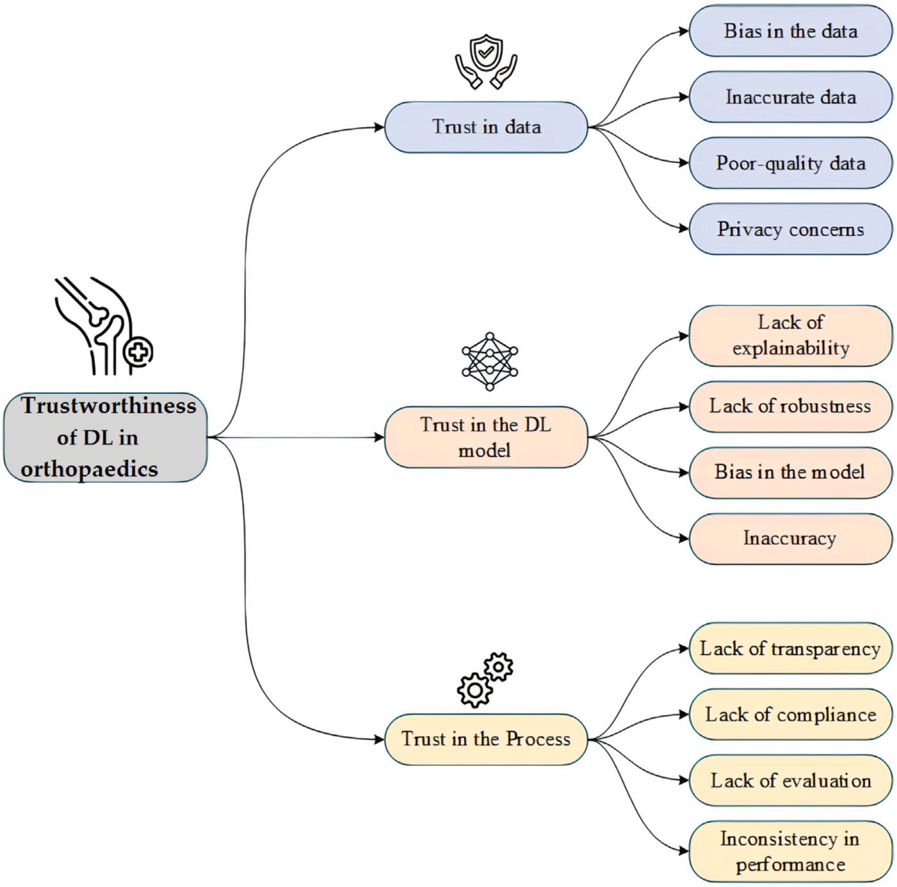


Fig. 21. Three primary elements determining the trustworthiness of DL in orthopaedics.

ceiling scores for the other measures analysed. Therefore, the authors recommend using the new SAS score to quantify the results of the TSA. However, future work is needed to perform external validations and quantify the reliability of this ML-based outcome tool.

The aforementioned studies and more recent work [268,269] have demonstrated the potential of ML to predict clinical outcomes after shoulder arthroplasty using a minimal set of characteristics of only 19 preoperative inputs. Future work in this area can further explore the use of ML to develop more accurate predictive models for clinical outcomes after shoulder arthroplasty. One potential avenue for future research could be the incorporation of more comprehensive and diverse data sources, such as imaging and patient-reported outcome measures, into predictive models. Additionally, further validation of predictive models using larger and more diverse patient populations could be performed to confirm their generalisability and effectiveness in clinical practise. Furthermore, DL models could be used to optimise surgical planning and decision making by predicting which surgical approach or implant would be best suited for a given patient based on their individual characteristics and predicted outcomes.

10. Requirements of building trustworthy deep learning for orthopaedics

A trustworthy DL model is a model that is accurate, fair, transparent, and free of bias [270–274]. Trustworthiness is critical in various spine orthopaedic tasks, including diagnosis, surgical planning, and postoperative follow-up, since DL is increasingly used to make decisions that impact patients' lives. To achieve trustworthiness in the DL orthopaedic model, it should produce accurate results, generalise well to new patients, and provide understandable explanations for its decisions [275–277]. /par

The three primary elements that determine the trustworthiness of DL in orthopaedics include the quality of the data, the quality of the DL model, and the trustworthiness of the process used to develop and deploy the model (see Fig. 21).

10.1. Trust in data

It is essential to have confidence in the data used to train a DL model, as the quality of the training input heavily influences the quality of the model output. Using low-quality data can lead to models that are biased or untrustworthy [278]. Quality can be maintained by ensuring that data are unbiased, accurate, high-quality, and protect privacy. Several factors degrade the trustworthiness of training data:

- Inaccurate data: When the data used to train the DL model are inaccurate, as may be the case when non-experts annotate the training dataset, the model will be less reliable and may lead to incorrect decisions. This is particularly pertinent in healthcare applications, where expert opinion is often required to label training data. There is the possibility of adverse patient outcomes due to DL-based decision making.
- Bias in data: A DL model may develop biased decision-making when trained on biased data. For example, a DL model trained on a dataset of orthopaedic medical images that lacks diversity in terms of race, gender, or age may show bias towards certain groups of people, resulting in inaccurate diagnoses or treatments for patients whose demographics were a minority in the training dataset.
- Poor-quality data: Using data that are not aligned with the task at hand or are not pertinent to the problem the model is being employed to address can result in suboptimal model performance.

An example is using data collected a decade ago for a current task, as it may no longer be valid or relevant, potentially leading to poor model performance. Another example would be to use a dataset that is not representative of the population or environment in which the model will be used. For example, a model trained on images of fracture detection from a single hospital may not perform well when applied to images from other hospitals due to variations in imaging modalities, resolution, or other factors.

- Privacy concerns: Using unprotected data to train a DL model can raise concerns about privacy. For example, suppose that a DL model is trained on a dataset of orthopaedic medical images that includes sensitive personal information such as details of patients, medical conditions, or personal identification numbers. In that case, it may result in privacy breaches if the data is not properly secured.

10.2. Trust in the DL model

Building trust in DL models requires addressing concerns such as explainability, accuracy, bias, and robustness. Ensuring that the model is easy to understand and explainable makes it easier to ensure that the model is making fair and unbiased decisions. To establish the reliability of a DL model, various elements must be taken into account [279–281], including:

- Level of explainability: When the decision-making process of a DL model is not/cannot be explained, it can be hard for humans to comprehend how it arrives at its outcomes. This can result in a lack of confidence in the model, as people may be unsure about how it is coming to its conclusions. This issue has been described earlier.
- Inaccuracy: DL may make flawed decisions and produce inaccurate outcomes, which can damage trust in the model and have negative effects if utilised for critical decision-making.
- Bias in the model: A DL model that is not fair can generate discriminatory or unjust decisions. For example, when a model is trained using biased data, it may replicate those biases when applied to new data, resulting in unfair outcomes.
- Lack of robustness: DL may not perform well or make errors when confronted with new data if they are not robust. This can lead to a lack of trust in the model, as individuals may question its dependability.

10.3. Trust in the process

It is essential to establish trust in the DL model and its evaluation process by carefully checking the evaluation process, complying with regulations and guidelines, and maintaining consistency in the performance of the model. It is necessary to overcome these issues; otherwise, it can result in a lack of trust in the model and the process.

- Lack of transparency: When the development and deployment processes for a DL model are not transparent, it can be challenging for people to comprehend the method of the model and accept how it is being used.
- Lack of compliance: It is crucial to verify that the method used to create and implement a DL model is consistent with applicable regulations and guidelines.
- Inconsistency in performance: A DL model that demonstrates inconsistent results over time or in various situations can be hard to trust.
- Lack of evaluation: Insufficient evaluation can make it difficult to determine the reliability and performance of a DL model.
  Splitting the data set into training and testing data sets is one of the key steps in the DL process. Avoiding bias against underrepresented target classes when training and testing a DL model is important. Different splitting strategies could mitigate data selection bias and ensure the diversity of the test data set to highlight the generalisability of the DL model.
  Stratified sampling is a data division technique that involves dividing a population into homogeneous subpopulations (strata) based on specific characteristics and then sampling each stratum using another probability sampling method [282] (see Fig. 22). This technique ensures that every characteristic of the population is adequately represented in the sample and helps to generalise and validate the study while avoiding research biases such as undercover bias. The DL algorithm must be tested on a large, diverse dataset to show the generalisability of the model and to ensure that the selection criteria are objective.

11. Deep learning and fusion techniques for orthopaedics

The information fusion technique combines various data forms or image modalities to make more reliable and accurate decisions. This could include data from electronic health records, medical devices, research studies, and other sources [283,284]. The aim of this process is to improve patient outcomes by providing healthcare professionals with results based on a more comprehensive understanding of the patient's condition. Fusion techniques can help with data scarcity issues and reduce the chance of overfitting. The orthopaedic field frequently generates multiple data modes for individual patients, including patient records, MRI scans, CT scans, and X-rays [285]. Therefore, the use of fusion techniques has become increasingly important in the analysis and interpretation of such data. As a result, implementing these techniques can enhance the outcomes and increase confidence in the final decision [286].

There are several fusion types; this section will discuss four techniques for orthopaedic tasks with DL fusion.

11.1. Feature fusion

This technique is based on the extraction of features using two or more models based on DL such as CNN, then the fusion of the extracted features, which will be used to train ML classifiers as shown in Fig. 23. Feature fusion can be achieved by concatenating, averaging or combining features in some other way [287]. The aim of feature fusion is to produce a more robust and informative representation of the data. By accomplishing this, this technique can enhance the performance of ML classifiers [288].

Dang et al. [289] developed a feature fusion algorithm to automatically detect Kashin–Beck disease (KBD) based on hand radiograph images. The KBD diagnosis method uses multi-feature fusion for classification. Two types of features are extracted from X-ray images using a DCNN and then combined and fed into a fully connected neural network (FCNN) to obtain diagnostic results. Experiments on a data set of 960 samples in KBD endemic areas of Tibet show that the multifunctional method achieved an average accuracy and sensitivity rate of 98.5% and 97.6% for diagnosis, which is 4.0% and 7.6% higher than the method using only global features. The proposed multi-feature fusion method substantially reduces large-scale screening costs and missed diagnosis rates in rural China.

Deep feature extraction and fusion are performed using pre-trained convolutional models Darknet-53 and Xception and hand-made features such as HOG and LBP for elbow by Malik et al. [165]. Principal component analysis (PCA) is used to select the best features, which are then supplied to the SVM, KNN, and NN classifiers. The proposed method is evaluated on 16,984 X-ray radiographs from the MURA dataset, achieving a precision of 97. 1% and a kappa score of 0.943%.

With the two aforementioned examples, the feature fusion technique improved the results in two orthopaedic tasks.

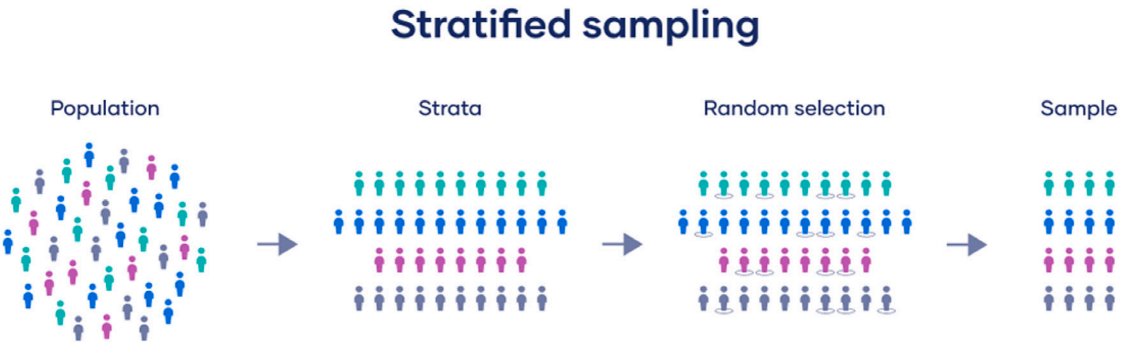


Fig. 22. Stratified sampling technique [282].

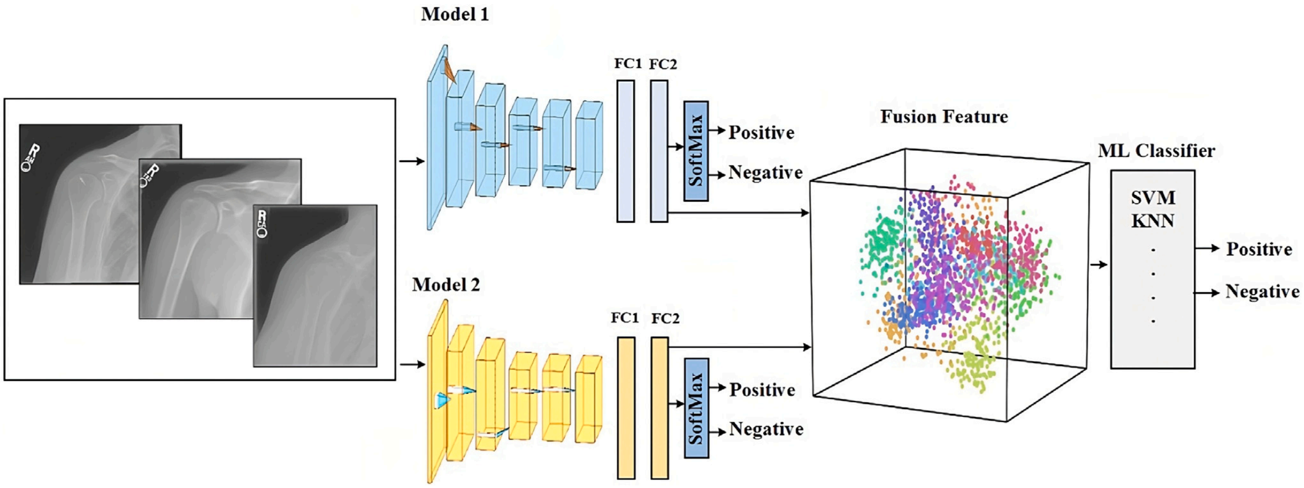


Fig. 23. Example of feature fusion in orthopaedics [169].

11.2. Image fusion

The image fusion technique aims to combine multiple images into one image that provides a more comprehensive and accurate patient description. This technique offers several advantages, such as improved information quality, an expanded range of operations, increased spatial and temporal coverage, reduced uncertainty, enhanced reliability, robust performance, and a more concise representation of the information [290,291].

Yoshii et al. [292] described the development of an image fusion system for 3D preoperative planning and fluoroscopy for the operative fixation of fractures. The study aimed to evaluate the reproducibility of preoperative planning in open reduction and internal fixation of distal radius fractures using the image fusion system and compare it with patients who did not use the same system. The results showed that the image fusion group had significantly smaller differences in plate-to-joint surface distances and distal screw choices than the control group. The study concludes that the image fusion system was useful in reproducing planned plate positions and distal screw choices in the osteosynthesis of distal radius fractures.

Image fusion can be helpful in segmentation tasks where two image modalities can be combined to produce a more precise result [293]. Another approach is to combine the segmentation results of different DL algorithms to achieve a more accurate segmentation result [293].

11.3. Decision fusion

Decision fusion techniques, as shown in Fig. 24, are vital to reduce overfitting and improve multimodal learning in the field of orthopaedics. The purpose of decision fusion is to combine the outputs or decisions of various algorithms, including DL models, to produce a final decision that is more accurate and reliable [294]. In orthopaedic applications, decision fusion can be particularly useful in situations where there are multiple modalities with multiple DL models, as well as in scenarios involving a single modality with multiple DL models. Decision fusion is a technique that helps integrate information from various sources, such as CT, MRI, clinical measurements, and patient records when dealing with multiple modalities and multiple DL models. This technique aggregates these inputs using methods such as majority voting or weighted voting. Decision fusion helps capture complementary information from each modality, reducing the risk of overfitting by ensuring that the final decision is not too dependent on any single modality or DL model. By combining different DL models that are trained on different modalities, decision fusion leverages the strengths of each model to enhance overall performance and robustness. Decision fusion can combine their outputs when working on a task involving a single type of data and multiple deep-learning models. This is particularly useful when the models are trained on different subsets of data or with different architectures. Decision fusion helps prevent overfitting and improve the generalisability of models by leveraging their diversity to capture different aspects of the data. The fusion techniques used in decision fusion enable the creation of a more comprehensive and reliable final decision.

11.4. Multi-modal fusion

The technique of multi-modal fusion involves combining data from different modalities or types of data. This method is known to improve the accuracy and reliability of decisions [295]. The main objective of

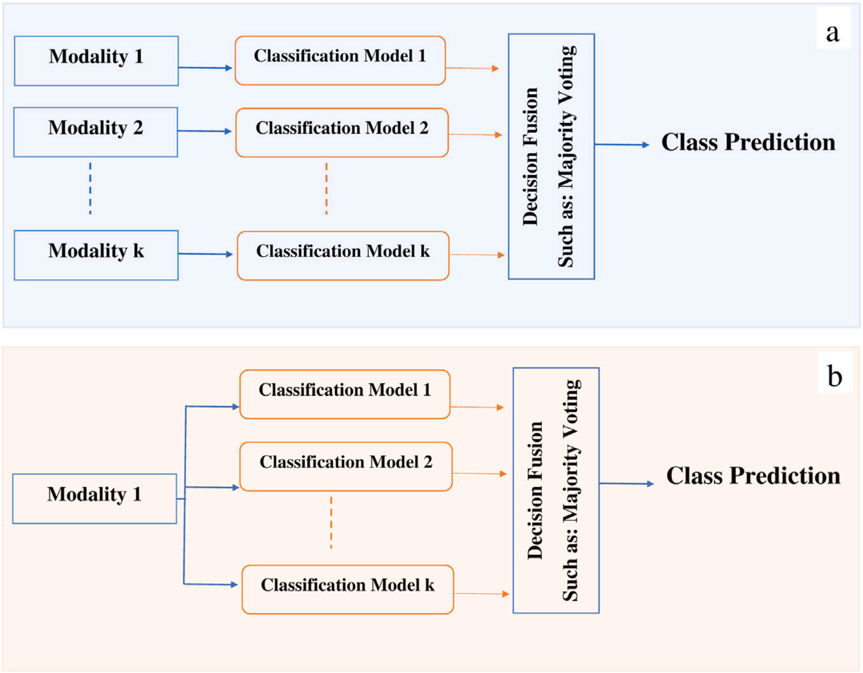


Fig. 24. Types of decision fusion in Orthopaedics: (a) regards to multiple modalities and multiple DL models; (b) regards to single modality and multiple DL models.

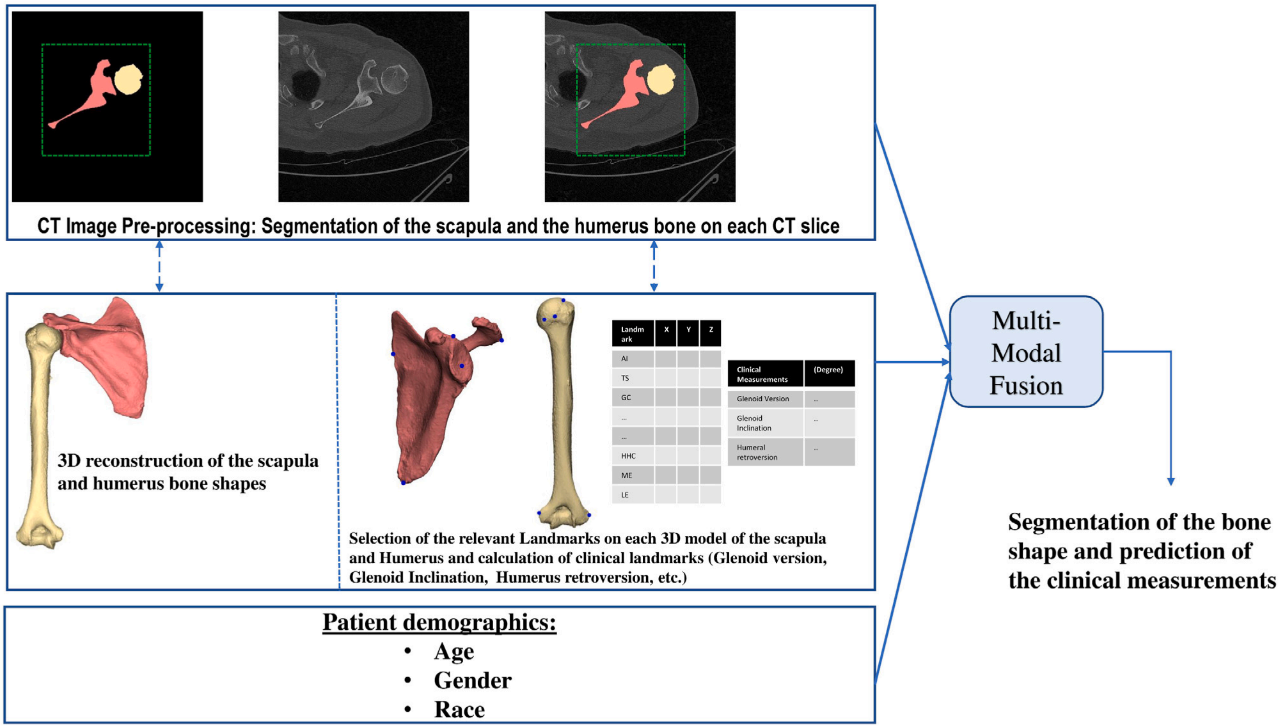
multimodal fusion is to build a more robust and informative representation of the data, which enhances the performance of ML-based models and decision-making processes.

Fig. 25 is a crucial illustration in the context of our research on orthopaedics. It demonstrates the practical use of multi-modal data fusion, a technique that combines information from various sources to enhance decision-making processes. This figure depicts the integration of three different inputs: CT scans, 3D models, 3D landmarks, clinical measurements, and patient records. These inputs are used together to make informed decisions about individual patients. Segmentation of the scapula and humerus bones, together with 3D reconstruction and landmarking of key points, highlights the complexity and depth of the data used in the orthopaedic analysis. This figure emphasises the importance of incorporating diverse datasets to fully understand orthopaedic conditions. Using multiple modalities, such as imaging scans, clinical measurements, and demographic data of the patient, we can obtain more accurate and reliable information. Ultimately, this approach leads to improved diagnostic accuracy and treatment planning, which contributes to better patient outcomes.

12. FDA approval requirements for DL application

The US FDA, Health Canada, and the Medicines and Healthcare Products Regulatory Agency (MHRA) of the United Kingdom (UK) have collaboratively established ten principles to assist in the development of Good ML Practise (GMLP) for medical devices that utilise AI/ML [296, 297]. These requirements aim to ensure that AI/ML-based medical devices are safe, effective, and of high quality. Section 10 of this article has already explained most of these requirements. These include:

- **Multi-Disciplinary Expertise Leveraging Throughout The Total Product Life Cycle:** Using multidisciplinary expertise through out the complete life cycle of a product is crucial. By comprehensively understanding how a model is intended to be integrated into the clinical workflow and its associated benefits and risks to patients, it is much easier to ensure the safety and effectiveness of medical devices that incorporate ML. This approach also ensures that the device addresses clinically significant needs over its lifespan.
- **The Implementation Of Good Software Engineering and Security Practises:** Implementation of good software engineering and security practises should be ensured during model design. This includes data quality assurance, data management, and cybersecurity practises. These practices are integrated into a risk management process that can effectively capture and communicate design decisions and ensure the authenticity and integrity of the data. Attention must be paid to the fundamentals of good software engineering practises.
- **Clinical Study Participants and Dataset Representation of the Intended Patient Population:** It is essential to ensure that the participants and data sets employed are representative of the patient population of interest. This requires data collection protocols that consider relevant features such as age, sex, race, and ethnicity, as well as use and measurement inputs. Adequate sample sizes should be used in clinical studies, training, and test data sets to manage any bias. It is also helpful to promote appropriate and generalisable performance in the intended patient population, assess usability, and identify situations where the model may fail. Broadly speaking, the study's results can be generalised to the population of interest by ensuring that the datasets and participants represent the intended patient population.
- **Independent Test Sets:** It is essential to check that training datasets are independent of test sets. This means that the data used for training the DL model should be separate and distinct from the data used to evaluate the model's performance. This principle is essential to promote unbiased and reliable evaluation of the model's performance and to avoid overfitting or underfitting, which can lead to poor generalisation of new and unseen data. To ensure independence, all factors that can cause dependence, such as those related to the patient, data acquisition, and site, must be considered and addressed accordingly.
- **Best Available Methods for Dataset selection:** The selected reference datasets should be established based on the best available methods to ensure that they represent the current standard in clinical practise. Reference data sets must be comprehensive and include a range of inputs to cover a variety of medical cases. Moreover, it is necessary to check that the datasets are high


CT Image Pre-processing: Segmentation of the scapula and the humerus bone on each CT slice
3D reconstruction of the scapula and humerus bone shapes
Selection of the relevant Landmarks on each 3D model of the scapula and Humerus and calculation of clinical landmarks (Glenoid version, Glenoid Inclination, Humerus retroversion, etc.)
Patient demographics:
• Age
• Gender
• Race
Multi-Modal Fusion
Segmentation of the bone shape and prediction of the clinical measurements

Landmark	X	Y	Z
AI			
TS			
GC			
...			
...			
HHC			
ME			
LE			

Clinical Measurements	(Degree)
Glenoid Version	..
Glenoid Inclination	..
Humeral retroversion	..

Fig. 25. Example of multi-model data fusion in Orthopaedics from our ongoing work.

quality, well-characterised, and annotated. The datasets should be updated on a regular basis to incorporate new and improved methods as they become available. Widely accepted reference datasets should be utilised to develop and test the DL model. This is essential to prove the generalisability and robustness of the model across the intended patient population.

- **Model Design Consideration of Available Data and Intended Device Use:** The design of the model is customised to fit the available data and allows the proactive mitigation of recognised risks such as overfitting, performance degradation and security vulnerabilities. The clinical risks and benefits related to the device are fully understood and used to determine meaningful performance objectives for testing, ensuring that the device can safely and effectively fulfil its intended purpose. This includes assessing the impact of global and local performance and variability/uncertainty in device inputs, outputs, target patient populations, and clinical usage conditions.
- **The Human-DL Team:** The performance of the human-DL team is prioritised in situations where the model includes human involvement. Human factors and interpretability of model outputs are considered, focussing on the performance of the Human-DL team rather than solely on the model's performance.
- **The Demonstration Of Device Performance During Clinically Relevant Conditions:** Testing involves creating statistically sound plans to produce clinically significant information that is separate from the training data set, reflecting the performance of the device in relevant clinical situations. This includes factors such as the intended patient population and important subgroups, the clinical environment and use by the Human-DL team, the measurement input, and possible confounders.
- **Information to Users:** Users receive clear and essential information to easily access relevant instructions suitable for healthcare providers or patients. Information includes the intended use and indications, the performance of the model for subgroups, data characteristics, inputs, limitations, interpretation of the user interface, and integration of the clinical workflow. Users are updated on device modifications and real-world performance monitoring, with transparent decision-making and a means to report concerns to developers.
- **Risk Management:** The performance of the deployed models is continuously monitored to ensure safety and improve efficiency in real world settings. Appropriate measures are also in place to manage the risks of overfitting, unintended bias, or model degradation during periodic or continuous post-deployment training. These measures are essential to ensure the safety and performance of the model as the Human-DL team uses it.

The FDA requirements mentioned above are not enough for DL applications in orthopaedics. The current FDA approval process for medical devices and software lacks a focus on health equity, as it mainly considers Caucasian populations and women as patient populations [298]. For example, medical imaging devices, pulse oximeters, and infrared thermometers have been implemented without being tested on a representative cohort of patients, resulting in readout error. This bias results in a reduced quality of care for marginalised groups. Extending the requirements of the pre-market approval pathway is necessary to advance health equity. The extension should require manufacturers to test their devices and software on diverse patient populations and provide information on the composition of patients who participated in the design and calibration of the device or software. The clinical investigation sections of the strict premarket pathway can also be improved in terms of study protocols, patient information, and study design. Efforts to improve representation in clinical testing are not unprecedented and should be expanded to ensure health equity [299].

13. Discussion

Since 2019, research studies on DL algorithms have been actively contributed in various medical fields, including orthopaedics, ophthalmology, dermatology, and cardiology. This trend is predicted to continue until the 'new winter' is reached, when AI development will reach its limit and plateau. So far, the application of DL methods in imaging studies of orthopaedic diseases has demonstrated exceptional results [300,301]. Several studies have stated that trained CNN models exhibit satisfactory classification results and diagnostic accuracy comparable to human experts in areas such as traumatology (fractures) and osteoarthritis. However, the assessment of small joints recorded fairly undesirable outcomes compared to large joints. When comparing binary diagnosis and multiclass classification, the precision of the first task consistently outperforms the accuracy of the second.

It is believed that these limitations can be addressed for two reasons. To address the relative weakness of nonbinary classification systems, a CNN model can be developed for medical image analysis to provide an accurate diagnosis and precise classification. The type of class needed for this process is relatively small. Specifically, Paoletti et al. [302] showed that using fewer class types in a deep-layer structured CNN model improved accuracy. The accuracy of multiclass classification is expected to enhance the development of a CNN model with increasing deep hyperparameters via medical image analysis.

On the other hand, medical images are highly refined data compared to images used to learn climate predictions or traffic conditions. As such, relevant noise-free image data can easily be obtained, such as the different heights of flying birds or traffic lights. Therefore, a proper data set can be produced to train the CNN model, even using simple data augmentations, such as the affine transformation. Overall, the advanced configuration of a CNN model and the accumulation of supplementary medical images is expected to increase the currently poor classification accuracy of osteoarthritis and fractures compared to the precision of diagnosis. Currently, the development of DL algorithms that are beneficial for segmentation is expected to improve the diagnosis and classification of joint-specific soft tissue [300].

Several notable studies have achieved high-level segmentation [303, 304]. U-Net is among the essential CNN semantic segmentation frameworks [305], with a robust design to recognise structural edges. In another study, Hiasa et al. [306] used a trained U-Net-based CNN model to segment the psoas major muscle and achieved an average intersection-over-union (IoU) of 86 6%. Thus, the U-Net-based CNN model is expected to be extensively applied to segment medical images. Interestingly, new U-Net-based CNN designs are being implemented steadily beyond the field of orthopaedics and have reported favourable results [307]. For example, Wang et al. [308] incorporated squeeze and excitation blocks (SE) into U-Net (SAR-U-Net) to perform zonal prostate segmentation. On the contrary, Yeung et al. [309] reported an improved U-Net through a trained dual attention-gated CNN (Focus U-Net) model and satisfactorily segmented the polyp colonoscopy image. Therefore, there are currently studies in orthopaedic surgery that have developed CNN models with high-level diagnosis and classification compared to human experts. More work is underway to improve the segmentation of medical images. For instance, Zhao et al. [310] proposed a femur segmentation method based on an improved U-Net network to address the challenges in segmenting the femur from CT bodies, including missed detection, false detection, and low segmentation accuracy. The proposed method introduced a residual module and an attention mechanism to enhance the features of small target femurs. Experimental results demonstrate that the proposed method achieves higher Intersection over Union (IoU), Recall, Precision, and F score than existing semantic segmentation networks such as U-Net, ResNet, and SegNet. The method can focus more on segmentation of small target femurs without affecting the segmentation of large target femurs, improving the overall segmentation performance of femur images.

Furthermore, a thorough investigation is required to determine the impact of data accumulation or the development of enhanced CNN to overcome such problems or verify that the limitation of a trained CNN model using image data is natural and unavoidable. Here, two approaches are presented. First, experts can address not only image data problems through the use of other information, such as demographic data of patients, the nature of the disorder, the level of pain, and a physical evaluation that impacts the diagnosis and classification of the disease. A report by Kim et al. [63] stated that a trained CNN model using additional information, such as demographic (age, BMI and sex), alignment and metabolic data that could influence knee osteoarthritis, achieved a higher statistically significant AUC. This addresses the current limitation of CNN models based solely on image analysis using DL algorithms. Although an enhanced CNN model is constructed and high-quality image data is gathered, they are potentially unable to match the level of experts who are able to include multiple factors in their decision-making. This technique can also increase trust in the final decision of the models by considering factors similar to those of experts. Multimodal fusion techniques can also help with issues by combining the outcome of the image and patient records, increasing trust and helping compensate for data scarcity [311–313].

In the authors' opinion, it is premature to exclude the possibility that CNN models would succeed at the level of experts in specific fields. However, the opposing views are outlined above, as CNN models evaluate images from a different perspective than human beings. For example, Langerhuizen et al. [43] included 23 scaphoid fracture data and 150 scaphoid fracture images that could only be verified by MRI analysis. Although the trained CNN model recorded a lower accuracy level than expert orthopaedic surgeons, the model identified five out of six occult scaphoid fractures that all orthopaedic surgeons missed. Hence, it is essential that there be a detailed discourse on image analysis models' ability to surpass those of human experts using DL in particular areas.

While ample room exists to enhance the current CNN models, the significance of studies conducted to date should not be undermined. Their potential contributions to clinical practise are significant. The currently available CNN model can be used to minimise the intensity of the task of expert readers as well as a reference to educate nonexpert medical professionals, such as specialists during training or medical students [314–316]. With the help of a developed CNN model, a paediatrician can use X-rays to approximate the bone age of patients without the supervision of an orthopaedic surgeon. More attractive and feasible studies that practically facilitate the interaction of patients and doctors will prevent clinical doctors from dealing with the issue of the accuracy of CNNs. For example, Mendes et al. [317] produced high resolution images by converting native medical CT scan images using Generative Adversarial Networks (GANs), and this study has the potential to be applied to enhance the resolution of MRI images [318]. This technology can offer high-quality magnetic resonance results to areas that would otherwise have restricted access to high-quality magnetic resonance imaging due to limited medical infrastructure or cost issues.

Several limitations have been highlighted following the review of DL approaches for orthopaedic diseases by image analysis. First, the CNN model is the only model approved by the US FDA to predict bone age in children and diagnose wrist fractures [258]. In comparison, the FDA has approved several models in other medical specialities as early as April 2018, beginning with a DL-based model for the automated diagnosis of diabetic retinopathy. Second, no prospective study has yet been reported in orthopaedics [319]. In 2020, a future and randomised control trial (RCT) based on the CONSORT-AI guidelines was proposed to enhance the quality of research and continue functional studies, which will be necessary [320]. Third, the most recently designed DL models are developed to perform a single task. Therefore, numerous DL algorithms are needed to assess each possible abnormality to ensure its usefulness in clinical practise. Several initiatives have been taken to address these drawbacks. For example, Grauhan et al. [321] proposed a CNN model to diagnose joint dislocation, osteoarthritis, and fractures through plain shoulder radiographs.

Ultimately, there is a requirement to minimise expert biases on a specific dataset. Traditionally, orthopaedic surgeons have used CT scans, magnetic resonance imaging, or ultrasounds to diagnose soft tissue diseases. However, DL algorithms frequently provide appropriate judgments that exceed human cognition. Previously, Kang et al. [322] introduced a trained CNN model using axillary lateral radiographs to diagnose subscapularis (SSC) tears and showed an acceptable accuracy

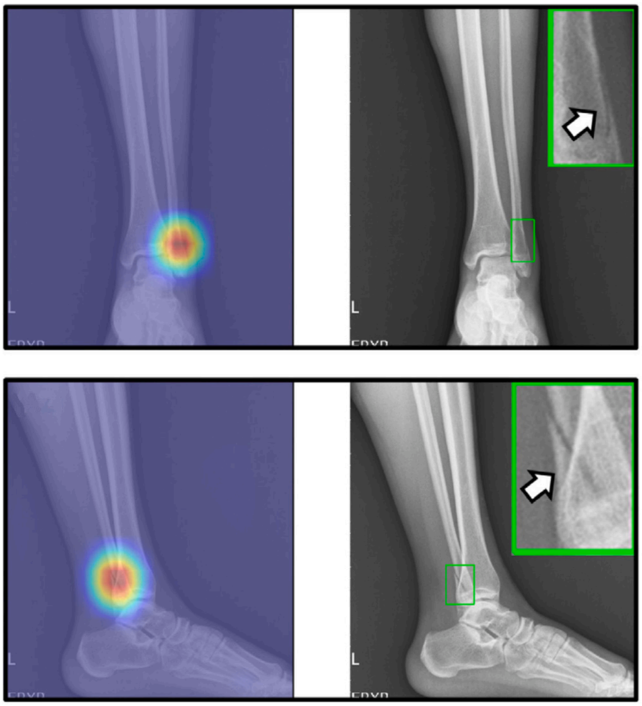


Fig. 26. Tiny fractures were detected by DL [324].

level. Therefore, orthopaedic surgeons are free to construct CNN models according to their imagination without bias.

The increasing use of black-box DL models in orthopaedics has raised concerns among stakeholders about the lack of transparency and interpretability of these models, which could result in unjustifiable or illegitimate decisions [196,197]. Supporting explanations of the model's output are necessary for the medical field. Interpretation techniques can be utilised to enhance the understanding of these models, which can be adopted after training or integration into the network. Post-training methods that utilise test images to evaluate the predictions of a trained network can be a time-saving and preferable option. One such technique is visualisation, which takes advantage of visual representations of network observations to describe predictions [323]. Visualisation techniques such as low-dimensional projections, heat maps, feature importance maps, and saliency maps can be used to understand network behaviour [276]. Incorporating visualisation techniques can improve the reliability and precision of model predictions in real-world scenarios.

In summary, the application of the DL model for image analysis is receiving a growing interest at a rapid pace, demonstrating a remarkable milestone in orthopaedics. The advanced innovation of CNN designs and the accumulation of high-resolution image data are expected to contribute to the development of highly advanced models. However, it is challenging to foresee the extent to which the development of DL models would outperform human experts' capacity [169]. DL can help clinical physicians identify complex orthopaedic problems and manage patients early to mitigate additional medical costs and impact on their quality of life (see Fig. 26) [324].

Surgeons aim to use DL methods for medical image analysis, but they seek trustworthy AI systems with precise, fair, and interpretable decision making processes.

Our review is a valuable resource for both researchers and surgeons who are interested in the potential applications of DL in orthopaedics. We discuss the existing challenges and highlight the untapped potential of DL to foster collaboration between researchers and surgeons. By working together, they can develop and implement trustworthy DL solutions in orthopaedics. This collaboration is crucial for ensuring the successful integration of DL technologies into clinical practice, leading to improved patient care and outcomes.

14. Conclusions and future directions

DL has gained significant attention in recent years in the field of orthopaedics. It has been demonstrated that DL can be applied to various orthopaedic tasks, including fracture detection, bone tumour diagnosis, implant recognition, and evaluation of the severity of osteoarthritis. DL has been shown to provide more accurate and efficient diagnoses than traditional methods while reducing the cost and time of diagnosis for patients and surgeons. This conceptual review presented the state of the art of DL in orthopaedics, including its applications, challenges, and potential solutions. Our work also highlighted the need to build trustworthy DL applications considering high accuracy, explainability, and fairness. DL can greatly improve the diagnosis and treatment planning in orthopaedics, and further research should focus on addressing the challenges and realising the full potential of DL in this field.

DL has numerous possibilities in orthopaedics to improve patient outcomes and advance the field. Suggested potential future directions of research include:

- More effort is needed to develop computer-assisted diagnosis using DL to assist physicians in diagnosing orthopaedic conditions and reduce diagnostic errors.
- Virtual 3D models of bones and joints with DL could be used to enable precision surgery by assisting surgeons in planning and performing procedures.
- Further development and implementation of predictive analytics using DL to predict patient outcomes and identify those at risk of complications.
- DL algorithms can be used to analyse patient movement and provide feedback to therapists, allowing the use of customised rehabilitation and physical therapy techniques.
- Wearable devices with integrated DL models to monitor patient activity and provide real-time feedback to healthcare providers.
- Patient education with personalised education programmes to help patients understand their condition and manage symptoms.
- DL models can identify abnormalities and diagnose orthopaedic conditions using medical images, reducing diagnostic errors and improving outcomes. More effort is required to provide DL models with high-quality large data to be trained well and produce accurate decisions.
- Described AI algorithms are being developed to increase transparency and interpretability, allowing healthcare providers to understand the decision-making process of algorithms. This will build trust and ensure the appropriate use of the technology.
- Assistant, operating, navigating, and guiding robots in ortho paedics will become more effective and efficient with advances in DL. DL models can analyse medical images and assist with surgical planning, improving accuracy and reducing complications. These robots can also be trained to navigate and guide during surgery, reducing surgical time and improving patient outcomes. The use of DL has the potential to revolutionise orthopaedic surgery, but more research and development are necessary to fully realise its potential.
- DL applications in orthopaedics should prioritise addressing trustworthy requirements such as fairness, accuracy, and privacy at an early stage to ensure their successful implementation.
- Federated learning is a DL technique that allows organisations or subgroups within an organisation to train and improve a shared global DL model collaboratively. However, the emergence of data fusion technology has presented new challenges for federated learning, such as multisource and heterogeneous data fusion.

Better utilisation of data and models in federated learning is necessary due to the increasing variety and quantity of data. Eliminating redundant data and combining various data sources can produce valuable new information. Future work should address challenges such as maintaining user privacy, designing universal models, and ensuring stability in data fusion results to facilitate the effective utilisation of data in federated learning with orthopaedics and other domains.

- To effectively use data and models in adversarial DL frameworks and applications, it is necessary to develop methodologies and algorithms to handle the increasing quantity and variety of data and systems. This requires focussing on processing heterogeneous data from multiple data fusion and intelligent systems in adversarial DL while ensuring universality and data fusion result stability. Developing advanced attack and defence mechanisms to explore weaknesses in modern deep learning architectures in data fusion and intelligent systems is a high-demand research task that requires prioritisation in orthopaedics and other domains.
- DL for orthopaedics should be explored more, and its potential application in areas such as spine surgery, bone healing, pain management, physical therapy, and prosthetics should be explored. For example, DL models could aid surgeons to perform spinal surgeries with greater precision and precision while predicting the healing times and outcomes of bone fractures. Additionally, DL could help identify patients at high risk of developing chronic pain after orthopaedic surgery, personalise physical therapy plans, and develop advanced prosthetics that better mimic the natural movement of the limb. Further research is needed to evaluate the potential benefits of DL in these areas.
- Finally, to improve the search methodology, several important improvements could be proposed. Firstly, expanding the search strategy by incorporating additional relevant keywords or variations can ensure a more comprehensive coverage of the literature. Second, considering the inclusion of supplementary databases or repositories dedicated to orthopaedics or deep learning research can further increase the scope of the study, potentially capturing valuable publications not indexed in the primary databases. These adjustments aim to broaden the search horizon and maximise the retrieval of pertinent literature.

Funding

The authors would like to acknowledge the support received through the following funding schemes of the Australian Government: Australian Research Council (ARC) Industrial Transformation Training Centre (ITTC) for Joint Biomechanics under grant IC190100020.

Laith Alzubaidi would like to acknowledge the support received through the Centre for Data Science First Byte Scheme, The Queensland University of Technology.

The authors would like to acknowledge the support received through the Akunah company to fund the resources and medical skills.

CRediT authorship contribution statement

Laith Alzubaidi: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Khamael AL-Dulaimi:** Data curation, Methodology, Validation, Writing – original draft. **Asma Salhi:** Data curation, Formal analysis, Methodology, Writing – original draft, Writing – review & editing. **Zaenab Alammar:** Investigation, Methodology, Writing – original draft, Writing – review & editing. **Mohammed A. Fadhel:** Data curation, Formal analysis, Methodology, Visualization, Writing – original draft, Writing – review & editing. **A.S. Albahri:** Data curation, Formal analysis, Methodology, Validation, Writing – original draft, Writing – review & editing. **A.H. Alamoodi:** Investigation, Methodology, Writing – original draft, Writing – review & editing, Formal analysis. **O.S. Albahri:** Formal analysis, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Amjad F. Hasan:** Methodology, Writing – review & editing, Validation, Writing – original draft. **Jinshuai Bai:** Data curation, Formal analysis, Methodology, Writing – original draft, Writing – review & editing. **Luke Gilliland:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Jing Peng:** Conceptualization, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Marco Branni:** Formal analysis, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Tristan Shuker:** Formal analysis, Validation, Writing – original draft, Writing – review & editing. **Kenneth Cutbush:** Data curation, Investigation, Supervision, Validation, Writing – original draft, Writing – review & editing. **Jose Santamaría:** Data curation, Formal analysis, Supervision, Validation, Writing – original draft, Writing – review & editing. **Catarina Moreira:** Data curation, Formal analysis, Methodology, Validation, Writing – original draft, Writing – review & editing. **Chun Ouyang:** Formal analysis, Methodology, Supervision, Writing – original draft, Writing – review & editing. **Ye Duan:** Formal analysis, Validation, Visualization, Writing – original draft, Writing – review & editing. **Mohamed Manoufali:** Writing – original draft, Writing – review & editing, Data curation, Formal analysis, Validation. **Mohammad Jomaa:** Methodology, Writing – original draft, Writing – review & editing. **Ashish Gupta:** Data curation, Formal analysis, Methodology, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Amin Abbosh:** Validation, Writing – review & editing. **Yuantong Gu:** Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing.

Declaration of competing 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.

References

[1] Mehdian R, Howard M. Artificial intelligence in trauma and orthopaedics. Artif Intell Med 2021;1–14.

[2] England N. Musculoskeletal conditions. 2023, https://www.england.nhs.uk/elective-care-transformation/best-practice-solutions/musculoskeletal/. [Accessed 1 January 2023].

[3] United States Bone and Joint Initiative. United States bone and joint initiative. 2023, https://www.boneandjointburden.org/. [Accessed 1 January 2023].

[4] Lee J, Chung SW. Deep learning for orthopedic disease based on medical image analysis: Present and future. Appl Sci 2022;12(2):681.

[5] Italia K, Launay M, Gilliland L, Nielsen J, Pareyon R, Hollman F, Salhi A, Maharaj J, Jomaa M, Cutbush K, et al. Single-stage revision reverse shoulder arthroplasty: Preoperative planning, surgical technique, and mixed reality execution. J Clin Med 2022;11(24):7422.

[6] Vaishya R, Scarlat MM, Iyengar KP. Will technology drive orthopaedic surgery in the future? Int Orthop 2022;1–3.

[7] Evans JT, Evans JP, Walker RW, Blom AW, Whitehouse MR, Sayers A. How long does a hip replacement last? A systematic review and meta-analysis of case series and national registry reports with more than 15 years of follow-up. Lancet 2019;393(10172):647–54.

[8] NJR. NJR centre. 2023, https://www.njrcentre.org.uk/. [Accessed 2 January 2023].

[9] Ren M, Yi PH. Artificial intelligence in orthopedic implant model classification: a systematic review. Skelet Radiol 2022;51(2):407–16.

[10] Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, Santamaría J, Fadhel MA, Al-Amidie M, Farhan L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. J Big Data 2021;8(1):1–74.

[11] Uysal F, Hardalaç F, Peker O, Tolunay T, Tokgöz N. Classification of shoulder X-ray images with deep learning ensemble models. Appl Sci 2021;11(6):2723.

[12] Alammar Z, Alzubaidi L, Zhang J, Santamaréa J, Li Y. A concise review on deep learning for musculoskeletal X-ray images. In: 2022 international conference on digital image computing: techniques and applications. DICTA, IEEE; 2022, p. 1–8.

[13] Nich C, Behr J, Crenn V, Normand N, Mouchère H, d'Assignies G. Applications of artificial intelligence and machine learning for the hip and knee surgeon: current state and implications for the future. Int Orthop 2022;1–8.

[14] Tiwari A, Yadav AK, Bagaria V. Application of deep learning algorithm in automated identification of knee arthroplasty implants from plain radiographs using transfer learning models: Are algorithms better than humans? J Orthop 2022.

[15] Yılmaz A. Shoulder implant manufacturer detection by using deep learning: Proposed channel selection layer. Coatings 2021;11(3):346.

[16] Levin JM, Lorentz SG, Hurley ET, Lee J, Throckmorton TW, Garrigues GE, MacDonald P, Anakwenze O, Schoch BS, Klifto C. Artificial intelligence in shoulder and elbow surgery: Overview of current and future applications. J Shoulder Elbow Surg 2024.

[17] Liu P, Zhang J, Liu S, Huo T, He J, Xue M, Fang Y, Wang H, Xie Y, Xie M, et al. Application of artificial intelligence technology in the field of orthopedics: a narrative review. Artif Intell Rev 2024;57(1):13.

[18] Lans A, Pierik RJ, Bales JR, Fourman MS, Shin D, Kanbier LN, Rifkin J, DiGiovanni WH, Chopra RR, Moeinzad R, et al. Quality assessment of machine learning models for diagnostic imaging in orthopaedics: a systematic review. Artif Intell Med 2022;132:102396.

[19] Xu R, Tang J, Li C, Wang H, Li L, He Y, Tu C, Li Z. Deep learning-based artificial intelligence for assisting diagnosis, assessment and treatment in soft tissue sarcomas. Meta-Radiology 2024;100069.

[20] Daggett SM, Cantarelli T, Gyftopoulos S, Krueger P, Ross AB. Cost-effectiveness analysis in diagnostic musculoskeletal radiology: A systematic review. Curr Probl Diagn Radiol 2022.

[21] Feldman V, Atzmon R, Dubin J, Bein O, Palmanovich E, Ohana N, Farkash U. Thousand shades of gray–the role of imaging display in diagnosis of occult scaphoid fractures–A pilot study. J Orthop 2022;34:327–30.

[22] Clementson M, Björkman A, Thomsen NO. Acute scaphoid fractures: guidelines for diagnosis and treatment. EFORT Open Rev 2020;5(2):96–103.

[23] Mascio A, Greco T, Maccauro G, Perisano C. Lisfranc complex injuries management and treatment: current knowledge. Int J Physiol Pathophysiol Pharmacol 2022;14(3):161.

[24] Grewal US, Onubogu K, Southgate C, Dhinsa BS. Lisfranc injury: a review and simplified treatment algorithm. Foot 2020;45:101719.

[25] Gyftopoulos S, Lin D, Knoll F, Doshi AM, Rodrigues TC, Recht MP. Artificial intelligence in musculoskeletal imaging: current status and future directions. AJR Amer J Roentgenol 2019;213(3):506.

[26] Chen P-T, Wu T, Wang P, Chang D, Liu K-L, Wu M-S, Roth HR, Lee P-C, Liao W-C, Wang W. Pancreatic cancer detection on CT scans with deep learning: a nationwide population-based study. Radiology 2023;306(1):172–82.

[27] Lin DJ, Schwier M, Geiger B, Raithel E, von Busch H, Fritz J, Kline M, Brooks M, Dunham K, Shukla M, et al. Deep learning diagnosis and classification of rotator cuff tears on shoulder MRI. Invest Radiol 2023.

[28] Germann C, Marbach G, Civardi F, Fucentese SF, Fritz J, Sutter R, Pfirrmann CW, Fritz B. Deep convolutional neural network–based diagnosis of anterior cruciate ligament tears: performance comparison of homogenous versus heterogeneous knee MRI cohorts with different pulse sequence protocols and 1.5-T and 3-T magnetic field strengths. Invest Radiol 2020;55(8):499.

[29] Caliva F, Namiri NK, Dubreuil M, Pedoia V, Ozhinsky E, Majumdar S. Studying osteoarthritis with artificial intelligence applied to magnetic resonance imaging. Nat Rev Rheumatol 2022;18(2):112–21.

[30] Kim H, Shin K, Kim H, Lee E-s, Chung SW, Koh KH, Kim N. Can deep learning reduce the time and effort required for manual segmentation in 3D reconstruction of MRI in rotator cuff tears? PLoS One 2022;17(10):e0274075.

[31] Meena T, Roy S. Bone fracture detection using deep supervised learning from radiological images: A paradigm shift. Diagnostics 2022;12(10):2420.

[32] Negrillo-Cárdenas J, Jiménez-Pérez J-R, Cañada-Oya H, Feito FR, Delgado-Martínez AD. Automatic detection of landmarks for the analysis of a reduction of supracondylar fractures of the humerus. Med Image Anal 2020;64:101729.

[33] FRACTURES. Types of fractures. 2023, https://www.orthopedic-institute.org/fracture-care/types-of-fractures/. [Accessed 17 January 2023].

[34] Kalmet PH, Sanduleanu S, Primakov S, Wu G, Jochems A, Refaee T, Ibrahim A, Hulst Lv, Lambin P, Poeze M. Deep learning in fracture detection: a narrative review. Acta Orthop 2020;91(2):215–20.

[35] Chung SW, Han SS, Lee JW, Oh K-S, Kim NR, Yoon JP, Kim JY, Moon SH, Kwon J, Lee H-J, et al. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. Acta Orthop 2018;89(4):468–73.

[36] Demir S, Key S, Tuncer T, Dogan S. An exemplar pyramid feature extraction based humerus fracture classification method. Med Hypotheses 2020;140:109663.

[37] Yamada Y, Maki S, Kishida S, Nagai H, Arima J, Yamakawa N, Iijima Y, Shiko Y, Kawasaki Y, Kotani T, et al. Automated classification of hip fractures using deep convolutional neural networks with orthopedic surgeon-level accuracy: ensemble decision-making with antero-posterior and lateral radiographs. Acta Orthop 2020;91(6):699–704.

[38] Urakawa T, Tanaka Y, Goto S, Matsuzawa H, Watanabe K, Endo N. Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network. Skelet Radiol 2019;48(2):239–44.

[39] Lee C, Jang J, Lee S, Kim YS, Jo HJ, Kim Y. Classification of femur fracture in pelvic X-ray images using meta-learned deep neural network. Sci Rep 2020;10(1):1–12.

[40] Lind A, Akbarian E, Olsson S, Nåsell H, Sköldenberg O, Razavian AS, Gordon M. Artificial intelligence for the classification of fractures around the knee in adults according to the 2018 AO/OTA classification system. PLoS One 2021;16(4):e0248809.

[41] Farda NA, Lai J-Y, Wang J-C, Lee P-Y, Liu J-W, Hsieh I-H. Sanders classification of calcaneal fractures in CT images with deep learning and differential data augmentation techniques. Injury 2021;52(3):616–24.

[42] Ozkaya E, Topal FE, Bulut T, Gursoy M, Ozuysal M, Karakaya Z. Evaluation of an artificial intelligence system for diagnosing scaphoid fracture on direct radiography. Eur J Trauma Emerg Surg 2020;1–8.

[43] Langerhuizen DW, Bulstra AEJ, Janssen SJ, Ring D, Kerkhoffs GM, Jaarsma RL, Doornberg JN. Is deep learning on par with human observers for detection of radiographically visible and occult fractures of the scaphoid? Clin Orthop Relat Res 2020;478(11):2653.

[44] Chen H-Y, Hsu BW-Y, Yin Y-K, Lin F-H, Yang T-H, Yang R-S, Lee C-K, Tseng VS. Application of deep learning algorithm to detect and visualize vertebral fractures on plain frontal radiographs. PLoS One 2021;16(1):e0245992.

[45] Yabu A, Hoshino M, Tabuchi H, Takahashi S, Masumoto H, Akada M, Morita S, Maeno T, Iwamae M, Inose H, et al. Using artificial intelligence to diagnose fresh osteoporotic vertebral fractures on magnetic resonance images. Spine J 2021;21(10):1652–8.

[46] Wang Y, Li Y, Lin G, Zhang Q, Zhong J, Zhang Y, Ma K, Zheng Y, Lu G, Zhang Z. Lower-extremity fatigue fracture detection and grading based on deep learning models of radiographs. Eur Radiol 2023;33(1):555–65.

[47] Liao Z, Liao K, Shen H, Van Boxel MF, Prijs J, Jaarsma RL, Doornberg JN, Van den Hengel A, Verjans JW. CNN attention guidance for improved orthopedics radiographic fracture classification. IEEE J Biomed Health Inf 2022.

[48] Jung J, Dai J, Liu B, Wu Q. Artificial intelligence in fracture detection with different image modalities and data types: A systematic review and meta-analysis. PLoS Digit Health 2024;3(1):e0000438.

[49] Oakden-Rayner L, Gale W, Bonham TA, Lungren MP, Carneiro G, Bradley AP, Palmer LJ. Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in patients in the emergency department: a diagnostic accuracy study. Lancet Digit Health 2022;4(5):e351–8.

[50] Wang X, Xu Z, Tong Y, Xia L, Jie B, Ding P, Bai H, Zhang Y, He Y. Detection and classification of mandibular fracture on CT scan using deep convolutional neural network. Clin Oral Invest 2022;1–9.

[51] Dupuis M, Delbos L, Veil R, Adamsbaum C. External validation of a commercially available deep learning algorithm for fracture detection in children. Diagn Interv Imaging 2022;103(3):151–9.

[52] Guan B, Yao J, Wang S, Zhang G, Zhang Y, Wang X, Wang M. Automatic detection and localization of thighbone fractures in X-ray based on improved deep learning method. Comput Vis Image Underst 2022;216:103345.

[53] Ashkani-Esfahani S, Yazdi RM, Bhimani R, Kerkhoffs GM, Maas M, DiGiovanni CW, Lubberts B, Guss D. Detection of ankle fractures using deep learning algorithms. Foot Ankle Surg 2022.

[54] Huang S-T, Liu L-R, Chiu H-W, Huang M-Y, Tsai M-F. Deep convolutional neural network for rib fracture recognition on chest radiographs. Front Med 2023;10.

[55] Cheng L-W, Chou H-H, Cai Y-X, Huang K-Y, Hsieh C-C, Chu P-L, Cheng I-S, Hsieh S-Y. Automated detection of vertebral fractures from X-ray images: A novel machine learning model and survey of the field. Neurocomputing 2024;566:126946.

[56] Schilcher J, Nilsson A, Andlid O, Eklund A. Fusion of electronic health records and radiographic images for a multimodal deep learning prediction model of atypical femur fractures. Comput Biol Med 2024;168:107704.

[57] Lee LS, Chan PK, Wen C, Fung WC, Cheung A, Chan VWK, Cheung MH, Fu H, Yan CH, Chiu KY. Artificial intelligence in diagnosis of knee osteoarthritis and prediction of arthroplasty outcomes: a review. Arthroplasty 2022;4(1):16.

[58] Xue Y, Zhang R, Deng Y, Chen K, Jiang T. A preliminary examination of the diagnostic value of deep learning in hip osteoarthritis. PLoS One 2017;12(6):e0178992.

[59] Üreten K, Arslan T, Gültekin KE, Demir AND, Özer HF, Bilgili Y. Detection of hip osteoarthritis by using plain pelvic radiographs with deep learning methods. Skelet Radiol 2020;49:1369–74.

[60] Tiulpin A, Thevenot J, Rahtu E, Lehenkari P, Saarakkala S. Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. Sci Rep 2018;8(1):1–10.

[61] Swiecicki A, Li N, O'Donnell J, Said N, Yang J, Mather RC, Jiranek WA, Mazurowski MA. Deep learning-based algorithm for assessment of knee osteoarthritis severity in radiographs matches performance of radiologists. Comput Biol Med 2021;133:104334.

[62] Pedoia V, Lee J, Norman B, Link TM, Majumdar S. Diagnosing osteoarthritis from T2 maps using deep learning: an analysis of the entire osteoarthritis initiative baseline cohort. Osteoarthr Cartil 2019;27(7):1002–10.

[63] Kim DH, Lee KJ, Choi D, Lee JI, Choi HG, Lee YS. Can additional patient information improve the diagnostic performance of deep learning for the interpretation of knee osteoarthritis severity. J Clin Med 2020;9(10):3341.

[64] Zhuang Z, Wang S, Si L, Xuan K, Xue Z, Shen D, Zhang L, Yao W, Wang Q. Local graph fusion of multi-view MR images for knee osteoarthritis diagnosis. In: Medical image computing and computer assisted intervention–MICCAI 2022: 25th international conference, Singapore, September 18–22, 2022, proceedings, part III. Springer; 2022, p. 554–63.

[65] Karnuta JM, Luu BC, Roth AL, Haeberle HS, Chen AF, Iorio R, Schaffer JL, Mont MA, Patterson BM, Krebs VE, et al. Artificial intelligence to identify arthroplasty implants from radiographs of the knee. J Arthroplasty 2021;36(3):935–40.

[66] Borjali A, Chen AF, Muratoglu OK, Morid MA, Varadarajan KM. Detecting total hip replacement prosthesis design on plain radiographs using deep convolutional neural network. J Orthop Res® 2020;38(7):1465–71.

[67] Kang Y-J, Yoo J-I, Cha Y-H, Park CH, Kim J-T. Machine learning–based identification of hip arthroplasty designs. J Orthop Transl 2020;21:13–7.

[68] Urban G, Porhemmat S, Stark M, Feeley B, Okada K, Baldi P. Classifying shoulder implants in X-ray images using deep learning. Comput Struct Biotechnol J 2020;18:967–72.

[69] Repository UML. UCI machine learning repository, shoulder implant X-ray manufacturer classification dataset. 2023, https://archive.ics.uci.edu/ml/datasets/Shoulder+Implant+Manufacture+Classification. [Accessed 28 January 2023].

[70] Sultan H, Owais M, Park C, Mahmood T, Haider A, Park KR. Artificial intelligence-based recognition of different types of shoulder implants in X-ray scans based on dense residual ensemble-network for personalized medicine. J Pers Med 2021;11(6):482.

[71] Sivari E, Güzel MS, Bostanci E, Mishra A. A novel hybrid machine learning based system to classify shoulder implant manufacturers. In: Healthcare. Vol. 10, MDPI; 2022, p. 580.

[72] Clement ND, Clement R, Clement A. Predicting functional outcomes of total hip arthroplasty using machine learning: A systematic review. J Clin Med 2024;13(2):603.

[73] Salman LA, Khatkar H, Al-Ani A, Alzobi OZ, Abudalou A, Hatnouly AT, Ahmed G, Hameed S, AlAteeq Aldosari M. Reliability of artificial intelligence in predicting total knee arthroplasty component sizes: a systematic review. Eur J Orthop Surg Traumatol 2024;34(2):747–56.

[74] Velasquez Garcia A, Bukowiec LG, Yang L, Nishikawa H, Fitzsimmons JS, Larson AN, Taunton MJ, Sanchez-Sotelo J, O'Driscoll SW, Wyles CC. Artificial intelligence–based three-dimensional templating for total joint arthroplasty planning: a scoping review. Int Orthop 2024;1–14.

[75] Leung K, Zhang B, Tan J, Shen Y, Geras KJ, Babb JS, Cho K, Chang G, Deniz CM. Prediction of total knee replacement and diagnosis of osteoarthritis by using deep learning on knee radiographs: data from the osteoarthritis initiative. Radiology 2020;296(3):584–93.

[76] Karaci A. Detection and classification of shoulder implants from X-ray images: YOLO and pretrained convolution neural network based approach. J Fac Eng Archit Gazi Univ 2022;37:283–94.

[77] Nadeem MW, Goh HG, Ali A, Hussain M, Khan MA, Ponnusamy Va. Bone age assessment empowered with deep learning: a survey, open research challenges and future directions. Diagnostics 2020;10(10):781.

[78] Calivá F, Kamat S, Martinez AM, Majumdar S, Pedoia V. Surface spherical encoding and contrastive learning for virtual bone shape aging. Med Image Anal 2022;77:102388.

[79] Lee H, Tajmir S, Lee J, Zissen M, Yeshiwas BA, Alkasab TK, Choy G, Do S. Fully automated deep learning system for bone age assessment. J Digit Imaging 2017;30:427–41.

[80] Liu C, Wang L, Lu W, Liu J, Yang C, Fan C, Li Q, Tang Y. Computer vision-aided bioprinting for bone research. Bone Res 2022;10(1):21.

[81] Nguyen QH, Nguyen BP, Nguyen MT, Chua MC, Do TT, Nghiem N. Bone age assessment and sex determination using transfer learning. Expert Syst Appl 2022;200:116926.

[82] Gaskin CM, Kahn MMSL, Bertozzi JC, Bunch PM. Skeletal development of the hand and wrist: a radiographic atlas and digital bone age companion. Oxford University Press; 2011.

[83] JM GHCNH, RH W, Marshall W, et al. Assessment of skeletal maturity and predication of adult height TW3 method. Gov Oppos 2001;36:27–47.

[84] Mansourvar M, Ismail MA, Herawan T, Gopal Raj R, Abdul Kareem S, Nasaruddin FH. Automated bone age assessment: motivation, taxonomies, and challenges. Comput Math Methods Med 2013;2013.

[85] Pietka E, Kaabi L, Kuo M, Huang H. Feature extraction in carpal-bone analysis. IEEE Trans Med Imaging 1993;12(1):44–9.

[86] Pietka E, Gertych A, Pospiech S, Cao F, Huang H, Gilsanz V. Computer-assisted bone age assessment: Image preprocessing and epiphyseal/metaphyseal ROI extraction. IEEE Trans Med Imaging 2001;20(8):715–29.

[87] Thodberg HH, Kreiborg S, Juul A, Pedersen KD. The BoneXpert method for automated determination of skeletal maturity. IEEE Trans Med Imaging 2008;28(1):52–66.

[88] Spampinato C, Palazzo S, Giordano D, Aldinucci M, Leonardi R. Deep learning for automated skeletal bone age assessment in X-ray images. Med Image Anal 2017;36:41–51.

[89] Halabi SS, Prevedello LM, Kalpathy-Cramer J, Mamonov AB, Bilbily A, Cicero M, Pan I, Pereira LA, Sousa RT, Abdala N, et al. The RSNA pediatric bone age machine learning challenge. Radiology 2019;290(2):498.

[90] Wibisono A, Saputri MS, Mursanto P, Rachmad J, Yudasubrata ATW, Rizki F, Anderson E, et al. Deep learning and classic machine learning approach for automatic bone age assessment. In: 2019 4th Asia-Pacific conference on intelligent robot systems. ACIRS, IEEE; 2019, p. 235–40.

[91] Iglovikov VI, Rakhlin A, Kalinin AA, Shvets AA. Paediatric bone age assessment using deep convolutional neural networks. In: Deep learning in medical image analysis and multimodal learning for clinical decision support: 4th international workshop, DLMIA 2018, and 8th international workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, proceedings 4. Springer; 2018, p. 300–8.

[92] Escobar M, González C, Torres F, Daza L, Triana G, Arbeláez P. Hand pose estimation for pediatric bone age assessment. In: Medical image computing and computer assisted intervention–MICCAI 2019: 22nd international conference, Shenzhen, China, October 13–17, 2019, proceedings, part VI 22. Springer; 2019, p. 531–9.

[93] Guo L, Wang J, Teng J, Chen Y. Bone age assessment based on deep convolutional features and fast extreme learning machine algorithm. Front Energy Res 2022;9:888.

[94] Li S, Liu B, Li S, Zhu X, Yan Y, Zhang D. A deep learning-based computer-aided diagnosis method of X-ray images for bone age assessment. Complex Intell Syst 2021;1–11.

[95] Cheng CF, Huang ET-C, Kuo J-T, Liao KY-K, Tsai F-J. Report of clinical bone age assessment using deep learning for an Asian population in Taiwan. Biomedicine 2021;11(3):50.

[96] Zulkifley MA, Mohamed NA, Abdani SR, Kamari NAM, Moubark AM, Ibrahim AA. Intelligent bone age assessment: an automated system to detect a bone growth problem using convolutional neural networks with attention mechanism. Diagnostics 2021;11(5):765.

[97] Nguyen QH, Nguyen BP, Nguyen MT, Chua MC, Do TT, Nghiem N. Bone age assessment and sex determination using transfer learning. Expert Syst Appl 2022;200:116926.

[98] Palaniswamy T. Hyperparameter optimization based deep convolution neural network model for automated bone age assessment and classification. Displays 2022;73:102206.

[99] Yang Z, Cong C, Pagnucco M, Song Y. Multi-scale multi-reception attention network for bone age assessment in X-ray images. Neural Netw 2023;158:249–57.

[100] Wang C, Wu Y, Wang C, Zhou X, Niu Y, Zhu Y, Gao X, Wang C, Yu Y. Attention-based multiple-instance learning for pediatric bone age assessment with efficient and interpretable. Biomed Signal Process Control 2023;79:104028.

[101] Upalananda W, Wantanajittikul K, Na Lampang S, Janhom A. Semi-automated technique to assess the developmental stage of mandibular third molars for age estimation. Aust J Forensic Sci 2023;55(1):23–33.

[102] Demirjian A, Goldstein H, Tanner JM. A new system of dental age assessment. Hum Biol 1973;211–27.

[103] Jabbar AJ, Abdulmunem AA. Bone age assessment based on deep learning architecture. Int J Electr Comput Eng 2023;13(2):2078.

[104] Rassmann S, Keller A, Skaf K, Hustinx A, Gausche R, Ibarra-Arrelano MA, Hsieh T-C, Madajieu YE, Nöthen MM, Pfäffle R, et al. Deeplasia: deep learning for bone age assessment validated on skeletal dysplasias. Pediatr Radiol 2024;54(1):82–95.

[105] Wu J, Mi Q, Zhang Y, Wu T. SVTNet: Automatic bone age assessment network based on TW3 method and vision transformer. Int J Imaging Syst Technol 2024;34(2):e22990.

[106] Laur O, Wang B. Musculoskeletal trauma and artificial intelligence: current trends and projections. Skelet Radiol 2022;51(2):257–69.

[107] Takeuchi N, Kozono N, Nishii A, Matsuura K, Ishitani E, Onizuka T, Mizuki Y, Kimura T, Yuge H, Uchimura T, et al. Prevalence and predisposing factors of neuropathic pain in patients with rotator cuff tears. J Orthop Sci 2023.

[108] Seida JC, LeBlanc C, Schouten JR, Mousavi SS, Hartling L, Vandermeer B, Tjosvold L, Sheps DM. Systematic review: nonoperative and operative treatments for rotator cuff tears. Ann Intern Med 2010;153(4):246–55.

[109] Moosmayer S, Lund G, Seljom US, Haldorsen B, Svege IC, Hennig T, Pripp AH, Smith H-Jr. At a 10-year follow-up, tendon repair is superior to physiotherapy in the treatment of small and medium-sized rotator cuff tears. J Bone Joint Surg 2019;101(12):1050–60.

[110] Longo UG, Carnevale A, Piergentili I, Berton A, Candela V, Schena E, Denaro V. Retear rates after rotator cuff surgery: a systematic review and meta-analysis. BMC Musculoskelet Disord 2021;22(1):749.

[111] Dyer J-O, Doiron-Cadrin P, Lafrance S, Roy J-S, Frémont P, Dionne CE, MacDermid JC, Tousignant M, Rochette A, Lowry V, et al. Diagnosing, managing, and supporting return to work of adults with rotator cuff disorders: Clinical practice guideline methods. J Orthop Sports Phys Therapy 2022;52(10):665–74.

[112] Yao J, Chepelev L, Nisha Y, Sathiadoss P, Rybicki FJ, Sheikh AM. Evaluation of a deep learning method for the automated detection of supraspinatus tears on MRI. Skelet Radiol 2022;1–11.

[113] Shim E, Kim JY, Yoon JP, Ki S-Y, Lho T, Kim Y, Chung SW. Automated rotator cuff tear classification using 3D convolutional neural network. Sci Rep 2020;10(1):15632.

[114] Lin C-C, Wang C-N, Ou Y-K, Fu J. Combined image enhancement, feature extraction, and classification protocol to improve detection and diagnosis of rotator-cuff tears on MR imaging. Magn Reson Med Sci 2014;13(3):155–66.

[115] Horiuchi S, Nozaki T, Tasaki A, Yamakawa A, Kaneko Y, Hara T, Yoshioka H. Reliability of MR quantification of rotator cuff muscle fatty degeneration using a 2-point dixon technique in comparison with the goutallier classification: validation study by multiple readers. Acad Radiol 2017;24(11):1343–51.

[116] Lee K, Kim JY, Lee MH, Choi C-H, Hwang JY. Imbalanced loss-integrated deep-learning-based ultrasound image analysis for diagnosis of rotator-cuff tear. Sensors 2021;21(6):2214.

[117] Kim JY, Ro K, You S, Nam BR, Yook S, Park HS, Yoo JC, Park E, Cho K, Cho BH, et al. Development of an automatic muscle atrophy measuring algorithm to calculate the ratio of supraspinatus in supraspinous fossa using deep learning. Comput Methods Programs Biomed 2019;182:105063.

[118] Taghizadeh E, Truffer O, Becce F, Eminian S, Gidoin S, Terrier A, Farron A, Büchler P. Deep learning for the rapid automatic quantification and characterization of rotator cuff muscle degeneration from shoulder CT datasets. Eur Radiol 2021;31:181–90.

[119] Medina G, Buckless CG, Thomasson E, Oh LS, Torriani M. Deep learning method for segmentation of rotator cuff muscles on MR images. Skelet Radiol 2021;50:683–92.

[120] Ro K, Kim JY, Park H, Cho BH, Kim IY, Shim SB, Choi IY, Yoo JC. Deep-learning framework and computer assisted fatty infiltration analysis for the supraspinatus muscle in MRI. Sci Rep 2021;11(1):1–12.

[121] Zech JR, Carotenuto G, Igbinoba Z, Tran CV, Insley E, Baccarella A, Wong TT. Detecting pediatric wrist fractures using deep-learning-based object detection. Pediatr Radiol 2023;1–10.

[122] Xuan P, Wu X, Cui H, Jin Q, Wang L, Zhang T, Nakaguchi T, Duh HB. Multi-scale random walk driven adaptive graph neural network with dual-head neighboring node attention for CT segmentation. Appl Soft Comput 2023;133:109905.

[123] Li J, Qian K, Liu J, Huang Z, Zhang Y, Zhao G, Wang H, Li M, Liang X, Zhou F, et al. Identification and diagnosis of meniscus tear by magnetic resonance imaging using a deep learning model. J Orthop Transl 2022;34:91–101.

[124] Javed Awan M, Mohd Rahim MS, Salim N, Mohammed MA, Garcia-Zapirain B, Abdulkareem KH. Efficient detection of knee anterior cruciate ligament from magnetic resonance imaging using deep learning approach. Diagnostics 2021;11(1):105.

[125] Couteaux V, Si-Mohamed S, Nempont O, Lefevre T, Popoff A, Pizaine G, Villain N, Bloch I, Cotten A, Boussel L. Automatic knee meniscus tear detection and orientation classification with mask-RCNN. Diagn Intervent Imaging 2019;100(4):235–42.

[126] Roblot V, Giret Y, Antoun MB, Morillot C, Chassin X, Cotten A, Zerbib J, Fournier L. Artificial intelligence to diagnose meniscus tears on MRI. Diagn Intervent Imaging 2019;100(4):243–9.

[127] Chang PD, Wong TT, Rasiej MJ. Deep learning for detection of complete anterior cruciate ligament tear. J Digit Imaging 2019;32:980–6.

[128] Flannery SW, Kiapour AM, Edgar DJ, Murray MM, Fleming BC. Automated magnetic resonance image segmentation of the anterior cruciate ligament. J Orthop Res® 2021;39(4):831–40.

[129] Li Z, Ren S, Zhou R, Jiang X, You T, Li C, Zhang W. Deep learning-based magnetic resonance imaging image features for diagnosis of anterior cruciate ligament injury. J Healthc Eng 2021;2021.

[130] Key S, Baygin M, Demir S, Dogan S, Tuncer T. Meniscal tear and ACL injury detection model based on AlexNet and iterative reliefF. J Digit Imaging 2022;35(2):200–12.

[131] Lee K-S, Jung S-K, Ryu J-J, Shin S-W, Choi J. Evaluation of transfer learning with deep convolutional neural networks for screening osteoporosis in dental panoramic radiographs. J Clin Med 2020;9(2):392.

[132] Yamamoto N, Sukegawa S, Kitamura A, Goto R, Noda T, Nakano K, Takabatake K, Kawai H, Nagatsuka H, Kawasaki K, et al. Deep learning for osteoporosis classification using hip radiographs and patient clinical covariates. Biomolecules 2020;10(11):1534.

[133] Sukegawa S, Fujimura A, Taguchi A, Yamamoto N, Kitamura A, Goto R, Nakano K, Takabatake K, Kawai H, Nagatsuka H, et al. Identification of osteoporosis using ensemble deep learning model with panoramic radiographs and clinical covariates. Sci Rep 2022;12(1):1–10.

[134] Nakamoto T, Taguchi A, Kakimoto N. Osteoporosis screening support system from panoramic radiographs using deep learning by convolutional neural network. Dentomaxillofac Radiol 2022;51(6):20220135.

[135] He Y, Pan I, Bao B, Halsey K, Chang M, Liu H, Peng S, Sebro RA, Guan J, Yi T, et al. Deep learning-based classification of primary bone tumors on radiographs: A preliminary study. eBioMedicine 2020;62.

[136] von Schacky CE, Wilhelm NJ, Schäfer VS, Leonhardt Y, Gassert FG, Foreman SC, Gassert FT, Jung M, Jungmann PM, Russe MF, et al. Multitask deep learning for segmentation and classification of primary bone tumors on radiographs. Radiology 2021;301(2):398–406.

[137] Eweje FR, Bao B, Wu J, Dalal D, Liao W-h, He Y, Luo Y, Lu S, Zhang P, Peng X, et al. Deep learning for classification of bone lesions on routine MRI. eBioMedicine 2021;68.

[138] Zhang S-C, Sun J, Liu C-B, Fang J-H, Xie H-T, Ning B. Clinical application of artificial intelligence-assisted diagnosis using anteroposterior pelvic radiographs in children with developmental dysplasia of the hip. Bone Joint J 2020;102(11):1574–81.

[139] Xu W, Shu L, Gong P, Huang C, Xu J, Zhao J, Shu Q, Zhu M, Qi G, Zhao G, et al. A deep-learning aided diagnostic system in assessing developmental dysplasia of the hip on pediatric pelvic radiographs. Front Pediatr 2022;9:1701.

[140] Atalar H, Üreten K, Tokdemir G, Tolunay T, Çiçeklidağ M, Atik OŞ. The diagnosis of developmental dysplasia of the hip from hip ultrasonography images with deep learning methods. J Pediatr Orthop 2023;43(2):e132–7.

[141] Pei Y, Yang W, Wei S, Cai R, Li J, Guo S, Li Q, Wang J, Li X. Automated measurement of hip–knee–ankle angle on the unilateral lower limb X-rays using deep learning. Phys Eng Sci Med 2021;44:53–62.

[142] Horng M-H, Kuok C-P, Fu M-J, Lin C-J, Sun Y-N. Cobb angle measurement of spine from X-ray images using convolutional neural network. Comput Math Methods Med 2019;2019.

[143] Alukaev D, Kiselev S, Mustafaev T, Ainur A, Ibragimov B, Vrtovec T. A deep learning framework for vertebral morphometry and cobb angle measurement with external validation. Eur Spine J 2022;31(8):2115–24.

[144] Ishikawa Y, Kokabu T, Yamada K, Abe Y, Tachi H, Suzuki H, Ohnishi T, Endo T, Ukeba D, Ura K, et al. Prediction of cobb angle using deep learning algorithm with three-dimensional depth sensor considering the influence of garment in idiopathic scoliosis. J Clin Med 2023;12(2):499.

[145] Ryu SM, Shin K, Shin SW, Lee SH, Seo SM, Cheon S-U, Ryu S-A, Kim M-J, Kim H, Doh CH, et al. Automated diagnosis of flatfoot using cascaded convolutional neural network for angle measurements in weight-bearing lateral radiographs. Eur Radiol 2023;33(7):4822–32.

[146] Ryu SM, Lee S, Jang M, Koh J-M, Bae SJ, Jegal SG, Shin K, Kim N. Diagnosis of osteoporotic vertebral compression fractures and fracture level detection using multitask learning with U-net in lumbar spine lateral radiographs. Comput Struct Biotechnol J 2023;21:3452–8.

[147] Ryu SM, Shin K, Shin SW, Lee SH, Seo SM, Cheon S-u, Ryu S-A, Kim J-S, Ji S, Kim N. Automated landmark identification for diagnosis of the deformity using a cascade convolutional neural network (FlatNet) on weight-bearing lateral radiographs of the foot. Comput Biol Med 2022;148:105914.

[148] Ryu SM, Shin K, Shin SW, Lee S, Kim N. Enhancement of evaluating flatfoot on a weight-bearing lateral radiograph of the foot with U-net based semantic segmentation on the long axis of tarsal and metatarsal bones in an active learning manner. Comput Biol Med 2022;145:105400.

[149] Lee JS, Shin K, Ryu SM, Jegal SG, Lee W, Yoon MA, Hong G-S, Paik S, Kim N. Screening of adolescent idiopathic scoliosis using generative adversarial network (GAN) inversion method in chest radiographs. PLoS One 2023;18(5):e0285489.

[150] Chen P. Knee osteoarthritis severity grading dataset. Mendeley Data 2018;1. http://dx.doi.org/10.17632/56rmx5bjcr, v1.

[151] Azcona D, McGuinness K, Smeaton AF. A comparative study of existing and new deep learning methods for detecting knee injuries using the mrnet dataset. In: 2020 international conference on intelligent data science technologies and applications. IDSTA, IEEE; 2020, p. 149–55.

[152] Karthik K, Kamath SS. A deep neural network model for content-based medical image retrieval with multi-view classification. Vis Comput 2021;37(7):1837–50.

[153] Soh S-E, Barker AL, Morello RT, Ackerman IN. Applying the international classification of functioning, disability and health framework to determine the predictors of falls and fractures in people with osteoarthritis or at high risk of developing osteoarthritis: Data from the osteoarthritis initiative. BMC Musculoskelet Disord 2020;21(1):1–8.

[154] Malik H, Jabbar J, Mehmood H. Wrist fracture—X-rays. Mendeley Data 2020.

[155] Jin L, Yang J, Kuang K, Ni B, Gao Y, Sun Y, Gao P, Ma W, Tan M, Kang H, et al. Deep-learning-assisted detection and segmentation of rib fractures from CT scans: Development and validation of FracNet. eBioMedicine 2020;62:103106.

[156] Lohchab V, Rathod P, Mahapatra PK, Bachhal V, Hooda A. Non-invasive assessment of knee osteoarthritis patients using thermal imaging. IET Sci Meas Technol 2022;16(4):242–9.

[157] Rajpurkar P, Irvin J, Bagul A, Ding D, Duan T, Mehta H, Yang B, Zhu K, Laird D, Ball RL, et al. Mura: Large dataset for abnormality detection in musculoskeletal radiographs. 2017, arXiv preprint arXiv:1712.06957.

[158] Kandel I, Castelli M, Popovič A. Musculoskeletal images classification for detection of fractures using transfer learning. J Imaging 2020;6(11):127.

[159] Kandel I, Castelli M, Popovič A. Comparing stacking ensemble techniques to improve musculoskeletal fracture image classification. J Imaging 2021;7(6):100.

[160] He M, Wang X, Zhao Y. A calibrated deep learning ensemble for abnormality detection in musculoskeletal radiographs. Sci Rep 2021;11(1):1–11.

[161] Liang S, Gu Y. Towards robust and accurate detection of abnormalities in musculoskeletal radiographs with a multi-network model. Sensors 2020;20(11):3153.

[162] Saif A, Shahnaz C, Zhu W-P, Ahmad MO. Abnormality detection in musculoskeletal radiographs using capsule network. IEEE Access 2019;7:81494–503.

[163] Fang L, Jin Y, Huang L, Guo S, Zhao G, Chen X. Iterative fusion convolutional neural networks for classification of optical coherence tomography images. J Vis Commun Image Represent 2019;59:327–33.

[164] Harini N, Ramji B, Sriram S, Sowmya V, Soman K. Musculoskeletal radiographs classification using deep learning. In: Deep learning for data analytics. Elsevier; 2020, p. 79–98.

[165] Malik S, Amin J, Sharif M, Yasmin M, Kadry S, Anjum S. Fractured elbow classification using hand-crafted and deep feature fusion and selection based on whale optimization approach. Mathematics 2022;10(18):3291.

[166] Alammar Z, Alzubaidi L, Zhang J, Li Y, Lafta W, Gu Y. Deep transfer learning with enhanced feature fusion for detection of abnormalities in x-ray images. Cancers 2023;15(15):4007.

[167] Alzubaidi L, Fadhel MA, Albahri A, Salhi A, Gupta A, Gu Y. Domain adaptation and feature fusion for the detection of abnormalities in X-Ray forearm images. In: 2023 45th annual international conference of the IEEE engineering in medicine & biology society. EMBC, IEEE; 2023, p. 1–5.

[168] Kumar K, Pailla B, Tadepalli K, Roy S. Robust MSFM learning network for classification and weakly supervised localization. In: Proceedings of the IEEE/CVF international conference on computer vision. 2023, p. 2442–51.

[169] Alzubaidi L, Salhi A, A.Fadhel M, Bai J, Hollman F, Italia K, Pareyon R, Albahri AS, Ouyang C, Santamaría J, Cutbush K, Gupta A, Abbosh A, Gu Y. Trustworthy deep learning framework for the detection of abnormalities in X-ray shoulder images. PLoS One 2024;19(3):e0299545.

[170] Li J, Pepe A, Gsaxner C, Luijten G, Jin Y, Ambigapathy N, Nasca E, Solak N, Melito GM, Memon AR, et al. MedShapeNet–A large-scale dataset of 3D medical shapes for computer vision. 2023, arXiv preprint arXiv:2308.16139.

[171] Jayakumar N, Hossain T, Zhang M. SADIR: Shape-aware diffusion models for 3D image reconstruction. In: International workshop on shape in medical imaging. Springer; 2023, p. 287–300.

[172] Li J, Pepe A, Luijten G, Schwarz-Gsaxner C, Kleesiek J, Egger J. Anatomy completor: A multi-class completion framework for 3d anatomy reconstruction. In: International workshop on shape in medical imaging. Springer; 2023, p. 1–14.

[173] Krieger K, Egger J, Kleesiek J, Gunzer M, Chen J. Multimodal extended reality applications offer benefits for volumetric biomedical image analysis in research and medicine. 2023, arXiv preprint arXiv:2311.03986.

[174] Luijten G, Gsaxner C, Li J, Pepe A, Ambigapathy N, Kim M, Chen X, Kleesiek J, Hölzle F, Puladi B, et al. 3D surgical instrument collection for computer vision and extended reality. Sci Data 2023;10(1):796.

[175] Aslani S, Jacob J. Utilisation of deep learning for COVID-19 diagnosis. Clin Radiol 2023;78(2):150–7.

[176] Sambasivan N, Kapania S, Highfill H, Akrong D, Paritosh P, Aroyo LM. "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In: Proceedings of the 2021 CHI conference on human factors in computing systems. 2021, p. 1–15.

[177] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Commun ACM 2017;60(6):84–90.

[178] Huh M, Agrawal P, Efros AA. What makes ImageNet good for transfer learning? 2016, arXiv preprint arXiv:1608.08614.

[179] Alzubaidi L, Duan Y, Al-Dujaili A, Ibraheem IK, Alkenani AH, Santamaría J, Fadhel MA, Al-Shamma O, Zhang J. Deepening into the suitability of using pre-trained models of ImageNet against a lightweight convolutional neural network in medical imaging: an experimental study. PeerJ Comput Sci 2021;7:e715.

[180] Krishnan R, Rajpurkar P, Topol EJ. Self-supervised learning in medicine and healthcare. Nat Biomed Eng 2022;1–7.

[181] Kumar K, Chakraborty S, Roy S. Self-supervised diffusion model for anomaly segmentation in medical imaging. In: International conference on pattern recognition and machine intelligence. Springer; 2023, p. 359–68.

[182] Liu C, Xie H, Zhang Y. Self-supervised attention mechanism for pediatric bone age assessment with efficient weak annotation. IEEE Trans Med Imaging 2020;40(10):2685–97.

[183] Vettoruzzo A, Bouguelia M-R, Vanschoren J, Rognvaldsson T, Santosh K. Advances and challenges in meta-learning: A technical review. IEEE Trans Pattern Anal Mach Intell 2024.

[184] Liu Q, Tian Y, Zhou T, Lyu K, Xin R, Shang Y, Liu Y, Ren J, Li J. A few-shot disease diagnosis decision making model based on meta-learning for general practice. Artif Intell Med 2024;147:102718.

[185] Alzubaidi L, Fadhel MA, Al-Shamma O, Zhang J, Santamaría J, Duan Y, R. Oleiwi S. Towards a better understanding of transfer learning for medical imaging: a case study. Appl Sci 2020;10(13):4523.

[186] Yang Q, Liu Y, Cheng Y, Kang Y, Chen T, Yu H. Federated learning. Synth Lect Artif Intell Mach Learn 2019;13(3):1–207.

[187] Li L, Fan Y, Tse M, Lin K-Y. A review of applications in federated learning. Comput Ind Eng 2020;149:106854.

[188] Wan S, Lu J, Fan P, Shao Y, Peng C, Chuai J, et al. How global observation works in federated learning: Integrating vertical training into horizontal federated learning. IEEE Internet Things J 2023.

[189] Bai J, Rabczuk T, Gupta A, Alzubaidi L, Gu Y. A physics-informed neural network technique based on a modified loss function for computational 2D and 3D solid mechanics. Comput Mech 2022;1–20.

[190] Li W, Bazant MZ, Zhu J. A physics-guided neural network framework for elastic plates: Comparison of governing equations-based and energy-based approaches. Comput Methods Appl Mech Engrg 2021;383:113933.

[191] Bhouri MA, Costabal FS, Wang H, Linka K, Peirlinck M, Kuhl E, Perdikaris P. COVID-19 dynamics across the US: A deep learning study of human mobility and social behavior. Comput Methods Appl Mech Engrg 2021;382:113891.

[192] Kissas G, Yang Y, Hwuang E, Witschey WR, Detre JA, Perdikaris P. Machine learning in cardiovascular flows modeling: Predicting arterial blood pressure from non-invasive 4D flow MRI data using physics-informed neural networks. Comput Methods Appl Mech Engrg 2020;358:112623.

[193] Buoso S, Joyce T, Kozerke S. Personalising left-ventricular biophysical models of the heart using parametric physics-informed neural networks. Med Image Anal 2021;71:102066.

[194] Alzubaidi L, Bai J, Al-Sabaawi A, Santamaría J, Albahri A, Al-dabbagh BSN, Fadhel MA, Manoufali M, Zhang J, Al-Timemy AH, et al. A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. J Big Data 2023;10(1):1–82.

[195] Roy S, Pal D, Meena T. Explainable artificial intelligence to increase transparency for revolutionizing healthcare ecosystem and the road ahead. Netw Model Anal Health Inform Bioinform 2023;13(1):4.

[196] Yang G, Ye Q, Xia J. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. Inf Fusion 2022;77:29–52.

[197] Van der Velden BH, Kuijf HJ, Gilhuijs KG, Viergever MA. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med Image Anal 2022;102470.

[198] Roy S, Meena T, Lim S-J. Demystifying supervised learning in healthcare 4.0: A new reality of transforming diagnostic medicine. Diagnostics 2022;12(10):2549.

[199] Teng Q, Liu Z, Song Y, Han K, Lu Y. A survey on the interpretability of deep learning in medical diagnosis. Multimedia Syst 2022;1–21.

[200] Li X, Xiong H, Li X, Wu X, Zhang X, Liu J, Bian J, Dou D. Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. Knowl Inf Syst 2022;64(12):3197–234.

[201] Chaddad A, Peng J, Xu J, Bouridane A. Survey of explainable AI techniques in healthcare. Sensors 2023;23(2):634.

[202] Alzubaidi L, Fadhel MA, Al-Shamma O, Zhang J, Santamaría J, Duan Y. Robust application of new deep learning tools: an experimental study in medical imaging. Multimedia Tools Appl 2022;81(10):13289–317.

[203] Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, p. 2921–9.

[204] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. 2017, p. 618–26.

[205] Ribeiro MT, Singh S, Guestrin C. "Why should i trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016, p. 1135–44.

[206] Tomsett R, Harborne D, Chakraborty S, Gurram P, Preece A. Sanity checks for saliency metrics. In: Proceedings of the AAAI conference on artificial intelligence. Vol. 34, 2020, p. 6021–9.

[207] Keshavan MS, Sudarshan M. Deep dreaming, aberrant salience and psychosis: connecting the dots by artificial neural networks. Schizophr Res 2017;188:178–81.

[208] Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9(11).

[209] Jakubovitz D, Giryes R, Rodrigues MR. Generalization error in deep learning. In: Compressed sensing and its applications. Springer; 2019, p. 153–93.

[210] Himeur Y, Al-Maadeed S, Kheddar H, Al-Maadeed N, Abualsaud K, Mohamed A, Khattab T. Video surveillance using deep transfer learning and deep domain adaptation: Towards better generalization. Eng Appl Artif Intell 2023;119:105698.

[211] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. 2014, arXiv preprint arXiv:1412.6572.

[212] Arnab A, Miksik O, Torr PH. On the robustness of semantic segmentation models to adversarial attacks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, p. 888–97.

[213] Albahri A, Hamid RA, Abdulnabi AR, Albahri O, Alamoodi A, Deveci M, Pedrycz W, Alzubaidi L, Santamaría J, Gu Y. Fuzzy decision-making framework for explainable golden multi-machine learning models for real-time adversarial attack detection in vehicular ad-hoc networks. Inf Fusion 2024;105:102208.

[214] Akhtar N, Jalwana M, Bennamoun M, Mian AS. Attack to fool and explain deep networks. IEEE Trans Pattern Anal Mach Intell 2021.

[215] Al-Shamma O, Fadhel MA, Hameed RA, Alzubaidi L, Zhang J. Boosting convolutional neural networks performance based on FPGA accelerator. In: International conference on intelligent systems design and applications. Springer; 2020, p. 509–17.

[216] AlBdairi AJA, Xiao Z, Alkhayyat A, Humaidi AJ, Fadhel MA, Taher BH, Alzubaidi L, Santamaría J, Al-Shamma O. Face recognition based on deep learning and FPGA for ethnicity identification. Appl Sci 2022;12(5):2605.

[217] Fadhel MA, Alzubaidi L, Gu Y, Santamaría J, Duan Y. Real-time diabetic foot ulcer classification based on deep learning & parallel hardware computational tools. Multimedia Tools Appl 2024;1–26.

[218] Gilbert F, Böhm D, Eden L, Schmalzl J, Meffert RH, Köstler H, Weng AM, Ziegler D. Comparing the MRI-based goutallier classification to an experimental quantitative MR spectroscopic fat measurement of the supraspinatus muscle. BMC Musculoskelet Disord 2016;17(1):1–7.

[219] Eckers F, Loske S, Ek ET, Müller AM. Current understanding and new advances in the surgical management of reparable rotator cuff tears: A scoping review. J Clin Med 2023;12(5):1713.

[220] Zhao M, Zhou Y, Chang J, Hu J, Liu H, Wang S, Si D, Yuan Y, Li H. The accuracy of MRI in the diagnosis of anterior cruciate ligament injury. Ann Transl Med 2020;8(24).

[221] Hegedus EJ, Goode A, Campbell S, Morin A, Tamaddoni M, Moorman CT, Cook C. Physical examination tests of the shoulder: a systematic review with meta-analysis of individual tests. Br J Sports Med 2008;42(2):80–92.

[222] Liu R, Pan D, Xu Y, Zeng H, He Z, Lin J, Zeng W, Wu Z, Luo Z, Qin G, et al. A deep learning–machine learning fusion approach for the classification of benign, malignant, and intermediate bone tumors. Eur Radiol 2022;32(2):1371–83.

[223] Jain N, Whitehouse S, Foley G, Yates E, Murray D. A review of orthopaedic classifications; are they justified in their use? In: Orthopaedic proceedings. Vol. 95, The British Editorial Society of Bone & Joint Surgery; 2013, 206–206.

[224] Mikelis F, Koletsi D. Scoping reviews in orthodontics: are they justified? Prog Orthod 2022;23(1):1–7.

[225] Siddiqi A, Horan T, Molloy RM, Bloomfield MR, Patel PD, Piuzzi NS. A clinical review of robotic navigation in total knee arthroplasty: historical systems to modern design. EFORT Open Rev 2021;6(4):252.

[226] Han P-f, Chen C-l, Zhang Z-l, Han Y-c, Wei L, Li P-c, Wei X-c. Robotics-assisted versus conventional manual approaches for total hip arthroplasty: A systematic review and meta-analysis of comparative studies. Int J Med Robot Comput Assist Surg 2019;15(3):e1990.

[227] Karunaratne S, Duan M, Pappas E, Fritsch B, Boyle R, Gupta S, Stalley P, Horsley M, Steffens D. The effectiveness of robotic hip and knee arthroplasty on patient-reported outcomes: a systematic review and meta-analysis. Int Orthop 2019;43:1283–95.

[228] Jacofsky DJ, Allen M. Robotics in arthroplasty: a comprehensive review. J Arthroplasty 2016;31(10):2353–63.

[229] Deckey DG, Verhey JT, Rosenow CS, Doan MK, McQuivey KS, Joseph AM, Schwartz AJ, Clarke HD, Bingham JS. Robotic-assisted total knee arthroplasty allows for trainee involvement and teaching without lengthening operative time. J Arthroplasty 2022;37(6):S201–6.

[230] Rajan PV, Khlopas A, Klika A, Molloy R, Krebs V, Piuzzi NS. The cost-effectiveness of robotic-assisted versus manual total knee arthroplasty: a Markov model–based evaluation. JAAOS J Am Acad Orthop Surg 2022;30(4):168–76.

[231] Murphy MP, Brown NM. CORR synthesis: When should the orthopaedic surgeon use artificial intelligence, machine learning, and deep learning? Clin Orthop Relat Res 2021;479(7):1497.

[232] Ghaednia H, Lans A, Sauder N, Shin D, Grant WG, Chopra RR, Oosterhoff JH, Fourman MS, Schwab JH, Tobert DG. Deep learning in spine surgery. In: Seminars in spine surgery. Vol. 33, Elsevier; 2021, 100876.

[233] Liow MHL, Chin PL, Pang HN, Tay DK-J, Yeo S-J. THINK surgical tsolution-one®(robodoc) total knee arthroplasty. SICOT-J 2017;3.

[234] Parratte S, Price AJ, Jeys LM, Jackson WF, Clarke HD. Accuracy of a new robotically assisted technique for total knee arthroplasty: a cadaveric study. J Arthroplasty 2019;34(11):2799–803.

[235] Siddiqi A, Smith T, McPhilemy JJ, Ranawat AS, Sculco PK, Chen AF. Soft-tissue balancing technology for total knee arthroplasty. JBJS Rev 2020;8(1):e0050.

[236] Bolam SM, Tay ML, Zaidi F, Sidaginamale RP, Hanlon M, Munro JT, Monk AP. Introduction of ROSA robotic-arm system for total knee arthroplasty is associated with a minimal learning curve for operative time. J Exp Orthop 2022;9(1):1–8.

[237] Mancino F, Cacciola G, Malahias M-A, De Filippis R, De Marco D, Di Matteo V, Gu A, Sculco PK, Maccauro G, De Martino I. What are the benefits of robotic-assisted total knee arthroplasty over conventional manual total knee arthroplasty? A systematic review of comparative studies. Orthop Rev 2020;12(Suppl 1).

[238] Shatrov J, Parker D. Computer and robotic–assisted total knee arthroplasty: a review of outcomes. J Exp Orthop 2020;7:1–15.

[239] Hönecke T, Schwarze M, Wangenheim M, Savov P, Windhagen H, Ettinger M. Noise exposure during robot-assisted total knee arthroplasty. Arch Orthop Trauma Surg 2022;1–7.

[240] Fu Y, Hu Y, Sundstedt V. A systematic literature review of virtual, augmented, and mixed reality game applications in healthcare. ACM Trans Comput Healthc (HEALTH) 2022;3(2):1–27.

[241] Choi SH, Park K-B, Roh DH, Lee JY, Mohammed M, Ghasemi Y, Jeong H. An integrated mixed reality system for safety-aware human-robot collaboration using deep learning and digital twin generation. Robot Comput-Integr Manuf 2022;73:102258.

[242] Jovanov E, Milenkovic A, Otto C, De Groen PC. A wireless body area network of intelligent motion sensors for computer assisted physical rehabilitation. J NeuroEng Rehabil 2005;2(1):1–10.

[243] Alamoodi A, Albahri O, Zaidan A, Alsattar H, Zaidan B, Albahri A, Ismail AR, Kou G, Alzubaidi L, Talal M. Intelligent emotion and sensory remote prioritisation for patients with multiple chronic diseases. Sensors 2023;23(4):1854.

[244] Hasan K, Biswas K, Ahmed K, Nafi NS, Islam MS. A comprehensive review of wireless body area network. J Netw Comput Appl 2019;143:178–98.

[245] Ruvio G, Cuccaro A, Solimene R, Brancaccio A, Basile B, Ammann MJ. Microwave bone imaging: a preliminary scanning system for proof-of-concept. Healthc Technol Lett 2016;3(3):218–21.

[246] Santos KC, Fernandes CA, Costa JR. Feasibility of bone fracture detection using microwave imaging. IEEE Open J Antennas Propag 2022;3:836–47.

[247] Sultan KS, Mahmoud A, Abbosh AM. Textile electromagnetic brace for knee imaging. IEEE Trans Biomed Circuits Syst 2021;15(3):522–36.

[248] Vaishya R, Patralekh MK, Vaish A, Agarwal AK, Vijay V. Publication trends and knowledge mapping in 3D printing in orthopaedics. J Clin Orthop Trauma 2018;9(3):194–201.

[249] Rouf S, Malik A, Raina A, Haq MIU, Naveed N, Zolfagharian A, Bodaghi M. Functionally graded additive manufacturing for orthopedic applications. J Orthop 2022;33:70–80.

[250] Maini L, Vaishya R, Lal H. Will 3D printing take away surgical planning from doctors? J Clin Orthop Trauma 2018;9(3):193.

[251] Morgan S, Barriga J, Dadia S, Merose O, Sternheim A, Snir N. Three dimensional printing as an aid for pre-operative planning in complex cases of total joint arthroplasty: A case series. J Orthop 2022;34:142–6.

[252] Goh GD, Yeong WY. Applications of machine learning in 3D printing. Mater Today: Proc 2022;70:95–100.

[253] Wang S, Chen X, Han X, Hong X, Li X, Zhang H, Li M, Wang Z, Zheng A. A review of 3D printing technology in pharmaceutics: Technology and applications, now and future. Pharmaceutics 2023;15(2):416.

[254] D'Alessio J, Christensen A. 3D printing for commercial orthopedic applications: advances and challenges. In: 3D printing in orthopaedic surgery. Elsevier; 2019, p. 65–83.

[255] Goh GD, Sing SL, Yeong WY. A review on machine learning in 3D printing: applications, potential, and challenges. Artif Intell Rev 2021;54(1):63–94.

[256] Li R, Peng Q. Deep learning-based optimal segmentation of 3D printed product for surface quality improvement and support structure reduction. J Manuf Syst 2021;60:252–64.

[257] Food U, Administration D, et al. Classification of products as drugs and devices and additional product classification issues. 2016, can be found under https://www.fda.gov/media/80384/download. [Accessed December 2020].

[258] Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med 2020;3(1):118.

[259] Groen AM, Kraan R, Amirkhan SF, Daams JG, Maas M. A systematic review on the use of explainability in deep learning systems for computer aided diagnosis in radiology: Limited use of explainable AI? Eur J Radiol 2022;110592.

[260] Farooq MS, Arooj A, Alroobaea R, Baqasah AM, Jabarulla MY, Singh D, Sardar R. Untangling computer-aided diagnostic system for screening diabetic retinopathy based on deep learning techniques. Sensors 2022;22(5):1803.

[261] Maqsood S, Damaševičius R, Maskeliūnas R. TTCNN: A breast cancer detection and classification towards computer-aided diagnosis using digital mammography in early stages. Appl Sci 2022;12(7):3273.

[262] Gayathri J, Abraham B, Sujarani M, Nair MS. A computer-aided diagnosis system for the classification of COVID-19 and non-COVID-19 pneumonia on chest X-ray images by integrating CNN with sparse autoencoder and feed forward neural network. Comput Biol Med 2022;141:105134.

[263] Kumar V, Roche C, Overman S, Simovitch R, Flurin P-H, Wright T, Zuckerman J, Routman H, Teredesai A. What is the accuracy of three different machine learning techniques to predict clinical outcomes after shoulder arthroplasty? Clin Orthop Relat Res 2020;478(10).

[264] Kunze KN, Krivicich LM, Clapp IM, Bodendorfer BM, Nwachukwu BU, Chahla J, Nho SJ. Machine learning algorithms predict achievement of clinically significant outcomes after orthopaedic surgery: a systematic review. Arthrosc: J Arthrosc Relat Surg 2022;38(6):2090–105.

[265] Franceschetti E, Gregori P, De Giorgi S, Martire T, Za P, Papalia GF, Giurazza G, Longo UG, Papalia R. Machine learning can predict anterior elevation after reverse total shoulder arthroplasty: A new tool for daily outpatient clinic? Musculoskelet Surg 2024;1–9.

[266] Kumar V, Roche C, Overman S, Simovitch R, Flurin P-H, Wright T, Zuckerman J, Routman H, Teredesai A. Using machine learning to predict clinical outcomes after shoulder arthroplasty with a minimal feature set. J Shoulder Elbow Surg 2021;30(5):e225–36.

[267] Roche C, Kumar V, Overman S, Simovitch R, Flurin P-H, Wright T, Routman H, Teredesai A, Zuckerman J. Validation of a machine learning–derived clinical metric to quantify outcomes after total shoulder arthroplasty. J Shoulder Elbow Surg 2021;30(10):2211–24.

[268] Kumar V, Allen C, Overman S, Teredesai A, Simovitch R, Flurin P-H, Wright T, Zuckerman J, Routman H, Roche C. Development of a predictive model for a machine learning–derived shoulder arthroplasty clinical outcome score. In: Seminars in arthroplasty: JSES. Vol. 32, Elsevier; 2022, p. 226–37.

[269] Baumgarten KM. Accuracy of blueprint in predicting range of motion one year after reverse total shoulder arthroplasty. J Shoulder Elbow Surg 2023.
[270] Huang D, Celi LA, O'Brien Z. Biases in machine learning in healthcare. Artif Intell Clin Med 2023;426.
[271] Kaur D, Uslu S, Rittichier KJ, Durresi A. Trustworthy artificial intelligence: a review. ACM Comput Surv 2022;55(2):1–38.
[272] Correa R, Shaan M, Trivedi H, Patel B, Celi LAG, Gichoya JW, Banerjee I. A systematic review of 'Fair'AI model development for image classification and prediction. J Med Biol Eng 2022;42(6):816–27.
[273] Albahri A, Duhaim AM, Fadhel MA, Alnoor A, Baqer NS, Alzubaidi L, Albahri O, Alamoodi A, Bai J, Salhi A, et al. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. Inf Fusion 2023.
[274] Alzubaidi L, Al-Sabaawi A, Bai J, Dukhan A, Alkenani AH, Al-Asadi A, Alwzwazy HA, Manoufali M, Fadhel MA, Albahri A, et al. Towards risk-free trustworthy artificial intelligence: Significance and requirements. Int J Intell Syst 2023;2023.
[275] Borjali A, Chen AF, Muratoglu OK, Morid MA, Varadarajan KM. Deep learning in orthopedics: how do we build trust in the machine? Healthc Transf 2020.
[276] Rasheed K, Qayyum A, Ghaly M, Al-Fuqaha A, Razi A, Qadir J. Explainable, trustworthy, and ethical machine learning for healthcare: A survey. Comput Biol Med 2022;106043.
[277] Albahri A, Jassim MM, Alzubaidi L, Hamid RA, Ahmed M, Al-Qaysi Z, Albahri O, Alamoodi A, Alqaysi M, Mohammed TJ, et al. A trustworthy and explainable framework for benchmarking hybrid deep learning models based on chest X-Ray analysis in CAD systems. Int J Inf Technol Decis Mak 2024.
[278] Liang W, Tadesse GA, Ho D, Fei-Fei L, Zaharia M, Zhang C, Zou J. Advances, challenges and opportunities in creating data for trustworthy AI. Nat Mach Intell 2022;4(8):669–77.
[279] Nazar M, Alam MM, Yafi E, Su'ud MM. A systematic review of human–computer interaction and explainable artificial intelligence in healthcare with artificial intelligence techniques. IEEE Access 2021;9:153316–48.
[280] Toreini E, Aitken M, Coopamootoo K, Elliott K, Zelaya CG, Van Moorsel A. The relationship between trust in AI and trustworthy machine learning technologies. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. 2020, p. 272–83.
[281] Banerjee P, Barnwal RP. Methods and metrics for explaining artificial intelligence models: A review. In: Explainable AI: Foundations, methodologies and applications. Springer; 2023, p. 61–88.
[282] Stratified. Stratified. 2023, https://www.scribbr.com/methodology/stratified-sampling/. [Accessed 8 March 2023].
[283] Ali F, El-Sappagh S, Islam SR, Kwak D, Ali A, Imran M, Kwak K-S. A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. Inf Fusion 2020;63:208–22.
[284] Alzubaidi L, Chlaib HK, Fadhel MA, Chen Y, Bai J, Albahri A, Gu Y. Reliable deep learning framework for the ground penetrating radar data to locate the horizontal variation in levee soil compaction. Eng Appl Artif Intell 2024;129:107627.
[285] Fadhel MA, Duhaim AM, Saihood A, Sewify A, Al-Hamadani MN, Albahri A, Alzubaidi L, Gupta A, Mirjalili S, Gu Y. Comprehensive systematic review of information fusion methods in smart cities and urban environments. Inf Fusion 2024;102317.
[286] Polinati S, Bavirisetti DP, Rajesh KN, Dhuli R. Multimodal medical image fusion based on content-based and PCA-sigmoid. Curr Med Imaging 2022;18(5):546–62.
[287] Dong C, Xu S, Dai D, Zhang Y, Zhang C, Li Z. A novel multi-attention, multi-scale 3D deep network for coronary artery segmentation. Med Image Anal 2023;85:102745.
[288] Zeng N, Wu P, Wang Z, Li H, Liu W, Liu X. A small-sized object detection oriented multi-scale feature fusion approach with application to defect detection. IEEE Trans Instrum Meas 2022;71:1–14.
[289] Dang J, Li H, Niu K, Xu Z, Lin J, He Z. Kashin-beck disease diagnosis based on deep learning from hand X-ray images. Comput Methods Programs Biomed 2021;200:105919.
[290] Zhang H, Xu H, Tian X, Jiang J, Ma J. Image fusion meets deep learning: A survey and perspective. Inf Fusion 2021;76:323–36.
[291] Azam MA, Khan KB, Salahuddin S, Rehman E, Khan SA, Khan MA, Kadry S, Gandomi AH. A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics. Comput Biol Med 2022;144:105253.
[292] Yoshii Y, Totoki Y, Sashida S, Sakai S, Ishii T. Utility of an image fusion system for 3D preoperative planning and fluoroscopy in the osteosynthesis of distal radius fractures. J Orthop Surg Res 2019;14:1–7.
[293] Tang F, Liang S, Zhong T, Huang X, Deng X, Zhang Y, Zhou L. Postoperative glioma segmentation in CT image using deep feature fusion model guided by multi-sequence MRIs. Eur Radiol 2020;30:823–32.
[294] Al-Timemy AH, Ghaeb NH, Mosa ZM, Escudero J. Deep transfer learning for improved detection of keratoconus using corneal topographic maps. Cogn Comput 2022;14(5):1627–42.
[295] Li J, Wang Q. Multi-modal bioelectrical signal fusion analysis based on different acquisition devices and scene settings: Overview, challenges, and novel orientation. Inf Fusion 2022;79:229–47.
[296] Food U, Administration D, et al. Good machine learning practice for medical device development: Guiding principles. 2021.
[297] Lyell D, Coiera E, Chen J, Shah P, Magrabi F. How machine learning is embedded to support clinician decision making: an analysis of FDA-approved medical devices. BMJ Health Care Inform 2021;28(1).
[298] Gichoya JW, Banerjee I, Bhimireddy AR, Burns JL, Celi LA, Chen L-C, Correa R, Dullerud N, Ghassemi M, Huang S-C, et al. AI recognition of patient race in medical imaging: a modelling study. Lancet Digit Health 2022;4(6):e406–14.
[299] Hammond A, Jain B, Celi LA, Stanford FC. An extension to the FDA approval process is needed to achieve AI equity. Nat Mach Intell 2023;1–2.
[300] Oeding JF, Williams III RJ, Camp CL, Sanchez-Sotelo J, Kelly BT, Nawabi DH, Karlsson J, Pearle AD, Martin RK, Jang SJ, et al. A practical guide to the development and deployment of deep learning models for the orthopedic surgeon: part II. Knee Surg Sports Traumatol Arthrosc 2023;1–9.
[301] Burns D, Abbas A, Toor J, Hardisty M. AI in orthopaedic surgery. Artif Intell Clin Med 2023;266.
[302] Paoletti ME, Moreno-Álvarez S, Haut JM. Multiple attention-guided capsule networks for hyperspectral image classification. IEEE Trans Geosci Remote Sens 2021;60:1–20.
[303] Yildiz Potter I, Yeritsyan D, Mahar S, Wu J, Nazarian A, Vaziri A, Vaziri A. Automated bone tumor segmentation and classification as benign or malignant using computed tomographic imaging. J Digit Imaging 2023;1–10.
[304] Riem L, Feng X, Cousins M, DuCharme O, Leitch EB, Werner BC, Sheean AJ, Hart J, Antosh IJ, Blemker SS. A deep learning algorithm for automatic 3D segmentation of rotator cuff muscle and fat from clinical MRI scans. Radiol: Artif Intell 2023;e220132.
[305] Jang SJ, Flevas DA, Kunze K, Anderson C, Fontana MA, Boettner F, Sculco TP, Baldini A, Sculco PK. Standardized fixation zones and cone assessments for revision total knee arthroplasty using deep learning. J Arthroplasty 2023.
[306] Hiasa Y, Otake Y, Takao M, Ogawa T, Sugano N, Sato Y. Automated muscle segmentation from clinical CT using Bayesian U-net for personalized musculoskeletal modeling. IEEE Trans Med Imaging 2019;39(4):1030–40.
[307] Schlemper J, Oktay O, Schaap M, Heinrich M, Kainz B, Glocker B, Rueckert D. Attention gated networks: Learning to leverage salient regions in medical images. Med Image Anal 2019;53:197–207.
[308] Wang J, Lv P, Wang H, Shi C. SAR-u-net: Squeeze-and-excitation block and atrous spatial pyramid pooling based residual U-net for automatic liver segmentation in computed tomography. Comput Methods Programs Biomed 2021;208:106268.
[309] Yeung M, Sala E, Schönlieb C-B, Rundo L. Focus U-net: A novel dual attention-gated CNN for polyp segmentation during colonoscopy. Comput Biol Med 2021;137:104815.
[310] Zhao J, Han J, Li J, Du G. Improved U-net network for segmentation on femur images. In: Advances in natural computation, fuzzy systems and knowledge discovery: proceedings of the ICNC-fSKD 2021 17. Springer; 2022, p. 50–60.
[311] Anastasio AT, Zinger BS, Anastasio TJ. A novel application of neural networks to identify potentially effective combinations of biologic factors for enhancement of bone fusion/repair. PLoS One 2022;17(11):e0276562.
[312] Qiu L, Zhao L, Hou R, Zhao W, Zhang S, Lin Z, Teng H, Zhao J. Hierarchical multimodal fusion framework based on noisy label learning and attention mechanism for cancer classification with pathology and genomic features. Comput Med Imaging Graph 2023;102176.
[313] Kumar S, Chaube MK, Alsamhi SH, Gupta SK, Guizani M, Gravina R, Fortino G. A novel multimodal fusion framework for early diagnosis and accurate classification of COVID-19 patients using X-ray images and speech signal processing techniques. Comput Methods Programs Biomed 2022;226:107109.
[314] Sapci AH, Sapci HA. Artificial intelligence education and tools for medical and health informatics students: systematic review. JMIR Med Educ 2020;6(1):e19285.
[315] Grunhut J, Wyatt AT, Marques O. Educating future physicians in artificial intelligence (AI): an integrative review and proposed changes. J Med Educ Curric Dev 2021;8:23821205211036836.
[316] Cussat-Blanc S, Castets-Renard C, Monsarrat P. Doctors in medical data sciences: A new curriculum. Int J Environ Res Public Health 2023;20(1):675.
[317] Mendes J, Pereira T, Silva F, Frade J, Morgado J, Freitas C, Negrão E, de Lima BF, da Silva MC, Madureira AJ, et al. Lung CT image synthesis using GANs. Expert Syst Appl 2023;215:119350.
[318] Jiang M, Zhi M, Wei L, Yang X, Zhang J, Li Y, Wang P, Huang J, Yang G. FA-GAN: Fused attentive generative adversarial networks for MRI image super-resolution. Comput Med Imaging Graph 2021;92:101969.
[319] Topol EJ. Welcoming new guidelines for AI clinical research. Nat Med 2020;26(9):1318–20.
[320] Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK, Ashrafian H, Beam AL, Chan A-W, Collins GS, Deeks ADJ, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Lancet Digit Health 2020;2(10):e537–48.

[321] Grauhan NF, Niehues SM, Gaudin RA, Keller S, Vahldiek JL, Adams LC, Bressem KK. Deep learning for accurately recognizing common causes of shoulder pain on radiographs. Skelet Radiol 2021;1–8.

[322] Kang Y, Choi D, Lee KJ, Oh JH, Kim BR, Ahn JM. Evaluating subscapularis tendon tears on axillary lateral radiographs using deep learning. Eur Radiol 2021;31(12):9408–17.

[323] Mall PK, Singh PK. Explainable deep learning approach for shoulder abnormality detection in X-Rays dataset. Int J Next-Gener Comput 2022;13(3).

[324] Cheng C-T, Hsu C-P, Ooyang C-H, Chou C-Y, Lin N-Y, Lin J-Y, Ku Y-K, Lin H-S, Kao S-K, Chen H-W, et al. Evaluation of ensemble strategy on the development of multiple view ankle fracture detection algorithm. Br J Radiol 2023;96(1145):20220924.