Intelligence-Based Medicine 5 (2021) 100033

Contents lists available at ScienceDirect

Intelligence-Based Medicine

journal homepage: www.journals.elsevier.com/intelligence-based-medicine

Artificial intelligence for diagnosis of fractures on plain radiographs: A scoping review of current literature

Clare Rainey [a,*], Jonathan McConnell [b], Ciara Hughes [a], Raymond Bond [c], Sonyia McFadden [a]

[a] *School of Health Sciences, Ulster University, Jordanstown, United Kingdom*
[b] *NHS Scotland, Greater Glasgow and Clyde, United Kingdom*
[c] *School of Computing, Ulster University, Jordanstown, United Kingdom*

ARTICLE INFO

Keywords:
Artificial intelligence
Fracture identification
Radiology
X-ray
Radiographic image interpretation
Plain radiography

ABSTRACT

Aim: To complete a scoping review of the literature investigating the performance of artificial intelligence (AI) systems currently in development for their ability to detect fractures on plain radiographic images.
Methods: A systematic approach was adopted to identify papers for inclusion in this scoping review and utilised the Preferred Reporting Items for Systematic Reviews and Meta-Analysis Statement (PRISMA). Following application of inclusion and exclusion criteria, sixteen studies were included in the final review.
Results: With the exception of one study, all studies report that AI models demonstrated an ability to perform fracture identification tasks on plain skeletal radiographs. Metrics used to report performance are variable throughout all reviewed studies and include area under the receiver operating characteristic curve (AUC), sensitivity and specificity, positive predictive value, negative predictive value, precision, recall, F1 score and accuracy. Reported performances for studies indicated AUC values range from AUC 0.78 (weakest) to the best performing system reporting AUC 0.99.
Conclusion: The review found a great variation in the AI model architectures, training and testing methodology as well as the metrics used to report the performance of the networks. A standardisation of the reporting metrics and methods would permit comparison of proposed models and training methods which may accelerate the testing of AI systems in the clinical setting. Prevalence agnostic metrics should be used to reflect the true performance of such systems. Many studies lacked any explainability for the algorithmic decision making of the AI models, and there was a lack of interrogation into the potential reasons for misclassification errors. This type of 'failure analysis' would have provided insight into the biases and the aetiology of AI misclassifications.

1. Background

1.1. Introduction

The use of radiology services in England has increased by 16% in the 5 years preceding the 2016/17 Diagnostic Imaging Dataset Annual Statistical Release [1]. Imaging activity counts for the year March 2019–March 2020 demonstrate that plain radiographic examinations make up the majority (52%) of imaging procedures undertaken [2]. Serious concerns have been uncovered in a number of hospital trusts in England, where the reporting backlog and adoption of the auto reporting system of working has led to incidents where pathologies went unreported, resulting in patient harm [1]. The most recent Care Quality Commission's Radiology Review (2018) [1] reported significant pressures on NHS trusts in England with 97% of trusts reporting an inability to meet increasing demands on radiology departments with the majority of backlogs in the reporting of plain radiographs. Artificial intelligence (AI) systems have been proposed to positively impact time efficiency within healthcare and, as such, the implementation of these systems has been prioritised in the NHS long term plan [3].

There are several significant drivers to the development of AI as a tool in the health care setting; namely, time constraints/efficiency, error avoidance or minimisation and workflow augmentation [4–6]. It is estimated that the implementation of an effective AI system for automated image reporting could reduce the time that radiologists spend reviewing images by 20% [7], and thus liberate 890,000 hours of radiologist time per annum in the UK [7]. This time can be spent doing non automatable tasks such as providing personalised patient care and more complex tasks where human input is essential [5].

[*] Corresponding author. School of Health Sciences, Ulster University, Jordanstown campus, Shore Road, Newtownabbey BT37 0QB, United Kingdom.
E-mail address: c.rainey@ulster.ac.uk (C. Rainey).

https://doi.org/10.1016/j.ibmed.2021.100033
Received 27 October 2020; Received in revised form 23 March 2021; Accepted 8 April 2021

1.2. Artificial intelligence in radiology

AI as a human adjunct in diagnosing pathology from radiographic images began in the 1960s [8]. A system was developed to convert images to numerical data, which was then stored on a computer that carried out statistical analysis.

In the 1980s, traditional computer aided detection (CAD) systems were beginning to be integrated into clinical radiology to detect human-programmed patterns in images to guide the clinician to areas requiring further attention [9,10]. Advances in computational power have permitted the development of increasingly more sophisticated applications of AI and CAD systems.

As computer processing power has increased, so follows the ability of the machine to accomplish increasingly more complicated and human-like tasks, such as the ability to learn from experience. This contrasts with older methods of CAD where systems are specifically programmed by human developers for feature detection. These are often referred to as symbolic reasoning, knowledge engineering or expert systems. Newer AI systems report higher accuracies [5,11] and more efficient training processes, as the AI learns from exposure to examples rather than human feature extraction and programming [12,13], although the success of these data driven algorithms rely on the availability of large volumes of data for training [12,14].

Many algorithms currently in development for image interpretation are based on Artificial Neural Network architectures (ANNs) [15,16]. These systems are inspired by the function of the human brain by using interconnected neurons or nodes which differentiate and make sense of different parts of the image. This form of AI can make predictions by either supervised or unsupervised learning [4,11]. In unsupervised learning, the system will identify similarities of features in images and allow for sorting of images into groups, for instance, grouping of patients with similar bone density [11]. Supervised learning is used when the AI is required to make diagnostic predictions based on human knowledge. In this case, the system or model is exposed to a large volume of examples, where the correct outcome or 'ground truth' label is known. The model then makes a series of decisions or predictions and receives feedback. ANNs are refined based on iterative feedback by assigning greater or lesser importance to particular nodes or artificial neurons by adjusting the 'weights' assigned to the neurons, using backpropagation [17]. This modulation will be tested again and adjusted to bring the AI prediction nearer to the ground truth label, usually the presence or absence of pathology or severity of a condition By determining the importance of various decisions based on a known outcome, the model can then learn the attributes of the input which were most significant in determining a particular outcome [9,18]. The ANN retains these weights and patterns of activation of the nodes if a correct prediction is made [18]. For example, an ANN might be exposed to a dataset of radiographic images where the outcome is known, for instance whether a fracture is present or not, and the algorithm learns based on the known diagnosis until an acceptable accuracy for fracture detection has been reached. This process is known as Supervised Machine Learning (ML), and encompasses, although is not limited to, ANNs. The exact reasoning by which the machine does this, however, is not clear due to a latent intermediate stage of processing. This stage takes place deep within the many layers of the system, hence the term 'deep learning' (DL). One type of ANN, which has been gaining attention recently in the field of computer vision and medical image interpretation, is the convolutional neural network (CNN). A CNN is a more sophisticated type of ANN which contains at least one convolutional layer, where weightings are shared between adjacent nodes. Although similar in structure to an ANN, these networks are proving to be particularly useful for image recognition tasks and are therefore able to be optimised and efficient for this purpose (see Table 1).

Trust and ethical issues exist due to the way ANNs and other DL models reach their decisions. These issues have been raised in a number of professional publications [19,20], and notably, in a joint statement by a worldwide radiology stakeholders' group [21]. These publications recognise the obvious benefits and necessity to incorporate AI into radiology but cautions that significant research should still be conducted into how AI should be utilised. They also emphasise the need for the clinicians and professionals involved in use and development of these systems to have an in-depth knowledge of their functionality.

Table 1
Introduction to concepts.

Artificial Intelligence (AI)	The ability of a computer to accomplish human-like tasks.
Machine Learning (ML)	ML is an AI system which is able to learn independently of human input by making a series of predictions or 'guesses' about an input and adjusts itself based on feedback from an established 'ground truth'.
Deep learning (DL)	DL is a subset of ML (and therefore AI) containing more processing layers – hence the term '*deep*'. Multiple layers allow for the accomplishment of more sophisticated tasks, e.g. the 2016 Alpha Go programme, natural language programming and image recognition.
Artificial Neural Networks (ANN)	An AI system inspired by the function of the human brain by the use of layers of interconnected nodes (artificial neurones)
Convolutional Neural Networks (CNN)	An advanced ANN, where neurones, and layers of neurones, can share information relating to the importance of detected features to other groups of neurones. This ability makes CNNs particularly good for complex for computer vison and image recognition tasks.
Support vector machines (SVM)	SVM are an older type of ML usually used in two-category classification tasks.
Training dataset	ML models are trained by exposure to multiple labelled examples, 'the training set' e.g. many images of a 'cat', 'dog', 'flower'.
Validation dataset	The validation set allows an initial impression of the performance of the model for fine-tuning of the model.
Test dataset	The test set is usually an unseen set of data, held-out from training and validation and used to provide final performance metrics of the model.
K-fold cross validation	Used for training and validation/testing using limited datasets by splitting the dataset into random number (k-) of groups (folds). Each fold will be used k times for training the model as well as validation/testing, therefore maximising the learning potential of the model.
Class balancing	Balanced classes have an equal number of desired outputs in each category. For example, in binary fracture classification (fracture/no fracture) an optimal training set would have a 1:1 split of fracture/no fracture for training, therefore maximising the ability of the ML to recognise both classes, although this does not usually replicate the real-world scenario.
Precision	Precision, or positive predictive value (PPV) is an indication of how many positive predications were actually positive. It is calculated using 'true' and 'false' positive (TP, FP) predictions: Precision = TP/TP + FP
Recall	Recall (or sensitivity) describes the ability of the model to correctly predict the presence of pathology and is calculated using 'true' positive (TP) and 'false' negative (FN) by the following equation: Recall = TP/TP + FN
Dice similarity coefficient (DSE) or F1-score	DSE or F1-score is metric used to describe the similarity between two response or outputs, in this case, AI predictions and ground truth. It is particularly useful in studies such as those described in this review as it takes both recall and precision into account and therefore is a suitable single metric which can accurately and efficiently report the performance of an ML on an imbalanced dataset.
Cohen's kappa	A prevalence agnostic metric used to quantify inter-rater agreement. The calculation takes into account the chance of any agreement occurring by chance.

1.3. Plain radiography in fracture identification

Fractures are a common reason for attendance at emergency departments around the world [22], although the use of AI to identify fractures on appendicular skeletal radiographs remains a relatively unexplored area. There were 2,489,052 hospital admissions for fractures in the ten-year period from 2004/2005 to 2013/2014 in England alone [22], which represents 47.84 per 10,000 of the population. This figure can be assumed to be much greater when patients who are not admitted to hospital and patients who are found not to have fractures are considered in the figures. This figure only considers patients who have been diagnosed as having a fracture and who have been admitted to hospital as a result of this fracture. The number of patients who have had radiographic imaging and have been found to have no fracture and those who have been diagnosed with a fracture but not admitted to hospital are not reflected in these statistics, therefore, the number of patients presenting for imaging for fractures can be assumed to be much greater than this.

Although there are no figures available for the number of radiographs taken to identify fractures, an indication can be gleaned from the number of patients presenting to minor injuries units in the UK where attendances have increased from 28% of total Emergency Department attendances in 2008/9 to 33% in 2017/18 in England [23]. Therefore, radiographic imaging for fracture identification contributes significantly to the workload of both radiologists and radiographers.

1.4. Reporting of artificial intelligence in medical imaging studies

As the field of AI in medical imaging grows, so follows the need for effective dissemination of results of studies which include details of the construction and performance of various models. The publication of detailed explanation and code availability, will allow for replication and validation of the proposed AI, permitting more efficient development into clinically useful tools and may improve clinicians' trust. However, as AI in medical imaging in its current format is still relatively new, a standardised system for the reporting of such studies has been lacking until very recently. To address this emerging issue, a Checklist for Artificial Intelligence in Medical Imaging (CLAIM) [24] has been produced, based on a modification of the Standards for Reporting of Diagnostic Accuracy Studies (STARD) [25] and is available at: https://pubs.rsna.org/doi/pdf/10.1148/ryai.2020200029. The use of this, or similar checklists should guide both AI in medical imaging researchers in the design and publication of findings and allow for robust review and comparison of the AI models proposed.

A robust review of the most recent developments and performances of AI systems is needed to provide a baseline of the state-of-the-art in AI to educate and inform those using and developing these systems for useful integration into the clinical workflow. This review aims to provide an insight into one area of development by providing a synthesis of the available literature on the performance of AI models to predict fractures on plain radiographic images.

2. Methods

2.1. Database search strategy

The search strategy for this review was designed in conjunction with a subject specialist librarian.

A search was conducted using broad search terms 'artificial intelligence' and 'computer aided diagnosis' was conducted on: Cochrane Library, PROSPERO, Ethos, ProQuest Dissertations, Google Scholar, JBI Database of Systematic Reviews and Implementation Reports. Results from this search were screened and none were found to match the search criteria and objectives of this paper.

A literature search was conducted in September 2019 and rerun in March and December 2020 to check for updates on the electronic academic databases Medline, Embase, CINAHL, Inspec and PubMed using the following key terms:

(artificial intelligence OR deep learn* OR machine learn*) AND (computer aided diagnosis OR clinical decision mak* OR automated diagnosis) AND (radiology OR radiography) with limits English language and human. A date range of 2016-present was applied to give an insight into the state-of-the-art of this rapidly evolving field and to attempt to ensure that the model architectures described in the literature were comparable. The first study in this field using 'modern' machine learning (ML) techniques was, to the best of the authors' knowledge, a study by Olczak et al., in 2017 [26], as cited in Chung et al., 2018 [27]. 2016 was then chosen as an assurance that any additional literature was identified.

To minimise the risk of introducing bias to this review, grey literature was sought from the following resources: Google Scholar, specialised databases (National Rehabilitation Information Centre, and the National Institute for Health Research Journals Library), and the International Clinical Trials Registry Platform. Hand searching of reference lists of articles and previous reviews was also performed to identify additional trials that were potentially eligible.

RefWorks Legacy® version was used to manage papers identified as a result of these searches. Duplicates were removed and all papers were screened for eligibility by reading titles and abstracts when the title did not adequately describe the study. The inclusion and exclusion criteria detailed in Table 2 were applied. Each remaining paper was read in full, and inclusion and exclusion criteria were applied again. This process is clarified in Fig. 1.

2.2. Data analysis

Each paper was read thoroughly, and data was extracted under the following headings: Anatomical area, pathology focus, determination of truth/reference standard, ML description/techniques, feature engineering detail, training set/method, test set/method, class balancing, performance metrics/results, methods to explain ML decision and misclassification explanation. Investigation into code availability was conducted by search of both the paper and any supplementary data provided.

3. Findings

3.1. Search results

Following searching of academic databases listed, 2786 papers were identified. An additional two papers were identified from grey literature and reference lists of included articles. 225 duplicate papers were removed. Following application of inclusion and exclusion criteria, by means of manual title and abstract screening of the remaining 2563 papers, 23 papers remained. All papers were read in full by the authors, and 13 papers were excluded for the reasons outlined in Fig. 1. At this stage, ten studies remained for full data analysis. A final inspection of grey literature and full search on all databases was conducted in December 2020 to identify any recent updates. A further seven papers were identified and included in this paper; however, two papers were unavailable

Table 2
Inclusion and exclusion criteria applied to search results.

Inclusion	Exclusion
Diagnostic imaging – conventional radiography All bone fracture diagnostic studies Recent publications – 2016-present Experimental study (with performance results)	Specialised imaging – CT, NM, MRI, mammography Non-diagnostic procedures – therapies and segmentation Artificial intelligence used for any other reason other than obtaining/assisting with diagnosis of fractures Studies published before 2016 Information only papers – no experimental results

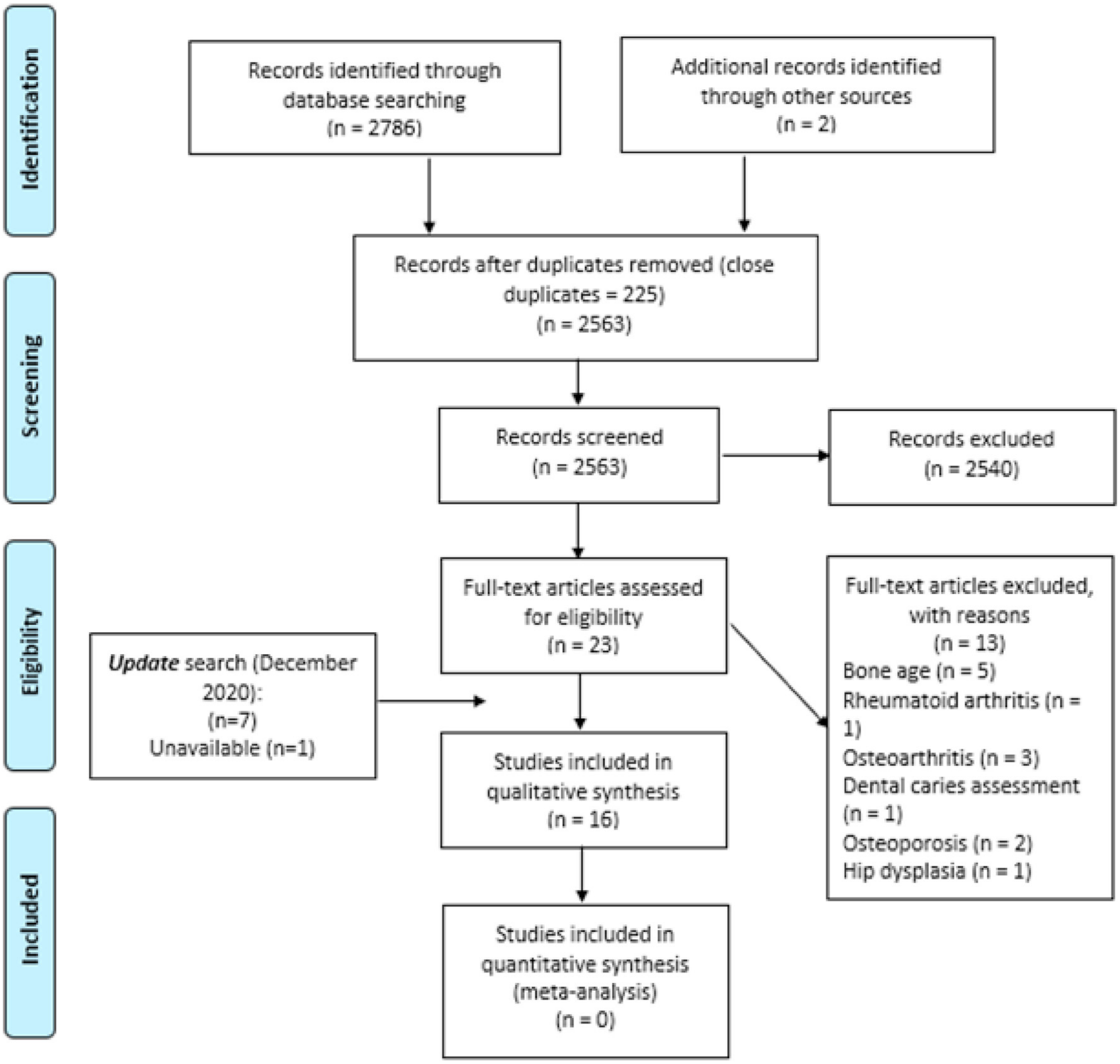
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 2786)
Additional records identified through other sources (n = 2)
Records after duplicates removed (close duplicates = 225) (n = 2563)
Records screened (n = 2563)
Records excluded (n = 2540)
Full-text articles assessed for eligibility (n = 23)
Full-text articles excluded, with reasons (n = 13)
Bone age (n = 5)
Rheumatoid arthritis (n = 1)
Osteoarthritis (n = 3)
Dental caries assessment (n = 1)
Osteoporosis (n = 2)
Hip dysplasia (n = 1)
Update search (December 2020): (n=7) Unavailable (n=1)
Studies included in qualitative synthesis (n = 16)
Studies included in quantitative synthesis (meta-analysis) (n = 0)

Fig. 1. Prisma flow chart.

due to institutional access issues. One of these papers has been located through a search on ARXIV and is included in this review. The one remaining paper is discussed in the recommendations section and details are in the reference list of this paper. There are, therefore, sixteen papers included in this review.

Extracted data is presented in full in Tables 3–5.

3.2. Anatomical area (Table 3)

All studies included were determining either the presence of a fracture or classification of fracture severity. Anatomical area varied across the studies. Four studies focused on the wrist and distal radius [29,33,35,38], eight on hip or pelvis fractures [28,30,31,34,37,39–41], one on proximal humerus [27], one on dental fractures [32], one paper focused on ankle fractures [36] and one study investigated a range of anatomical areas [26], as listed in Table 3, with fracture detection as the pathological focus.

3.3. Pathology focus (Table 3)

Featured models used both binary detection and multi-class classification as outcomes, i.e. presence of fracture, expressed as a binary prediction (fracture/no fracture) and classification of fracture severity and location of fracture as multi-class problems. Binary classes used to predict hip, wrist, ankle and hand were reported in the studies, while fracture location and multi-class discrimination was determined for hip and shoulder radiographs. One study used only a region of interest box to identify the area of abnormality on orthopantomographic images, rather than textual diagnoses [32].

3.4. Prediction classes – description/number (Table 3)

Twelve of the sixteen studies reported the ability of a convolutional neural network to predict the presence or absence of a fracture [28–31, 33–38,40,41]. Five studies included some discrimination of the location or severity of any identified fractures [27,34,37,39,41]. One of these studies classified fractures according to a well-known scale for proximal humerus fractures [27].

All others focussed on proximal femur/hip fractures. Two studies required the model to decide between three classes, one relating to the presence of a fracture and two choices of classification according to a well know classification system [34,39]. One of these studies further subclassified one of the categories [39]. A further study which also required the model to make a choice of three outputs required the AI to determine presence or absence of a fracture with the additional option of 'missing' [26], although the meaning of this is unclear. The study with the largest number of output options investigated an AI model's ability to classify the type of proximal femur abnormality by presence and location of any

Table 3
Anatomical and pathological focus of AI model.

Author/Country/ Year	Anatomical area	Pathology focus	Prediction classes – description/number	Determination of truth/reference standard
Badgeley et al, 2019. USA [28]	Pelvis/hip	Hip fracture	Two - Fracture/no fracture	Inferred from patient's clinical notes: radiologist comment: 'acute fracture' or 'no acute fracture'.
Bluthgen et al, 2020. Switzerland [29]	Wrist (distal radius) single and multi-view comparison	Distal radius fracture/no fracture	Two - defect (0 - fracture); intact (1 - no fracture).	TRAINING SET - radiology reports + confirmation by two radiology residents (3rd and 5th year) using electronic healthcare record, CT scans and images. EXTERNAL SET (MURA) - interpreted by radiology residents as above for fracture/no fracture. AREA AGREEMENT – Region of interest drawn by radiologists (agreement with each other if Dice Similarity Coefficient (DCE) = 0.7) and deep learning system agree if overlap (does not state by how much).
Cheng et al, 2019. Taiwan [30]	Pelvis	Hip fracture	Two – fracture/no fracture	Diagnosis from trauma registry. Computed Tomography (CT), clinical course and other imaging used to determine equivocal cases.
Chung et al, 2018. South Korea [27]	Proximal humerus (shoulder) single view	Fracture - detection and classification.	Five - Neer's classification of proximal humerus fractures - four types of fracture + normal = five classifications	Two shoulder orthopaedists and one radiologist (musculoskeletal specialist). When no agreement from independent reports, CT and other imaging is checked. If still no agreement, image excluded.
Damien et al, 2019. Lebanon [31]	Pelvis fracture - iliopectineal line only	Iliopectineal line disruption	Two - Positive/negative for fracture. Feature extraction, size of connected components, number of connected components (or 'parts' of pelvis)	Unclear - Radiopaedia diagnoses assumed
Fukuda et al, 2019. Japan [32]	Dental OPG	Vertical root fractures	Region of interest indication by ML	Two oral and maxillofacial radiologist and one endodontist set regions of interest containing fractures.
Gan et al, 2019. China [33]	Distal radius	Fractures – distal radius (wrist)	Initial region of interest – localisation of distal radius only. Two – fracture/no fracture on final regions of interest.	Radiology report plus verification from two senior orthopaedists (in cases of no agreement, consensus was obtained from a third senior orthopaedist and corresponding CT scans)
Jiménez-Sanchez et al (2019) [34]	Proximal femur	Fractures – localisation and classification	1) Binary classification – Fracture/no fracture 2) Discrimination – classes A, B and no fracture	Three clinical experts provided class labels and localisation of the head of femur using region of interest boxes: one trauma surgeon, one senior radiologist and one trauma surgery resident (5th year) evaluated a ***split*** of the dataset each.
Kim and McKinnon, 2018. UK [35]	Wrist (lateral)	Fractures - distal radius or ulna	Two - Fracture/no fracture	Radiology report and registrar verification
Kitamura et al, 2019. USA [36]	Single and multi-view ankle.	Fractures	Two - Fracture/no fracture	Radiology reports reviewed by radiologist and 4th year radiology resident.
Krogue et al, 2020. USA [37]	Hip	Fractures	Binary task: fractured/not fractured Classification task: Fractured (undisplaced femoral neck, displaced femoral neck fractures, intertrochanteric fractures) Unfractured Containing hardware (previous internal fixation, arthroplasty) Localisation by bounding box	Review by two orthopaedic residents. In cases of uncertainty, computed tomography, magnetic resonance imaging and post-surgical imaging was used as confirmation.
Lindsay et al, 2018. USA [38]	Wrist	Fractures	Two - fracture/no fracture	One or more senior sub-specialist orthopaedic surgeons using a bounding box to locate pathology.
Olczak et al, 2017. Sweden [26]	Hand, scaphoid, wrist and ankle.	Fractures, body part, laterality and exam view	Three for pathology- yes/no/missing fracture. Side: Left/Right. View: distal, Antero-posterior, oblique, proximal, radial, lateral, ulnar, missing. Body part: finger, thumb, scaphoid, hand, wrist, ankle, missing.	Automated language extraction from radiologists' reports, along with 'multiple visits'. Other information (laterality, body part, view) from DICOM headings.
Tanzi et al, 2020. Italy, Sweden [39]	Proximal femur	Classification of fractures	1) Three-class discrimination (unfractured, type A fracture, type B fracture) 2) Sub-classification of the type A fracture – A1, A2 and A3.	Senior trauma surgeon and specialist orthopaedist reviewed all images in initial dataset.
Urakawa et al, 2019. Japan [40]	Inter trochanteric hip fractures	Detection of fracture	Two: fracture/no fracture	Single orthopaedic surgeon using antero-posterior hip radiographs (in 91.7% of cases), lateral hip radiographs (50 patients), computed tomography (seven patients) and magnetic resonance scan results (90 patients) and surgical intervention.
Yu et al, 2020. USA [41]	Proximal femoral fractures	Detection of fracture	Two: fracture/no fracture + localisation: transcervical, intertrochanteric, subtrochanteric.	Musculoskeletal radiologist from antero-posterior radiographs and confirmed by either computed tomography or operative report. Localisation by classifying the fracture in one of three pre-specified areas

Table 4
AI development, training and testing methods.

Author/ Country/Year	ML description/ techniques	Feature engineering detail	Training set/method	Training set demographics	Test set/method (Sample size/cases)	Class balancing
Badgeley et al, 2019. USA [28]	Inception V3 pretrained on ImageNet	299 × 299 pixels	23,602 hip radiographs Train:test 3:1. Optimised parameters from the ImageNet challenge. Penultimate layer is removed, leaving 2048 image feature scores which are used in subsequent unsupervised models. Dimension reduction techniques were used to 'visualise the distribution of image variation' (p31). t-Distributed Stochastic Neighbor Embedding (t-SNE) projected the image feature vector into a 2 d plane with the R package. (50 dimensions, perplexity 30, theta 0.5, initial momentum 0.5, final momentum 0.8, learning rate 200)	Female 66% Mean age 61 Mean body mass index 28 Number of scanners 11	Train:test 3:1 of 23,602 hip radiographs. Confounding factors removed for testing - patient trait, scanner type and 'other information'	Full dataset: 3% fracture n = 779. Test set: 3% fracture No attempts to artificially balance classes.
Bluthgen et al, 2020. Switzerland [29]	Manufacturing CNN. Retrained for fracture detection. ViDi image analysis software (deep learning).	MURA dataset has a maximum pixel height of 512 pixels. Internal dataset resized to match this. Aspect ratio maintained.PNG format. Single images used for training. Images (AP and lateral) together on same image for test.	524 radiographs. Grid search plug-in used to test hyperparameter combinations. Two best combinations used for training: Hyperparameters selected and best 2 chosen to train: Model 1–85 pixels Model 2–60 pixels. Both: 150 epochs, contrast 50%, aspect ratio 10%, rotation 10%, shear 20%, scale 1-%, sampling density 5, luminance 40%. 'Data augmentation' had a positive effect on model performance in the validation stage.	Not explicitly stated	INTERNAL and EXTERNAL (MURA) sets. Internal set: 100 radiographs, 42% fracture. External set: 200 radiographs (AP and lateral) - 100 cases. 50% fracture (50 cases/ 50 cases)	Total training set: 524 radiographs, 166 fracture (31.7% fracture). Test set: external test set balanced: 50 fracture 50 no fracture. Internal test set: 42% fracture. No attempts to artificially balance classes in training data set. External data set intentionally balanced.
Cheng et al, 2019. Taiwan [30]	DenseNet 121	Whole images. Resized to 512 × 512 pixels with 8-bit greyscale colour.	Pretrained on limb radiographs - 25,505 (90% train, 10% valid). Retrained on pelvic radiographs - 3605 (80% train, 20% valid). Batch size 8. Adam optimiser used. Initial learning rate of 10^{-3} Final model trained on 60 epochs. Data augmentation in training: zoom 10%, horizontal flip, vertical flip, rotation 10°	Mean age with fracture: 72.34 Mean age without hip fracture: 44.88 Gender with hip fracture: 42% male Gender without hip fracture: 68.2 male	100 pelvis radiographs: 25 femoral neck fracture, 25 intertrochanteric fracture, 50 no fracture.	PELVIS training model: 1975 fracture/1630 no fracture. Test set: balanced test set: 50 no fracture; 50 fracture. No attempts to artificially balance classes in training set.
Chung et al, 2018. South Korea [27]	Pretrained Microsoft ResNet-152, fine-tuned on images from dataset.	Cropped images to include humeral head making up approximately 50% of image size. 256 × 256 pixel (downsized)	Full dataset - 1891 AP shoulder radiographs **from 7 different hospitals**. Training on 9/10 of dataset + remnants. Repeated three times. Caffe 9 used on Ubuntu 16.04 with NVIDIA GTX 1070. ResNet fine-tuned final layers: 'base lr: 0.0001; max: 3 epochs; step: 2 epochs; gamma: 0.1; weight decay: 0.00001; train batch size: 24; 1 epoch.' (suppl. data) Classification training: Non fracture radiographs removed to reduce overfitting.	Total dataset: 1891 patients (591 men, mean age: 65	1/10 of total dataset. CNN v human study: 3 groups of readers: 28 GPs; 11 general orthopaedists; 19 shoulder orthopaedists. 14 images for testing	Full dataset: 1376 fracture cases, 515 normal from dataset of 1891 images. 346, 514, 269, 247 and 515 for greater tuberosity, surgical neck, three-part fractures, four-part fractures and no fracture classes respectively. No attempts to artificially balance classes apparent in either training or test data sets.

(continued on next page)

Table 4 (*continued*)

Author/ Country/Year	ML description/ techniques	Feature engineering detail	Training set/method	Training set demographics	Test set/method (Sample size/cases)	Class balancing
Damien et al, 2019. Lebanon [31]	Support Vector Machines (SVM). Neural network (limited information): 2 hidden layer, six neurons per layer.	Radiologist or surgeon selects Region of interest (ROI) demarcating iliopectineal line. Image denoising and edge detection performed, then smoothed.	Dataset from Radiopaedia ®radiographs. Neural network: 75 images for training +11 for validation. 100 for SVM SVM - radial basis function used as kernel. NN – two hidden layers, six neurones in each layer, hyperbolic tangent as activation function.	Not explicitly stated		% of training and testing images in each class is unclear. No attempts to artificially balance classes apparent.
Fukuda et al, 2019. Japan [32]	Digits v 5.0 training system - customised DetectNet	.JPEG images downsized to 900 × 900 for ML	300 Orthopantomographic images Total number of teeth not stated. Trained on Ubuntu 16.04 operating system, GEForce 1080Ti GPU (Nvidia) over 1000 epochs using Adam solver with an initial learning rate of 0.0001. Five models created and tested with test set for each (five-fold cross validation).	50% male/female Mean age 66.05	Five-fold cross validation. Four parts of dataset train and validation. Repeated five times, changing TEST dataset each time.	300 OPG - 330 fractured teeth in total. Total teeth not stated. At least one vertical root fracture per OPG. Test set: demographics of test data set not explicitly stated No attempts to artificially balance classes in training or test data sets.
Gan et al, 2019. China [33]	Inception V4: authors reported that this ML has achieved ‘state of the art results in recent image classification contests’ (p. 396)	Whole radiograph.JPEG images, resized to 600 × 800 pixels. Resultant region of interest containing distal radius resized to 200 × 200	Training dataset: 2040 antero-posterior wrist radiographs: 1341 with fracture, 699 no fracture for region of interest identification. Resultant region of interest radiographs + augmentation: 6120 images: 4023 with fracture, 2097 no fracture for final testing and validation (15% for validation) For diagnostic CNN: Google open source TensorFlow 1.11.0 on Ubuntu 16.04. NVIDIA Titan X. ‘Optimiser, stochastic gradient descent; batch size, 100; dropout, 0.5; 20,000 iterations; initial learning rate, 0.001; learning rate decay type, fixed.’ (Suppl data)	Entire dataset: 1366 men, 974 females. With fracture: 56% men, 44% female. Without fracture: 63% men, 37% female. Mean age 48 (mean age with fracture: 48, without: 48)	300 antero-posterior wrist radiographs: 150 with fracture, 150 no fracture	1341 fracture/699 no fracture in initial dataset. Following augmentation: 4023 fracture/2097 no fracture in final (region of interest) dataset used for training. Augmentation was not intended to balance classes Balanced test dataset (150 #/150 no #).
Jiménez-Sanchez et al (2019), Spain, Germany and France [34]	For classification task: ResNet, pretrained on ImageNet. For localisation: AlexNet.	ResNet; radiographs downsized to 224 × 224 pixels AlexNet: radiographs downsized to 227 × 227	Initial dataset: 780 subsequently sampled pelvis radiographs of patients with proximal femur fractures. 4% of patients had antero-posterior projections. The remainder had antero-posterior and lateral projections. Most cases had one non-fractured proximal femur. Train:validate:test 70:10:20% Training on a Linux based workstation (16 GB RAM, Intel Xenon CPU at 3.5 GHz, 64 GB GeForce GTX 1080). Stochastic gradient descent for optimisation. Models trained until convergence (Classification and localisation: 80 and 200 epochs respectively). Batch size 64. Momentum 0.9 for all models. Learning rate initialised 1×10^{-2} for classification and 1×10^{-8}	69% female. Mean age 75.7 ± 13.2	1) Binary classification task: not fractured/ abnormal: 115/115 2) Discrimination task: Classes A, B and no fracture: 55; 60; 55.	Data augmentation: translation, scaling and rotation Training set of pelvises with at least one fractured hip. Pelvis radiographs (two femora) were parted in two (one femur each image) resulting in 780 fractured femora and 567 not fractured for two class problem. Three-class problem (type A, B and not fractured): 327, 453 and 567 respectively. No explicit attempt to further balance of training dataset. Intentionally class balanced ***test*** sets: 1) Binary classification task: not fractured/ abnormal: 115/115 (images from three class problem plus additional

(*continued on next page*)

Table 4 (*continued*)

Author/ Country/Year	ML description/ techniques	Feature engineering detail	Training set/method	Training set demographics	Test set/method (Sample size/cases)	Class balancing
			for localisation. Decay varied.			55 not fractured images) 2) Discrimination task: Classes A, B and no fracture: 55; 60; 55.
Kim and McKinnon, 2018. UK [35]	Inception V3 network trained on ImageNet.	.JPEG images at 'most appropriate' windowing as determined by radiologist. Annotations removed.	Transfer learning. Lateral wrist radiographs: 695 fracture/695 no fracture. Data amplified (non-identical copies): 5560 #/5552 no #. 80:10:10 train:validation:test with 100 kept for final test. Retrained top layer of Inception V3. Initial learning rate of 0.02, learning decay by a factor of 0.67 after every 1800 iterations.	Not explicitly stated	Final testing: 100, unseen. 50 fracture/50 no fracture. Consecutive set until fracture/no fracture number is reached.	Data amplified (non-identical copies): 5560 fracture/5552 no fracture. Incidental balanced datasets – no explicit attempts to artificially balance data. Balanced test dataset: 50 fracture 50 no fracture.
Kitamura et al, 2019. USA [36]	Five networks; Inception V3, Resnet 101 layer, Resnet (drop/aux), Xception (drop/aux). DE NOVO training. All five used together for best performance. De Novo programmed.	Resized to 300 × 300 pixels. One greyscale channel.	298 fracture/298 normal *examinations* (single or multiple views). Trained on single views: 689 abnormal views/752 normal views = 1441 total views. Augmentation of images for generalisation (rotation, flipping, brightness, contrast variation). Models trained on GEForce 1080 GTX GPU. All five models converged after 2000 epochs. Learning rate 4e-6 and 6e-6. L_2 decay rate between 0.4 and 0.9. Dropout rate kept at 0.5. Convergence of training via Softmax cross entropy loss, determined as converged when loss values plateaued (2000 epochs).	Not explicitly stated	Test and validation: 40 normal/40 abnormal with three views each: 240 total images.	Trained on single views: 689 abnormal views/ 752 normal views = 1441 total views from 298 fracture examinations/298 normal examinations. Test set intentionally balanced: 40 fracture examinations, 40 normal examinations with three views for each case.
Krogue et al, 2020. USA [37]	DenseNet 169 for fracture classification with final Softmax layer for each class RetinaNet object detection (with ResNet architecture) for bounding box detection	Resized to 224 × 224 pixels replicated into three channels.	1849 individual hip images. Data augmentation on training set. Initialised with ImageNet pretrained weights. Trained with Adam. Learning rate 0.00001, batch size 25, learning rate decay 0.9, training stopped after 10 epochs with no improvement.	Mean patient age: 74.6years 62% female in initial dataset (2004 full radiographs).	Validation set: 739 Test set: 446, including randomisation of classes for equal distribution of classes in each dataset.	Fractured: 47.9% (including subclassifications of fracture type) Unfractured: 52.1% Proportion of classes kept the same for all datasets No attempts to artificially balance classes.
Lindsay et al, 2018. USA [38]	DCNN: extension of U-net architecture	Rotation, cropping and aspect ratio: 1024 × 512 pixel	Pretraining - 100,855 other body parts. 90% of 31,490 wrists. 10% validation. Training stopped after no improvement over five epochs. Rotations, cropping, mirroring, lighting and contrast adjustments made to images to make model more robust. Two stage training: 1) bootstrapping on pre-training dataset (random initialisation of parameters) 2) Adam used. Training stopped when model performance had not improved after five epochs.	Not explicitly stated	Set 1–3500 wrist radiographs from wrist dataset. Set 2–1400 PA/ lateral wrist radiographs sampled over three-month period.	Not stated – no indication of balance of classes in either train or test datasets.
Olczak et al, 2017. Sweden [26]	Five networks chosen from Caffe library: BVLC Reference CaffeNet (8 layer), VGG CNN S Network (8 layers), VGG CNN (16 and 19 layers),	Images cropped and resized to 256 × 256 pixels.	Entire dataset: 256,458 images with 56% fracture: 70% train, 20% valid, 10% test. All networks pretrained on ImageNet and converted to Torch7.	Not explicitly stated – entire dataset 256,458, 56% fracture, 43% no fracture, 1% 'missing'.	Prediction compared with two radiologists (full view images, other views and radiologists' report) in 400 images chosen from the test set. It is unclear if the	No attempts to artificially balance classes in this very large dataset. Dataset contained 43% with no fracture, 56% with fracture and 1%

(*continued on next page*)

Table 4 (*continued*)

Author/ Country/Year	ML description/ techniques	Feature engineering detail	Training set/method	Training set demographics	Test set/method (Sample size/cases)	Class balancing
	Network-in-network (14 layers). VGG 16 layer exhibits best performance in fracture detection. Retrained for 13 epochs.		Final fully connected layer replaced with outcomes for the study. Each outcome had its own fully connected layer in parallel, using ConcatTable. Stochastic gradient descent – batch size, one. Learning rate adapted at the end of each epoch. 13 epochs in total. Best performing network used for testing.		dataset used for testing the model is larger than this.	missing this information (unclear meaning in the paper). Information regarding any attempt to balance the test set is not apparent
Tanzi et al, 2020. Italy, Sweden [39]	Three networks initially evaluated: ResNet, VGG16 and Inception for best performance: 1) Fine-tuned Inception V3 with last layer replaced with a Softmax layer (for classification). Pretrained on ImageNet 2) Multistage cascade CNNs (three Inception V3 plus binary network) for hierarchical classification discrimination.	Each individual hip joint cropped to 299 × 299 pixels	Retrospective dataset: antero-posterior cropped hip images: 1133 unbroken femurs, 570 type A, 750 type C and 4 type C (excluded due to low numbers). 80% for training and validation: 455 type A, 600 type B and 907 *broken.* Data augmentation of final dataset. *Validation* by 5-fold cross validation. Keras neural network library (in Python) on TensorFlow, Ubuntu 16.04.5 LTS with GeForce GTX 1080Ti. Initially, higher weights applied to classes with fewer images. Batch size 32, Adam optimiser, learning rate 0.0001, beta values of 09 and 0.999. Sparse categorical crossentropy used to calculate loss. The model was run for 150 epochs with early stopping patience of 10 epochs.	Median age: 81 67.5% female	20% of images from each class: 115 type A, 150 type B, 226 *broken.*	Compensation for unbalanced classes by a function applied to assign higher weight to classes with fewer images. Test dataset: 20% of images for each class: 115 type A, 150 type B and 226 unbroken, therefore retaining the prevalence from the initial dataset.
Urakawa et al, 2019. Japan [40]	VGG-16	Each individual proximal femur (femoral head + greater and lesser trochanters) cropped to 300 × 300 pixels.	Retrospective dataset: 3346 hip images (1773 fractured, 1573 not fractured). Train:validation:test: 2678:334:334. Data augmentation resulted in 132500 images for training. TensorFlow VGG-16. ImageDataGenerator used to augment 50 images per iteration. L2 regularisation (weight decay 0.001). Early stopping on validation set (not training set). Adam optimiser. Exponential learning rate scheduling: initial learning rate: 0.0001, decay steps: 265 iterations, decay rate: 0.8. Best performance at 1457 iterations. These weights used for testing.	Of initial dataset (prior to exclusion criteria applied n = 1773): 286 men, 1487 women; mean age 85 (range: 29–104)	334 cropped radiographs	Not specifically stated Training: 1408 with fracture, 1270 with no fracture. Individual hip images. Test set: 180 fracture, 154 no fracture images. No attempts to artificially balance classes.
Yu et al, 2020. USA [41]	Inception V3 pretrained on ImageNet, with: 1) top layer of the network (1000 nodes) replaced by fully connected layer (1024 nodes), terminated with final Softmax layer with two classifiers – fracture/no fracture	Manual cropping to region of interest with proximal femur centred. Pixel size of regions not stated.	Retrospective dataset: 307 fractured pelvis images: 610 normal and 451 fractured individual proximal femora. Train:validation:test: 3:1:1 Training set: 367 normal image, 111 group 1 localisation, 130 group 2 localisation and 30 group 3 localisation. 20 fold cross validation. Cross dash entropy loss	Fracture group: 151 men, 156 women. Mean age = 69.4 (range: 21–97) 'Normal' group: 155 men, 155 women. Mean age = 62 (range: 18–95) 155 right hip, 152 left hip.	20% of both fracture and no fracture groups. 20-fold cross-validation for binary classification task.	Balanced classes (***patient-wise***) for training: 60% patients in the fracture group and 60% patients in the no fracture group. One additional previous radiograph, per patient, for included patients were added to augment the dataset. Additional 'normal' included from

(*continued on next page*)

Table 4 (*continued*)

Author/ Country/Year	ML description/ techniques	Feature engineering detail	Training set/method	Training set demographics	Test set/method (Sample size/cases)	Class balancing
	(each with ReLU activation) and, 2) final Softmax layer with four nodes: Normal, and three fracture classes for localisation.		function using stochastic gradient descent optimiser. Learning rate 0.001. Learning rate decay 0.5. Batch size 8. Drop out rate 0.5. Model initialised using pre trained weights. Weights of final layer initialised using a gaussian distribution. Model trained for 200 epochs. Models converged at approximately 80 epochs.			the source (Electronic Medical Record) to up-sample 'normal' group. Training on 60% of normal and 60% of fracture – intentionally balanced dataset in a 1:1 ratio fracture:no fracture. Test data set: 123 normal; 37 subcapital; 43 intertrochanteric; 10 subtrochanteric. No data augmentation in validation and test sets

fracture, hardware from previous surgery or arthroplasty [37]. The output from one study used a region of interest (ROI) box only to predict dental root fractures [32].

3.5. Reference standard (Table 3)

Two studies determined truth from radiologist reports already available in the clinical notes or trauma registry only [28,30]. The means of verification of truth is unclear in one study [26], where the authors state that automated language extraction applied to radiologists' reports, along with 'multiple visits' (p.582). Ten studies obtained ground truth references from either consensus diagnosis from several experts in the field or verified the report accompanying the images [27,29,30,32–38]. Two studies used the opinion of one expert as ground truth; in one study by inspection of single projection radiographs and computed tomography or operative report [41], and the other by inspection of all patient images and scans [40]. One study used images from Radiopaedia®, and reference standard determination is not explicitly stated, but assumed to be diagnoses from the webpage.

Six studies also required the experts to provide ROI indication for the pathological area on the image [29,32–34,37,38].

3.6. ML description/techniques (Table 4)

All studies included used convolutional neural networks (CNNs) to achieve desired output of either fracture detection or classification with the exception of one study which reported the use of a Support Vector Machine (SVM) in addition to a CNN to delineate the iliopectineal line on pelvic radiographs [31]. SVMs are a different type of machine learning and are used usually in classification tasks [17,31] (Table 1). There was little commonality in the types of networks chosen for training on the specific tasks. Networks reported in the papers include the InceptionV3 network, which is 43 layers deep, DenseNet (121 layers), ResNet (152 layers), DetectNet and another used a U-Net model. A further three studies used a combination of CNNs to determine the best performing networks [26,36,39]. These included one study using a combination of VVG networks (with differing numbers of layers from 8 to 19 layers), Network-in-Network (14 layers) and CaffeNet (8 layer) [26] and another using Inception V3, ResNet 101 (drop/aux) and Xception models, individually and together for best performance [36]. One study used an Inception V3 network in a cascade for hierarchical multi-class discrimination [39].

3.7. Feature engineering (Table 4)

Most studies state that the images used to train and test the systems have been downsized to the dimensions required by the AI model. These ranged from 224 × 224 pixels for proximal femur [34,37] to 900 × 900 pixels for segmented regions on orthopantomographic images [32]. One study used images resized to 1024 × 512 [38].

3.8. Training set/method (Table 4)

The dataset used to educate the AI system is usually referred to as the 'training set'. The size of the training set varied considerably between studies depending on whether the AI model was trained from scratch or by transfer learning. Transfer learning is the process by which an AI system is trained on a dataset of images general dataset to the final task, for example the ImageNet database of common objects [28,34,35,41]. The parameters and initial weights are set for image recognition tasks in general and then more efficiently refined for the eventual task by exposure to a further dataset of images specific to the desired task, for example wrist radiographs. The largest dataset used for training was a CNN study that pretrained the system with 100,855 body part radiographs (foot, elbow, shoulder, knee, spine, femur, ankle, humerus, pelvis, hip, and tibia) [38]. This system was with fine-tuned and validated using 31,940 wrist radiographs, which was the focus of the study.

The study with the smallest training set determined the angulation of the iliopectineal line as a determinant of fracture [31]. A total of 75 radiographs obtained from an online radiology reference resource (Radiopaedia.org) were used to train the neural network, although it should be noted that this study was mainly investigating the use of an SVM and, therefore, not directly comparable to the other studies.

Ten studies provided demographic information on the composition of the datasets used for training in the form of patient sex and mean age, therefore allowing assessment of any potential bias present in training [27,28,30,32–34,37,39–41].

3.9. Test set/method (Table 4)

The size of the datasets used for testing were highly variable. The study by Olczak et al. [26], which included four anatomical regions, tested the AI model on 25,645 images, the highest number of test images from the included studies. The SVM study by Damien et al. [31] tested its algorithm on only 14 images. The remaining studies have test sets ranging from 100 to 3900 images. All studies used unseen test sets, except for three studies where the full dataset was used for training and testing with k-fold cross-validation, with two using five iterations of training and testing (k = 5) [32,39] and one where k = 20 [41] (for description of k-fold cross validation see Table 1).

3.10. Class balancing techniques (Table 4)

Class balancing describes the correction of the prevalence of any class in the dataset. Imbalanced classes can occur in many real-life scenarios, such as detection of fraud and disease state. This means that for any

Table 5
Performance metrics, results and explainability of the AI decision.

Author/Country/ Year	Performance metrics/results	Methods to explain ML decision	Misclassification explanation	Code availability
Badgeley et al, 2019. USA [28]	Best model: 0.78. Following removal of confounding factors Area Under Curve (AUC): 0.52.	None	Not specific, however when only image data remained the diagnostic ability of the model decreased to no better than chance.	Yes: https://github.com/mbadge/hipsMultimodal
Bluthgen et al, 2020. Switzerland [29]	Best performing model AUC: 0.95, 0.94, 0.96. Sensitivity: 86 (64–97), 90 (70–99), 90 (70–99), specificity: 97 (82–100), 90 (73–98), 97 (82–100), for AP vs lateral vs combined views respectively Region of Interest (ROI) plotted by radiologists and Machine Learning (ML); agreement if regions overlapped by 70%. Internal set (# cases n = 21): radiologist/ML agreement: model 1: 100%, 88%, and 94% and model 2: 94%, 87% and 89% (projections: antero-posterior, lateral, combined). External set: # cases (n = 50): radiologist/ML agreement: 91%, 92% and 88% (model 1) and 100%, 89% and 93% (model 2). AUC (combined views): model 1, internal dataset: 0.95; model 2, internal dataset: 0.96. Model 1 external dataset: 0.87, model 2 external dataset: 0.89. AUC 0.8 on external dataset for single AP projection using model 1.	Heat maps from deep learning system- peak activation region only and consensus region of interest (ROI) confirmation from at least two radiologists and a radiology resident. There does not seem to be any quantification of agreement between AI and radiologists and registrar, although it is stated that agreement is counted as 'correct' if there in some overlap with the radiologist/resident determined ROI containing fracture.	False negative results were 'uncommon in their extent' (p.5) or markedly displaced.	Not explicitly stated
Cheng et al, 2019. Taiwan [30]	MODEL: AUC 0.98. accuracy: 91%, sensitivity: 98%, specificity: 84%, false negative: 2%, F1: 0.916. PRIMARY PHYSICIANS: sensitivity range 84–100%, specificity range 46–94%. EXPERTS (2x radiologists, 4x orthopaedic surgeons: mean sensitivity: 99.35, mean specificity: 87.7%	Heat maps (grad-CAM). 95.9% of the class discriminative regions contained the fracture, as determined by the authors.	Heatmaps examined-two from the test set of 100 radiographs activated at wrong site but proposed rationale for this is unclear.	Not explicitly stated
Chung et al, 2018. South Korea [27]	Top 1 accuracy (i.e. predicted the correct 1 out of 5 possible options) of 96% (95% CI 94–97%). Model sensitivity 0.99 and spec 0.97. AUC 1.00 (CI 0.995–0.998) for discerning fracture from normal.	None	None	Not explicitly stated
Damien et al, 2019. Lebanon [31]	Accuracy: 92.9%. Sensitivity 80%. Specificity 99%. Support Vector Machine (SVM): Accuracy: 91.3%. Sensitivity: 94.2%. Specificity: 87.5%	None	None	Not explicitly stated
Fukuda et al, 2019. Japan [32]	ML ROI - taken as correct if 'sufficiently include the root of the tooth #'. Recall: 0.75, precision (positive predictive value): 0.93, F measure (2 (recall + precision)/(recall + precision)): 0.83, expressed as MEAN of the 5 models.	Region of interest boxes around tooth with vertical root fracture	Yes - potential explanation given - teeth without endodontic treatment were misclassified in 58.3% of misclassified cases. Recall rates were low for maxillary incisors.	Not explicitly stated
Gan et al, 2019. China [33]	Identification of region of interest (distal radius) by Faster R–CNN: 'Intersection of the union' (area of overlap/area of union) average = 0.87. Accuracy for fracture identification: Inception V4 (IV4): 93%; Orthopaedists (O): 94%; Radiologists (R): 84%. Sensitivity for fracture identification: IV4: 90%; O: 93%; R: 81%. Specificity for fracture identification: IV4: 96%; O: 95%; R: 87%. Youden index: IV4: 0.86; O: 0.87; R: 0.68.	None (except identification of the distal radius region of interest by the Faster R–CNN – 100% success rate)	Yes, the 15 images which did not detect a confirmed fracture were reviewed. Five lacked the usual fracture traits (fracture lines and fragments) and the fracture was only apparent on the corresponding lateral radiographs.	Not explicitly stated
Jiménez-Sanchez et al (2019), Spain, Germany and France [34]	AlexNet – identification of region of interest 100%. ResNet-50 performance: Accuracy; Precision; Recall and F1 score (in %) listed respectively: Full radiographs: 83%, 78%, 83%, 84%. *Manual localisation* (regions of interest provided by experts): 93%, 93%, 94%, 94%.	Regions of interest for fracture prediction were examined for binary prediction and discrimination tasks, with 93.82 and 88.35% agreement respectively.	No specific explanation for misclassifications offered.	Not explicitly stated

(*continued on next page*)

Table 5 (*continued*)

Author/Country/Year	Performance metrics/results	Methods to explain ML decision	Misclassification explanation	Code availability
	AUC for fracture detection 0.9807; for classification: 0.9475 on manual ROI. *Automatic localisation* (AlexNet): 93%, 94%, 93%, 93%.			
Kim and McKinnon, 2018. UK [35]	AUC (ROC) 0.954. Sens: 0.954, spec: 0.88. ROC on ML vs. verified report	None	None	Not explicitly stated
Kitamura et al, 2019. USA [36]	Best (all 5 models developed used together): accuracy: 0.81, sensitivity: 0.80, specificity: 0.830, positive predictive value (PPV): 0.82, negative predictive value (NPV): 0.81.	None	None	Yes – available on the corresponding author's GitHub (for convenience: https://github.com/GeneKitamura)
Krogue et al, 2020. USA [37]	Binary accuracy, sensitivity, specificity, AUC and Cohen's kappa: 93.8%, 92.7%, 95.0%, 0.973, 0.877 Multiclass accuracy: 90.4% over all classes. Cohen's kappa: 0.862. Multiclass sensitivity, specificity and AUC: No fracture: 94.5%, 92.6%, 0.972 Intertrochanteric fracture: 93.3%, 96.9%, 0.984 Femoral neck fracture (displaced): 87.5%, 98.9%, 0.991 Femoral neck (nondisplaced): 46.2%, 97.8%. 0.868 Arthroplasty: 96.9%, 100%, 1.00 Open reduction, internal fixation: 100%, 100%, 1.00 Hip region detection by RetinaNet in all images with intersection-over-union (ratio of overlap:combined area) of 0.92 with manually labelled regions. No statistical difference in binary and multiclass fracture detection was reported between manual and automatically generated bounding boxes. Human observers (model-quality images) v. model: model performed statistically significantly better. Human observers (full quality images) v. model: Model performed better, but only statistically significantly better in the 'resident' group.	Heat maps: found to "indicate high importance to cortical outlines" p.8 (ARXIV document)	Explanation of fracture type (multiclass) misclassifications-if misclassified the model usually predicted some other fracture type. Localisation errors in six radiographs where the hip was only partially contained in image.	Not explicitly stated
Lindsay et al, 2018. USA [38]	Set 1 - AUC 0.967. Set 2 - AUC 0.975. Model used to determine effect of ML on non-specialist clinicians (ED clinicians (MD)) and physician assistants (PA)). A dataset of 266 radiographs used: clinicians before and after model predictions respectively: MDs: sensitivity from 82.7% to 92.5%; specificity from 87.4% to 94.1%. PAs: sensitivity from 78% to 89.9%; specificity from 87.5% to 93.6%. Model average: sensitivity 93.9%; specificity 94.5%. Qualitatively, the model was 'generally able' to locate fracture in the same location as the subspecialist orthopaedic surgeons.	Heatmaps (dense conditional probability map)	None	Not explicitly stated
Olczak et al, 2017. Sweden [26]	Fracture detection accuracy 83% (95% CI: 80–87%) for best performing network, (VGG-16 layer).	None	Manual review of misclassification: fracture visible on another examination in the series (not on tested image).	Not explicitly stated
Tanzi et al, 2020. Italy, Sweden [39]	Using **three** Inception V3 networks: **Accuracy** (over five-folds): Broken/unbroken: 0.91 A/B: 0.87 A1/2/3: 0.61 Average accuracy (for three classes): 0.86 Average accuracy (for five classes): 0.80 **Addition of further training for last two networks with A1,2 and 3**	Grad-CAM heat maps: differentiation of focus for type A and type B fractures identified.	Inspection of poor performing discriminations by specialists identified issues with discrimination of A1 and A2 fracture – additional training and binary network to improve performance as described under 'performance metrics/results' heading.	Not explicitly stated

(continued on next page)

Table 5 (*continued*)

Author/Country/ Year	Performance metrics/results	Methods to explain ML decision	Misclassification explanation	Code availability
	fracture training set + additional binary network (optimal performance) (precision, recall, F1 score, respectively): Unbroken:0.93, 0.90, 0.91 B: 0.85, 0.83, 0.84 A1: 0.49, 0.54, 0.51 A2: 0.5, 0.55, 0.51 A3: 0.73, 0.73, 0.73			
Urakawa et al, 2019. Japan [40]	VGG-16 was compared with five orthopaedic surgeons on 334 cropped images: Accuracy, sensitivity, specificity and AUC respectively: VGG-16: 95.5, 93.9, 97.4, 0.984 Orthopaedic surgeons: 92.2, 88.3, 96.8, 0.969	None stated	None stated	Not explicitly stated
Yu et al, 2020. USA [41]	Binary classification sensitivity, specificity, accuracy and AUC: 97.1%, 96.7%, 96.9%, 0.9944 Multiclass classification sensitivity, specificity and accuracy for: Normal: 95.8%, 94.3%, not stated Subcapital: 84.1%, 92.8%, 91.3% Intertrochanteric: 76.8%, 94.5%, 90.9% Subtrochanteric: 20%, 99.1%, 95.4% Binary classification *(human readers: MSK radiologists and radiology residents)* sensitivity, specificity, accuracy: 100%, 98.4%, 99.2% Multiclass classification *(human readers: MSK radiologists and radiology residents)* sensitivity, specificity and accuracy for: Subcapital: 83.1%, 99%, 95.5% Intertrochanteric: 97%, 92.9%, 93.9% Subtrochanteric: 66.7%, 100%, 98.5%	Activation maps (heat maps)	All heat maps agreed with ground truth with the exception of the subtrochanteric classification, suggested due to the training set not being large enough to cover all fracture morphologies.	Not explicitly stated

dataset gathered in these cases, there is likely to be a majority and minority class. This is true in fracture identification and many other medical imaging cases. Training an ML on imbalanced datasets will result in the model being biased to the majority class. This is obviously highly undesirable in medical imaging, where misclassification of a positive case will have significant consequences.

Class balancing techniques can be adopted to ensure there are equal numbers of images in each prediction class for training. This is important when training the algorithm so that the AI system can equally learn the patterns in each class equally and learn to discriminate. There are a number of methods to correct class imbalance. Data scientists can often intentionally under sample the majority class, apply weights to the algorithm to penalise the majority class or artificially up-sample the underrepresented class by creating synthetic cases using techniques such as Synthetic Minority Oversampling Technique (SMOTE).

There were limited attempts to artificially balance classes.

Of the papers included in this review, only one study reported the used of perfectly balanced classes, examination-wise, for training, i.e., the number of images in each class for training were not perfectly balanced (689 fracture, 752 no fracture apparent) [36]. Classes were balanced intentionally by perusal of radiology reports to achieve balance.

The greatest discrepancy between classes was reported in a study where the proposed model was trained using a dataset with only 3% images in the fracture class [28], although the identification of fracture was only one focus of this large study. Interestingly, this study tested the model on both balanced and imbalanced datasets and reported a significantly higher area under the precision-recall curve for the balanced dataset, therefore indicating that the model is able to correctly detect the fracture class better in the balanced dataset. The remainder of the studies had more equally balanced classes, ranging from 31.7% fracture [29] of a small training set, n = 166, to one study with equally balanced classes for training [36].

In five studies the fracture class was greater than the no fracture class [28,29,32,33,40], although there were five classification classes in one of these studies [27]. Cases across the five classification categories in this study were balanced: 346, 514, 269, 247 and 515 for greater tuberosity, surgical neck, three-part fractures, four-part fractures, and no fracture classes respectively. Non-fracture classes were removed for specific training in classification of fracture severity in this study [27]. One study used a compensation mechanism for training a dataset with unbalanced classes by assigning greater weight to the lesser-represented group [39]. One study balanced classes by patient pathology in a hip fracture study, but as individual hips were isolated for compilation of the final dataset, this actually resulted in imbalanced classes [41]. The study describes how further 'normal' cases were then intentionally identified from the Electronic Medical Record to increase the minority class, which in this study was the 'normal' class.

Class balancing the test set is also harmless, as metrics such as sensitivity and specificity are 'prevalence agnostic', however metrics such as accuracy are biased to disease prevalence (the dominant class - referred to as the 'accuracy paradox'). It could be argued, however, that it would be helpful for the ML to be tested on a dataset replicating the clinical scenario, where there are likely to be imbalanced classes to gain true understanding of the model performance. Reporting metrics should be chosen carefully to give an accurate measure of the performance of the ML on imbalanced datasets. This is further discussed in section 5.

Of the studies included in this review, five studies used intentionally balanced datasets for testing [29,30,33,34,36]. Only one study had very imbalanced test dataset [28]. One study used a balanced, external dataset (MURA) to test the generalisability of the model [29]. Another study used a prospective sample from the clinical environment, although these examinations were obtained from the same hospital as the training images and there is no information on the balance of classes in either the testing or training datasets [38].

3.11. Performance metrics/results (Table 5)

The oldest study [26] compared the performance of 5 networks and found a VGG-16-layer network to have the highest accuracy of 83% (95% CI 80–87%). Three studies were published in 2018 [27,35,38] and each used different CNNs for different anatomical areas. One study reported top-1 accuracy, which represents the ability of the AI to select the correct classification from a number of available options. In this case, five classification options for proximal humeral fracture were presented and the AI was able to correctly classify in 96% of cases [27]. The remaining two studies published in 2018 focussed on detection of wrist fractures using different network architectures [35,38]. Both studies reported performance by area under the receiver operating characteristic curve (AUC), sensitivity and specificity. Both reported AUC exceeding 0.95, sensitivities of 0.939 [38] and 0.954 [35] and specificities of 0.945 [38] and 0.88 [35].

Five studies published in 2019 focused on determining hip or pelvis fractures from pelvic radiographs [28,30,31,34,40]. One study used a DenseNet 121-layer network to determine and characterise proximal femur fractures with three prediction classes, including normal and reported AUC of 0.98, accuracy of 91%, sensitivity and specificity of 98% and 84% respectively and F_1 score of 0.916 [30]. Some of the same metrics were used to report the results from a study using a VGG16 model to predict hip fractures with AUC, accuracy, sensitivity and specificity reported as 0.984, 95.5%, 93.9% and 97.4% respectively, although F_1 score was not used as a reporting metric study [40]. Another 2019 study described a ResNet50 model which was trained to detect hip fractures on cropped images, with regions delineated both manually by an expert, and automatically by an AlexNet model. Results reported indicated that the model performed equally well on both sets of cropped images with accuracy, precision, recall and F_1 score for the manually cropped images of 93%, 93%, 94%, 94% and automatically localised images of 93%, 94%, 93% and 93% [34]. In the same year, another study reported less positive results. In this study an Inception V3 network was used to determine proximal femoral fracture in a two-class problem (fracture/no fracture) and found that AUC dropped from 0.78 to 0.52 when all ‘confounding variables’ were removed from the images [28]. This was despite the study using a pretrained network which was retrained on a dataset of over 20,000 pelvis radiographs. However, more promising results were reported using an Inception V4 network on a different anatomical area for binary classification of fractures on cropped radiographs of the distal radius. Sensitivity, specificity and accuracy were reported as 90%, 96% and 93% respectively [33].

A further study published in 2019 adopted a different methodology to quantifyiliopectineal line disruption to determine fracture using an SVM and CNN as a classifier to determine fracture with reported accuracy, sensitivity and specificity of 92.9%, 80% and 99% respectively [31]. However, detail of the neural network used in this study is not stated.

The most recent studies have reported promising results using a range of models: a ViDi v.2 manufacturing CNN [29], DenseNet 169 [37], and two studies reported results using an Inception V3 model [39,41], although one study maximised the results by using the model in cascade with an additional binary network for further discrimination between classes [39]. Three of the four studies reported area under the receiver operating curve as a performance metric with results for binary classification [29,37,41] ranging from 0.80, on an external dataset of wrist radiographs [29] to 0.994 in a study using an Inception V3 to predict hip fractures using regions of interest cropped by experts [41]. The remaining study reported accuracy, precision, recall and F_1 scores for comparable binary tasks [39], detailed in full in Table 5.

3.12. Methods to explain ML decision (Table 5)

Eight studies reported some method of AI explanation: six studies by heatmap [29,30,37–39,41] and two studies by a region of interest (ROI) bounding box [32,34] with high agreement in all cases.

3.13. Misclassification explanation (Table 5)

Eight studies make some attempt to offer explanation for misclassifications [26,29,30,32,33,37,39,41]. Full detail is presented in Table 5. One study investigated the effect of the removal of ‘confounding variables’, such as those variables relating to the patient and ‘hospital process’, for example, patient age, sex and body mass and scanner type, scanner model, scan priority and time of day of the scan, from hip radiographs. They found that when these confounding factors are removed, the performance of the AI dropped from AUC 0.78 AI performed no better than chance (AUC 0.52) [28]. A study using an AI model to identify teeth with vertical root fractures reported that the model misclassified more often in teeth which have no endodontic treatment and that recall rates were low for maxillary incisors, although an explanation for this is not offered [32]. In another study, the misclassified images were examined, along with other images in the imaging series, and it was discovered that when the AI found an image to incorrectly contain fracture, the fracture may have been evident in another image in the series [26]. One study reported that the AI misclassified on two images from the test set of 100 images by inspection of heatmaps produced but an explanation for this is not proffered [30]. Studies also reported a lack of ability of the AI to discriminate between fracture subclasses [37,39], and misclassification due to the usual fracture traits not being visible on the particular projection presented to the AI [33].

3.14. Code availability

Only two studies made their code available to the reader [28,36]. The availability of code, along with transparent experimental methodology is essential to be able to replicate the study and to test model generalisability on other datasets.

3.15. Clinical integration and prospective sampling

No studies have been integrated into the clinical workflow for testing. Three studies used a k-fold cross validation method, as described previously, for training and testing [32,39,41]. All studies, except for those already mentioned using k-fold cross validation, used entirely unseen test sets, taken from the entire dataset before training. One study compared the model performance on an unseen internal and external dataset in testing [29] and only one study obtained a prospective sample over a three-month period in testing [38]. In this study, images were acquired from a set date onward, rather in retrospect from the hospital database. This study found that there was little difference in the model’s ability to detect fracture on a test set retrained from the training set and a prospective sample with AUC of 0.97 and 0.98, respectively.

It is clear from these findings that many variations exist in both the systems being used, the training and validation methods and the process by which data from these studies are articulated.

4. Discussion

Reported results demonstrate that machine learning based on artificial neural networks can detect fractures from radiographic images with impressive accuracies. Studies included in this review indicate that this is achieved using a variety of AI model types and training/testing methods. Studies included also varied in the methods to determine a reference standard for the images used for both training and testing.

Each study reported the model performance using some combination of AUC, accuracy, sensitivity, specificity, precision, recall, Cohen’s kappa and F_1 score (see Table 1). The most commonly described metric was AUC, with only four studies not reporting some AUC results [26,31,32,39]. AUC and Receiver Operating Characteristics (ROC) are metrics commonly used to assess the performance of ML systems and other classification tasks.

It is imperative that reporting metrics used to report ML performance should be explainable, understood by the end-user and should be appropriate to the task to accurately reflect the performance of the model. When the classes in the training dataset and the prevalence of the outcomes in the eventual population dataset is balanced, reporting metrics which may already be familiar to clinicians can be used to evidence model performance, for example, sensitivity, specificity, and accuracy. This is not usually the case in pathology identification and in many medical applications. The disease class is usually the minority class. Misclassification of this class would obviously be very undesirable and the choice of a model which was unable to detect pathology would be useless. If standard reporting metrics were used it would be possible to report a high accuracy for a model which had a propensity to predict all ‘no pathology’ (majority class) outcomes, which would therefore be highly specific but essentially not fit for purpose.

As discussed, some studies trained and tested the algorithm on balanced, or almost balanced datasets. This is an ideal situation in training, as the model will ‘learn’ to identify both classes equally, however, when the model is eventually applied to the clinical setting it will have to perform well on a naturally imbalanced dataset. The reported accuracy, sensitivity and specificity metrics used to report the model performance are an indication of how well the model performs in the laboratory only. One study tested their algorithm on a prospective clinical dataset, reporting accuracy, sensitivity, and specificity but there was no indication of the balance of classes in this test dataset, therefore these metrics may not permit full assessment of the model performance.

In clinical ML tasks, where there is likely to be a majority and minority class, it is imperative to report findings using metrics which incorporate allowances for the imbalanced prevalence of the target population to give an accurate representation of the ML performance and for comparison between different models for the same task. For this purpose, precision, recall, F_β and AUC have been recommended in the literature [42,43].

Precision, recall and F_β incorporate true positive predictions, where the ML predicted pathology in agreement with the reference standard; true negative predictions, where the ML predicted that there was no pathology in agreement with the reference standard; false positive predictions where the ML predicted pathology where the reference standard did not, and false negative, where the ML predicted no pathology, where the reference standard indicated that there was pathology.

Recall (or sensitivity) describes the ability of the model to correctly predict the presence of pathology and is calculated by the following equation [32]:

$$\text{Recall} = \text{TP} / \text{TP} + \text{FN}$$

Precision, or positive predictive value (PPV) can be used to report the ability of the model to identify pathology as a proportion of all positives i.e., it is an indication of how many positive predications were actually positive, therefore giving an indication of the number of disease cases which were misinterpreted.

$$\text{Precision} = \text{TP} / \text{TP} + \text{FP}$$

From these metrics, F_β can be calculated as a single measure to represent the model performance. F_β is simply the harmonic mean of precision and recall [30,43]. The value of β will determine the weighting of recall in the calculation. For tasks such as those used in pathology identification, where it is important for the model to be able to identify both the presence and absence of pathology correctly, a score of one is used.

$$F_1 = 2\ (\text{precision x recall} / \text{precision} + \text{recall})$$

These metrics provide an interpretable overview of the overall performance of the model and are useable in all scenarios as these metrics are prevalence agnostic, however they are based on the use of prediction classes, rather than the more usual prediction score output provided by ML systems. A suitable threshold value to provide a positive prediction class needs to be decided to provide this. Additional information may therefore be gained by the use of a metric capable of analysing full prediction scores. These scores can be plotted in a Receiver Operating Characteristic curve (ROC). Inspection of this graph will allow the best choice of threshold value for determining prediction class to be chosen by determining acceptable balance between sensitivity and specificity for the specific task. The area under the ROC curve (AUC) allows for direct comparison of different models (or choice of parameters) in a single metric, which is suitable for use in moderately imbalanced datasets. Reporting of F_1 and AUC as a minimum will provide simple, comparable single metrics which will be interpretable by clinicians and data analysts alike, providing accurate reporting of the performance of the model with both balanced and imbalanced datasets and therefore improve confidence in critique of proposed models as they are presented in the clinical setting. Three recent studies, two published in 2019 [32,34], and one in 2020 [39], reported the performance of their model using precision, recall and F_1 [32,34]. One study reported F_1 only, along with sensitivity, specificity, accuracy, and AUC [30].

Cohen’s kappa (see Table 1) has also been proposed in some studies to provide a measure of inter-rater agreement and will give an indication of the agreement of the model prediction and the reference standard, although has not been extensively used in the included studies.

As mentioned, most studies reported the model performance using AUC. The best performing model which reported performance using AUC was in a study using AI to predict hip fractures. The authors (Yu et al., 2020) quoted performances of 0.99 for a binary classification task using an Inception V3 model trained, validated and tested by 20-fold cross validation [41]. The training set used in this study was balanced, patient wise, for fracture/no fracture by intentional oversampling of the minority class. The test set was not augmented. The determination of the reference standard in this study was by computed tomography and/or an operative report, therefore providing additional information than given by a report on the plain radiographic images alone. Regions of interest (individual hips) were manually cropped by experts prior to interpretation by the AI. It should be noted, however, that part of this study involved a multi-class discrimination task with less promising results reported (Table 5). Heatmaps provided confirmation of the area of the image the AI deemed most important in determining its prediction, therefore adding to the reliability of these diagnoses.

Lindsay et al. [38] tested their model on both a proportion of the initial dataset, and a prospective sample of all wrist radiographs acquired from the same clinical setting from which the training set was acquired, although no information is given on the balance of classes in this dataset or its similarity to the training dataset. The authors reported performance using sensitivity, specificity and AUC which give an indication of the overall model performance, although may not give an overall impression if the test dataset was heavily imbalanced.

One study reported results which were in contrast with other included studies. Badgeley et al. [28] also used an Inception V3 model to predict fracture on pelvic radiographs before and after removal of ‘confounding factors’, described in section 3.13. The model performed well (AUC 0.78) on a dataset of 23,602 whole pelvis radiographs with a 3:1 training:test split, yet, following removal of patient trait details, scanner type and “other factors” the diagnostic accuracy dropped dramatically, and the system performed no better than chance (AUC 0.52). It should be noted, however, that despite this study using a large training dataset, there were no class balancing attempts. The incidence of fracture in the entire dataset was 3% (n = 779) and labels were inferred from the patient’s clinical notes, which the authors acknowledge as a limitation of the study.

The quality of labels, or reference standards, used for training are of paramount importance (Table 4). A system will only ever perform to the standard of the ground truth label that it is trained with [35,38]. In ten studies the reference standard was obtained from more than one clinician with experience in their field or by expert verification of an established diagnosis [27,29,32–39]. For example, in the study by Chung et al. [27], two subspecialised shoulder orthopaedists and one specialist musculoskeletal radiologist labelled the images. Additional information from other modalities was applied when the reports did not concur to achieve a match. In the dental study by Fukuda et al. [32], oral and maxillofacial radiologists provided a region of interest around any fractured teeth on orthopantomographic images. However, there are some studies where the diagnosis is taken from single radiologist report made at the time of the examination [26,28,30]. This offers no indication of the reliability of the report provided, particularly as reports are usually generated in response to a clinical question and additional information from the image may be missed, although in one of these studies, other imaging and clinical course were investigated in equivocal cases [30]. A system trained on images labelled by multiple experts and determining diagnosis from differing sources and eventual patient outcome should, in theory, perform best on unseen images, although this can only be assessed when the training methodologies are comparable. It is proposed that there are limitations in even the best human generated reference standard as the model may be able to detect more subtle indicators from the images which are imperceptible to the human eye [35]. The model with best performance reported from these studies used diagnosis from initial imaging, verified by a musculo-skeletal radiologist, following review of additional imaging or operative report, therefore confirming the initial diagnosis [41].

In order for an AI model to be useful in the clinical setting, the model must have been exposed to sufficient inputs from different x-ray equipment, clinical setting, devices and acquisition techniques. All reported models were tested on unseen datasets or by k-fold cross-validation, but in many of the studies reviewed, the training and testing images were obtained from the same hospital, which calls into question the capabilities of the model to be generalisable to any clinical setting. To investigate this, one study used an external dataset (MURA) to test its model and found that it did not perform as well on this dataset as on the internal dataset, where images from the same hospital as the training set was used, as noted in Table 3 [29].

Despite studies reporting impressive performances and transparent methodologies, AI systems using neural networks are approached with caution [44,45]. This is due, in part, to a lack of clarity in how the system determines its diagnosis and any failures being incomprehensible to clinician end-users and ML experts alike, due to the complexity and size of the parameters in the algorithm [47]. Most studies did not make any attempt to offer explanation for any misclassifications of the AI models used, although a number of studies used heatmaps, as described in section 3.12 and 3.13, to visually represent the region the model used to form its prediction (Table 5) [29,30,37–39,41]. Of these studies, all stated that the heatmaps demonstrated the model's agreement with the fracture region determined as 'ground truth'. These system augmentations can affect how the human engages with and trusts the machine. This can be called 'human-computer interaction'. The end-users of such systems, clinicians, need to be comfortable with their interaction and with the functionality of these machines. This is particularly important when using the most modern types of ML, as described in this review. The need for 'interpretable' and explainable AI (XAI) has driven the development of means to provide the user with interfaces which provide information on how the system has determined its predictions [47]. Visual representations in the form of 'saliency maps', 'heat maps' and other novel visualisation methods [45,48] are one way of gaining insight into the rationale for the decision, by highlighting the pixels which the algorithm found most important. Through the use of heat maps and other forms of explainable AI feedback, our interaction with these systems will hopefully become more natural and acceptable, even to the non-expert [49].

From this review it is clear that without standardisation of both reporting metrics, benchmark datasets and high-quality labels, an assessment of the best performing variables, such as training methods, ground truth determinations and AI model types and architectures cannot take place. One study tested their model on an open access dataset [29], however, no studies used any open access datasets for the training of their models. There remains a dearth of large, publicly available datasets for use on training AI, in large part due patient privacy and permission concerns. The use of high-quality datasets, with reliable reference standards will eliminate bias introduced by the acquisition of data from one clinical centre and allow for accurate comparison of the models [50,51]. To the authors knowledge, there is only one publicly available dataset for plain musculoskeletal radiographs (MURA) [14].

Clarity regarding the predicted performance of the models in situations mimicking the 'real world' scenario using simple, reliable reporting metrics along with end-user acceptable feedback and explanation will assist in allocation of appropriate trust and implementation of these systems into useful clinical application.

The availability of code and transparent reporting of methodologies used to train, validate and test the datasets, including specifics of hardware, system and network requirements are essential to replicate the studies in different settings and therefore permit the testing of the validity and generalisability of the models [50,51].

5. Limitations and strengths of this review

Due to the wide variability of methodologies and performance metrics reported, a full systematic review and meta-analysis could not be carried out, as the authors had initially intended. Many papers made it through the initial search, leaving 2563 papers for inspection by the authors. This demonstrates that the search criteria may have been too broad and there is the potential that human fatigue would result in important papers being missed, although this is not thought to be the case. However, automation of the process of extraction of relevant studies could be useful when large numbers of studies are identified for review and, in particular, in studies not limited to one area of practice, such as this one.

One study [52] was not available for inclusion in this study, due to institutional restrictions and limited access to the British Library resources during the Covid-19 pandemic.

The team working on this review collectively bring many years of research experience from differing backgrounds in both clinical and academic research in medical imaging and radiography, health science and healthcare informatics.

This review is conducted through the lens of clinical applicability of AI systems with insight into the computer science principles behind AI systems development.

6. Recommendations

A larger-scale review should be conducted to establish the state-of-the-art of AI systems used for fracture identification in all relevant radiographic imaging modalities, for example, including, but not limited to computed tomography, magnetic resonance imaging and nuclear medicine.

A further review using the literature described here, with particular focus on programming specifics may of additional use to developers of AI systems for fracture detection purposes, with the inclusion of the omitted study described above [52].

Only one study [26] investigated the ability of a range of networks to identify fracture on multiple anatomical areas. The authors, however, do not report any findings suggesting a correlation between the performance of a particular network and anatomical area. Future studies investigating this and identifying any networks which may perform better on specific anatomical areas/regions, would be useful in directing efficient development of anatomy-specific AI systems.

Further studies should investigate if different AI models or specific modifications to existing AI models would detect different types or locations of fracture for example, the study by Tanzi et al. (2020) [39] modified a cascade of three Inception V3 models by the addition of a binary network to better discriminate between two classes which the model was unable to discern.

Many of the studies reviewed here used re-sized images. Research should be undertaken to investigate the effect of using full scale images for AI interpretation as this would more accurately replicate the clinical situation (as per recommendations by Krogue et al. [37]).

7. Conclusion

As medical AI systems develop, the need to assess the impact in the clinical setting is of paramount importance due to the low level of error tolerance in this setting. The need to further develop systems to integrate into the radiology workflow should be the focus of further studies. This cannot begin until the 'best' systems to use and methods of testing are transparent. Analysis of the systems currently being produced will allow focussed research and development. This is not possible without a standardised system of reporting, permitting assessment of the performance of models currently being developed. Standardised reporting of all aspects of the study (based on, for example, the CLAIM checklist [24]) with transparent methodologies, code availability and understandable, appropriate and uniform reporting metrics will permit study replication, robust systematic reviews and meta-analyses. This may enhance the trust of the end users of these systems to and provide more focussed direction for development of clinically useful systems.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Clare Rainey: Conceptualization, Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing, Writing – original draft, Data curation, Formal analysis, Validation. **Jonathan McConnell:** Conceptualization, Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing. **Ciara Hughes:** Conceptualization, Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing. **Raymond Bond:** Conceptualization, Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing. **Sonyia McFadden:** Conceptualization, Investigation, Methodology, Project administration, Resources, Supervision, 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] Care Quality Commission. Radiology review: a national review of radiology reporting within the NHS in England. Available at: https://www.cqc.org.uk/sites/default/files/20180718-radiology-reporting-review-report-final-for-web.pdf; 2018. accessed 24th June 2020.

[2] NHS. Diagnostic imaging dataset statistical release. Available at: https://www.england.nhs.uk/statistics/wp-content/uploads/sites/2/2020/07/Provisional-Monthly-Diagnostic-Imaging-Dataset-Statistics-2020-07-23.pdf; 2020. accessed 2nd July 2020.

[3] Stephens S. Reform health conference: unlocking the promise of digital health. London, 5th June 2019. Available at: https://reform.uk/events/health-conference-unlocking-promise-digital-health. [Accessed 10 June 2019].

[4] Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine learning for medical imaging. Radiographics 2017;37(2):505–15.

[5] Liew C. The future of radiology segmented with artificial intelligence: a strategy for success. Eur J Radiol 2018;102:152–6.

[6] SECTRA. The radiologists' handbook for future excellence: 2020. Available at: https://medical.sectra.com/resources/the-radiologists-handbook-for-future-excellence-2020/; 2020. accessed: 3rd June 2020.

[7] NHS. The Topol Review: preparing the healthcare workforce to deliver the digital future. Available at: https://www.hee.nhs.uk/our-work/topol-review; 2019. accessed 24th June 2020.

[8] Lodwick G, Keats TE, Dorst JP. The coding of roentgen images for computer analysis as applied to lung cancer. Radiology 1963;81(2) [online] Available: https://pubs.rsna.org/doi/10.1148/81.2.185. Accessed 15th June 2019.

[9] Savadjiev P, Chong J, Dohan A, Vakalopoulou M, Reinhold C, Paragios N, Gallix B. Demystification of AI-driven medical image interpretation: past, present and future. Eur Radiol 2019;29(3):1616–24.

[10] Wong SH, Al-Hasani H, Alam Z, Alam A. Artificial intelligence in radiology: how will we be affected? Eur Radiol 2019;29(1):141–3.

[11] Hirchmann A, Cyriac J, Stieltjes B, Richiardi J, Omoumi P. Artificial intelligence in musculoskeletal imaging: review of current literature, challenges and trends. Semin Muscoskel Radiol 2019;23:304–11. 03.

[12] Shen D, Wu G, Heung-Il Suk. Deep learning in medical image analysis. Annu Rev Biomed Eng 2017;19:221–48.

[13] Chan HP, Samala RK, Hadjiiski LM, Zhou C. Deep learning in medical image analysis. In: Lee G, Fujita H, editors. Deep learning in medical image analysis. Springer Nature Switzerland; 2020. p. 3–21.

[14] Rajpurkar P, Irvin J, Bagul A, Ding D, Duan T, Mehta H, Yang B, Zhu K, Laird D, Ball RL, Langlotz C, Shpanskaya K, Lungren M, Ng AY. MURA: large dataset for abnormality detection in musculoskeletal radiographs. Available at: https://arxiv.org/abs/1712.06957; 2018. accessed: 15th May 2020.

[15] Anwar SM, Majid M, Qayyum A, Awais M, Alnowami M, Khan MK. Medical image analysis using convolutional neural networks: a review. J Med Syst 2018;42(11):226–39.

[16] Tang A, Tam R, Cadrin-Chenevert A, Guest W, Chong J, Barfett J, Cheplev L, Cairns R, Mitchell R, Cicero M, Poudrette MG, Jaremko JL, Reinhold C, Gallix B, Gray B, Geis R. Canadian association of radiologists white paper on artificial intelligence in radiology. Can Assoc Radiol J 2018;69:120–35.

[17] Erickson BJ. Deep learning and machine learning in imaging: basic principles (Chapter 4). In: Ranschaert ER, editor. Artificial intelligence in medical imaging. Springer Nature Switzerland; 2019. p. 39–46.

[18] Kohli M, Prevedello LM, Filice RW, Geis JR. Implementing machine learning in radiology practice and research. Am J Roentgenol 2017;208(4):754–60.

[19] HCPC. Proposed changes to the HCPC standards of proficiency (radiographers). Available at: https://www.hcpc-uk.org/globalassets/consultations/2020/standards-of-proficiency/radiographers/table-of-proposed-changes—radiographers.pdf; 2020. accessed 23rd June 2020.

[20] International Society of Radiographers and Radiological Technologists and the European Federation of Radiographer Societies. Artificial intelligence and the radiographer/radiological technologist profession: a joint statement of the international society of radiographers and radiological technologists and the European federation of radiographer societies. Radiography 2020;26:93–5.

[21] Geis JR, Brady A, Wu CC, Spencer J, Ranshaert E, Jaremko JL, Langer SG, Kitts AB, Birch J, Shields WF, van den Hoven van Genderen R, Kotter E, Gichoya JW, Cook TS, Morgan MB, Tang A, Safdar NM, Kohli M. Ethics of artificial intelligence in radiology: a summary of the joint European and North American multi-society statement. J Am Coll Radiol 2019;293(2):1–6.

[22] Jennison T, Brinsden M. Fracture admission trends in England over a ten-year period. Ann R Coll Surg Engl 2019;101:208–14.

[23] NHS. Diagnostic imaging dataset statistical release. Available: https://www.england.nhs.uk/statistics/wp-content/uploads/sites/2/2019/08/provisional-monthly-diagnostic-imaging-dataset-statistics-2019-08-22.pdf; 2019. Accessed 15th June 2019.

[24] Mongan J, Moy L, Khan CE. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. Radiology: Artif Intell 2020;2(2):1–6. https://pubs.rsna.org/doi/pdf/10.1148/ryai.2020200029.

[25] Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HCW. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. BMJ Br Med J (Clin Res Ed) 2003;326(7379):41.

[26] Olczak J, Fahlberg N, Maki A, Ali SR, Jilert A, Stark A, Sköldenberg O, Gordon M. Artificial intelligence for analyzing orthopedic trauma radiographs: deep learning algorithms—are they on par with humans for diagnosing fractures? Acta Orthop 2017;(6):581.

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

[28] Badgeley MA, Zech JR, Oakden-Rayner L, Glicksberg BS, Liu M, Gale W, McConnell MV, Percha B, Snyder TM, Dudley JT. Deep learning predicts hip fracture using confounding patient and healthcare variables. NPJ Digital Medicine 2018;31.

[29] Blüthgen C, Becker AS, Vittoria de Martini I, Meier A, Martini K, Frauenfelder T. Detection and localization of distal radius fractures: deep learning system versus radiologists. Eur J Radiol 2020;126:108925.

[30] Cheng CT, Chou C, Liao CH, Chung IF, Ho T, Lee T, Chang C, Chen C. Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. Eur Radiol 2019;29(10):5469–77.

[31] Damien P, Nader RB, Yaacoub C, Lahoud J. Iliopectineal line fracture detection for computer-aided acetabular fracture classification. In: 2019 ninth international conference on image processing theory, tools and applications (IPTA), image processing theory, tools and applications (IPTA), 2019 ninth international conference; 2019. p. 1–5.

[32] Fukuda M, Inamoto K, Shibata N, Ariji Y, Yanashita Y, Kutsuna S, Nakata K, Katsumata A, Fujita H, Ariji E. Evaluation of an artificial intelligence system for detecting vertical root fracture on panoramic radiography. Oral Radiol 2019 (online) [available at: https://link.springer.com/article/10.1007/s11282-019-00409-x. accessed 23rd June 2020.

[33] Gan K, Xu D, Lin Y, Shen Y, Zhang T, Hu K, Zhou K, Bi M, Pan L, Wu W, Liu Y. Artificial intelligence detection of distal radius fractures: a comparison between the convolutional neural network and professional assessments. Acta Orthop 2019; 90(4):394–400.

[34] Jiménez-Sánchez A, Kazi A, Albarqouni S, Kirchoff C, Biberthaler P, Navab N, Kirchoff S, Mateus D. Precise proximal femur fracture classification for interactive training and surgical planning. Int J Comput Assist Radiol Surg 2020;15:847–57.

[35] Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. Clin Radiol 2018;73(5):439–45.

[36] Kitamura G, Chung CY, Moore BE. Ankle fracture detection utilizing a convolutional neural network ensemble implemented with a small sample, de novo training, and multiview incorporation. J Digit Imag 2019;32(4):672–7.

[37] Krogue JD, Cheng KV, Hwang KM, Toogood P, Meinberg EG, Geiger EJ, Zaid M, McGill KC, Patel R, Sohn JH, Wright A, Darger BF, Padrez KA, Ozhinsky E, Majumdar S. Automatic hip fracture identification and functional subclassification with deep learning. Radiol Artif Intell 2020;2(2). https://doi.org/10.1148/ryai.2020190023 [accessed via ARXIV for this review)].

[38] Lindsey R, Daluiski A, Chopra S, Lachapelle A, Mozer M, Sicular S, Hanel D, Gardner M, Gupta A, Hotchkiss R, Potter H. Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci U S A 2018;115(45):11591–6.

[39] Tanzi L, Vezzetti E, Moreno R, Aprato A, Audisio A, Massè A. Hierarchical fracture classification of proximal femur X-ray images using a multi-stage deep learning approach. Europaen J Radiol 2020;133:109373. https://doi.org/10.1016/j.ejrad.2020.109373.

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

[41] Yu JS, Yu SM, Erddal BS, Demirer M, Gupta V, Bigelow M, Salvador A, Rink T, Lenoberl SS, Prevedello LM, White RD. Detection and localisation of hip fractures on anteroposterior radiographs with artificial intelligence: proof of concept. Clin Radiol 2020;3(75):1–9. 237e.

[42] Stephens LM, Mortazavi BJ, Deo RC, Curtis L, Kao DP. Recommendations for reporting machine learning analyses in clinical research. Circulation: Cardiovascular Quality and Outcomes 2020;13(10):e006556. https://doi.org/10.1161/CIRCOUTCOMES.120.006556.

[43] England JR, Cheng PM. Artificial intelligence for medical image analysis: a guide for authors and reviewers. Cardiopulmonary Imaging: A Review 2018;212:513–9.

[44] Fazal MI, Patel ME, Tye J, Gupta Y. The past, present and future role of artificial intelligence in imaging. Eur J Radiol 2018;105:246–50.

[45] Yang F, Huang Z, Scholtz J, Arendt DL. How do visual explanations foster end users' appropriate trust in machine learning? ?Assoc Comput Mach 2020:189–201.

[47] Ryes M, Meier R, Pereirs S, Silva C, Dahlweid FM, Tengg-Kobligk H, Summers RM, Weist R. On the interpretability of artificial intelligence in radiology: challenges and opportunities. Radiology: Artif Intell 2020;2(3):e109943.

[48] Kumar D, Wong A, Taylor GW. Explaining the unexplained: a Class-Enhanced Attentive Response (CLEAR) approach to understanding deep neural networks. Available at: https://ieeexplore.ieee.org/Xplore/home.jsp; 2018. Accessed: 10th August 2019.

[49] Alqaraawi A, Schuessler M, WeiB P, Costanza E, Berthouze N. Evaluating saliencey map explanations for convolutional neural networks: a user study. 2020. ArXiv: 2002.00772v1 [Accessed: 18th March 2020].

[50] Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, Topol E, Ioannidis JPA, Collins GS, Maruthappu M. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. Br Med J 2020;368:m689.

[5][1]] Langlotz CP, Allen B, Erickson BJ, Kalpathy-Cramer J, Bigelow K, Cook TS, Flanders AE, Lungren MP, Mendelson DS, Rudie JD, Wang G, Kandararpa K. A roadmap for foundational research on artificial intelligence in medical imaging: from the 2018 MIH/RSNA/ACR/The Academy Workshop. Radiology 2019;291(3).

[52] Thian YL, Li Y, Jagmohan P, Sia D, Vhan VEY, Tan RT. Convolutional neural networks for automated fracture detection and localization on wrist radiographs. Radiol Artif Intell 2019;1(1). https://doi.org/10.1148/ryai.2019180001.