North American Spine Society Journal (NASSJ) 15 (2023) 100236

Contents lists available at ScienceDirect

North American Spine Society Journal (NASSJ)

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

Systematic Reviews/Meta-Analyses

The use of deep learning in medical imaging to improve spine care: A scoping review of current literature and clinical applications

Caroline Constant, DMV, MSc, MENG, DACVS-LA, DECVS[a,b,c,*], Carl-Eric Aubin, PhD, ScD(hc), PEng[b], Hilal Maradit Kremers, MD, MSc[a], Diana V. Vera Garcia, MD[a], Cody C. Wyles, MD[a,e], Pouria Rouzrokh, MD[a,d], Annalise Noelle Larson, MD[a,e]

[a] Orthopedic Surgery AI Laboratory, Mayo Clinic, 200 1st St Southwest, Rochester, MN, 55902, United States
[b] Polytechnique Montreal, 2500 Chem. de Polytechnique, Montréal, QC H3T 1J4, Canada
[c] AO Research Institute Davos, Clavadelerstrasse 8, CH 7270, Davos, Switzerland
[d] Radiology Informatics Laboratory, Mayo Clinic, 200, 1st St Southwest, Rochester, MN, 55902, United States
[e] Department of Orthopedic Surgery, Mayo Clinic, 200, 1st St Southwest, Rochester, MN, 55902, United States

ARTICLE INFO

Keywords:
Artificial intelligence
Machine learning
Deep learning
Spine
Imaging
Clinical care
Review

ABSTRACT

Background: Artificial intelligence is a revolutionary technology that promises to assist clinicians in improving patient care. In radiology, deep learning (DL) is widely used in clinical decision aids due to its ability to analyze complex patterns and images. It allows for rapid, enhanced data, and imaging analysis, from diagnosis to outcome prediction. The purpose of this study was to evaluate the current literature and clinical utilization of DL in spine imaging.

Methods: This study is a scoping review and utilized the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology to review the scientific literature from 2012 to 2021. A search in PubMed, Web of Science, Embased, and IEEE Xplore databases with syntax specific for DL and medical imaging in spine care applications was conducted to collect all original publications on the subject. Specific data was extracted from the available literature, including algorithm application, algorithms tested, database type and size, algorithm training method, and outcome of interest.

Results: A total of 365 studies (total sample of 232,394 patients) were included and grouped into 4 general applications: diagnostic tools, clinical decision support tools, automated clinical/instrumentation assessment, and clinical outcome prediction. Notable disparities exist in the selected algorithms and the training across multiple disparate databases. The most frequently used algorithms were U-Net and ResNet. A DL model was developed and validated in 92% of included studies, while a pre-existing DL model was investigated in 8%. Of all developed models, only 15% of them have been externally validated.

Conclusions: Based on this scoping review, DL in spine imaging is used in a broad range of clinical applications, particularly for diagnosing spinal conditions. There is a wide variety of DL algorithms, database characteristics, and training methods. Future studies should focus on external validation of existing models before bringing them into clinical use.

Introduction

Despite extensive research to improve spine care, spinal disorders remain prevalent [2]. A large body of evidence demonstrates the negative impact of spinal disorders on individuals and society, resulting in disabilities and considerable economic losses [3]. The annual cost of spine care has risen in the last decade [3,5], suggesting a need for innovation to improve the care of patients with spinal conditions.

FDA device/drug status: Not applicable.

Author disclosures: ***CC:*** Nothing to disclose. ***CEA:*** Grants: Natural Sciences and Engineering Resources Council of Canada (Industrial Research Chair Program With Medtronic of Canada) (C, Paid directly to institution/employer), (F, Paid directly to institution/employer); Canada First Research Excellence Fund (C, Paid directly to institution/employer), (I, Paid directly to institution/employer); Consulting: Medtronic (E). ***HMK:*** Scientific Advisory Board: CSR (A, Paid directly to institution/employer), Grants: NIH: (I, Paid directly to institution/employer). ***DVVG:*** Nothing to disclose. ***CCW***: Nothing to disclose. ***PR:*** Nothing to disclose. ***ANL:*** Nothing to disclose.

* Corresponding author: AO Research Institute Davos, Clavadelerstrasse 8, Davos CH 7270, Switzerland. Tel.: (41) 79-9106976.
E-mail address: caroline.constant@aofoundation.org (C. Constant).

https://doi.org/10.1016/j.xnsj.2023.100236
Received 20 May 2023; Accepted 14 June 2023
Available online 19 June 2023

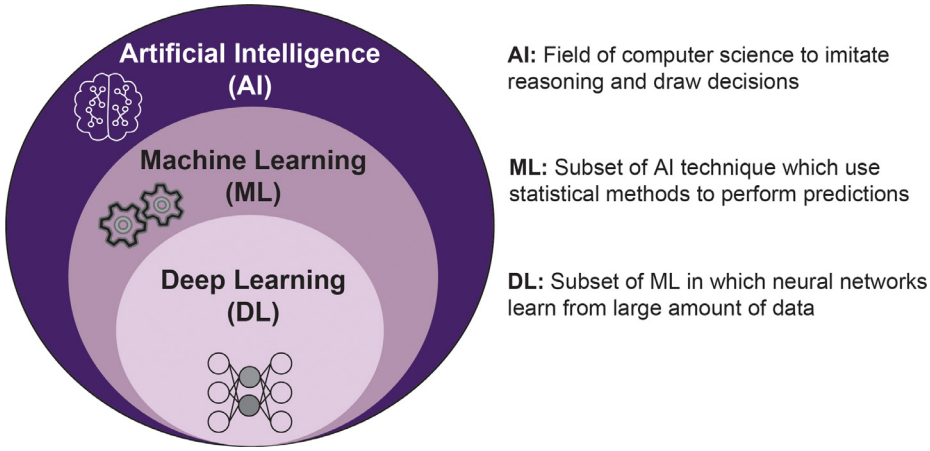

Fig. 1. Overview of artificial intelligence, machine learning and deep learning.

Medical imaging is critical for clinical decision-making and is integral to determining treatment indications and surgical outcomes. Various imaging modalities can be used for accurate detection and diagnosis of spinal pathologies [91]. However, despite following similar diagnostic standards, experienced radiologists can arrive at different diagnoses and measurements with error rates estimated to be 3% to 5% [7]. Much higher discrepancy rates are reported in neuroradiology, with variable reads in up to 21% of imaging studies [8]. Therefore, a key challenge in the diagnosis of spine pathology is improving the workflow to diagnose diseases quickly, automatically, and accurately. Further, medical care costs are increasing [3], and radiologists who train for 5 or more years after medical school represent an expensive and valuable resource. AI tools to augment the performance of radiologists and provide a low-cost tool to prevent errors hold significant promises for improved quality and cost savings. Reliable treatment planning, precise intervention, and accurate therapy are required when treating spine conditions to achieve optimal outcomes. While modern imaging techniques inform perioperative case management [9], poor outcomes are still frequent. Thus, the development of scalable, perioperative automated assistance tools could help address this challenge.

Within health care, artificial intelligence (AI) algorithms are increasingly used for complex tasks, including remote patient monitoring, medical diagnosis and imaging, risk assessment, virtual assistance, hospital management, and drug discovery [10]. AI is a field of computer science that attempts to build enhanced "intelligence" into computer systems by implementing algorithms that apply rules to imitate reasoning and draw decisions (Fig. 1).[136] Within AI, machine learning (ML) is a promising field for improving patient-specific spine care as it allows computers to learn without being explicitly programmed. ML can perceive important imaging trends that the average practitioner may not perceive [11,12]. To do so, ML uses provided data or previous experience to develop predictive models to determine subtle patterns and predict outcomes from a collection of statistical techniques [24].[137] In other words, ML techniques are based on available data with specific features (input data), which are used to train a machine (computer) to perform (to learn) the desired task generating a specific output (output data). The medical field has recently seen a fast improvement in ML techniques, specifically through deep learning (DL), an advanced form of machine learning capable of feature extraction to perform several tasks precisely developed to help clinicians [13]. In most circumstances, DL is based on neural networks (NN), network architectures formed of several layers, called hidden layers, containing multiple units, called artificial neurons, interconnected by mathematical relations called synapses. In each unit, a mathematical sum resulting from inputs' multiplication by a weight and inputs' summing to a bias term is processed by a linear or nonlinear activation function. In this context, the hidden layers help the network refine the input-output synapses between the units. DL attracts great interest for clinical application and image analysis in radiology partly due to its outstanding performance in image recognition, classification, and segmentation tasks [10,14,15].

Narrative and systematic reviews have recently been completed on AI applications in the spine [16–20]; yet they did not systematically assess all published studies on the subject. Previous reviews have demonstrated that DL techniques are robust and scalable for spine care applications. However, no review has comprehensively mapped and assessed the quality of the clinical applications of DL combined with medical imaging for spinal diseases research. Thus, given the recent technological developments, we undertook a review of DL techniques for spine imaging applications to inform data scientists and clinicians on the methods and applications of big data in spine. Furthermore, we aim to highlight the challenges and limitations of DL techniques, identify gaps in the field, and outline potential opportunities for further research. Thus, this scoping review aims to broadly review and systematically evaluate the current progress in DL and how it has been applied to medical imaging and clinical applications intended for clinical spine care.

Materials and methods

Search strategy

A systematic literature review using a scoping review approach and following Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines [21] and PRISMA-ScR extension for Scoping Reviews guidelines [22] was carried out on October 15, 2021, and repeated on January 1, 2022. The systematic research was constructed to identify studies describing DL in medical imaging for clinical spine applications. The final search terms and additional methodology details used are shown in *Appendix A – Search Strategy*. First, the literature search was conducted through the health-related research database Medline. Next, the information technology database IEEE Xplore was searched. Lastly, databases that index both fields, including Web of Science and Embase, were searched for relevant literature.

Eligibility criteria

Studies on DL and medical imaging dealing with applications intended for spine clinical care were selected. This review did not consider other AI methods based on fundamental ML techniques, preferring only DL-based approaches. Publications in peer-reviewed journals after 2012 were included. The beginning timepoint was selected as the first scalable convolutional neural network, significantly improving the state-of-the-art natural-image classification results [23]. Additionally, non-peered-review references published after 2012, such as case reports, proceedings, or abstracts, and book chapters, were included. Studies that

Table 1
Extracted characteristics from the studies included in the review.

Characteristic	Description
(1) Author's data	• Country of the authors' affiliations: the affiliation country from the majority of the authors or the corresponding author • Authors' fields of expertise: health fields, data science fields, or both • Status with industry: if one of more author was affiliated with an industrial partner
(2) Year	• The year it was published based on Medline, IEEE Xplore, Web of Science, or Embased databases
(3) Study type and design	• Study type: classification of primary studies into basic, clinical and epidemiological research; and subclassification into interventional or observational [1] • Study design: classification of studies into retrospective or prospective nature and further categorization into study design being cross-sectional, cohort, descriptive, case-control, or case-series type study designs using a described classification algorithms [4]
(4) Area of spine care focus	• Clinical application type: either diagnostic tools, clinical decision support tools, automated clinical/instrumentation assessment, clinical outcome prediction, combined or others • Studied anatomy: either cervical, thoracic, lumbar, sacral, combiner • Studied disease: if there was specific pathology that the study targeted • Studied surgery: if there were particular surgeries or procedures that the study targeted • Study imaging modality: either standard radiograph, dual-energy x-ray absorptiometry (DEXA), CT, MRI including which sequence(s), US, interventional imaging (fluoroscopy, O-arm, guided navigation), combined, or others
(5) Number of subjects and images included	• Overall size: number of subjects, images[a], or both included in the complete study • Overall diseased subjects: number of patients, images, or both included in the general study diagnosed with at least one spine condition • Categorized size: overall subjects, images, or both included in the general study binned into categories of <100, 100-1000, 1000-10000, 10000-100000, and >100000
(6) Size of the dataset used for DL development and validation	• Dataset size: number of subjects, images, or both included in the DL development phase (when applicable) • Dataset diseased subjects: number of patients, images, or both included in the DL development phase diagnosed with at least one spine condition (when applicable) • Categorized dataset size: dataset subjects: number of patients, images, or both included in the DL development phase (when applicable) binned into categories of <100, 100–1,000, 1,000–10,000, 10,000–100,000, and >100,000
(7) Origin of the dataset	• Either single-center, multicentric, public registry or dataset, synthetic images, or combined
(8) Whether the dataset is publicly available, part of a registry, or institutional data	• Availability and information of origin and access of datasets from publicly available and part of a clinical registry or database datasets
(9) DL method and architecture used	• DL methodology: either CNN, long short-term memory networks (LSTM), Recurrent neural network (RNN), Generative Adversarial Networks (GAN), Radial basis function networks (RBFN), Multilayer Perceptrons (MLP), Self-organizing maps (SOM), deep belief network (DBM), restricted Boltzmann machines (RBM), or other • DL task type: classification, regression, segmentation, object detection, image generation, or other • DL architecture: architecture and backbone family (ex: DenseNet, VGGs, etc) • Number of pipeline(s) and DL architecture(s) used or tested • Other ML techniques used or tested in the study
(10) DL training and validation	• Training: split of the dataset into training, validation, testing, and use of cross-validation • External validation: if external validation of the completed DL pipeline was performed
(12) Evaluation of performances	• Performance metrics used to validate their pipeline • If they used external data to validate their pipeline

proposed solely technical applications without direct specific clinical implications, such as image segmentation, image quality improvement, and image reconstruction alone, were excluded. Animal experiments, reviews, correspondences, expert opinions, and editorials were also excluded. Two reviewers (XX, XX) independently reviewed all studies, reaching a consensus on all included studies. A third reviewer resolved the disagreements in the inclusion process (XXX).

Data collection and analysis

Data from all included studies were collected into a standardized data extraction sheet (Table 1). In addition, the key findings, clinical deployment, expected clinical benefits and value, economic implications, and ethical considerations of implementation in the health care system were recorded. To analyze the data, a narrative review synthesis method with descriptive statistics was selected to capture the extensive range of research investigating DL for spine clinical care. It should be noted that a meta-analysis was not appropriate for this review, given the broad range of spinal conditions, DL techniques, and types of data used in the studies identified. All retrieved manuscripts and abstracts were included in the qualitative synthesis.

Results

Overview of studies characteristics

The search strategies identified 1,475 records, with 365 of these studies meeting the criteria for inclusion (Fig. 2 and *Appendix B – List of Included Studies*). The top 3 affiliated countries were the United States (17%), China (16%), and Canada (12%). Most studies were authored by multidisciplinary teams (61%), including experts from both medicine (most frequently neuro- or spine surgery and radiology) and engineering (most commonly electrical engineering, computer science, and/or data science), with the remaining articles authored by either medicine (19%) or engineering (14%) experts only, or could not be retrieved (9%). Most studies did not receive or disclose a significant contribution from an industrial partner (79%).

Study type and design

The study type could be identified for 352 published studies and was primarily classified as clinical research (64%; Fig. 3). The information available in the remaining 13 studies was insufficient to classify them. The division of the studies into retrospective or prospective investigations was impossible in 96 studies (26%), meaning that we could not distinguish if the data were collected explicitly for the study or from an existing data source. Data were collected retrospectively for the remaining studies (n = 269) (88%).

The study design could be identified for 318 of the published studies and was predominantly categorized as cross-sectional (47%) and descriptive studies without comparison groups (43%), and less frequently as cohort studies (9%). The information available in the remaining 47 studies (13%) was insufficient to determine the study design. For the particular case of clinical research for the diagnostic, prognostic, and

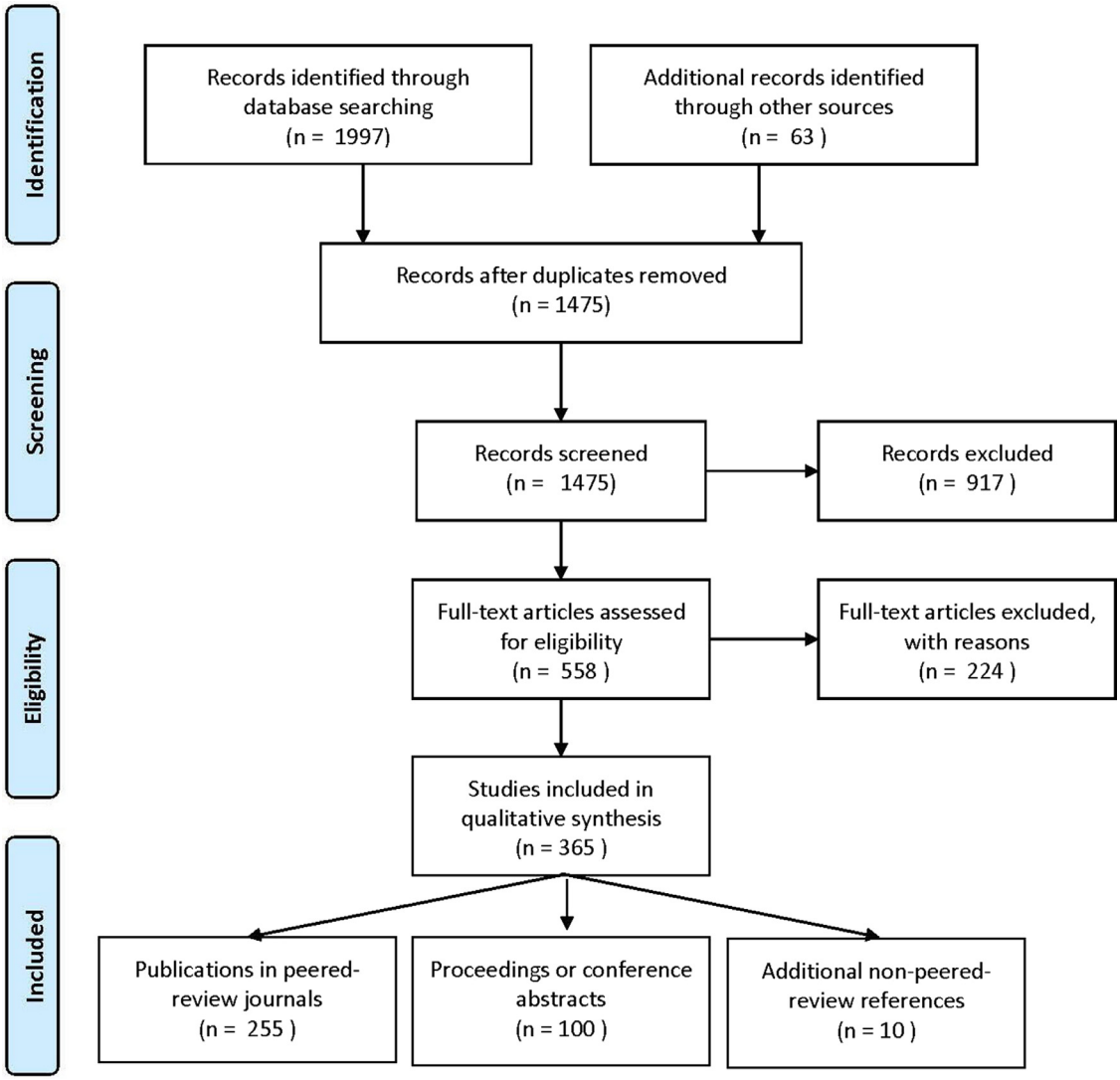

Fig. 2. The Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) statement flowchart of the preformed for the review of the current stateof-the-art progress and utilization of DL in the field of medical imaging for spine care.

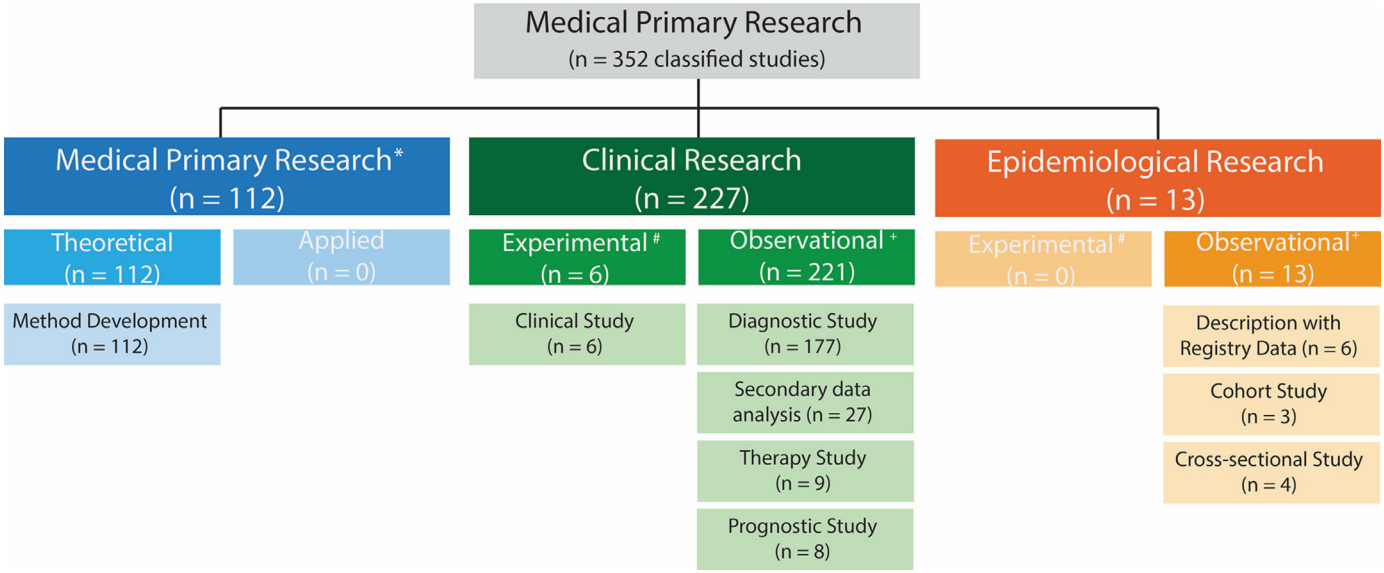

Fig. 3. Classification of the types of studies from 352 published studies focusing on DL in the field of medical imaging for spine care according to the classification schemes for studies in medical research by Rohrig et al.[1] * , *sometimes known as experimental research; #, analogous term to interventional; +, analogous term to noninterventional or nonexperimental*

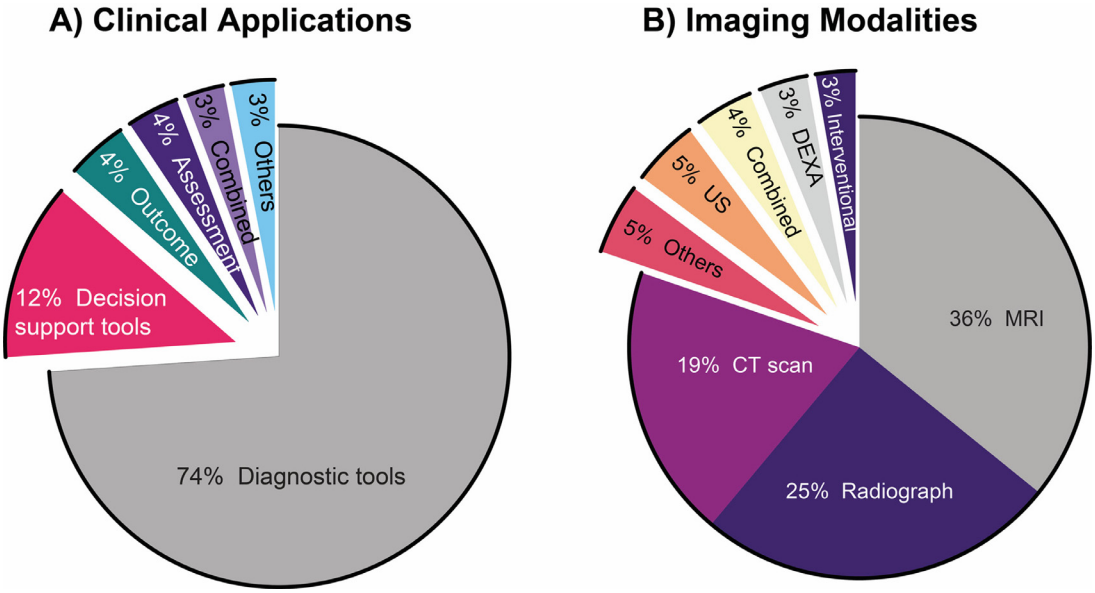

Fig. 4. Clinical applications (A) and imaging modalities (B) related to the published studies investigated in this review focusing on DL in the field of medical imaging for applications intended for spine clinical care. The clinical application types included diagnostic tools, clinical decision support tools, automated clinical/instrumentation assessment, clinical outcome prediction, combined or others. The imaging modalities used included magnetic resonance imaging (MRI), radiograph, computed tomography (CT scan), ultrasound (US), dual-energy x-ray absorptiometry (DEXA), intervention imaging (fluoroscopy, O-arm, guided navigation), and others.

predictive test accuracy studies, the top 3 study designs were cross-sectional studies (56%), descriptive studies without comparison groups (26%), and cohort studies (8%).

Spine clinical care and imaging focus

The clinical application of DL in spine image analysis was most frequently related to developing or validating new diagnostic tools (74%; Fig. 4**A**) and primarily included studies that aimed to identify or diagnose spinal conditions. Three main themes arose from these diagnostic studies: (1) detection of diseases by developing prediagnosis screening tools or anomaly detection, (2) predicting the diagnosis of new patients based on a training dataset of prior diagnoses, and (3) differentiating between spinal conditions with similar imaging features or symptomatology. Other frequently identified clinical themes were clinical decision support, assessment, and outcome prediction studies, mainly aimed at predicting the progression of spinal conditions, exploring treatment possibilities, or supporting clinical opportunities for such conditions. Two main themes were identified among studies examining clinical decision support tools: (1) identifying preoperative factors to provide personalized and timely treatment or surgical interventions and (2) supporting procedures such as injections under imaging guidance and surgical navigation. Studies investigating prognosis primarily focused on using DL to predict the development of complications following spinal surgery. Other clinical applications found in the reviewed studies included public health investigations, which used large epidemiological or public datasets to monitor or screen for spinal conditions in the general population, describe average spinal measurements, or estimate disease prevalence; and clinical administration tools, which included studies that aimed at improving administrative processes in clinical work and healthcare organizations such as automated report generation. MRI was the most common imaging modality used in the reviewed studies (36%; Fig. 4**B**), and the most frequently used MRI sequences were T2-weighted alone (53%), a combination of T1-weighted and T2-weighted (26%), and T1 weighted alone (13%). Most studies used DL techniques for spinal conditions or clinical care targeting a single spinal region (47%). The majority of the remaining studies investigated more than 1 spinal region (35%). In contrast, only a minority of studies explored

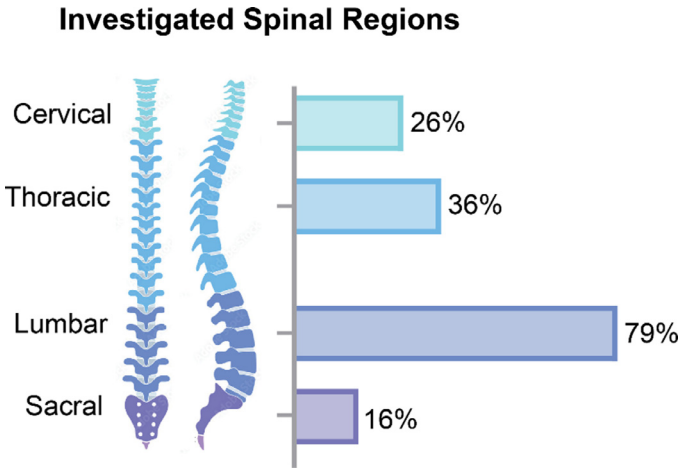

Fig. 5. Distribution of the frequency of investigation of spinal regions targeted by the published studies investigated in this review focusing on DL in the field of medical imaging for applications intended for spine clinical care. *Note: sum is not equal to 100% as 49% of reviewed studies investigated multiple spinal regions.*

DL techniques on the whole spine (14%) or did not provide enough information to identify a specific spine region (5%). The lumbar spinal segment was the most routinely studied region and was included alone or in combination with other spinal regions (s) in 79% of all investigations (Fig. 5).

Through analysis of the data, 5 main domains of spinal conditions were identified, with the top 3 being inflammatory and degenerative conditions (26%), spinal deformity and alignment problems (22%), and fractures (14%; Fig. 6). Among the studies that investigated inflammatory or degenerative spine conditions, MRI was routinely used (82%), and improvement or automation of the diagnosis was the main cited objective (89%). Automated tissue classification and measurements were the most investigated DL pipeline outcomes (67% and 18%, respectively). Among the studies targeting spinal alignment problems, radiographic images were generally used (69%), with the main objective of

Table 2
Registries and clinical databases used in the studies focusing on deep learning in the field of medical imaging for spine care investigated in this review with the corresponding summary when available.

Registry / database name	Official title	Responsible party, owner or sponsor	Resources
ALSPAC [36]	Avon longitudinal study of parents and children	University of Bristol, Bristol, United Kingdom	http://www.bristol.ac.uk/alspac/researchers/access/
AO CSM-I	Surgical treatment of cervical spondylotic myelopathy	AO Innovation Translation Center (AO Clinical Investigation and Publishing Documentation), Dübendorf, Switzerland	https://clinicaltrials.gov/ct2/show/NCT00565734
AO CSM-NA	AOSPINE assessment of surgical techniques for treating cervical spondylotic myelopathy	AOSpine North America Research Network, Pennsylvania, USA	https://clinicaltrials.gov/ct2/show/NCT00285337
CSORN [37]	Canadian spine outcomes and research	Canadian Spine Society, Markdale, Canada	https://www.csorncss.ca/
Genodisc Project	Disc-degeneration linked pathologies: novel biomarkers and diagnostics for targeting treatment and repair	University of Oxford, Oxford, United Kingdom	https://cordis.europa.eu/project/id/201626/reporting
GESPIC [38]	German spondyloarthritis inception cohort	Charite University, Berlin, Germany	https://clinicaltrials.gov/ct2/show/NCT01277419
Hangzhou lumbar spine study [39]	Hangzhou lumbar spine study: a study focusing on back health in a Chinese population	First Affiliated Hospital of Zhejiang University, Hangzhou, China	Not found
H-PEACE [40]	Health and prevention enhancement	Seoul National University Hospital, Seoul, South Korea	http://en-healthcare.snuh.org/HPEACEstudy
LumbSeg [41,42]	Lumbar vertebra segmentation CT image datasets	*Not found*	Not found
Manitoba BMD Registry [43]	The Manitoba BMD Registry	Manitoba Bone Density Program Committee, Manitoba, Canada	https://www.gov.mb.ca/health/primarycare/providers/chronicdisease/bonedensity/research.html
MDCS [36]	Malmö Diet and Cancer Study	*University of Lund, University Hospital, Malmo, Sweden*	*Not found*
MIDICAM	Cervical spondylotic myelopathy: Application of spinal diffusion-based microstructural imaging (DMI) and phase-contrast MRI	University Medical Center Neurozentrum, Freiburg im Breisgau, Germany	https://www.drks.de/drks_web/navigate.do?navigationId=trial.HTML&TRIAL_ID=DRKS00012962 German registry of clinical trials, number: DRKS00012962
MPP [44]	Malmö preventive medicine project	University of Lund, University Hospital, Malmo, Sweden	Not found
MrOs [45]	Osteoporotic fractures in men Study (MrOS) from Database of Genotypes and Phenotypes (dbGaP)	University of California San Francisco, California, USA	https://sfcc.ucsf.edu/news/osteoporotic-fractures-men-mros-study-group-publish-two-articles
MySPINE	Functional prognosis simulation of patient-specific spinal treatment for clinical use	Fundacio Institut de Bioenginyeria de Catalunya, Barcelona, Spain	https://cordis.europa.eu/project/id/269909
NFBC1966	Northern Finland Birth Cohort 1966	University of Oulu, Oulu, Finland	http://www.oulu.fi/nfbc/
NHANES 2011-2012	National Health and Nutrition Examination Survey 2011–2012 Database	National Center for Health Statistics, Maryland, USA	https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/overview.aspx?BeginYear=2011
NHANES II	Second National Health and Nutrition Examination Survey Database	National Center for Health Statistics, Maryland, USA	https://wwwn.cdc.gov/nchs/nhanes/nhanes2/default.aspx
OSTPRE [46]	Kuopio osteoporosis Risk factor and prevention study	Kuopio University Hospital and University of Eastern Finland, Kuopio, Finland	https://sites.uef.fi/kmru/ostpre/
OSTPRE-FPS [47]	OSTPRE fracture prevention study	Kuopio University Hospital, Kuopio, Findland	https://clinicaltrials.gov/ct2/show/NCT00592917
PROOF [48]	Multicountry Registry of Clinical Characteristics	AbbVie, Cham, Switzerland	http://www.chictr.org.cn/showprojen.aspx?proj=10022
ROAD	Research on osteoarthritis/osteoporosis Against disability	University of Tokyo, Tokyo, Japan	Not found
SCI database	*Not found*	Orange image diagnostic center, USA	Not found
SDSG	Spinal Deformity Study Group database	Not found	Not found
TRACK-SCI [45,49]	Transforming research and clinical knowledge in spinal cord injury	University of California, San Francisco, California, USA	https://clinicaltrials.gov/ct2/show/NCT04565366
TwinsUK registry	TwinsUK registry from	King's College London, United Kingdom	www.twinsuk.ac.uk
Wakayama Spine study [50]	Wakayama Spine study	Wakayama MedicalUniversity School of Medicine, Wakayama, Japan	Not found
Whiplash [51]	Neuromuscular mechanisms underlying poor recovery from whiplash injuries	Northwestern University, Ilinois, USA	ClinicalTrials.gov Identifier: NCT02157038

improving measurement accuracy for sagittal or coronal balance (85%). Spinal alignment measurements were the most commonly investigated pipeline outcome (66%). Studies addressing spinal fractures consistently aimed to improve the accuracy and speed of diagnosis of vertebral fractures and detection of fractured levels (90%). To do so, CT scans and radiographs were the most common imaging modality used (58% and 20%, respectively), and automated vertebral status classification and fracture detection with or without specific anatomical localization were the most commonly studied pipeline outcomes (50% and 32%, respectively).

The studies targeting surgical conditions had multiple objectives, but primarily provided clinical decision support tools and automated instrumentation assessment (56% and 50%, respectively) and commonly used ultrasound or fluoroscopy as imaging modalities (33% and 27%,

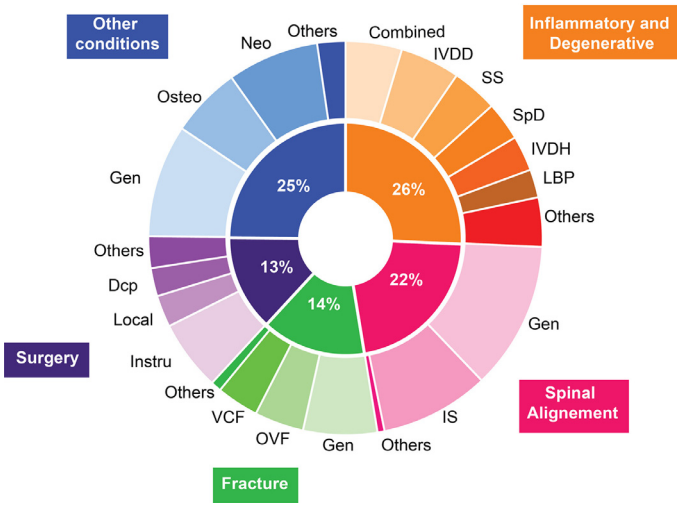

Fig. 6. Spinal diseases and conditions examined in the published studies investigated in this review focusing on DL in the field of medical imaging for applications intended for spine clinical care. Inflammatory and degenerative conditions included intervertebral disc degeneration (IVDD), spinal stenosis (SS), spondylitis and spondyloarthritis (SpD), intervertebral disc herniation (IVDH), and lower back pain complex (LBP). Spinal alignment conditions included measurement applicable to general alignment problems (Gen) and idiopathic scoliosis (IS). Fracture assessment included general vertebral bone assessment (Gen), osteoporotic vertebral fracture (OVF), and vertebral compression fracture. Investigation targeting surgical procedures included spinal instrumentation (Instru), local analgesia or anesthesia procedures (Local), and spinal decompression surgery (DCP). Among other conditions, general spinal assessment (Gen), osteoporosis (Osteo), and neoplastic diseases (Neo) have been studied.

respectively). While less common, neoplasia was another recurrent condition investigated (8%), with most studies aimed at improving diagnosis accuracy and lesion delimitation (65%) using mainly MRI imaging (62%). To this aim, automated lesion detection or delimitation and tissue classification were the most commonly investigated pipeline outcomes (50% and 35%, respectively). Six percent of studies included in this review examined osteoporosis and mainly intended to detect bone properties abnormalities or diagnose osteoporosis automatically (80%), primarily using CT scans and DEXA as imaging modalities (48% and 33%, respectively). Within those, the primary pipeline outcomes were bone status classification and bone properties measurements, such as bone mineral density (42% and 38%, respectively). The overall publication rate for presented abstracts published as full-length articles in peer-reviewed journals was 6%.

Subjects, images, and datasets

Only 41% of the reviewed studies clearly reported the number of subjects and medical images included in the research. The number of subjects included was reported more often than the number of images (33% vs. 19%). Most studies included 100 to 1000 subjects or images (Fig. 7) with a mean of 945 and 9504, respectively. When the studies targeted a particular spinal condition, subjects with a relevant positive diagnosis were regularly enrolled (73%) with a mean enrollment ratio of 79% and 71% of all subjects and images, respectively. Seventy-five percent of studies reported whether their included subjects or images had surgical implants, with most stating the exclusion of them if patients had spinal implants (75%). Only 30% of studies reported the health status of the enrolled subjects or the presence of other spinal diseases on the images included, with most studies not including data about other conditions or diseases (55%). For the studies that included the DL training phase, the mean of subjects and images in the DL development phase were 763 and 4,286, respectively.

The origin of the data could be retrieved for 288 of the published studies and was primarily part of institutional data (58%), registries or clinical databases (11%; Table 2), public datasets (4%; Table 3), or a combination (4%). When institutional data, registries, or clinical databases were used, the data originated more commonly from a single center rather than multiple centers (68% vs. 32%).

DL method and architecture

In 84% of the included studies, images were used as input through DL computer vision tasks. In contrast, the others used extracted features from images or collected data from medical imaging results (e.g., disc height) as input. Computer vision tasks were, in general, performed by 3 different DL approaches (1) landmark detection, often combined with prior structure detection, (2) structure segmentation, or (3) shape model matching. To this end, the preferred DL technique was CNN which was used in 77% of included studies. CNN techniques were investigated mainly for classification (43%), measurement tasks by structure segmentation or landmark detection (26%), and detection tasks (20%). For most classification tasks, convolutional and pooling layers were first used to extract features from the input images, followed by fully connected layers for output feature classification. Measurement tasks using landmark detection were usually performed similarly or through segmentation tasks where the fully connected layer of the CNN was replaced with

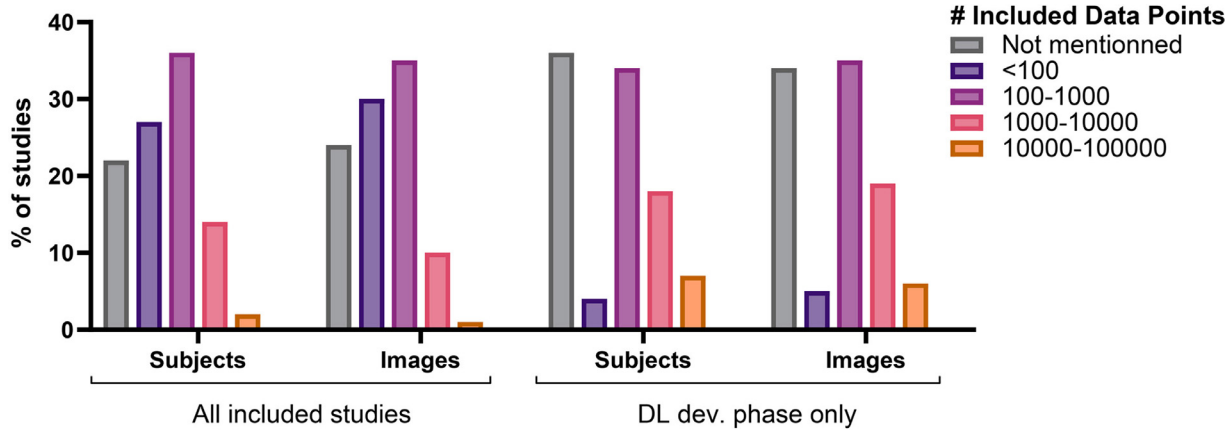

Fig. 7. Number of subjects and images included in the published studies investigated in this review focusing on DL in the field of medical imaging for applications intended for spine clinical care. The number of data points is categorized and presented for the overall subjects and images comprised in all included studies and for the subjects and images comprised only in the DL development phase of the studies, when applicable.

Table 3
Publicly available datasets used in the studies focusing on deep learning in the field of medical imaging for spine care investigated in this review with the corresponding summary when available.

Public dataset name	Official title	Responsible party, owner or sponsor	Resources
[76]CSI 2014 workshop dataset	localization and identification challenge of the CSI 2014 Workshop.	University of Washington in St Louis, Missouri, USA	*The link provided in the publication is no longer working* (http://research.microsoft.com/spine/)
[77]DeepLesion	*Not found*	*Not found*	https://www.kaggle.com/datasets/kmader/nih-deeplesion-subset
[78]ILSVRC	ImageNet Video dataset (ILSVRC)	Stanford Vision Lab, Stanford University, Princeton University	https://image-net.org/challenges/LSVRC/index.php
IoMT Spine Dataset	Internet of Medical Things (IoMT) platform Spine Dataset	Not found	http://spineweb.digitalimaginggroup.ca/spineweb.
MICCAI [55,56]	Testing set A of the MICCAI Challenge on Vertebral Fracture Analysis	Medical Image Computing and Computer Assisted Intervention Workshop & Challenge (MICCAI 2016)	Upon request at: http://spineweb.digitalimaginggroup.ca/dataset.php
MS Annotated Spine CT Database	MS Annotated Spine CT Database	*Not found*	No longer working (http://research.microsoft.com/en-us/projects/spine/)
PAM50 [54,57]	Unbiased multimodal MRI template of the spinal cord and the brainstem	Polytechnique Montreal, Canada and Aix-Marseille Université, France	https://github.com/sct-data/PAM50/releases/download/r20191029/20191029_pam50.zip
SpineWeb Dataset [6]	Dataset 16: 609 spinal anterior-posterior x-ray images	Stanford Vision Lab, Stanford University, Princeton University London Health Sciences Center, Ontario, Canada	https://image-net.org/challenges/LSVRC/index.phpUpon request (http://spineweb.digitalimaginggroup.ca/dataset.php)
SpiSeg [52]	Spine and vertebrae segmentation datasets	University of California, Irvine, School of Medicine, California, USA	Upon request at: http://spineweb.digitalimaginggroup.ca/dataset.php
TCIA [57]	Cancer Imaging Archive	Washington University School of Medicine, Missouri, USA and Frederick National Laboratory for Cancer Research, Frederick, MD, USA	No longer working (http://research.microsoft.com/en-us/projects/spine/)https://cloud.google.com/healthcare-api/docs/resources/public-datasets/tcia
UCI Vertebral Column Dataset	UCI Repository of Machine Learning Databases	UC Irvine Machine Learning Repository	http://archive.ics.uci.edu/ml/machine-learning-databases/00212/
VerSe'20[56]	Large Scale Vertebrae Segmentation Challenge	Medical Image Computing and Computer Assisted Intervention (MICCAI 2020)	https://zenodo.org/record/3759104#.Y3UE0nbMInI
xVertSeg.v1[53]	Segmentation and Classification of Fractured Vertebrae	University of Ljubljana, Faculty of Electrical Engineering, Slovenia	http://lit.fe.uni-lj.si/xVertSeg/

up-sampling layers (encoder-decoder architecture), resulting in the conversion of landmarks into segmentable heatmaps.

Sixty-four percent of studies adapted an existing DL architecture, most commonly based on U-Net (21%) and ResNet (16%), all of which operated on image data (Fig. 8 and Table 4). Studies using extracted features from images or collected data from medical imaging results as input preferably used MLP network structures, which were used in 21% of included studies. Other DL methodologies were reported, including long short-term memory layer, radial basis function network, and self-organizing map, but were only investigated in 2% of the included studies. Most studies investigated only one DL architecture (57%) with a mean ± SD of 1.8 ± 1.3 DL architectures investigated per study. Nevertheless, the performance of several DL pipelines was investigated in most studies (91%) with a mean ± SD of 2.6 ± 2.8 pipelines or models investigated per study. Sixteen percent of studies also surveyed other ML techniques as an alternative to DL in their pipelines or as comparison results, with support vector machine and K-nearest neighbor algorithms being the most frequently reported.

DL training and validation of studies with DL dev

A DL model was developed and internally validated in 92% of included studies, while a pre-existing DL model was externally investigated in 8%. The dataset split into training, validation, and testing was mentioned in 155 (47%) development studies. When available, the mean number of subjects and images used in the DL development phase was 763 and 4,286, respectively. The mean proportions of data split into the training, validation, and testing datasets were 0.75, 0.18, and 0.20, respectively. Of the development studies, 26% described the prevention of overfitting using cross-validation. Only 15% of development studies externally validated the completed DL pipeline on a data set distinct from the training dataset. Of these, 43% performed external validation using data from a completely different origin, such as from a foreign country or other hospitals. Internal validation was the only validation technique in the remaining studies (85%).

Evaluation of performances

A large variety of performance metrics were retrieved from the studies to evaluate the similarity between the DL prediction and the ground truth (Table 5). For DL pipelines with classification tasks as outputs (binary as well as multi-class problems), the measurement of performance, whenever reported, generally included sensitivity and specificity of the technique (58%) and area under the curve (51%). The DL pipelines with detection tasks as outputs mainly reported the precision (48%) for localizing the object's position in the image and recall (56%) when judging whether objects belonging to certain classes appear in regions of interest. The DL pipelines with measurement tasks outputs frequently used various error calculations to evaluate the model performances, principally the mean absolute error (20%) and standard error (20%).

Only 8 peer-reviewed research articles (4%) adhered to a reporting checklist, including 3 guidelines for transparent reporting of predictive or AI models [25–27] and one for reporting diagnostic accuracy studies [28]. A minority of studies (7%) provided the code used for model de-

Table 4
Overview of common CNN architectures used in the studies focusing on deep learning in the field of medical imaging for spine care investigated in this review with the variants reported for the spine and corresponding short summary.

Backbone	Year	Description	Variants reported for the use in spine	Ressources
AlexNet [23]	2012	An architecture developed for image classification that launched DL development. The network contains a set of convolutional and max-pooling layers ended by 3 fully connected layers.	AlexNet. *Note: also used in in some of the R-CNN architectures*	https://github.com/deep-diver/AlexNet
VGG [58]	2014	Backbone widely used for computer vision and computer sciences tasks. VGG is composed of convolutional, max-pooling, and fully connected layers. It uses smaller kernels to create deeper networks.	VGG-11, VGG-16, VGG-19, VGG-M, VGG-Net, VGG-Net16, FCN-VGG-16	https://github.com/pytorch/vision/blob/master/torchvision/models/vgg.py
U-Net [59]	2015	Fully convolutional network Popular architecture for biomedical image semantic segmentation. It comprises a contracting path ("traditional" CNN) and an expansive path.	3D U-Net, BiLuNet, Co-U-Net, DC-U-Net, Deeplab V3+, Deep-U-Net, fuse-U-Net, MDR2-U-Net, MDR2-U-Net, Residual U-Netstacked hourglass network (SHN)	https://lmb.informatik.uni-freiburg.de/people/ronneber/u-net/u-net-release-2015-10-02.tar.gz
ResNet [60]	2015	Backbone widely used for object detection and image segmentation that introduced skip connected. ResNet is made of residual NN consisting of skip-connections or recurrent units between blocks of pooling and convolutional layers. Many versions with different deepness.	FR-ResNet, Multi ResNet, ResNet-12, ResNet-18, ResNet-34, ResNet-50, ResNet-101, ResNet-ST-50, ResNet-XT-50. *Note: also used in some of the Faster R-CNN architectures.*	https://github.com/pytorch/vision/blob/master/torchvision/models/resnet.py
Inception-v2-v3[61]	2015	Network with a low number of parameters to be able to be used on lower-performance machines. It was modified from Inception V1 (blocks of inception containing sets of convolutional layers) by replacing nn convolutional kernels into 3×3 or 1×1 using a concatenation method.	Inception-V3	https://github.com/weiaicunzai/pytorchcifar100/blob/master/models/inceptionv3.py
Inception-ResNet-V2[62]	2015	ResNet and inception architecture combine using skip-connections between blocks of layers, called residual connections.	Inception-ResNet-V2, Inception-ResNet-V3. *Note: also used in some Faster R-CNN architectures.*	https://github.com/zhulf0804/Inceptionv4_and_Inception-ResNetv2.PyTorch
YOLO [63,64]	2015	YOLO, short for "You Only Look Once" is from a series of object detection models capable of detecting multiple objects simultaneously. It is based on CNN and is used to predict classes as well as object localization.	YOLOV2, YOLOV3, HoloYOLO	https://github.com/ultralytics/yolov5
DenseNet [60,65]	2018	CNN built on the idea of ResNet but with a lower number of connections of L/(L+1)/2, with L being the number of layers. The feature maps of all previous layers are used as input in the next layer, making DenseNet well-suited for smaller datasets.	DenseNet-14, DenseNet-22, DenseNet-26, DenseNet-48, DenseNet-121, DenseNet-169, DenseNet-201, DMML-Net, FC-DenseNet, BMDC-Net	https://github.com/liuzhuang13/DenseNet

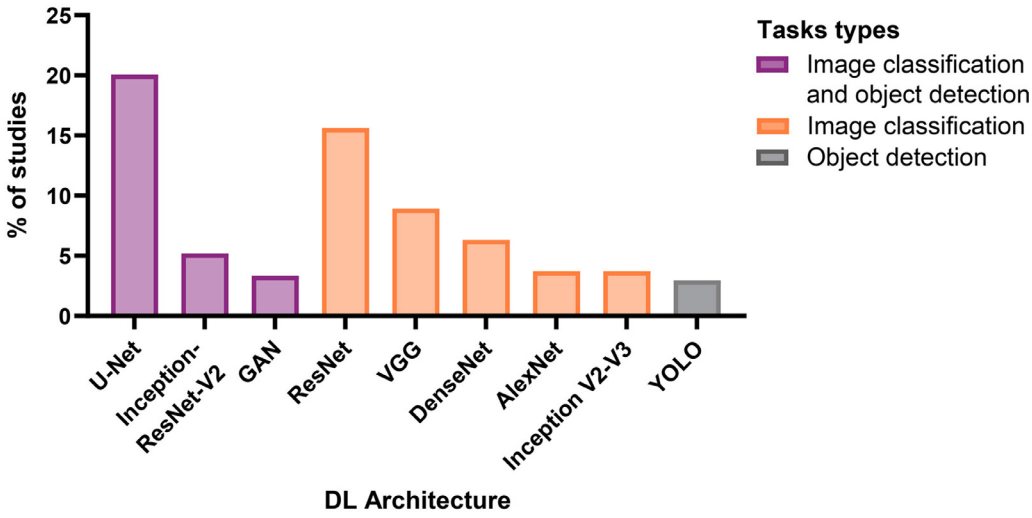

Fig. 8. Distribution of the most commonly investigated DL network architectures in the studies included in the review investigating DL network structure for computer vision tasks in medical imaging for spine care (n = 307 studies).

Table 5
List of the main performance metrics with corresponding equations reported in the studies focusing on deep learning in the field of medical imaging for spine care investigated in this review.

Metric group	Abbr.	Name	Equation	Description	% use
Probabilistic	Se	Sensitivity [66,67]	$Se = \frac{TP}{TP+FN}$	True positive detection capabilities. Math equivalent: Recall, TPR	49%
	Sp	Specificity [66,67]	$Sp = \frac{TN}{TN+FP}$	Capabilities for correctly identifying true negative classes. Math equivalent: TNR	41%
	Acc	Accuracy [66]	$Acc = \frac{TP+TN}{TP+TN+FP+FN}$	Total number of correct predictions, compared to the total number of predictions	48%
	ROC	Receiver operator characteristics [66]	*Line plot showing performance with different discrimination thresholds*	Line plot of the diagnostic ability of a classifier through TPR against FPR	10%
	AUC	Area under receiver operator characteristics [68]	$AUC = \frac{1}{2}(\frac{FP}{FP+TN} + \frac{FN}{FN+TP})$	Area under the simple trapezoid	36%
	PPV	Positive predictive value [69]	$PPV = \frac{TP}{TP+FP}$	Amount of true diagnosis with respect to true positive test. Math equivalent: precision	11%
	KAP	Cohen Kappa Coefficient [66]	$fc = \frac{(TN+FN)(TN+FP)+(FP+TP)(FN+TP)}{TP+TN+FN+FP}$ $KAP = \frac{(TN+FN)-fc}{(TP+TN+FN+FP)-fc}$	Measure of agreement between annotated and predicted Classifications or predicted and ground truth segmentation	
F-measure based metrics	DSC	Dice Similarity Coefficient or Sorensen-Dice Index [66,67,69]	$DSC = F1 = \frac{2TP}{2TP+FP+FN}$	Amount of pixel overlap over the total number of pixels in predicted and ground truth segmentation	24%
	F1	F1 score			
	JACC IoU	Jacca rd Index [66,67,69] Intersection-over-Union	$JAC = IoU = \frac{TP}{TP+FP+FN}$	Amount of pixel overlap divided by their union	8%
Spatial overlap and distance	PREC	Precision [69]	$PREC = \frac{\lvert A \cap B \rvert}{\lvert B \rvert}$	Amount of predicted pixel overlap over with respect to ground truth	14%
	AHD	Average Hausdorff distance/Max. Symmetric Surface Distance [66,70]	$d(A,B) = \frac{1}{N}\sum_{a \in A}\min_{b \in B}\lvert\lvert a-b \rvert\rvert$ $AHD(A,B) = \max(d(A,B), d(B,A))$ in which A and B represent the ground truth and predicted segmentation, respectively, and ||a-b|| represents a distance function like Euclidean distance	Spatial distance average over predicted and ground truth points	4%

Notes: TP, true positive; TN, true negative; FP, false positive; FN, false negative; and A and B represent the ground truth and predicted segmentation.

velopment, final trained models, or both (Table 6). In addition, 5% of studies used proprietary DL models or commercial prototypes.

Discussion

This scoping review synthesizes the recent literature on the use of DL techniques combined with medical imaging for spine clinical applications, highlighting current research and applications in clinical practice. Most of the included studies were observational clinical research and used existing datasets. For the most part, studies focused on the benefits of DL in combination with medical imaging to improve the detection and diagnosis of spinal conditions, such as inflammatory conditions, degenerative disease, and spinal deformity. Various DL methods and architectures were explored, and some studies proposed novel ones. Many DL approaches showed promising performance, demonstrating the potential of DL in the management of spinal conditions to improve the efficiency of clinical care and research.

Overall quality of the studies

The quality and robustness of the DL models and possible clinical implications heavily rely on the quality of the research and the input datasets, which governs the extent to which its findings can be trusted. The broad objective of this scoping review was to include studies from peer-reviewed journals and other sources such as preprint servers and conference proceedings. The peer review process is designed to critically assess the relevance of new research as well as to insure the appropriate study design and data analysis. Therefore, peer-reviewed studies are usually assumed to have at least a minimum acceptable quality. While peer-reviewed publication of completed results remains the primary goal of most medical research, a substantial number of abstracts will not be published in the medical literature. Therefore, including only articles published in peer-review journals would have limited our ability to broadly review and comprehensively map the clinical applications of DL combined with medical imaging since 30% of the included studies were obtained from outside this publication process. Nevertheless, only 6% of included conference abstracts went on to peer-reviewed publication, highlighting the need to use caution before integrating the results of these abstracts into research or clinical practice. Thus, the contribution of conference abstracts in creating a solid body of evidence required for clinical implementation remains uncertain.

Clear and transparent reporting is crucial in assessing a study's quality. Nevertheless, only 4% of the included peer-reviewed articles adhered to a reporting checklist. The quality assessment of the studies included in this review was beyond the scope of this investigation but would be an interesting area for future research. Nevertheless, the data extracted from the included studies allowed us to make a general quality assessment of the eligible studies based on checklists developed explicitly for AI in medical imaging based (CLAIM, MI-CLAIM).[5251] Our review suggests that many studies would be classified as "incomplete" for numerous checklist items, reflecting potential methodological limitations. In many cases, the study design was incomplete; 60% of studies did not detail the cohorts' characteristics. Similar observations can be made for data and optimization items (21% did not describe the origin of the data, 47% did not report the split of the dataset) and reproducibility and transparency reporting (93% failed to provide the code used for model development or final trained models). Similar to our experience, previous medical studies focusing on AI have also noted a lack

Table 6
A short list of available code or DL platforms used in the methodology or provided as a result in the published studies investigated in this review focusing on DL in the field of medical imaging for applications intended for spine clinical care.

Name	Summary	Implementation
SpineCube [29]	Intelligent agent for diagnosing scoliosis and evaluating the severity of scoliosis.	https://github.com/js3611/Deep-MRI-Reconstruction
[71,72]	Direct automated quantitative measurement of the spine by cascade amplifier regression network with manifold regularization	https://github.com/pangshumao/CARN
[73]	Testing DL model for automated detection of vertebral fractures of the lumbar spine	https://links.lww.com/CORR/A505
VFADL [74]	Source code for automated identification of vertebral fractures at VF assessment performed with dual-energy x-ray absorptiometry	https://github.com/DougUC/VFADL-PUBLIC/blob/master/VFADL.ipynb
BMDC-Net [75]	Method for the qualitative detection of BMD (normal bone mass, low bone mass, and osteoporosis) via diagnostic CT slices	https://github.com/tangchao1010/classification-of-BMD
DMML-Net [76]	Deep multiscale multitask learning network to directly localize all lumbar organs with bounding boxes and grade all lumbar organs with crucial differential diagnoses (normal and abnormal).	https://github.com/zhyhan/DMML-Net/tree/master
[77]	Fully automated algorithm for the detection of bone marrow edema lesions in patients with axial spondyloarthritis	https://github.com/krzysztofrzecki/bone-marrow-oedema-detection
[78]	DL model for detection of cervical spinal cord compression in MRI scans.	https://github.com/zamirmerali/dcm-mri
MBNET [79]	Multi-task deep neural network with supervised learning applied for 2 tasks, semantic segmentation and parameter inspection for the diagnosis of lumbar vertebrae	https://github.com/LuanTran07/BiLUnet-Lumbar-Spine
LEN-LCN [80]	Implementation code of automated Landmark Estimation and Correction Network to estimate landmarks on lateral X-rays.	https://github.com/LuanTran07/BiLUnet-Lumbar-Spine
DeepSeg [81]	Fully-automatic framework for segmentation of the spinal cord and intramedullary multiple sclerosis lesions from conventional MRI data using CNN	https://github.com/spinalcordtoolbox/spinalcordtoolbox/tree/master/spinalcordtoolbox/deepseg
[82,83]	Scripts for image segmentation using CNN to segment bones in US images automatically	https://github.com/SlicerIGT/aigt
VerteSeg [84]	Code for automatic segmentation of vertebrae from sagittal IDEAL (Iterative Decomposition of water and fat with Echo Asymmetric and Least-squares estimation) spine MR images	https://github.com/zhoji/verteseg
[85]	Deep CNN model to classify osteopenia and osteoporosis using lumbar spine X-ray images.	https://github.com/zhang-de-lab/zhang-lab/tree/master/osteoporosis
[86]	Model developed using NiftyNet for neck muscle segmentation.	https://github.com/kennethaweberii/Neck_Muscle_Segmentation
[87,88]	Automatic landmark estimation and spinal curvature estimation for adolescent idiopathic scoliosis	https://github.com/zc402/Scoliosis
SpineAI [89]	Implementation code to automatically detect and classify lumbar spinal stenosis on MRI images.	https://github.com/NUHS-NUS-SpineAI/SpineAI-Detect-Classify-LumbarMRI-Stenosis
[90]	Model for identifying fresh VCF from digital radiography.	https://github.com/TXVision/DR_Fracture_Classification
[91]	Code for lumbar spine hanging protocol label lumbar spine views/positions, detect hardware and rotate the lateral views to straighten the image.	https://github.com/GeneKitamura/L_spine_hanging_protocol
Anduin [56,92]	Freely available research tool to segment vertebrae in a CT scan and to assess various bone measures in clinical CT.	anduin.bonescreen.de
Spinal Cord Toolbox [93,94]	Open-source set of command-line tools dedicated to the processing and analysis of spinal cord MRI data.	https://github.com/spinalcordtoolbox/spinalcordtoolbox
Nora Imaging [95]	Web-based framework using CNN for medical image analysis	http://www.nora-imaging.org
NiftyNet [96]	Open source CNN platform for medical image analysis.	http://niftynet.io
Modified NiftyNet [97]	Monai and NiftyNet version of the code to generate the multi-organ segmentation of the head and neck area	https://github.com/elitap/NiftyNet
DLTK [98]	NN toolkit written in Python to enable fast prototyping with a low entry threshold for medical imaging	https://github.com/DLTK/DLTK
V-Net [99]	3D image segmentation based on a volumetric, fully CNN.	https://github.com/faustomilletari/VNet
SegNet [100]	Deep fully CNN architecture for semantic pixel-wise segmentation	http://mi.eng.cam.ac.uk/projects/segnet/
SpineNet [101]	CNN backbone with scale-permuted intermediate features and cross-scale connections learned on an object detection task by Neural Architecture Search.	https://github.com/lucifer443/SpineNet-Pytorch/tree/a7059eff295dcee16d719b381f80af8eb3fe42f6
SpineTK [102]	Code to train a network for doing MR, CT, and X-ray image annotation, including landmark annotation of 6 keypoints on individual vertebral bodies for vertebral height measurement	https://github.com/abhisuri97/SpineTK
[103]	Source code for deep residual learning for multi-class robotic tool segmentation	https://github.com/warmspringwinds/tf-image-segmentation

of data reporting and poor model transparency [18,30]. Implementing a standardized mandatory checklist into the DL peer-review process, as it is currently done with the STROBE checklist for human observational studies [31], could help enhance the quality of the published studies and improve model reproducibility and comparison [18]. In turn, improving the quality of the research and robustness of the DL models may accelerate their implementation into clinical practice.

Datasets and DL reliability

Data quality and availability are significant determinants of models' performance and reliability and have been recognized as a fundamental challenge to developing DL for medical imaging [32]. In the current review, issues similar to previously reported limitations regarding the data's quality [32] were raised, including imbalanced data, lack of adequately annotated data, and limited confidence intervals. While the correct sample size required to train a DL model to perform adequately is challenging to estimate in advance, the reported datasets seem very limited compared to datasets for general computer vision tasks, which typically range from a hundred thousand to millions of annotated pictures [33]. One likely explanation for the difference in dataset size is the limited number of samples and patients currently available in the public databases for medical imaging tasks compared to public databases available for general computer vision tasks. Nevertheless, the studies included in this review commonly reported good DL performance despite potential issues related to data quantity. Still, it remains unclear how well the final DL models perform their task regarding over-fitting to their training datasets.

DL model performance

Most of the studies included in this review (91%) evaluated several DL pipelines or multiple types or DL models. The subsequent identification of the best model was usually based on comparing their performance using various metrics, predominantly calculated by comparing DL predictions against reference data obtained from human observers, as it is common practice in AI [20]. The predominantly used performance metrics were probabilistic measurements, including accuracy, sensitivity and specificity, and AUC. However, these metrics have considerable limitations and cannot be considered reliable in some situations frequently encountered in the reviewed studies. The main limitation of their use was the label imbalance observed in most datasets that included diseased patients. Using accuracy as an indicator of performance with an unbalanced dataset may artificially improve the performances due to this sensitivity of accuracy to the prevalence of positive diagnosis in a dataset and the tendency of this performance metric to favor the majority class [34]. These studies were then prone to positive-negative class bias and misleading models' performances in such situations due to limited pre-test probability assessment [35]. Acknowledging potential bias or study limitations is vital for accurate result interpretation, especially when claims are made regarding clinical care. Nevertheless, it is unclear if the studies included in the review accounted for the label class imbalance, specifically the number of healthy and diseased or positives and negatives in the dataset.

Clinical implementation and ethics

Caution is needed when developing DL methodologies for clinical practice. Before clinical implementation, external validation and replication of the DL models' performances should be completed. Compared to internal validation, external validation allows a more robust demonstration of the clinical utility of the methodology. Nevertheless, only a minority of studies (8%) investigated pre-existing DL models, and few of the remaining studies (15%), which were developing DL models, externally validated the completed DL pipeline on a dataset distinct from their training dataset. Although a tool may appear promising in a particular setting, they are unlikely to perform the same after being deployed into different spine clinical care settings, particularly if employed across different patient populations. Nevertheless, very few studies provided information or demonstrated the use of DL techniques in real-world situations, suggesting that further consideration and research are required to test such models' clinical utility and applicability. For all the reasons mentioned earlier, the field of DL combined with medical imaging for spine clinical applications does not appear ready for widespread clinical recognition and remains in its development phase. As such, DL does not replace other research or analytic approaches; instead, it can potentially add value to the available tools for spine clinical care research. Partnerships between clinicians and data science experts are essential to ensure the clinical utility of the DL models developed for healthcare.

Future research directions and conclusions

The DL and medical imaging for spine clinical care is an emerging reseach field with exciting recent developments with the potential to improve patient care. This review of 365 studies showed that problem-specific DL models could significantly improve the detection and diagnosis of spinal conditions on medical imaging. While it is evident that DL is unlikely to replace radiologists or other health experts in the near future, it holds the potential to be an efficient tool to decrease the clinical burden of radiologists and clinicians. Though less frequently investigated, research into other applications of DL, such as clinical decision support, assessment, and outcome prediction, has demonstrated initial positive results. Nevertheless, the available studies on these topics are currently limited, and further research is required to identify additional benefits of DL for spine clinical care. The analysis of the included studies highlighted the following needs for further studies or improvement: 1) commitment to data and model transparency, 2) reproducibility and generalizability improvement of DL models, 3) performing external and comprehensive validation of the proposed methodologies on different datasets, and 4) establishing of DL ethics guidelines at all levels of DL development. In addition, efforts to improve research methodologies and the impact of DL on patients should be better considered. Furthermore, standard imaging protocols, agreed-upon datasets to perform DL models' benchmarking, standardized performance metrics, and unbiased accuracy indicators appeared to be lacking in the current literature. Such standards would improve the quality of AI research and allow for better clinical implementation and further advances in the field of DL for spinal imaging.

The generation of large spine datasets combined with DL tools accessible to researchers and clinicians is also needed to support the development of novel DL applications and improve the current spine clinical care models. More work is needed to define best practices with DL tools to guide clinical-decision making process for spine clinical care and to facilitate eventual clinical implementation.

Declarations of Competing Interest

The authors have no financial or professional relationships to declare.

Funding

This work was supported by the following relationship. Nevertheless, all relationships are nonfinancial except professor or student salary but not directly related to this work. Work supported by the TransMedTech Institute and its main financial partner, the Apogee Canada First Research Excellence Fund, the NSERC/Medtronic Industrial Research Chair in Spine Biomechanics, Quattrone-Foderaro Grant in AI Innovation in Orthopedic Surgery 2021, Mayo Foundation, the Mayo Clinic, and the AO Foundation.

Acknowledgments

The author would like to show gratitude to Dr Marchionatti E., for her help with the research methodology and technical assistance.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.xnsj.2023.100236.

References

[1] Rohrig B, du Prel JB, Wachtlin D, et al. Types of study in medical research: part 3 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2009;106:262–8.
[2] Andersson G, Frymoyer J. The adult spine: principles and practice. The adult spine: principles and practice. Philadelphia: Lippincott-Raven 1997:93–141.
[3] Lo J, Chan L, Flynn S. A systematic review of the incidence, prevalence, costs, and activity and work limitations of amputation, osteoarthritis, rheumatoid arthritis, back pain, multiple sclerosis, spinal cord injury, stroke, and traumatic brain injury in the United States: a 2019 update. Arch Phys Med Rehabil 2021;102:115–31.
[4] Diebo BG, Shah NV, Boachie-Adjei O, et al. Adult spinal deformity. Lancet 2019;394:160–72.
[5] Ma VY, Chan L, Carruthers KJ. Incidence, prevalence, costs, and impact on disability of common conditions requiring rehabilitation in the United States: stroke, spinal cord injury, traumatic brain injury, multiple sclerosis, osteoarthritis, rheumatoid arthritis, limb loss, and back pain. Arch Phys Med Rehabil 2014;95:986–995.e981.
[6] Kim GU, Chang MC, Kim TU, et al. Diagnostic modality in spine disease: a review. Asian Spine J 2020;14:910–20.
[7] Brady AP. Error and discrepancy in radiology: inevitable or avoidable? Insights Imaging 2017;8:171–82.
[8] Briggs GM, Flynn PA, Worthington M, et al. The role of specialist neuroradiology second opinion reporting: is there added value? Clin Radiol 2008;63:791–5.

[9] Sun J, Wu D, Wang Q, et al. Pedicle screw insertion: is O-arm–based navigation superior to the conventional freehand technique? A systematic review and meta–analysis. World Neurosurg 2020;144:e87–99.
[10] Hosny A, Parmar C, Quackenbush J, et al. Artificial intelligence in radiology. Nat Rev Cancer 2018;18:500–10.
[11] Chang M, Canseco JA, Nicholson KJ, et al. The role of machine learning in spine surgery: the future is now. Front Surg 2020;7:54.
[12] Samuel AL. Some studies in machine learning using the game of checkers. IBM J Res Dev 2000;44:206–26.
[13] Schmidhuber J. Deep learning in neural networks: an overview. Neural networks 2015;61:85–117.
[14] Chartrand G, Cheng PM, Vorontsov E, et al. Deep learning: a primer for radiologists. Radiographics 2017;37:2113–31.
[15] Erickson BJ, Korfiatis P, Akkus Z, et al. Toolkits and libraries for deep learning. J Digital Imaging 2017;30:400–5.
[16] Huang J, Shlobin NA, DeCuypere M, et al. Deep learning for outcome prediction in neurosurgery: a systematic review of design, reporting, and reproducibility. Neurosurgery 2022;90:16–38.
[17] Azimi P, Yazdanian T, Benzel EC, et al. A review on the use of artificial intelligence in spinal diseases. Asian Spine J 2020;14:543–71.
[18] Smets J, Shevroja E, Hügle T, et al. Machine learning solutions for osteoporosis-a review. J Bone Miner Res 2021;36:833–51.
[19] Ong W, Zhu L, Zhang W, et al. Application of artificial intelligence methods for imaging of spinal metastasis. Cancers (Basel) 2022;14:4025.
[20] Vrtovec T, Ibragimov B. Spinopelvic measurements of sagittal balance with deep learning: systematic review and critical evaluation. Eur Spine J 2022;31:2031–45.
[21] Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Int J Surg 2010;8:336–41.
[22] Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169:467–73.
[23] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Commun ACM 2017;60:84–90.
[24] Mathes T, Pieper D. An algorithm for the classification of study designs to assess diagnostic, prognostic and predictive test accuracy in systematic reviews. Syst Rev 2019;8:226.
[25] Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Bmj 2015;350:g7594.
[26] Schwendicke F, Singh T, Lee JH, et al. Artificial intelligence in dental research: Checklist for authors, reviewers, readers. J Dent 2021;107:103610.
[27] Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 2016;18:e323.
[28] Bossuyt PM, Reitsma JB, Bruns DE, et al. [Reporting studies of diagnostic accuracy according to a standard method; the Standards for Reporting of Diagnostic Accuracy (STARD)]. Ned Tijdschr Geneeskd 2003;147:336–40.
[29] Yang J, Zhang K, Fan H, et al. Development and validation of deep learning algorithms for scoliosis screening using back images. Commun Biol 2019;2:390.
[30] Ferizi U, Honig S, Chang G. Artificial intelligence, osteoporosis and fragility fractures. Curr Opin Rheumatol 2019;31:368–75.
[31] von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med 2007;4:e296.
[32] Minaee S, Kafieh R, Sonka M, et al. Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning. Med Image Anal 2020;65:101794.
[33] Apostolopoulos ID, Mpesiana TA. Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks. Phys Eng Sci Med 2020;43:635–40.
[34] Tecle N, Teitel J, Morris MR, et al. Convolutional Neural Network for Second Metacarpal Radiographic Osteoporosis Screening. J Hand Surg 2020;45:175–81.
[35] Branco P, Torgo L, Ribeiro RP. A survey of predictive modeling on imbalanced domains. ACM Computing Surveys (CSUR) 2016;49:1–50.
[36] Manjer J, Carlsson S, Elmståhl S, et al. The Malmö Diet and Cancer Study: representativity, cancer incidence and mortality in participants and non-participants. Eur J Cancer Prev 2001;10:489–99.
[37] Evaniew N, Cadotte DW, Dea N, et al. Clinical predictors of achieving the minimal clinically important difference after surgery for cervical spondylotic myelopathy: an external validation study from the Canadian Spine Outcomes and Research Network. J Neurosurg Spine 2020:1–9.
[38] Rudwaleit M, Haibel H, Baraliakos X, et al. The early disease stage in axial spondylarthritis: results from the German Spondyloarthritis Inception Cohort. Arthritis Rheum 2009;60:717–27.
[39] Hu XJ, Chen LH, Battié MC, et al. Methodology and cohort profile for the Hangzhou Lumbar Spine Study: a study focusing on back health in a Chinese population. J Zhejiang Univ Sci B 2018;19:547–58.
[40] Lee C, Choe EK, Choi JM, et al. Health and Prevention Enhancement (H-PEACE): a retrospective, population-based cohort study conducted at the Seoul National University Hospital Gangnam Center, Korea. BMJ Open 2018;8:e019327.
[41] Stern D, Likar B, Pernuš F, et al. Parametric modelling and segmentation of vertebral bodies in 3D CT and MR spine images. Phys Med Biol 2011;56:7505–22.
[42] Ibragimov B, Likar B, Pernuš F, et al. Shape representation for efficient landmark-based segmentation in 3-d. IEEE Trans Med Imaging 2014;33:861–74.
[43] Leslie WD, Caetano PA, Macwilliam LR, et al. Construction and validation of a population-based bone densitometry database. J Clin Densitom 2005;8:25–30.
[44] Berglund G, Nilsson P, Eriksson KF, et al. Long-term outcome of the Malmö preventive project: mortality and cardiovascular morbidity. J Intern Med 2000;247:19–29.
[45] Shardell M, Parimi N, Langsetmo L, et al. Comparing analytical methods for the gut microbiome and aging: gut microbial communities and body weight in the osteoporotic fractures in men (MrOS) study. J Gerontol A Biol Sci Med Sci 2020;75:1267–75.
[46] Rikkonen T, Salovaara K, Sirola J, et al. Physical activity slows femoral bone loss but promotes wrist fractures in postmenopausal women: a 15-year follow-up of the OSTPRE study. J Bone Miner Res 2010;25:2332–40.
[47] Salovaara K, Tuppurainen M, Kärkkäinen M, et al. Effect of vitamin D(3) and calcium on fracture risk in 65- to 71-year-old women: a population-based 3-year randomized, controlled trial–the OSTPRE-FPS. J Bone Miner Res 2010;25:1487–95.
[48] Poddubnyy D, Sieper J, Akar S, et al. Characteristics of patients with axial spondyloarthritis by geographic regions: PROOF multicountry observational study baseline results. Rheumatology (Oxford) 2022;61:3299–308.
[49] Tsolinas RE, Burke JF, DiGiorgio AM, et al. Transforming Research and Clinical Knowledge in Spinal Cord Injury (TRACK-SCI): an overview of initial enrollment and demographics. Neurosurg Focus 2020;48:E6.
[50] Ishimoto Y, Yoshimura N, Muraki S, et al. Prevalence of symptomatic lumbar spinal stenosis and its association with physical performance in a population-based cohort in Japan: the Wakayama Spine Study. Osteoarthritis Cartilage 2012;20:1103–8.
[51] Modarresi S, MacDermid JC, Suh N, et al. How is the probability of reporting various levels of pain 12 months after noncatastrophic injuries associated with the level of peritraumatic distress? Clin Orthop Relat Res 2022;480:226–34.
[52] Yao J, Burns JE, Forsberg D, et al. A multi-center milestone study of clinical vertebral CT segmentation. Comput Med Imaging Graph 2016;49:16–28.
[53] Ibragimov B, Korez R, Likar B, et al. Segmentation of pathological structures by landmark-assisted deformable models. IEEE Trans Med Imaging 2017;36:1457–69.
[54] De Leener B, Fonov VS, Collins DL, et al. PAM50: Unbiased multimodal template of the brainstem and spinal cord aligned with the ICBM152 space. Neuroimage 2018;165:170–9.
[55] Wu H, Bailey C, Rasoulinejad P, et al. In: Automatic landmark estimation for adolescent idiopathic scoliosis assessment using BoostNet. Quebec, Canada: Springer; 2017. p. 127–35.
[56] Sekuboyina A, Bayat A, Husseini ME, et al. Verse: a vertebrae labelling and segmentation benchmark. arXiv org e-Print archive 2020;73:102166. arXiv preprint arXiv: 2001.09193.
[57] Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 2013;26:1045–57.
[58] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 2014. https://arxiv.org/abs/1409.1556.
[59] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. Medical image computing and computer-assisted intervention – MICCAI 2015. MICCAI 2015. Lecture notes in computer science, 9351. Cham: Springer; 2015. doi:101007/978-3-319-24574-4_28.
[60] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV: IEEE Computer Society; 2016. p. 770–8.
[61] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas, NV: IEEE Computer Society; 2016. p. 2818–26.
[62] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning Thirty-first AAAI conference on artificial intelligence. San Francisco, CA: AAAI Press; 2017.
[63] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV: IEEE Computer Society; 2016. p. 779–88.
[64] Redmon J, Farhadi A. Yolov3: an incremental improvement. arXiv preprint arXiv:180402767 2018. https://arxiv.org/abs/1804.02767.
[65] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 4700–8.
[66] Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging 2015;15:29.
[67] Popovic A, de la Fuente M, Engelhardt M, et al. Statistical validation metric for accuracy assessment in medical image segmentation. Int J Computer Assisted Radiol Surg 2007;2:169–81.
[68] Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. 2020. arXiv preprint arXiv:201016061 https://arxiv.org/abs/2010.16061.
[69] Yeghiazaryan V, Voiculescu I. Family of boundary overlap metrics for the evaluation of medical image segmentation. *J Med Imaging* (Bellingham) 2018;5:015006.
[70] Heimann T, Bv Ginneken, Styner MA, et al. Comparison and evaluation of methods for liver segmentation from CT datasets. IEEE Trans Med Imaging 2009;28:1251–65.
[71] Pang S, Su Z, Leung S, et al. Direct automated quantitative measurement of spine by cascade amplifier regression network with manifold regularization. Med Image Anal 2019;55:103–15.
[72] Pang S, Leung S, Nachum IB, et al. Direct automated quantitative measurement of spine via cascade amplifier regression network. arXiv preprint arXiv:180605570 2018. https://arxiv.org/abs/1806.05570.
[73] Li YC, Chen HH, Horng-Shing Lu H, et al. Can a deep-learning model for the automated detection of vertebral fractures approach the performance level of human subspecialists? Clin Orthop Relat Res 2021;479:1598–612.

[74] Derkatch S, Kirby C, Kimelman D, et al. Identification of vertebral fractures by convolutional neural networks to predict nonvertebral and hip fractures: a Registry-based cohort study of dual X-ray absorptiometry. Radiology 2019;293:405–11.
[75] Tang C, Zhang W, Li H, et al. CNN-based qualitative detection of bone mineral density via diagnostic CT slices for osteoporosis screening. Osteoporos Int 2021;32:971–9.
[76] Han Z, Wei B, Leung S, et al. Automated pathogenesis-based diagnosis of lumbar neural foraminal stenosis via deep multiscale multitask learning. Neuroinformatics 2018;16:325–37.
[77] Rzecki K, Kucybała I, Gut D, et al. Fully automated algorithm for the detection of bone marrow oedema lesions in patients with axial spondyloarthritis – Feasibility study. Biocybernetics Biomed Eng 2021;41:833–53.
[78] Merali Z, Wang JZ, Badhiwala JH, et al. A deep learning model for detection of cervical spinal cord compression in MRI scans. Sci Rep 2021;11:10473.
[79] Tran VL, Lin H-Y, Liu H-W. MBNet: a multi-task deep neural network for semantic segmentation and lumbar vertebra inspection on X-ray images, in computer vision – ACCV 2020 2021;635-651.
[80] Yang G, Fu X, Xu N, et al. A Landmark estimation and correction network for automated measurement of sagittal spinal parameters. In: Yang H, Pasupa K, Leung, ACS, Kwok JT, Chan JH, King I, editors. Neural information processing. ICONIP 2020. Communications in computer and information science, vol 1332. Cham: Springer; 2020.
[81] Gros C, De Leener B, Badji A, et al. Automatic segmentation of the spinal cord and intramedullary multiple sclerosis lesions with convolutional neural networks. Neuroimage 2019;184:901–15.
[82] Ungi T, Greer H, Sunderland KR, et al. Automatic spine ultrasound segmentation for scoliosis visualization and measurement. IEEE Trans Biomed Eng 2020;67:3234–41.
[83] Wu V, Ungi T, Sunderland K, et al. Automatic segmentation of spinal ultrasound landmarks with U-net using multiple consecutive images for input, 2021.
[84] Zhou J, Damasceno PF, Chachad R, et al. Automatic vertebral body segmentation based on deep learning of dixon images for bone marrow fat fraction quantification. Front Endocrinol (Lausanne) 2020;11:612.
[85] Zhang B, Yu K, Ning Z, et al. Deep learning of lumbar spine X-ray for osteopenia and osteoporosis screening: a multicenter retrospective cohort study. Bone 2020;140:115561.
[86] Weber KA 2nd, Abbott R, Bojilov V, et al. Multi-muscle deep learning segmentation to automate the quantification of muscle fat infiltration in cervical spine conditions. Sci Rep 2021;11:16567.
[87] Zhang C, Wang J, He J, et al. Automated vertebral landmarks and spinal curvature estimation using non-directional part affinity fields. Neurocomputing 2021;438:280–9.
[88] Wu H, Bailey C, Rasoulinejad P, et al. In: Automatic landmark estimation for adolescent idiopathic scoliosis assessment using BoostNet, 2017; 2017. p. 127–35.
[89] Hallinan JTPD, Zhu L, Yang K, et al. Deep learning model for automated detection and classification of central canal, lateral recess, and neural foraminal stenosis at lumbar spine MRI. Radiology 2021;300:130–8.
[90] Chen W, Liu X, Li K, et al. A deep-learning model for identifying fresh vertebral compression fractures on digital radiography. Eur Radiol 2021;32:1496–505.
[91] Kitamura G. Hanging protocol optimization of lumbar spine radiographs with machine learning. Skeletal Radiol 2021;50:1809–19.
[92] Loffler MT, Jacob A, Scharr A, et al. Automatic opportunistic osteoporosis screening in routine CT: improved prediction of patients with prevalent vertebral fractures compared to DXA. Eur Radiol 2021;31:6069–77.
[93] De Leener B, Lévy S, Dupont SM, et al. SCT: Spinal Cord Toolbox, an open-source software for processing spinal cord MRI data. Neuroimage 2017;145:24–43.
[94] Lemay A, Gros C, Zhuo Z, et al. Automatic multiclass intramedullary spinal cord tumor segmentation on MRI with deep learning. Neuroimage Clin 2021;31:102766.
[95] Wolf K, Reisert M, Beltrán SF, et al. Spinal cord motion in degenerative cervical myelopathy: the level of the stenotic segment and gender cause altered pathodynamics. J Clin Med 2021;10:3788.
[96] Gibson E, Li W, Sudre C, et al. NiftyNet: a deep-learning platform for medical imaging. Comput Methods Programs Biomed 2018;158:113–22.
[97] Tappeiner E, Pröll S, Hönig M, et al. Multi-organ segmentation of the head and neck area: an efficient hierarchical neural networks approach. Int J Comput Assist Radiol Surg 2019;14:745–54.
[98] Pawlowski N, Ktena SI, Lee MC, et al. Dltk: state of the art reference implementations for deep learning on medical images. arXiv preprint arXiv:171106853 2017.
[99] Milletari F, Navab N, Ahmadi SA; 2016 V-Net: fully convolutional neural networks for volumetric medical image segmentation:565–71. arXiv: 1606.04797.
[100] Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell 2017;39:2481–95.
[101] Du X, Lin TY, Jin P, et al. SpineNet: learning scale-permuted backbone for recognition and localization. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA: IEEE Xplore; 2020. p. 11589–98.
[102] Jamaludin A, Windsor R, Ather S, et al. Machine learning based berlin scoring of magnetic resonance images of the spine in patients with ankylosing spondylitis from the measure 1 study. Ann Rheum Dis 2020:40–1.
[103] Pakhomov D, Premachandran V, Allan M, et al. In: Deep residual learning for instrument segmentation in robotic surgery; 2019. p. 566–73.