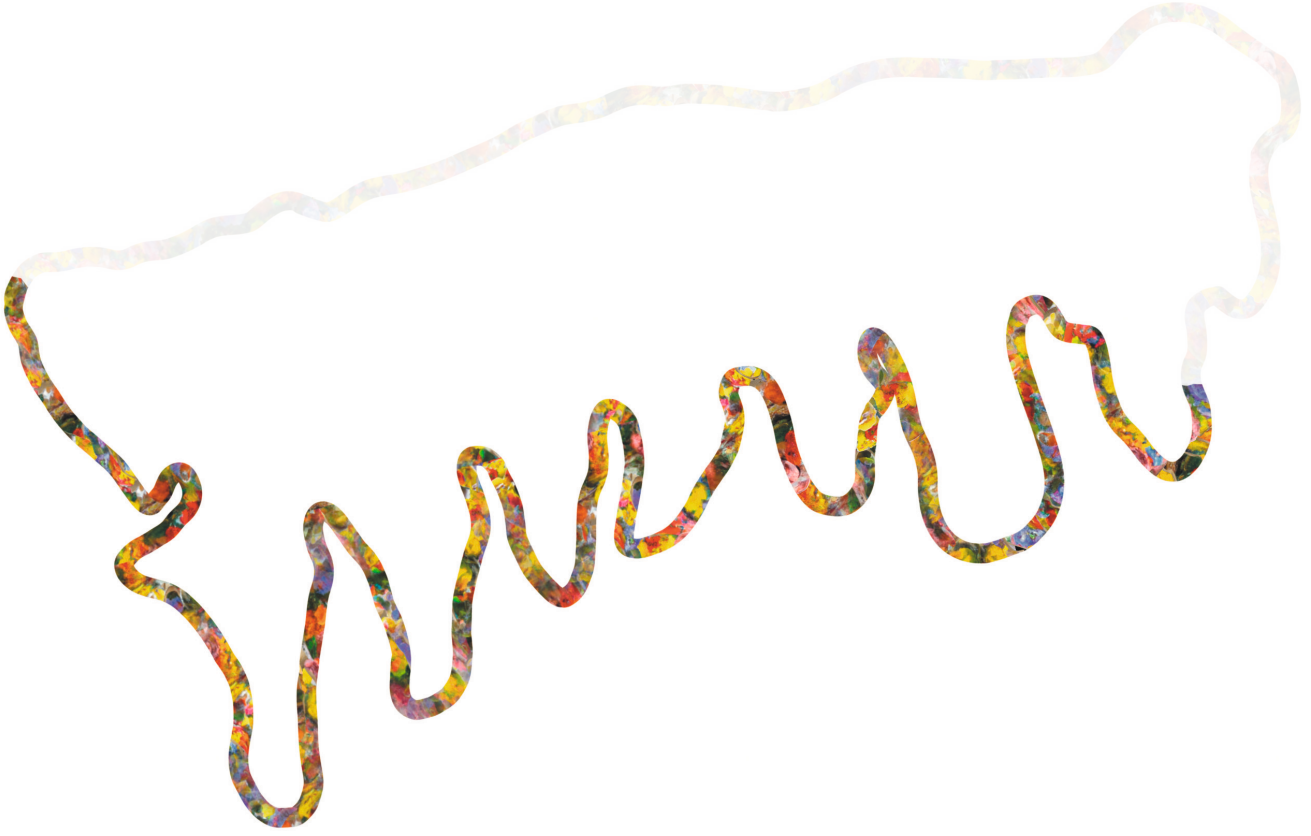
Outcome after anterior cervical discectomy: from inferential statistics to Machine Learning

Goedmakers, C.M.W.



Chapter 8

Machine Learning for Image Analysis in the Cervical Spine: Systematic Review of the Available Models and Methods

C.M.W Goedmakers[1,2], L.M. Pereboom[3], J.W. Schoones[4], M.L. de Leeuw den Bouter[5], R.F. Remis[6], M. Staring[7,8], C.L.A. Vleggeert-Lankamp[1,9]

Brain and Spine. 2022;2:101666.

[1]Department of Neurosurgery Haaglanden Medical Centre and HAGA teaching hospitals, the Hague, the Netherlands, [2]Computational Neuroscience Outcomes Center, Department of Neurosurgery, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, [3]Faculty of Mechanical, Maritime and Materials Engineering (3mE), Delft University of Technology, Delft, The Netherlands, [4]Walaeus Library, Leiden University Medical Center, Leiden, the Netherlands, [5]Delft Institute of Applied Mathematics, Department of Numerical Analysis, Delft University of Technology, Delft, The Netherlands, [6]Circuits and Systems Group, Microelectronics Department, Delft University of Technology, Delft, The Netherlands, [7]Department of Radiology, Leiden University Medical Center, Leiden, the Netherlands, [8]Intelligent Systems Department, Delft University of Technology, Delft, The Netherlands, [9]Department of Neurosurgery, Spaarne Gasthuis Haarlem/Hoofddorp, the Netherlands

Abbreviations

AAM = Active Appearance Model

ANN = Artificial Neural Network

ASM = Active Shape Model

ASM-M = Active Shape Model Mahalanobis Distance-Based

ASM-RRF = Active Shape Model random regression forest-based

ASM-RCF-AM = Active Shape Model Random Classification Forest-based with ArgMax

Aver. = Average

C = Cervical

COG = Center of Gravity

(C)RBM = (Conditional) Restricted Boltzmann Machines

CSF = Cerebrospinal fluid

DNN = Deep Neural Network

DC = Dice Coefficient

DO = Dice Overlap

DR = Detection Rate

FAST = FMRIB Automated Segmentation Tool, Part of FMRIB Software Library (FSL)

GC = Graphical Cut

GHT = Generalized Hough Transform

GLM = Grey-Level Model

GLV = Gray-Level Values

GM = Graphical Model

GT = Ground Truth

HMM = Hidden Markov Model

HD = Hausdorff Distance

HT = Hough Transform

IR = Identification Rate

IQR = Interquartile Rage

J-CNN = Joint learning model Convolutional Neural Network

KDE = Kernel Density Estimation

kNN = k-Nearest Neighbours

LE = Localization Error

MASD = Mean Absolute Surface Distance

MDCP = Mean Distance to the Closest Point

MRF = Markov Random Field

MSE = Root Mean Square Error

NLM = National Library of Medicine

NHANES II = Second National Health and Nutrition Examination Survey

PCA = Principe Component Analysis
QM = Quantitative Morphometry
RANSC = Random Sample Consensus
RCF = Random Classification Forrest
R-CNN = Region Based Convolutional Neural Networks
SC = Spinal Canal
Sens = Sensitivity
SiFC = Sparse intervertebral fence composition
SP = Shape Prior
Spec = Specificity
SRF = Structured Regression Forest
SSAE = Stacked Sparse Autoencoder
SSM = Statistical Shape Model
SVM = Support Vector Machine
TDCN = Transformed Deep Convolutional Neural Network
VolHOG = Histograms of oriented gradients for volumetric data
W(S) = Whole Spine

Introduction

Neck pain is the number four cause of physical disability worldwide, and it can be an important symptom in identifying degenerative pathologies of the cervical spine. In most cases, acute neck pain resolves without invasive treatment, but in nearly 50% of patients, the pain returns or develops a chronic nature. With the current ageing population and the relatively high prevalence of neck pain and spine disease, there is increasing demand on radiological image analysis in healthcare [18]. However, the analysis of those visualizations is time-consuming and is subject to significant interobserver variability [60]. Automating parts of the radiological image analysis process can support clinicians in providing a more accurate and consistent image assessment with increased time efficiency.

Over the last decade, the application of artificial intelligence (AI) in medical research has become increasingly popular. Machine Learning (ML) techniques show promise in computer aided diagnostics (CAD), specifically for clinical tasks related to detection and segmentation, as well as classification and prediction [23, 57, 59, 65, 67]. A ML algorithm is able to "learn," which means in this context that the algorithm can improve performance by previous experience or provided data to give a valid result for never-before-seen data, without being explicitly programmed to do so [32].

The majority of the available literature on image analysis concentrates on the thoracic and lumbar spine, while the cervical spine is studied less often. The difference can be partly attributed to the lower incidence of neck pain in the general population, compared to (lower) back pain [54]. Nevertheless, the neck is an essential part of the body with several vital anatomical structures whose functioning can be visualized using radiological imaging. Additionally, considering the relative novelty of the subject matter no systematic reviews have been published, while this could significantly improve the quality of future research on this topic.

Therefore, we aim to create the first overview of the available Machine Learning methods for image analysis of the cervical spine, while weighing and discussing their risks and benefits and providing recommendations for future research in this field. We will divide the systematic review into two sections, one focusing on ML for segmentation and the other on applying ML to automate the study different properties, such as segment mobility and curvature, of the cervical spine on radiological imaging. The overview provided in this systematic review may function as a reference for all authors conducting research on computer aided diagnostics of cervical spine disease.

Methods

Literature search

The initial literature search was performed in PubMed, EMBASE and Web of Science, on December 18th, 2020. Two of the authors (CG, LP) separately evaluated the articles by title, abstract and full text, when necessary, to select the studies that met the predefined selection criteria. As the topic of this review touches both the medical, and the technical research field, both points of view had

to be highlighted. Therefore, an additional literature search was performed in the Google Scholar, Scopus, SPIE Digital Library and IEEE Explore databases, to obtain as many articles as possible from both medical and technical journals. The search strategies used in the different databases were based on the search string as shown in Figure 1.

(("surgical procedures, operative"[majr] OR "operative surgical procedure"[tiab] OR "operative surgical procedures"[tiab] OR "surgery"[tiab] OR "surgical"[tiab] OR "surgeries"[tiab] OR "spinal surgery"[tiab] OR "spinal surgeries"[tiab] OR "operative procedures"[tiab] OR "operative procedure"[tiab] OR "Diagnostic Imaging"[majr] OR "Spine/diagnostic imaging"[majr] OR "Spinal Diseases/diagnostic imaging"[majr] OR "radiology"[tiab] OR "radiography"[tiab] OR "radiolog*"[tiab] OR "radiograph*"[tiab] OR "magnetic resonance imag*"[tiab] OR "MRI"[tiab] OR "MR"[tiab] OR "CT"[tiab] OR "computer tomograph*"[tiab] OR "computed tomograph*"[tiab] OR "x-ray imag*"[tiab]) AND ("Artificial intelligence"[majr] OR "artificial intelligence"[ti] OR "Machine Learning"[ti] OR "Algorithms"[majr] OR "Algorithm*"[ti] OR "computer-aided classification"[ti] OR "template matching"[ti] OR "interest point detection"[ti] OR "learning based"[ti] OR "Active learning"[ti] OR "Automatic segmentation"[ti] OR "segmentation"[ti] OR "detection"[ti] "computer-aided classification"[ti] OR "Deep Learning"[ti] OR "deep neural network*"[ti] OR "Hierarchical segmentation"[ti] OR "Machine Learning"[ti] OR "AI"[ti]) AND ("spine"[majr] OR "spine"[ti] OR "vertebral column"[ti] OR "vertebral columns"[ti] OR "spinal column"[ti] OR "spinal columns"[ti] OR "vertebra"[ti] OR "vertebrae"[ti] OR "intervertebral disk*"[ti] OR "intervertebral disc*"[ti]))

(Performed: 20-12-2020)

Figure 1. The search strategy used to perform the systematic search in the medical databases

Studies were included when they reported on a form of automated radiologic image analysis focusing on the human cervical spine or whole spine including the cervical vertebrae.

Studies were excluded if they met any of the following criteria: (1) Publications not written in English; (2) Conference abstracts; (3) Narrative reviews; (4) Cadaver studies without proven clinical application; (5) Phantom studies without proven clinical application; (6) Studies that describe a protocol without any form of analysis; (7) Studies on the thoracic or lumbar spine; (8) Studies on radiation dose, artifact reduction, sequence analysis or robotic surgery; (9) Studies on image processing without segmentation, landmarking or any other measurement on the spine involved.

Any discrepancy in selection between the reviewers was resolved in open discussion (CG, LP), and, if needed, a third reviewer was asked to make a final decision (CVL). Reference screening and citation tracking were performed on the identified articles. This systematic review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses: the PRISMA Statement [45].

Quality assessment

The methodological quality of all studies was assessed separately by two reviewers (CG, LP), using a version of the Modified New Castle – Ottawa Quality assessment scale for Cohort Studies [63].

If there was no consensus about the assessment, a third reviewer (CVL) was consulted. The New Castle – Ottawa scale was manually adjusted to better fit human to model comparison studies with a technical nature.

The items reviewed in the assessment were: 1.1 Representativeness of cohort; 1.2 Model selection, development and implementation; 1.3 Comparison made; 1.4 Ground truth assessment and Data extraction; 2. Applicability and Generalizability (data variability, semi-/fully-automatic, different modalities); 3.1 Outcome Assessment (clear split, ground truth objectified); 3.2 Outcome reporting (different outcome measures, uncertainty metrics reported); 3.3 Sharing (data or code sharing). All items could be awarded a maximum of 1 point, except for 'Applicability and Generalizability', for which a maximum of 2 points could be given. Studies could maximally be awarded 9 points. Studies were then divided into low (7-9 points), intermediate (5-6 points) or high (4 or less points) risk of bias.

Data Extraction

Data extraction for all included articles was performed by two reviewers separately (CG, LP) and any controversies were resolved by a third reviewer (CVL). From each article the following information was collected: year of publication, image modality, spine region, model description, degree of automation, number of images included, train to test set distribution and description of how the ground truth was acquired. The determination of the ground truth can be done by either one or more clinical experts and can be provided in different formats; i.e. bounding boxes, vertebra centers, or complete pixel-wise segmentations. Only outcomes that were mentioned in the text or tables of a publication were included into the analysis, as extracting outcomes from graphs was deemed too imprecise and time-intensive.

In order to compare model performance, commonly reported outcome measures were extracted from each publication. Outcomes were divided in either the internal comparison group; when the model's performance was compared to the ground truth, or the external comparison group; when the model's performance was compared to model performance from previous publications.

Outcomes of articles in the segmentation category were reported in five major groups:

- Accuracy: Accuracy, Identification Rate (IR), Detection Rate (DR)
- Error (mm): Localization error (LE), Mean Distance Closest Point (MDCP),
    - Mean Absolute Surface Distance (MASD), Point-to-Surface error,
    - Hausdorff Distance (HD), Center of Gravity (COG)
- Overlap: Dice Overlap (DO), Dice Coefficient (DC), Dice Index (DI)
- Time: Runtime, Efficiency
- Other: Precision, Sensitivity, Specificity

The aims of studies included into the second category, cervical spine analysis, can be divided up into five broad categories: 1. Biomechanical analysis; 2. CVM stage; 3. Clinical prognosis/prediction;

4. Image registration/Planning; 5. Clinical / Radiological Feature Detection. Additional variables collected for the second category articles were aim, included vertebrae and key points.

Results

Article selection

Through searching PubMed, EMBASE and Web of Science, using the predefined search strategy, 956 records could be identified. 654 remained after duplicates were removed. An added search in Google Scholar, Scopus, SPIE Digital Library and IEEE Explore yielded an additional 28 publications. The 682 unique records were screened for title and abstract, after which a total of 506 articles could be excluded. The full-texts of the remaining 176 articles were screened, and 125 did not fit all in- and exclusion criteria and were therefore removed. The remaining 51 articles were included in this systematic review and, based on their primary aim, divided into the two main categories; 1. Segmentation (n = 32) and 2. Cervical Spine Analysis (n = 19). The first category was then divided into two subcategories; 1.1 Conventional Machine Learning Segmentation (n = 20) and 1.2 Deep Learning Segmentation (n = 12) (Figure 2).

Where articles in the first subcategory focus more on the conventional Machine Learning methods for segmentation, studies in the second category deploy the relatively newer, neural networks. In the second category studies were included that did not necessarily focus on segmentation but in some other way analyzed the cervical spine and its radiologic characteristics.

The increasing popularity of Machine Learning for image analysis of the cervical spine is clearly illustrated when the number of included publications in this study is plotted against the year of publication in total and per subcategory (Figure 3, Figure 4). The majority of the included articles (n = 39) is published within the last 5 years.

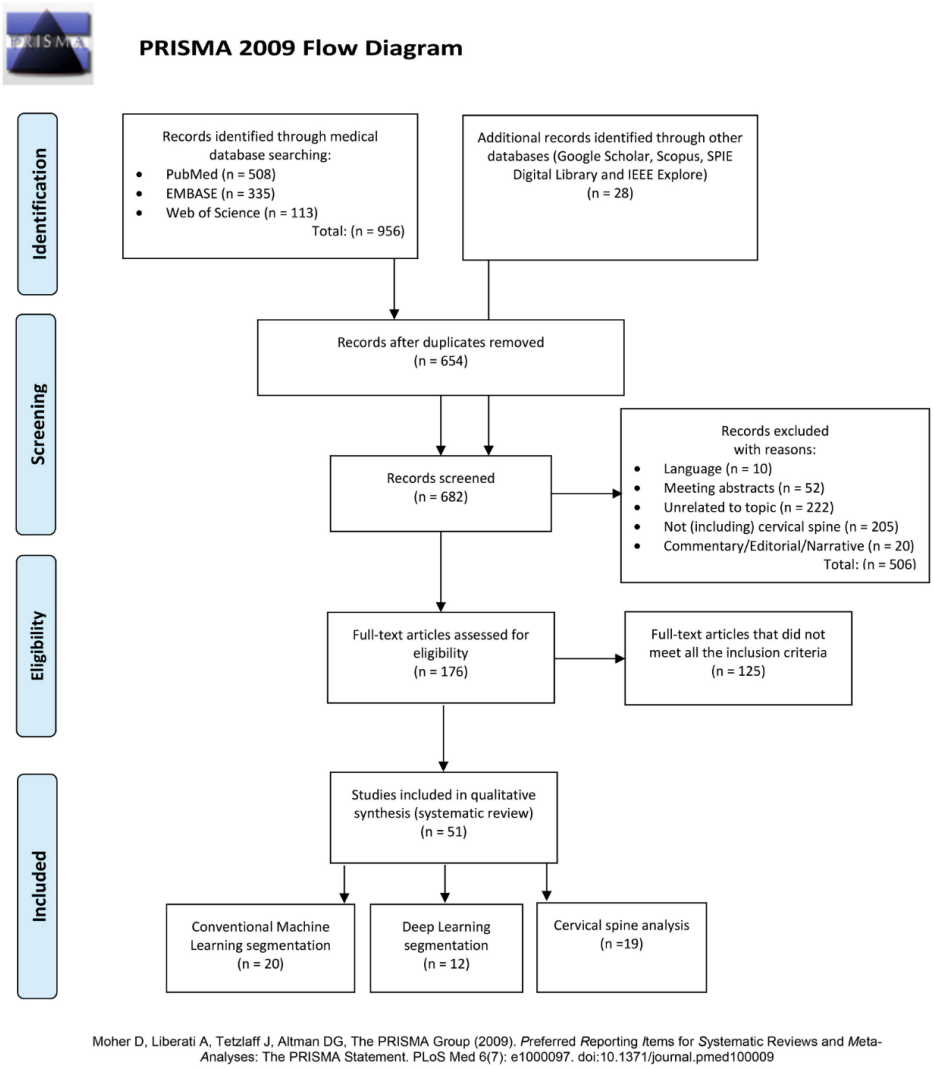


Figure 2. Flowchart illustrating the inclusion and exclusion process of articles

Quality Assessment

In the Conventional Machine Learning Segmentation group there was one study included with a high risk of bias, eleven studies with an intermediate risk of bias and eight with a low risk of bias (Appendix A). In the Deep Learning Segmentation group three studies showed intermediate risk of bias and nine a low risk of bias, while there were no studies included with a high risk of bias (Appendix B). Lastly, in the Cervical Spine Analysis group there was one study with a high, 14 with an intermediate and 4 with a low risk of bias (Appendix C).

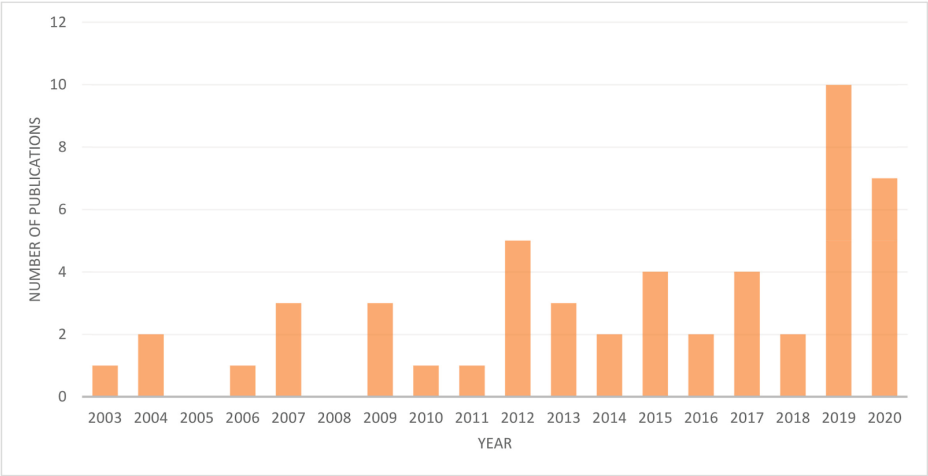


Figure 3. Number of publications plotted per year

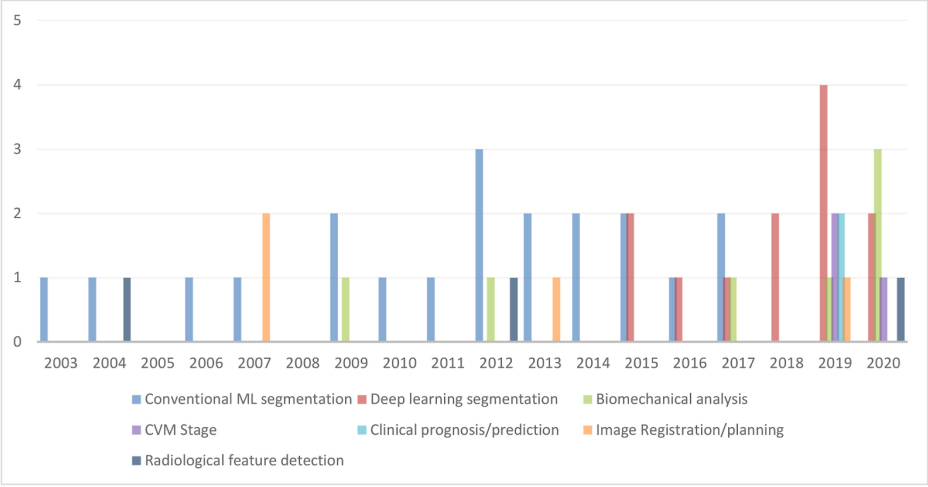


Figure 4. Number of publications per subcategory per year

In general it can be observed that the more recently studies were published, the more likely they were to have a decreased risk of bias. Therefore, the percentage of low risk of bias studies in the Deep Learning Segmentation group is higher than in the Conventional Machine Learning Segmentation group, as the latter includes more recent studies. The same pattern - a decreased risk of bias over time - can be observed in the Cervical Spine Analysis group.

Qualitative synthesis

Conventional Machine Learning Segmentation Techniques

The total number of included studies involving Conventional Machine Learning segmentation techniques is 20, of which 6 studies focused on X-ray images, 6 on MR imaging and 8 studies on CT imaging. The major part, consisting of 14 studies, involved two-dimensional models. The remaining 6 studies used three-dimensional models, of which 4 studies used CT imaging and 2 studies used MRI. The number of images used per study varied widely by image modality. The range of the number of included X-ray images was between 66 (Lahrmam, 2014 [38]) and 10024 (Xu, 2012 [64]). The range of included MR images and CTs was diffusely reported, as publications did not only use different numbers of scans but also different numbers of slices, sometimes differentiating per spine region. The number of studies with semi- or fully-automated methods was the same (n = 10).

The highest accuracy for MR imaging were reported by Weiss (2006) [62]; 96% for the initial model and 100% for the modified model, for the whole spine and the cervical spine, respectively. The study included the entire spine; the vertebrae and intervertebral disks and the ground truth consisted of 'independent assignments' of neurologists. In total, 50 MR images were included, 27 were used for the initial model and 23 MR images were used for the modified model. Image volumes are enhanced with a tophat filter, the program assigns the threshold values and applies a median spatial filter to the search regions. Voxels exceeding threshold values are then subjected to additional constraints and the centroids of these voxel clusters are then connected. 3D linear interpolation and Gaussian filters were applied, the longest disc chains were then analyzed in clusters, which obtained the above mentioned accuracies [20].

The best performing methods are VolHOG for MR images (Daenzer, 2014) [19], Modified GHT and K-means clustering with the use of X-ray imaging (Lahrmam, 2014 [38]) and a statistical and Gaussian shape model in combination with a principal component in combination with CT imaging (Clogenson, 2015 [17]). The research of Daenzer (2014) approached the cervical vertebra detection with a proposed novel machine learning method based on new radiological features, combined with a linear SVM. An accuracy of 98.1% was achieved with the baseline model and improved to 99.1% with the VolHOG. In addition, various levels of artificial noise are used during the performance analysis of the algorithm.

The ground truth in these studies is based on manually determined datapoints by (clinical) experts. All studies reported an internal comparison, comparing the performance of their model to the ground truth, and 9 studies additionally reported some form of external comparison, with earlier publications, published in the years 2003 to 2009. Of all Conventional Machine Learning segmentation studies, 7 studies reported segmentation results for only the whole spine, while 8 reported results for specifically the cervical spine. In 4 studies, the segmentation results were reported for both the cervical spine specifically and the whole spine. Clogenson (2015) [17] is an exception, just focusing on vertebra C2, which decreases external validity as compared to the other included studies.

Table 1. Conventional Machine Learning Segmentation articles overview

Author (Year)	Modality	Spine Region	2D/ 3D	Semi-/Fully-Automatic	N = / Split	Ground Truth	Comparison	
							Internal	External
Zamora (2003)[66]	X-ray	Cervical Spine	2D	Semi	+ Unclear: * NHANESII data from NLM* 100 images Test	Morphometric points placed by three expert radiologists	+ Stages: * GHT * GHT + ASM * GHT + ASM + DM	-
Burnett (2004) [10]	CT	Whole Spinal canal	2D	Fully	+ 5 CT's total* 557 slices	* Contours drawn by six dosimetrists* Performance judged by two radiation oncologists	* Manual contours vs. automated contours in random 100 axial CT images* Each image of 5 patients rated as successful/ unsuccessful or requiring further edit	-
Weiss (2006) [62]	MRI	Whole Spine	2D	Semi	+ 50 MRI total* 27 for initial* 27 + 23 additional for modified	Neuroradiologists independent assignments	* Initial algorithm concordance with neuroradiologist assignments * Modified algorithm concordance with neuroradiologist assignments	-
Schmidt (2007)[51]	MRI	Whole Spine	2D	Fully	+ 30 MRI total* 29 train/1 test* Additional test set with 37 new MRI's	Annotated ground truth data by expert	* Geometry model vs. Rotational invariant model, with and without including appearance * Combined model vs. new set of 37 MRI's* Semi vs. Fully	-
Klinder (2009)[36]	CT	Whole spine	3D	Fully	+ 74 CT total* 10 train* 64 test.	* Corrected mesh ground truth & Voxel based independent ground truth. * Verification by clinical expert	* Detection vs. Identification vs. Segmentation* Rate for individual vertebrae	-

Table 1. Conventional Machine Learning Segmentation articles overview (continued)

Author (Year)	Modality	Spine Region	2D/ 3D	Semi-/Fully-Automatic	N = / Split	Ground Truth	Comparison	
							Internal	External
Huang (2009) [31]	MRI	Cervical Spine	2D	Fully	+ 22 MRI total* 17 MRI for quantitative * 10 Thor/ Lumbar* 5 Cervical* 2 Whole body	Manually marked pixelwise ground truth	* Standard Adaboost vs. Proposed Adaboost without RANSC-based refinement vs. proposed Adaboost with RANSAC-Based refinement	* Proposed method vs. Active contour (snake) vs. Level set segmentation
Banik (2010) [8]	CT	Whole Spine and Spinal Canal	2D	Semi	+ 39 CT total+ 13 patients total + 458 slices total * Vertebral column 13 CT from 6 patients * Spinal canal 3 CT from 3 patients	Contours manually annotated by an expert radiologist	* Vertebral column segmentation vs. Expert radiologist * Spinal canal segmentationvs. Expert radiologist	-
Giulietti (2011)[25]	MRI (EPI)	Spinal Canal	2D	Semi	+ Unclear* 7 patients total	Manually annotated by two experienced operators both trained in spinal segmentation	* Proposed method vs. GT	* Proposed method vs. FSL tool FAST (conventional segmentation algorithm, slightly adapted)
Chen (2012) [15]	X-ray	Cervical Spine	2D	Semi	+ Unclear* 120 X-rays total	Orthodontic Novices and Orthodontic experts (ON and OE), compared rather than setting GT	* ON vs. CACVL (x- and y-axis)* ON vs. OE (x- and y-axis)* OE vs. CACVL (x- and y-axis)	
Glocker (2012)[26]	CT	Cervical Spine and Whole Spine	2D	Semi	+ 200 CT total+ 2595 vertebra * 116 cervical vertebrae * 50:50 split* Training set used for training step 1 and 2, Test set for remaining	Expert radiologist annotation	* Performance per stage reported (Stage 1: Regression Forest, Stage 2: HMM) * GT annotations compared to total model per region (Cervical, Thoracic, Lumbar and Whole Spine)	

Table 1. Conventional Machine Learning Segmentation articles overview (continued)

Author (Year)	Modality	Spine Region	2D/ 3D	Semi-/Fully-Automatic	N = / Split	Ground Truth	Comparison	
							Internal	External
Xu (2012)[64]	X-ray	Cervical Spine	2D	Fully	+ Unclear* 10024 Cervical X-ray total* 60 samples in each AAM training set * 100 images were selected from test samples randomly	Coordinates control points obtained from segmentation results compared to points marked manually by doctors	* Pixel error comparison between different parts experiment: * Conventional Machine Learning AAMs (Part I) vs.* Improved AAMs (Part II) vs. * Improved AAMs + detection of initial position (Part III)	
Glocker (2013)[27]	CT	Cervical Spine and Whole Spine	2D	Semi	+ 200 CT scans total* 2595 vertebra total * 116 cervical vertebrae * Split 50:50 * Training set used for training step 1 and 2,* Test set for remaining * 224 high pathological cases, also 50:50 test/ train split.	Manual annotations, not further specified by whom. Glocker (2012) by expert radiologist	* Localization errors and Identification rates compared to GT annotations.	* Localization errors and Identification rates compared to Glocker (2012) method
Mirzaalian (2013)[44]	CT	Cervical Spine and Whole Spine	3D	Fully	+ 7 CT scans total * Each 22 vertebra * 154 spatially normalized, single vertebra volumes* Split training: test, 6:1	Manually annotated meshes, not further specified by whom	* GT vs. GC + SP GT vs. SSM + ML GT vs. SSM + ML + Norm	* SSM + ML + Norm with perturbed vertebra box information (5 mm in XYZ-translations and XYZ scales and 2° in orientation)
Larhmam (2014)[38]	X-ray	Cervical Spine	2D	Semi	+ 66 X-ray images total * Validation: 20 annotated X-rays from Jolimont DB* Training: 5 annotated X-rays from NHANES-II* Test: 45 X-rays NHANES-II	Manually annotated 6 landmarks for total 100 vertebra by 1 radiologist	* DR for different datasets: NHANES II vs. Jolimont DB* Mean DR	* DR for different publications vs. previous work from same authors vs. proposed method

Table 1. Conventional Machine Learning Segmentation articles overview (continued)

Author (Year)	Modality	Spine Region	2D/ 3D	Semi-/Fully-Automatic	N = / Split	Ground Truth	Comparison	
							Internal	External
Daenzer (2014)[19]	MRI	Cervical Spine	3D	Fully	+ 42 MRI images total* 2 disjointed datasets * 21 (T2) MR images for each dataset* 50:50 split	C3-C7 manually segmented on a per voxel basis by a team consisting of computer scientist and an anatomist	* (Proposed) VolHOG vs. Baseline (case by-ten-fold cross validation (CV)) default detector using pseudo-3D-HOG descriptor vector [without Rician noise]	* VolHOG with different Rician noise levels
Clogenson (2015)[17]	CT	C2 Vertebra	3D	Semi	+ 92 CT images total * Without pathology* 31 scans test set	7 landmarks placed, by an anatomy expert; surgeon, radiologist	* Distance error SSM compared to GT mesh landmark placement	* 2 Previously developed methods: Mirzaalian (2013) Klinder (2005)
De Leener (2015)[20]	MRI	Spinal Canal and Cervical Spine	2D/ 3D	Fully	+ Unclear* Training unclear* Test set 17 MRI images	Both spinal column + vertebral level both manually annotated by an expert	* Accuracy of the Intervertebral Disks identification (%) * MSE global detection accuracy after correcting for the wrong identifications	* Proposed method vs. Active surface for SC T1w* Proposed method vs. Active surface for SC T2w* Proposed method for CSF T2w
Al Arif (2016) [2]	X-ray	Cervical Spine	2D	Fully	+ 90 X-ray images total * 81 training * 9 test * 10-fold cross-validation scheme	All 90 images manually demarcated by expert 'radiographers', clicked on 20 points along the vertebra boundary	* ASM-M * ASM-RRF * ASM-RCF-AM * ASM-RCF-KDE (The proposed classification forest-based framework with kernel density-based prediction)	-

Table 1. Conventional Machine Learning Segmentation articles overview (continued)

Author (Year)	Modality	Spine Region	2D/ 3D	Semi-/Fully-Automatic	N = / Split	Ground Truth	Comparison	
							Internal	External
Mehmood (2017) [43]	X-ray	Cervical Spine	2D	Semi	+ Unclear * Total 150 X-ray * C3-C7	Centers of the 150 photos are manually annotated, not further specified by whom	Proposed method compared to manual annotations with output: * Accuracy * Per vertebra - Accuracy per vertebra - Distance error per vertebra Proposed method using 5 atlases compared to manual labels with output: * DI/DC * Mean error distance * Hausdorff distance	Accuracy * Larhman (2012) * Larhman (2013) * Lecron (2012) * Benjelloun (2012) (auto) (semi-auto)
Hanaoka (2017)	CT	Wohole Spine	3D	Fully	+ Unclear * 50:50 on validation thoracolumbar set * Cervical unclear * Training done on set of 50 CT's whole torso CTs (including C1-5)	Training datasets manually labeled by an expert		Not including cervical

Table 2. Conventional Machine Learning Segmentation articles extracted outcomes

Author (Year)	Model Features	Internal				External	
		Accuracy (%)	Time	Error (mm)	Other	Accuracy / Other	Error (mm)
Zamora (2003) [66]	+ Hierarchical:* ASMs with GLV/GLM* Customized GHT* Customized DM	* GHT: 65%* GHT + ASM: 75%* GHT + ASM + DM: 75%	-	-	-	-	-
Burnett (2004) [10]	* DM* Fourier descriptors * Wavelet based Edge detection algorithm* Chamfer matching	* Automated contours: 93% vs. Manual contours: 69% * Automated successful in 91%, further editing in 7% and unsuccessful in 2%	-	-	-	-	-
Weiss (2006) [62]	* Threshold values* Median spatial filter* Three-dimensional linear interpolation* Gaussian filters * Disc constraints, count discs, longest chains, analyze clusters	* Initial algorithm: 96% * Modified algorithm: 100%	* Runtime: 1 min 47 sec (± 20 sec) central processing unit time (range: 58 sec-52 min 47 sec)	-	-	-	-

Table 2. Conventional Machine Learning Segmentation articles extracted outcomes (continued)

Author (Year)	Model Features	Internal				External	
		Accuracy (%)	Time	Error (mm)	Other	Accuracy / Other	Error (mm)
Schmidt (2007)[51]	* Probabilistic GM, object recognition framework* Fully interconnected model * No model assumptions * Truncated Gaussian distributions, randomized classification trees, branching tests at tree nodes	* Geometry: DR: 91% incl. appearance 94%* Rotational: DR: 95% incl. appearance 97%* Semi-automatic Rotational DR: 94% Geometry DR: 94%	-	* Geometry: 6.2 (IQR 4.6-6.5) incl. appearance: 5.5* Rotational: 5.8 (IQR 4.8-6.3) incl. appearance: 5.1* Deep Learning: 7.8 (IQR: 5.3- 9.5)* Semi-automatic: Geometry: 5.8 (IQR: 3.9-5.5) Rotational: 5.7 (IQR: 4.2-5.3)	-	-	-
Klinder (2009) [36]	* Geometric modelling* GHT (template matching)* Shape constrained DM	* Vertebra detection: 92%* Vertebra identification: 95%	* Detection: 13,3 sec* Identification: 36, 5 min* Segmentation: 179.5 sec	* Segmentation: 1.12 ± 1.04			
Huang (2009)[31]	* Deep Learning iterative normalized-cut* AdaBoost-based detection * Refinement robust curve fitting* Energy minimization process	DR: * Standard Adaboost 85.11%* Proposed without RANSC- based refinement 89.36% * Proposed with RANSAC-Based refinement 97.87%	-	-	-	Whole spine DR:* Proposed method: 96% * Active contour (Snake) 81% * Level set segmentation: 83%	-

Table 2. Conventional Machine Learning Segmentation articles extracted outcomes (continued)

Author (Year)	Model Features	Internal				External	
		Accuracy (%)	Time	Error (mm)	Other	Accuracy / Other	Error (mm)
Banik (2010)[8]	* Fuzzy segmentation* Hough Transform	-	-	* Vertebral column: HD 3.2 ± 2.4 MDCP 0.7 ± 0.6* Spinal canal: HD 1.6 ± 0.5 MDCP 0.6 ± 0.1	-	-	-
Giulietti (2011)[25]	* K-means clustering* Gaussian Smoothed* Non-linear noise reduction	* Proposed: 82% ± 1	* Proposed efficiency: 0.67±0.02	-	* Proposed: Sens 89% ± 1 Spec 75% ± 2	* FAST: Accuracy 73% ± 1 Sens 82% ± 1, Spec 64% ± 2Efficiency: 0.53 ± 0.02	-
Chen (2012)[15]	* Landmark calculation* Fast marching method* Parabolic curve fitting * 'Computer aided cervical vertebra landmarking' (CACVL)	* 162 p-values* Authors conclude based on p-values per vertebra: "CACVL has the same or higher accuracy, better repeatability and less work load than manual landmarking methods"	-	-	-	-	-

Table 2. Conventional Machine Learning Segmentation articles extracted outcomes (continued)

Author (Year)	Model Features	Internal				External	
		Accuracy (%)	Time	Error (mm)	Other	Accuracy / Other	Error (mm)
Glocker (2012)[26]	* Regression forests* Probabilistic graphical models* Hidden Markov Model* Discriminative regression* Generative model	IR: * Cervical 72% * Whole 81%	* Training a single tree takesabout 3 minutes* Testing a whole forest on a scan takes less than 1 sec. * The HMM refinement takesabout 5-15 sec. * Total, the localization and identification of all vertebrae in one test image is achieved in less than 2 min	Localization error: * Cervical Stage 1:Median Mean Std25.97 30.74 18.64 Stage 26.87 10.85 12.49* Whole Spine Stage 1:Median Mean Std15.91 18.35 11.32 Stage 25.31 9.50 10.55	-	-	-
Xu (2012) [64]	* Haar-like features* Adaboost* Multi-resolution AAM* Parallel cascade Structure	-	-	Error in Pixel (!)Part Min Average Max I 8.652 27.597 52.336II 3.682 4.792 6.007III 4.035 19.018 146.925	-	-	-

Table 2. Conventional Machine Learning Segmentation articles extracted outcomes (continued)

Author (Year)	Model Features	Internal				External	
		Accuracy (%)	Time	Error (mm)	Other	Accuracy / Other	Error (mm)
Glocker (2013)[27]	* Sparse centroid annotations transformed into dense probabilistic labels * Randomized classification forests * Feature extraction through supervised, hierarchical clustering training data * Objective function which favors compact clusters of image points having equal labels	*Proposed method: Normal vs. Pathological CTIR Cervical 78% vs 80%IR Whole 76% vs 70 %	-	* Proposed method LENormal vs. - Pathological CT Median Mean StdCervical6.3-5.9 7.7-7.0 4.4- 4.7 Whole Spine7.6-8.8 11.5-12.4 14.1-11.2	-	* Performance normal CT see results Glocker (2012)* Pathological CT: IR Cervical 54% IR Whole 51%	* Normal CT see results Glocker (2012) * Pathological CT:LE Median, Mean, Std: C: 11.5, 17.0, 17.7 W: 14.8, 20.9, 20.0
Mirzaalian (2013)[44]	* Statistical Shape Model* Graphical Cut with Shape prior* ML to capture local appearance-related prior shape information* Probabilistic boosting-tree classifier	-	* Runtime: Estimated at 2 minutes	* Cervical Point-to-surface GC + SP: 7.9 ± 7.9 SSM + ML: 2.2 ± 0.7 SSM + ML + Norm: 1.4 ± 0.4* Whole Spine Point-to-surface GC + SP: 6.1 ± 6.7 SSM + ML: 1.8 ± 0.6 SSM + ML + Norm: 1.6 ± 0.7	-	-	* Point-to-surface:C = 1.6 ± 0.6W = 1.6 ± 0.7

Table 2. Conventional Machine Learning Segmentation articles extracted outcomes (continued)

Author (Year)	Model Features	Internal				External	
		Accuracy (%)	Time	Error (mm)	Other	Accuracy / Other	Error (mm)
Larhmam (2014)[38]	* Modified GHT* K-means clustering	* DR: NHANES II 99.1%, Jolimont DB total: 96% Mean total: 97.5%	-	* Mean RMS angle error 6.70 degrees (!)	-	* DR: Proposed:97.5%, Casciaro: 83% Klinder: 92% Dong: 92.4% Prior work author: 89%	-
Daenzer (2014)[19]	* New features with linear SVM for classification. * Algorithm for bivariate gradient orientation histogram generation from three-dimensional raster image data. * VolHOG	DR:* Baseline: 98.1%* VolHOG: 99.1%	* Runtime: 4.1 ± 1.32 min	Aver. COG Distance (mm) * Baseline: 1.72 ± 0.81* VolHOG: 1.64 ± 0.70	Aver. DO * Baseline: 0.80 ± 0.07* VolHOG: 0.81 ± 0.06Precision * Baseline: 0.9653* VolHOG: 0.9859	* DR with Rician noise 0: 99.05% Rician noise 1: 94.29% Rician noise 2: 91.43%	-
Clogenson (2015)[17]	* SSM* Gaussian Shape model* PCA	-	* Runtime: 30 sec	Distance error:* Proposed method: 0.9 ± 0.12	-	Runtime:* Mirzaalian 2 min* Kindler: 3 min	Distance error:* Mirzaalian 1.4 ± 0.4* Kindler 0.81 ± 0.97

Table 2. Conventional Machine Learning Segmentation articles extracted outcomes (continued)

Author (Year)	Model Features	Internal				External	
		Accuracy (%)	Time	Error (mm)	Other	Accuracy / Other	Error (mm)
De Leener (2015)[20]	* Iterative propagation * Tubular DM* Elliptical HT* Vesselness filter* Contrast level-based local mesh structures orientation adaption	* Proposed for intervertebral disc identification: 98.3%	* Runtime: Overall acquisition 22 min	* MSE for intervertebral disc identification 1.05 ± 1.45 mmHD* CSF T2w: Proposed method 289 ± 0.95	DC* CSF T2w Proposed method 0.91 ± 0.02	DC* SC T1w: Proposed method 0.91 ± 0.01 vs. Active surface 0.88 ± 0.04* SC T2w: Proposed method D.C. 0.91 ± 0.03 VS Active surface: DC 0.87 ± 0.03	HD* SC T1w: Propose method 1.79 ± 0.28 VS Active surface 2.05 ± 0.48* SC T2w: Proposed method 1.68 ± 0.36 VS Active surface 2.19 ± 0.53
Al Arif (2016)[2]	* ASM* RCF* KDE	* ASM-M 76.00%* ASM-RRF 78.24%* ASM-RCF-AM 80.00%* ASM-RCF-KDE 83.33%	-	Error (Median/ Mean ± Std): * ASM-M 0.8019 / 0.8582 ± 0.3437* ASM-RRF 0.6933 / 0.7704 ± 0.3766* ASM-RCF-AM 0.7054 / 0.8060 ± 0.3998* ASM-RCF-KDE0.6896 / 0.7688 ± 0.3965	-	-	-

Table 2. Conventional Machine Learning Segmentation articles extracted outcomes (continued)

Author (Year)	Model Features	Internal				External	
		Accuracy (%)	Time	Error (mm)	Other	Accuracy / Other	Error (mm)
Mehmood (2017)[43]	* Localization using GTH* Clustering Fuzzy C Means * Centroid Detection	* Proposed method total: 93.76%* Per vertebra: - C3 96.74 - C4 96.65 - C5 95.51 - C6 95.33 - C7 84.55	-	Distance Error* C3 7.7043* C4 7.6779* C5 7.8887* C6 8.9402* C7 13.0709	-	Accuracy* Larhman (2012) 89%* Larhman (2014) 97.5%* Lecron (2012) 81.60%* Benjelloun (2012) (auto) 64.5% (semi-auto) 89%	-
Hanaoka (2017)[28]	* SSM* Multi-atlas-based* Landmark-guided diffeomorphic demons algorithm	-	* Runtime: 15 min	* Mean error distance 0.59 ± 0.14 * HD 5.30 ± 2.14	* DC0.90 ± 0.02	-	-

Deep Learning Segmentation Techniques

There was a total number of 12 studies included that proposed Deep Learning segmentation techniques, of which two studies focused on MR imaging, 8 studies focused on CT imaging, and just one study used X-ray imaging. The study of Cai (2016) [13] involved both MRI and CT imaging. The majority, consisting of 7 studies, involved three-dimensional models, of which one study (Jakubicek, 2019) [33] combined 2D and 3D. The remaining 4 studies used two-dimensional models. The number of images used per study varied again widely per image modality, comparable to the Conventional Machine Learning segmentation studies. The range of the number of included CT images was between 41 (Bae, 2019) [6] and 392 (Jakubicek, 2019) [33]. The range of used MR images was slightly smaller but comparable; from 60 (Cai, 2016) [13] to 245 MRI images (Forsberg, 2017) [24]. However, the interpretation of this range is difficult as publications, like in the conventional Machine Learning group, did not only use different numbers of scans but also different numbers of slices, sometimes differentiating per spine region.

Almost all studies deployed fully automated methods. Only one study used a semi-automated approach (Forsberg, 2017) [24], which then also achieves highest detection accuracies (98.8 – 99.8 %). Forsberg (2017) focused on both the cervical and lumbar spine, creating two separated training and configuration pipelines, both having the same CNN setup. The CNN uses fully connected layers, drop-out rate of 0.5, a categorical cross-entropy cost function and Nesterov momentum accelerated Stochastic gradient descent (SGD). The included MR images, together with the annotated spine labels, are focused on either the lumbar or cervical part of the spine. The dataset was originated from an image archive. The missed detections were mainly concerning partly visible vertebrae on the available images. This research showed promising results for labeling and detection by a CNN, focusing on both the cervical and lumbar spine [26].

The highest segmentation accuracy was achieved by the SpineCNN from Jakubicek (2019) [33] (93.3 %). Thereby, the best performing methods are CNN based methods for both CT and MR images. The study presents a fully automated approach based on 130 CT scans, which includes two CNNs and a spine tracing algorithm, among which a fine-tuned AlexNet and a VGG-16 R-CNN. A population approach was used to increase robustness. The novel combination of the CNN and the tracing, results in almost 90% of correctly identified spinal centerlines within 20 seconds of computing time [24]. The only study focusing on X-ray imaging (Al Arif, 2018) [1] used a 6-layered FCN, with an accuracy of center localization of 93.7 %.

Similar to the Conventional Machine Learning segmentation studies, the ground truth in the Deep Learning segmentation publications is based on manually determined ground truth by (clinical) experts. The majority of the included studies regarding Deep Learning segmentation methods used both internal and external comparison of their results (n = 9). Results were reported for the whole spine and cervical spine, in 2 and 4 studies, respectively. In 6 studies, half of the total number of studies related to Deep Learning segmentation methods included, the results were reported for both the cervical part and the whole spine.

Table 3. Deep Learning Segmentation articles overview

Author (Year)	Modality	Spine Region	2D/ 3D	Semi-/Fully-Automatic	N = / Split	Ground Truth	Comparison	
							Internal	External
Suzani (2015) [56]	CT	Cervical Spine and Whole Spine	3D	Fully	+ 224 CT total* 50:50 split* 112 images each * Two-fold cross-validation	Expert annotations	* Detection rates and localization errors for different regions of the vertebral column	* Compared with detection rates and localization errors of Glocker (2013) and Glocker (2012)
Chen (2015) [14]	CT	Cervical Spine and Whole Spine	3D	Fully	+ 302 CT total* 242 CT training * 60 CT testing	Manually, not further specified by whom	* J-CNN compared with 'standard' CNN and GT	* Compared with IR, Mean error and Std Glocker (2013)
Cai (2016) [13]	CT, MRI	Cervical Spine and Whole Spine	2D	Fully	+ 60 MRI total* 90 CT total * 6 MRI Training* 4 CT Training* 54 MRI Testing* 86 CT Testing	Manually selected and labeled by one-click annotation	* Proposed vs. Baseline MR* Proposed vs. Baseline CT	* Regression + HMM vs. Dense Prob. Label vs. Proposed
Forsberg (2017)[24]	MRI	Cervical Spine	2D	Semi	+ 245 MRI total * 223/221 T1-/T2- * 2321 spine labels * Random split * Training : Validation : Test* 60:20:20 = 147:49:49	Navigational support annotations, manual quality assurance step performed	* T1 vs. T2 compared for sensitivity, precision accuracy, localization error, labeling accuracy	* DR for Huang (2009), Klinder (2009), Zhan (2015) compared* Extensive comparison table (including localization error and labeling rate) available in full publication.
Liu (2018) [40]	CT	Cervical Spine	3D	Fully	+ 60 CT total* 40 Training set* 20 Test set	Manually labelled by 'professional physicians'	* Proposed method DI, Mean absolute surface distance, HD* Proposed method vs. GT per vertebra, see full publication	* Proposed method compared to Wang (2015), Klinder (2009), Casto-Mateos (2015), Huang (2013), Korez (2015)

Table 3. Deep Learning Segmentation articles overview (continued)

Author (Year)	Modality	Spine Region	2D/ 3D	Semi-/Fully-Automatic	N = / Split	Ground Truth	Comparison	
							Internal	External
Al Arif (2018)[1]	X-ray	Cervical Spine	2D	Fully	296, train 124, test 172 * Dataset1 138 total * 90:10, Test:Train * 124 : 14 * Dataset2 158 total * All added to test	Manual annotation of vertebral boundaries, not specified by whom	Performance for:* Global Localization * Center Localization * Fully automatic segmentation network UNet vs. UNet-s	-
Jakubicek (2019)[33]	CT	Whole Spine	2D/ 3D	Fully	+ 392 CT total* 4:2:3 split * Train : Validation : Test * 174 Train* 87 Validation* 130 Test	Positions determined by experts	* The algorithm is compared in centerline detection to the GT* The performance of CordCNN and SpineCCN were compared	-
Bae (2019)[6]	CT	Cervical Spine	2D	Fully	+ 41 CT total * 80:20 split * Train:Validation, * (n=17 set) 14 train, 3 validation * (n=24 healthy controls set) 19 train set, 5 validation * Validation performed on both sets for different models* Two different sets from two different hospitals	Provided by 2 human experts, kappa calculated	* Intra-validation (which seems like validation) and extra-validation (which seems like test set performance)* Models functioned once as training and once as test set. * Comparison with lumbar model* Extra-validation results reported in Outcome as this seems test set performance.	

Table 3. Deep Learning Segmentation articles overview (continued)

Author (Year)	Modality	Spine Region	2D/ 3D	Semi-/Fully-Automatic	N = / Split	Ground Truth	Comparison	
							Internal	External
Rak (2019) [49]	MRI	Whole Spine	3D	Fully	* 64 MRI total* Training : Test * 4 : 1 = 51 : 13	Falsely detected vertebrae were corrected manually, i.e. user-specified vertebra center was used, not further specified by whom	* Model performance on dataset 1 (whole spine images) T1 vs. T2 weighted MRI images	* Comparison added with performance reported in different publications, however omitted from 'Outcomes' section because mainly Thoracolumbar images.
Wang (2019) [61]	CT	Cervical Spine and Whole Spine	3D	Fully	+ 98 CT total * 50:50 split* Two sets n=49 * Randomly selecting 31 from 63 normal and 18 cases from 35 abnormal CT. * Both ways tested	Centroids annotated by two experts	* Cervical vs. whole spine Localization error for SRF step and refinement step* Cervical vs. whole spine Identification Rate	* In Graphs CNN based and RF+HMM based approaches error compared per vertebra.* Not reported in 'Outcomes'
Chen (2020) [16]	CT	Cervical Spine and Whole Spine	3D	Fully	+ 303 CT total* 242 Training * 60 Hold-out evaluation	Vertebrae centroids coordinates not further specified by whom	* IR and mean error for proposed method ascompared to GT* For full model performance ablation studies; see full publication	* Proposed method compared in IR and mean error with: Glocker et al. (2013) Chen et al. (2015)Yang et al. (2017)Liao et al. (2018)

Table 3. Deep Learning Segmentation articles overview (continued)

Author (Year)	Modality	Spine Region	2D/ 3D	Semi-/Fully-Automatic	N = / Split	Ground Truth	Comparison	
							Internal	External
Jakubicek (2020)[32]	CT	Cervical Spine and Whole Spine	3D	Fully	+ 421 CT total * 4 different	Visual evaluation of the centerline detection by medical expert and classified as correct or incorrect	* Vertebra identification accuracy for proposed method compared with former kernel-PCA model	Proposed method compared in IR, mean ADE and TPR with: Klinder et al. 2009 Ma et al. 2010 Hanaoka et al. 2010 Stern et al. 2010 Glocker et al. 2012 Glocker et al. 2013 Kelm et al. 2013 Cai et al. 2015 Chu et al. 2015 Cheng et al. 2016 Chen et al. 2015 Yang et al. 2017 Liao et al. 2018 Sekuboyina et al. 2018

Table 4. Deep Learning Segmentation articles extracted outcomes

Author (Year)	Model Features	Internal				External	
		Accuracy (%)	Time	Error (mm)	Other	Accuracy / Other	Error (mm)
Suzani (2015)[56]	* Canny edge detector* 6-layered DeepNN * Feed forward* Stochastic gradient descent	* C: 96.0* WS: 96.0	* Runtime: Overall runtime less than 3 seconds	* C: 17.1 ± 8.7* WS: 18.2 ± 11.4	* C: 97.8% / 95.0% / No dice / 91.2%* WS: 97.2% / 95.0% / No dice / 94.4%	* Glocker (2012) WS 93.9% / 92.9% 94.7% / 93.7% /* No efficiency / Dice	* Glocker (2013) WS 12.4 ± 11.2* Glocker (2012) WS 20.9 ± 20.0
Chen (2015)[14]	* Random Forrest * J-CNN * Shape Regression model* Weights initialized with Gaussian distribution	Proposed: * C: IR 91.84* WS: IR 84.16CNN: * C: IR 84.00* WS: IR 77.13	-	Proposed:* C: 5.12 ± 8.22* WS: 8.82 ± 13.04CNN: * C: 6.80± 10.90* WS: 10.41±13.82	-	* Glocker (2013) IR C: 88.76%IR WS: 74.04%	* Glocker (2013) C: 6.81 ± 10.02WS: 13.20 ± 17.83
Cai (2016) [13]	* TDCN* Feature fusion by (C)RBM)* SVM	* Proposed WS: MRI/CT 98.1/96* Baseline WS:MRI/CT 97.8/86	-	* Proposed: WSMRI: 3.23 ± 2.09CT: 3.81 ± 2.98* Baseline: WSMRI: 3.1 ± 2.1CT: 3.81 ± 2.98	-	-	C : (Median, Mean, ± Std)* Regression + HMM: 6.5, 8.2, ± 6.1* Dense Prob. Label: 6.3, 7.7, ± 4.4Proposed: 3.7, 3.2, ± 2.8

Table 4. Deep Learning Segmentation articles extracted outcomes (continued)

Author (Year)	Model Features	Internal				External	
		Accuracy (%)	Time	Error (mm)	Other	Accuracy / Other	Error (mm)
Forsberg (2017)[24]	* CNN* Fully connected layers * 50% drop-out* Categorical cross-entropy as cost function* SGD Nestorov momentum	Detection accuracy* T1: 98.8 (421/422)* T2: 99.8 (408/409) Labeling accuracy* T1: 96.0 (409/426)* T2: 96.6 (394/408)	-	* LE T1: 1.18 ± 0.81 T2: 1.24 ± 1.01	* T1: Sens, Precision 99.1%, 99.8%* T2: Sens, Precision 99.8%, 100%	* Huang (2009) DR: 97.9/98.1%* Klinder (2009) DR: 92%* Zhan (2015) DR: 98.0%/98.8%	* See full publication
Liu (2018) [40]	* DM* VGG-like CN* Adaboost* SiFC	-	* 48 min to obtain the final cervical segmentation results for a single CT	Proposed method: * HD 2.94* MASD 0.70	Proposed method: DI* 95.47 ± 0.80	* Klinder -* Wang DI 92.79 ±1,55* Castro-Mateos DI 90 ± 5,1* Huang DI 94 ± 2.1* Korez DI 94.4 ± 2.1	* Klinder MASD 0.97 * Wang -* Castro-Mateos MASD 0.82 HD 5.5. * Huang HD 10.06 ± 1.71 * Korez MASD 0.3 HD 2.94

Table 4. Deep Learning Segmentation articles extracted outcomes (continued)

Author (Year)	Model Features	Internal				External	
		Accuracy (%)	Time	Error (mm)	Other	Accuracy / Other	Error (mm)
Al Arif (2018)[1]	* 6-layered FCN* Global localization* Center localization* UNet-s with updated shape-aware loss function	Center localization * DR 93,73Pixel-wise accuracy (mean, median, Std)* UNet: 97.71, 96.69, ± 3.04 * UNet-S: 97.92, 97.01 ± 2.79	* Producing a localization result for any image under a second	Center localization: * Average error: 1.81 mmPoint to curve error(mean, median, Std) * UNet: 0.43, 0.62, ± 0.81 * UNet-s: 0.44, 0.55, ± 0.40	Global Localization* Sensitivity: 96%* Specificity: 96%DC (mean, median, Std) * UNet: 0.952, 0.938, ±0.048 * UNet-s: 0.957, 0.944, ±0.044* Complete Fully automatic: 0.84	-	-
Jakubicek (2019)[33]	* CNN, fine-tuned AlexNet (for SpineCNN spine-ends detection)* VGG-16 faster R-CNN (for CordCNN centerline delimitation)	Spine Centerline * DR 89.4 (for population size 20)Classification Accuracy* SpineCNN 93.3* CordCNN 58.3	* Runtime: 16.5 sec	SpineCNN* 3.9 ± 8.4CordCNN* 9.4 ± 6.3	-	-	-
Bae (2019) [6]	* 2D CNN* 2D U-net for 3D	-	-	* Model 1. MASD: 1.05 ± 0.63 HD: 29.17 ± 19.74* Model 2. MASD 0.38 ± 0.17 HD: 20.85 ± 7.11	* Model 1. DC: 88.67 ± 5.82* Model 2. DC: 93.15 ± 3.09	-	-

Table 4. Deep Learning Segmentation articles extracted outcomes (continued)

Author (Year)	Model Features	Internal				External	
		Accuracy (%)	Time	Error (mm)	Other	Accuracy / Other	Error (mm)
Rak (2019) [49]	* Deep Learning Graph Cut based on patch formulation * Patch incorporates appearance and shape information derived from CNN (U-net)* Fully convolutional* Star-convexity constraints (1 block) * Deep Learning non-overlap constraints on encoding swaps (1 block)	-	* Runtime: 1.35 ±0.08 s and 0.90 ±0.03 s per vertebra on consumer hardware. * A complete whole spine segmentation took 32.4 ±1.92 s on average.	-	DC* T1: 93.6 ± 3.0 * T2: 94.0 ± 2.3(All displacements, all folds)	-	-
Wang (2019)[61]	* Deep SSAE * contextual features* SRF* Mean KDE Otsu method * MRF, instead of CNN.	IR:* Cervical 86.6* Whole spine 82.2	Runtime* Proposed less than 56 s in one test image* RF-HMM +/- 110 s * J-CNN +/- 87 s	LE (Median, Mean, Std)* SRF Cervical 8.74, 11.21, ± 9.45* SRF Whole Spine 12.14, 14.41, ± 9.51* Refinement C 6.56, 8.54, ± 7.45* Refinement WS 11.27, 10.08, ± 7.97	-	-	-

Table 4. Deep Learning Segmentation articles extracted outcomes (continued)

Author (Year)	Model Features	Internal				External	
		Accuracy (%)	Time	Error (mm)	Other	Accuracy / Other	Error (mm)
Chen (2020)[16]	* FCN* HMM	IR* Proposed: 94.67 (WS) 89.5 (C)	-	* Proposed: 2.56 ± 3.15 (WS) 2.50 ± 3.66 (C)	-	IR* Glocker: 74.00%(WS) 88.8% (C) * Chen: 84.16% (WS) 91.8% (C)* Yang: 85.00% (WS) 92.0% (C) * Liao: 88.30% (WS) 95.1% (C)	* Glocker13.20 ± 17.8 (WS)6.81 ± 10.02 (C) * Chen8.82 ± 13.04 (WS) 5.12 ± 8.22 (C)* Yang 8.60 ± 7.80 (WS) 5.60 ± 4.0 (C) * Liao 6.47 ± 8.56 (WS) 4.48 ± 4.56 (C)
Jakubicek (2020)[32]	*CNN (AlexNet) *R-CNN	IR * Proposed: 94.6 (D1) 76.7 (D2) 90.9 (D3) 83.2 (D4)	Computation time: IdentCNN 5 sec. IdentCNN + DP 7 sec.	Mean ADE 2.34 ± 0.84 (D1) 6.75 ± 4.17 (D2) 5.08 ± 3.95 (D3) 3.84 ± 2.82 (D4)	TPR [%] 98.2 (D1) 91.3 (D2) 92.4 (D3) 96.7 (D4)	IR *Yang 2017: 85.0 *Liao 2018: 88.3 *Sekuboyina 2018: 88.5 NB. Only 3 most recent results are shown. See Table 1 (Jakubicek) for total overview	*Yang 2017: 8.60 ± 7.80 *Liao 2018: 6.47 ± 8.56 *Sekuboyina 2018: 6.20 ± 4.10 NB. Only results for 3 most recent publications are shown. See full publication for complete overview.

Cervical Spine Analysis

There was a total number of 19 studies included involving cervical spine analysis, of which four studies focused on MR imaging, two studies focused on CT imaging, and the majority of the studies used X-ray imaging (n=11). The aims of the included studies could be further divided into five subgroups; 1. Biomechanical analysis (n = 7), 2. CVM stage (n = 3), 3. Clinical prognosis/prediction (n = 2), 4. Image registration/Planning (n = 4), 5. Clinical/Radiological Feature detection (n = 3).

Two studies included both CT and MR imaging, of which du Bois (2007) [21] included PET imaging as well. The use of two-dimensional and three-dimensional visualizations were equal (n=9), and the remaining study of Kage (2020) [35] used a combination of 2D and 3D imaging. Most studies included vertebrae C2-C6, of which several expanded with inclusion of the vertebrae C1, C7 or T1. Other studies used a smaller area of the spine, vertebrae C2-C4, which had the aim to determinate the CVM stage (Kok, 2019 [37] and Amasya, 2020 [4]). The study by Dzyubachyk (2013) [22] was the only one to include the entire spine in the analysis model, with the aim to create an automated reconstruction of the complete spine, based on multistation 7T MR images. The authors applied intensity inhomogeneity correction and used coherent local intensity clustering (CLIC) and fuzzy-c-means-clustering. The performance of the model by Dzyubachyk (2013) [22] was validated based on 18 different datasets, which showed a mean registration error of 0.53 millimeters, which was lower than the MR image pixel size and showed thereby sufficient accuracy.

A wide range of methods was deployed. The best method for radiological feature detection is a CNN model, while the SVM model gave the best result in terms of clinical classification. The ANN approach was reported to work best for CVM stage determination and the FE model, in combination with X-ray imaging, is the most-used method for biomechanical analysis of the spine. In the included spine analysis studies, the amount of fully and semi-automated methods was 7 and 12, respectively.

Table 5. Cervical Spine Analysis articles overview and extracted outcomes.

Author (Year)	Modality	Category	Aim	Model Features	2D/ 3D	Semi- / Fully- Automatic	N = / Included Vertebrae	Key Points
Benjelloun (2009)[9]	X-ray	Biomechanical Analysis	* Determine vertebral motion induced by their movement between two or several positions	* Automatic corner points of interest point detection* Harris corner detector * Face contour detect	2D	Semi	n = 100(C2-C6)	* Difference between manual corner detection and proposed method in range 0-4 degrees.
Lecron (2012)[39]	X-ray	Biomechanical Analysis	* Determine cervical spine mobility by automated angle calculation between adjacent vertebra	* Scale-invariant Feature Transform * Speeded-up Robust Features* Coupled multi-class SVM* ASM/SSM	2D	Fully	n = 49 (C3-C7)	* Vertebræ are successfully detected in 89.8% of cases and it is demonstrated that SURF slightly outperforms SIFT. * ASM for segmentation, statistical shape model specific to the vertebral level improves the result.* Angular errors of cervical spine mobility are presented; errors remain within the inter-operator variability of the reference method.
Balkovec (2017)[7]	X-ray	Biomechanical Analysis	* Determine precise time-course details about individual vertebrae and intervertebral motion	* Iterative Template Matching Algorithm* Video fluoroscopy / Digital Motion)	2D	Semi	n = 3 (C2-C6)	* Errors in intervertebral angular and shear displacements no greater than 0.4° and 0.055 mm, respectively. * Aberrant intervertebral motions in the cervical spine were typically found to correlate with patient-specific anatomical features such as disc height loss and osteophytes.

Table 5. Cervical Spine Analysis articles overview and extracted outcomes. (continued)

Author (Year)	Modality	Category	Aim	Model Features	2D/ 3D	Semi- / Fully-Automatic	N = / Included Vertebrae	Key Points
Kage (2020) [35]	X-ray	Biomechanical Analysis	* Quantifying segmental motion in spine	* 2D/3D Shape-matching algorithm	2D/3D	Semi	n = 1 (C4-C6)	* System's overall RMSE ranged between 0.21–0.49mm and 0.42–1.80°. The RMSE associated with RSA ranged between 0.14–0.69mm and 0.96–2.33° for bead centroid identification and 0.25–1.19mm and 1.69–4.06° for dynamic bead tracking.
Nikkhoo (2019)[47]	X-ray	Biomechanical Analysis - FE model	* Investigate spine biomechanics associated with typical cervical disorders	* Finite Element Model * Based on 30 X-ray parameters * 3D reconstruction lower spine region	3D	Semi	n = 6 (C3-C7)	* Severe disc alteration (Grade 3) presented a significant decrease in the ROM and intradiscal pressure (flexion, extension, and axial rotation) on the C5-C6 and slightly increase on the adjacent levels. * Maximum stress in Annulus Fibrosus (AF) and facet joint forces increased for Grade 3 for both altered and adjacent levels.

Table 5. Cervical Spine Analysis articles overview and extracted outcomes. (continued)

Author (Year)	Modality	Category	Aim	Model Features	2D/ 3D	Semi- / Fully- Automatic	N = / Included Vertebrae	Key Points
Nikkhoo (2020)[46]	X-ray	Biomechanical Analysis - FE model	* Investigate the biomechanical impact of laminectomy on cervical intersegmental motion and load sharing	* Finite Element Model* Geometrically specific	3D	Semi	n = 10(C3-C7)	* Post laminectomy increased the intersegmental ROM, disc stress, and intradiscal pressure at the upper cervical levels during sagittal plane motion and axial rotation, while the lower levels experienced the opposite trend, as compared with intact models. * No significant changes were observed in facet joint forces after surgery
Srinivasan (2020)[55]	X-ray	Biomechanical Analysis - FE model	* Effect of Heterotopic Ossification after arthroplasty on the adjacent levels and change in range of motion (ROM)	* Finite Element Model * Strain Energy Density (SED)	3D	Semi	n = 7(C2-T1)	* The Bryan disc significantly reduced ROM at the implanted level in flexion. However, in extension, ROM increased at the implanted level and decreased slightly at the adjacent levels. After HO, ROM drastically reduced at the implanted level in both extension and flexion, and showed a minor increase in the adjacent levels, indicating that biomechanical behavior of high-grade HO is similar to a fused segment, thereby reducing the intended and initial motion preservation.

Table 5. Cervical Spine Analysis articles overview and extracted outcomes. (continued)

Author (Year)	Modality	Category	Aim	Model Features	2D/ 3D	Semi- / Fully- Automatic	N = / Included Vertebrae	Key Points
Makaremi (2019)[42]	X-ray	CVM Stage	* Determine the Cervical Vertebra Maturation Degree	* 6-layered CNN (supervised classification)* 5 convolution, normalization, max- pooling and dropout layers * 1 combination of dense, normalization and dropout layers * Stoachastic Gradient Descent (SGD)* Adapative Moment Estimation (Adam) optimization	2D	Fully	n = 2470(C2-C7)	* The results show the performances of the proposed method in different contexts with different number of images for training, evaluation and testing and different pre-processing of images. The proposed model and method are validated by cross validation.
Kok (2019) [37]	X-ray	CVM Stage	* Determine the Cervical Vertebra Maturation Degree	* k-nearest neighbors (k-NN) * Naïve Bayes (NB) * Decision tree (Tree)* Artificial Neural Networks (ANN)* SVM * RF * Logistic Regression (Log.Regr.)	2D	Semi	n = 300 (C2-C4)	* According to the average rank of the algorithms in predicting the CSV classes, ANN was the most stable algorithm with its 2.17 average rank. * kNN and Log.Regr. algorithms had the lowest accuracy values. SVM-RF-Tree and NB algorithms had varying accuracy values. ANN could be the preferred method for determining CVS.

Table 5. Cervical Spine Analysis articles overview and extracted outcomes. (continued)

Author (Year)	Modality	Category	Aim	Model Features	2D/ 3D	Semi- / Fully- Automatic	N = / Included Vertebrae	Key Points
Amasya (2020)[4]	X-ray	CVM Stage	* Determine the Cervical Vertebra Maturation Degree	* 1-layered ANN * Softmax acitivation function	2D	Semi	n = 647(C2-C4)	* Interobserver agreement was lower than Agreement between ANN and observers combined. Average of 58.3% agreement was observed between the ANN model and the human observers. * Developed ANN model performed close to, if not better than, human observers in CVM analysis.
Hopkins (2019)[30]	MRI	Clinical Prognosis / Prediction	* Model 1: Predicting Cervical Spondylotic Myelopathy (CSM) * Model 2: Predicting CSM severity (mJOA score)	* 7-layered 'deep' neural network	2D	Semi	n = 28 (C2-T1)	* Model 1: The mean cross-validated accuracy of the trained model was 86.50% (95% confidence interval, 85.16%e87.83%) with a median accuracy of 90.00%. Average sensitivity, specificity, positive predictive value, and negative predictive value were 90.25%, 85.05%, 81.58%, and 91.94%, respectively. * Model 2: The mJOA model predicted scores, with a mean and median error of e0.29 mJOA points and e0.08 mJOA points, respectively, mean error per batch was 0.714 mJOA points.

Table 5. Cervical Spine Analysis articles overview and extracted outcomes. (continued)

Author (Year)	Modality	Category	Aim	Model Features	2D/ 3D	Semi- / Fully-Automatic	N = / Included Vertebrae	Key Points
Jin (2019) [34]	MRI (DTI)	Clinical Prognosis / Prediction	* Evaluate the potential of AI in the analysis of DTI for the prognosis of myelopathy	* Logistic regression (LR) * K-nearest neighbors (KNN)* Radial basis function* Kernel SVM * VGG16	3D	Semi	n = 75(C2-C7)	The accuracy of the classifications reached 74.2% ± 1.6% for LR, 85.6% ± 1.4% for KNN, 89.7% ± 1.6% for RBF-SVM, and 59.2% ± 3.8% for the deep leaning model. The RBF-SVM algorithm achieved the best accuracy, with sensitivity and specificity of 85.0% ± 3.4% and 92.4% ± 1.9% respectively.
du Bois (2007)[21]	CT, MRI, PET	Image Registration / Planning	* Develop a pipeline to register multimodal images of the neck * Estimate the displacement field generated by articulated bodies for patient undergoing radiotherapy	* Linear elastic biomechanical finite element model* Simultaneous Perturbation Stoachstic Approximation (SPSA) for optimization	3D	Semi	n = 7 (C2-C7)	* Significant decreases in the mean, min and max errors. Mean errors before registration [3.88–5.96 mm] decrease after registration [1.91– 3.31 mm] and variances decreases from [1.26–1.4 mm] to [0.65–1.11 mm]. The minimum errors before registration [2.3–3.88 mm] become [0.2–1.17 mm] and the maximum errors [5.59–8.47 mm] decrease after registration [3.17–4.9 mm].

Table 5. Cervical Spine Analysis articles overview and extracted outcomes. (continued)

Author (Year)	Modality	Category	Aim	Model Features	2D/ 3D	Semi- / Fully- Automatic	N = / Included Vertebrae	Key Points
Pekar (2007) [48]	MRI	ImageRegistration / Planning	* Automate planning of MRI scans of the spine	* Anatomy recognition algorithm* Eigen analysis of the image Hessian* Template matching * Triangulated surface mesh and orientation vector * Robust multi-variate median approach* Rigid registration of the detected set of landmarks with the training set atlas	3D	Fully	n = 15(C2-C7)	* Results for detection of disc candidates: true detections 95.3%. * Visual evaluation of the validation study demonstrates the seemingly robust results for automated planning vs. manual planning.
Dzyubachyk (2013)[22]	MRI	Image Registration / Planning	* Automate the reconstruction of the complete spine from multi station 7 T MR	* Intensity inhomogeneity correction* 3D gradient- correlation (GC) * Coherent local intensity clustering (CLIC) * Fuzzy-c-means- clustering	3D	Fully	n = 23(Whole Spine)	* In all the test cases, the algorithm was able to produce correct reconstruction of the spine volume. The resulting mean registration error (0.53 mm) is found to be lower than the pixel size.

Table 5. Cervical Spine Analysis articles overview and extracted outcomes. (continued)

Author (Year)	Modality	Category	Aim	Model Features	2D/ 3D	Semi- / Fully-Automatic	N = / Included Vertebrae	Key Points
Rashad (2019)[50]	MRI, CT	Image Registration / Planning	* Automate elastic image registration, enables elastic MRI-to-CT image co-registration	* Elastic image registration* Elastic fusion of 3D MRI data* Deformable registration* Rigid fusion	3D	Fully	n = 10(C2-T2)	* Elastic fusion of 3D MRI data showed the highest image registration accuracy (target registration error of 3.26 mm with 95% confidence).* Elastic fusion of 2D axial MRI data (<4.75 mm with 95% c.) was more reliable than for 2D sagittal sequences (<6.02 mm with 95% c.). * The Deep Learning method enables elastic MRI-to-CT image co-registration for cervical indications with changes of the head position.
Schmitz (2004)[52]	CT	Clinical / Radiological Feature Detection	* Determine regional variations in the thickness of human cervical spine endplates with high spatial resolution	* Cubic interpolation algorithm* Mathlab gradient function change in density* Voxel wise statistical comparison* Statistical Parametric Mapping (SPM)* Gaussian random fields	2D	Fully	n = 6 (C4-C7)	* Anterior and medial aspects of superior endplates were shown to be significantly thinner than lateral and dorsal parts. Superior endplates were found to be thicker than inferior endplates.

Table 5. Cervical Spine Analysis articles overview and extracted outcomes. (continued)

Author (Year)	Modality	Category	Aim	Model Features	2D/ 3D	Semi- / Fully- Automatic	N = / Included Vertebrae	Key Points
Tan (2012) [58]	CT	Clinical / Radiological Feature Detection	* Automate vertebral height measurement	* Level set based * Triangular meshes * Marching vertebral bodies* Level set evolving on triangular meshes	3D	Semi	n = 1(C1-C7)	* The method has high precision, with a coefficient of variation of only 0.197% and Bland-Altmann 95% limits of agreement of [-0.11, 0.13] mm. * For local heights (anterior, middle, posterior) the algorithm was up to 4.2 times more precise than a manual mid-sagittal plane method.
Shin (2020) [53]	X-ray	Clinical / Radiological Feature Detection	* To analyze the temporal trends in cervical curvature across sex and age groups using an automated Deep Learning system (DLS).	* U-net* CNN * Polynomial regression * Harrison posterior tangent method	2D	Fully	n = 13 691(C2-C7)	* This study suggests a significant, increasing loss of normal cervical lordotic curvature for both sexes and young adults that is greater in progressively younger cohorts and women. * The performance of DLS segmentation for the anterior vertebral line had a pixel-wise accuracy of 96.67%

Quantitative synthesis

It was considered to pool accuracy rates in the Conventional Machine Learning and Deep Learning segmentation groups, however it was found that outcomes in the included studies were too heterogeneously reported for doing so. Authors chose to report different outcome metrics and the majority did not report on uncertainty metrics (confidence intervals, standard errors, standard deviations or p-values) with their primary outcome. Pooling the data would therefore require statistical imputation for the majority of the uncertainty metrics. Subsequently, this means that heterogeneity tests, such as the I 2, were not performed, as data could not be pooled.

Discussion

In this systematic review an overview was provided of the literature on the available Machine Learning techniques for automated image analysis of the cervical spine on radiological imaging. The results of the included studies show a wide variety of possibilities in Machine Learning methods, depending on the aim of the application and the available modalities. In segmentation models, Deep Learning methods show promising results with the application of (fully automatic) CNN models using X-ray, CT or MR imaging. Regarding cervical spine analysis, the biomechanical properties are most often studied using finite element models. The application of artificial neural networks and support vector machine models looks promising for other classification purposes.

Most of the published work on image analysis of the spine focusses on the (thoraco-) lumbar spine. This can be explained by the higher prevalence of lumbar spine pathology, as compared to cervical spine pathology. However, this study, focusing on the cervical spine, is the first of its kind and we therefore believe it can be used as a reference study for all researchers aiming to use radiological image analysis for the cervical spine, as well as other diseases in the neck area.

Unfortunately, results in this systematic review were too heterogeneously reported and therefore pooling the results was not possible. Reporting outcomes clearly and homogenously is an important requirement to compare performance among publications. The authors of this review want to plead for more consistent reporting of outcomes, i.e. the same set of outcome variables for every segmentation, classification or prediction study in order to increase the external validity and reproducibility of these type of studies. Several guidelines that describe the appropriate reporting process for Machine Learning studies have been published [29, 41]. However, after reviewing the vast amount of data from the included studies in this systematic review it can be concluded additional guidelines for reporting specifically on image analysis studies using machine learning, are needed. Apart from the recommendation to report a minimal of accuracy (in percentages from 0-100%) and error (in mm), reporting uncertainty metrics (confidence intervals, standard errors or standard deviations) with the primary outcome metrics should be required, as it is essential in order to unify the reporting process and aids pooling of results from future studies. Another essential recommendation is for authors to share code. The majority of publications included in

this systematic review did not share their code. Creating an academic environment in which code sharing is promoted is essential to keep improving the work in this field.

The concept of 'Grand Challenges' presents a promising alternative to current comparative research on the topic of image analysis, by eliminating a range of biases. The aim of these public challenges is to let participants apply their algorithms to the provided Grand Challenge task, using the public test set of images provided by the challenge organizers. In a Grand Challenge organized for analysis of breast histology images, a total of 64 submitted algorithms improved the state-of-the-art in classification of microscopy images to an accuracy of 84% [5].

This systematic review demonstrates a solid body of evidence describing effective segmentation of the cervical spine, with CNN achieving highest accuracy combined with the lowest computing times. Additionally, publications on the different applications for cervical spine analysis show high potential for Machine Learning for several classification and prediction tasks. However, the possibilities for implementation are far-reaching and several newer applications still deserve more attention in future research, including; automated detection, localization and classification of degenerative changes, specifically in the cervical spine. On thoracolumbar CT machine learning was used for automated detection of sclerotic metastases and detection, localization and classification of traumatic vertebral body fractures [11, 12], something that has not been done for the cervical spine yet. On thoracolumbar lateral X-rays the intervertebral disc height measurements were conducted for 1186 participants using machine learning [3], while the study included in this review on the same topic for the cervical spine showed results for only 1 patient [58].

The challenges in future research are not just in focusing on the cervical vertebrae or increasing the numbers of images, but also in the integration of different models into one fully automated pathway. Incorporating both radiological and clinical parameters into a fully-automatic model and implementing those into the clinical workflow is the end goal. As was established in this review, the detection and segmentation of the cervical spine have achieved sufficient attention in research, but it is the clinically important classification and prediction tasks, and combining those with detection and segmentation into a fully automatic structure, what future research should focus on.

References

1. S. Al Arif, K. Knapp, G. Slabaugh, Fully automatic cervical vertebrae segmentation framework for X-ray images, Comput Methods Programs Biomed 157 (2018) 95-111.
2. S.M.M.R. Al Arif, M. Gundry, K. Knapp, G. Slabaugh, Improving an Active Shape Model with Random Classification Forest for Segmentation of Cervical Vertebrae, Springer International Publishing, Cham, 2016, pp. 3-15.
3. B.T. Allaire, M.C. DePaolis Kaluza, A.G. Bruno, E.J. Samelson, D.P. Kiel, D.E. Anderson, M.L. Bouxsein, Evaluation of a new approach to compute intervertebral disc height measurements from lateral radiographic views of the spine, European Spine Journal 26(1) (2017) 167-172.
4. H. Amasya, E. Cesur, D. Yıldırım, K. Orhan, Validation of cervical vertebral maturation stages: Artificial intelligence vs human observer visual analysis, Am J Orthod Dentofacial Orthop 158(6) (2020) e173-e179.
5. G. Aresta, T. Araújo, S. Kwok, S.S. Chennamsetty, M. Safwan, V. Alex, B. Marami, M. Prastawa, M. Chan, M. Donovan, G. Fernandez, J. Zeineh, M. Kohl, C. Walz, F. Ludwig, S. Braunewell, M. Baust, Q.D. Vu, M.N.N. To, E. Kim, J.T. Kwak, S. Galal, V. Sanchez-Freire, N. Brancati, M. Frucci, D. Riccio, Y. Wang, L. Sun, K. Ma, J. Fang, I. Kone, L. Boulmane, A. Campilho, C. Eloy, A. Polónia, P. Aguiar, BACH: Grand challenge on breast cancer histology images, Med Image Anal 56 (2019) 122-139.
6. H.J. Bae, H. Hyun, Y. Byeon, K. Shin, Y. Cho, Y.J. Song, S. Yi, S.U. Kuh, J.S. Yeom, N. Kim, Fully automated 3D segmentation and separation of multiple cervical vertebrae in CT images using a 2D convolutional neural network, Computer Methods and Programs in Biomedicine 184 (no pagination) (2019).
7. C. Balkovec, J. Veldhuis, J.W. Baird, G. Wayne Brodland, S.M. McGill, Digital tracking algorithm reveals the influence of structural irregularities on joint movements in the human cervical spine, Clinical Biomechanics 56 (2018) 11-17.
8. S. Banik, R.M. Rangayyan, G.S. Boag, Automatic segmentation of the ribs, the vertebral column, and the spinal canal in pediatric computed tomographic images, J Digit Imaging 23(3) (2010) 301-22.
9. M. Benjelloun, S. Mahmoudi, Spine localization in X-ray images using interest point detection, J Digit Imaging 22(3) (2009) 309-18.
10. S.S. Burnett, G. Starkschalla, C.W. Stevens, Z. Liao, A deformable-model approach to semi-automatic segmentation of CT images demonstrated by application to the spinal canal, Med Phys 31(2) (2004) 251-63.
11. J.E. Burns, J. Yao, T.S. Wiese, H.E. Muñoz, E.C. Jones, R.M. Summers, Automated detection of sclerotic metastases in the thoracolumbar spine at CT, Radiology 268(1) (2013) 69-78.
12. J.E. Burns, J. Yao, H. Muñoz, R.M. Summers, Automated Detection, Localization, and Classification of Traumatic Vertebral Body Fractures in the Thoracic and Lumbar Spine at CT, Radiology 278(1) (2016) 64-73.
13. Cai, M. Landis, D.T. Laidley, A. Kornecki, A. Lum, S. Li, Multi-modal vertebrae recognition using Transformed Deep Convolution Network, Comput Med Imaging Graph 51 (2016) 11-9.
14. H. Chen, C. Shen, J. Qin, D. Ni, L. Shi, J.C.Y. Cheng, P.-A. Heng, Automatic Localization and Identification of Vertebrae in Spine CT via a Joint Learning Model with Deep Neural Networks, Springer International Publishing, Cham, 2015, pp. 515-522.
15. L. Chen, Z. Lan, X. Xu, J. Lin, H. Hu, Accuracy and repeatability of computer aided cervical vertebra landmarking in cephalogram, J Huazhong Univ Sci Technolog Med Sci 32(1) (2012) 119-123.

16. Y. Chen, Y. Gao, K. Li, L. Zhao, J. Zhao, Vertebrae Identification and Localization Utilizing Fully Convolutional Networks and a Hidden Markov Model, IEEE Transactions on Medical Imaging 39(2) (2020) 387-399.
17. M. Clogenson, J.M. Duff, M. Luethi, M. Levivier, R. Meuli, C. Baur, S. Henein, A statistical shape model of the human second cervical vertebra, International Journal of Computer Assisted Radiology and Surgery 10(7) (2015) 1097-1107.
18. S.P. Cohen, Epidemiology, Diagnosis, and Treatment of Neck Pain, Mayo Clinic Proceedings 90(2) (2015) 284-299.
19. S. Daenzer, S. Freitag, S. von Sachsen, H. Steinke, M. Groll, J. Meixensberger, M. Leimert, VolHOG: a volumetric object recognition approach based on bivariate histograms of oriented gradients for vertebra detection in cervical spine MRI, Med Phys 41(8) (2014) 082305.
20. B. De Leener, J. Cohen-Adad, S. Kadoury, Automatic Segmentation of the Spinal Cord and Spinal Canal Coupled With Vertebral Labeling, IEEE Trans Med Imaging 34(8) (2015) 1705-18.
21. A. du Bois d'Aische, M. De Craene, X. Geets, V. Gregoire, B. Macq, S.K. Warfield, Estimation of the deformations induced by articulated bodies: Registration of the spinal column, Biomedical Signal Processing and Control 2(1) (2007) 16-24.
22. O. Dzyubachyk, B.P. Lelieveldt, J. Blaas, M. Reijnierse, A. Webb, R.J. van der Geest, Automated algorithm for reconstruction of the complete spine from multistation 7T MR data, Magn Reson Med 69(6) (2013) 1777-86.
23. A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542(7639) (2017) 115-118.
24. D. Forsberg, E. Sjöblom, J.L. Sunshine, Detection and Labeling of Vertebrae in MR Images Using Deep Learning with Clinical Annotations as Training Data, J Digit Imaging 30(4) (2017) 406-412.
25. G. Giulietti, P.E. Summers, D. Ferraro, C.A. Porro, B. Maraviglia, F. Giove, Semiautomated segmentation of the human spine based on echoplanar images, Magn Reson Imaging 29(10) (2011) 1429-36.
26. B. Glocker, J. Feulner, A. Criminisi, D.R. Haynor, E. Konukoglu, Automatic localization and identification of vertebrae in arbitrary field-of-view CT scans, Med Image Comput Comput Assist Interv 15(Pt 3) (2012) 590-8.
27. B. Glocker, D. Zikic, E. Konukoglu, D.R. Haynor, A. Criminisi, Vertebrae localization in pathological spine CT via dense classification from sparse annotations, Med Image Comput Comput Assist Interv 16(Pt 2) (2013) 262-70.
28. S. Hanaoka, Y. Masutani, M. Nemoto, Y. Nomura, S. Miki, T. Yoshikawa, N. Hayashi, K. Ohtomo, A. Shimizu, Landmark-guided diffeomorphic demons algorithm and its application to automatic segmentation of the whole spine and pelvis in CT images, Int J Comput Assist Radiol Surg 12(3) (2017) 413-430.
29. B.J. Heil, M.M. Hoffman, F. Markowetz, S.-I. Lee, C.S. Greene, S.C. Hicks, Reproducibility standards for machine learning in the life sciences, Nature Methods 18(10) (2021) 1132-1135.
30. B.S. Hopkins, K.A. Weber, 2nd, K. Kesavabhotla, M. Paliwal, D.R. Cantrell, Z.A. Smith, Machine Learning for the Prediction of Cervical Spondylotic Myelopathy: A Post Hoc Pilot Study of 28 Participants, World Neurosurg 127 (2019) e436-e442.
31. S.H. Huang, Y.H. Chu, S.H. Lai, C.L. Novak, Learning-based vertebra detection and iterative normalized-cut segmentation for spinal MRI, IEEE Trans Med Imaging 28(10) (2009) 1595-605.
32. R. Jakubicek, J. Chmelik, J. Jan, P. Ourednicek, L. Lambert, G. Gavelli, Learning–based vertebra localization and labeling in 3D CT data of possibly incomplete and pathological spines, Computer Methods and Programs in Biomedicine 183 (2020) 105081.

33. R. Jakubicek, J. Chmelik, P. Ourednicek, J. Jan, Deep-learning-based fully automatic spine centerline detection in CT data, Annu Int Conf IEEE Eng Med Biol Soc 2019 (2019) 2407-2410.

34. R. Jin, K.D. Luk, J.P.Y. Cheung, Y. Hu, Prognosis of cervical myelopathy based on diffusion tensor imaging with artificial intelligence methods, NMR Biomed 32(8) (2019) e4114.

35. C.C. Kage, M. Akbari-Shandiz, M.H. Foltz, R.L. Lawrence, T.L. Brandon, N.E. Helwig, A.M. Ellingson, Validation of an automated shape-matching algorithm for biplane radiographic spine osteokinematics and radiostereometric analysis error quantification, PLoS One 15(2) (2020) e0228594.

36. T. Klinder, J. Ostermann, M. Ehm, A. Franz, R. Kneser, C. Lorenz, Automated model-based vertebra detection, identification, and segmentation in CT images, Med Image Anal 13(3) (2009) 471-82.

37. H. Kök, A.M. Acilar, M.S. İzgi, Usage and comparison of artificial intelligence algorithms for determination of growth and development by cervical vertebrae stages in orthodontics, Prog Orthod 20(1) (2019) 41.

38. M.A. Larhmam, M. Benjelloun, S. Mahmoudi, Vertebra identification using template matching modelmp and K-means clustering, Int J Comput Assist Radiol Surg 9(2) (2014) 177-87.

39. F. Lecron, M. Benjelloun, S. Mahmoudi, Cervical spine mobility analysis on radiographs: a fully automatic approach, Comput Med Imaging Graph 36(8) (2012) 634-42.

40. X. Liu, J. Yang, S. Song, W. Cong, P. Jiao, H. Song, D. Ai, Y. Jiang, Y. Wang, Sparse intervertebral fence composition for 3D cervical vertebra segmentation, Phys Med Biol 63(11) (2018) 115010.

41. W. Luo, D. Phung, T. Tran, S. Gupta, S. Rana, C. Karmakar, A. Shilton, J. Yearwood, N. Dimitrova, T.B. Ho, S. Venkatesh, M. Berk, Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View, J Med Internet Res 18(12) (2016) e323.

42. M. Makaremi, C. Lacaule, A. Mohammad-Djafari, Deep Learning and Artificial Intelligence for the Determination of the Cervical Vertebra Maturation Degree from Lateral Radiography, Entropy 21(12) (2019) 24.

43. A. Mehmood, M.U. Akram, A. Tariq, Vertebra localization and centroid detection from cervical radiographs, 2017 International Conference on Communication, Computing and Digital Systems (C-CODE), 2017, pp. 287-292.

44. H. Mirzaalian, M. Wels, T. Heimann, B.M. Kelm, M. Suehling, Fast and robust 3D vertebra segmentation using statistical shape models, Annu Int Conf IEEE Eng Med Biol Soc 2013 (2013) 3379-82.

45. D. Moher, A. Liberati, J. Tetzlaff, D.G. Altman, P.G. The, Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement, PLOS Medicine 6(7) (2009) e1000097.

46. M. Nikkhoo, C.H. Cheng, J.L. Wang, C.C. Niu, M. Parnianpour, K. Khalaf, The Biomechanical Response of the Lower Cervical Spine Post Laminectomy: Geometrically-Parametric Patient-Specific Finite Element Analyses, Journal of Medical and Biological Engineering. (2020).

47. M. Nikkhoo, C.-H. Cheng, J.-L. Wang, Z. Khoz, M. El-Rich, N. Hebela, K. Khalaf, Development and validation of a geometrically personalized finite element model of the lower ligamentous cervical spine for clinical applications, Computers in Biology and Medicine 109 (2019) 22-32.

48. V. Pekar, D. Bystrov, H.S. Heese, S.P. Dries, S. Schmidt, R. Grewer, C.J. den Harder, R.C. Bergmans, A.W. Simonetti, A.M. van Muiswinkel, Automated planning of scan geometries in spine MRI scans, Med Image Comput Comput Assist Interv 10(Pt 1) (2007) 601-8.

49. M. Rak, J. Steffen, A. Meyer, C. Hansen, K.D. Tönnies, Combining convolutional neural networks and star convex cuts for fast whole spine vertebra segmentation in MRI, Comput Methods Programs Biomed 177 (2019) 47-56.

50. A. Rashad, M. Heiland, P. Hiepe, A. Nasirpour, C. Rendenbach, J. Keuchel, M. Regier, A. Al-Dam, Evaluation of a novel elastic registration algorithm for spinal imaging data: A pilot clinical study, Int J Med Robot 15(3) (2019) e1991.
51. S. Schmidt, J. Kappes, M. Bergtholdt, V. Pekar, S. Dries, D. Bystrov, C. Schnörr, Spine detection and labeling using a parts-based graphical model, Inf Process Med Imaging 20 (2007) 122-33.
52. B. Schmitz, T. Pitzen, T. Beuter, W.I. Steudel, W. Reith, Regional variations in the thickness of cervical spine endplates as measured by computed tomography, Acta Radiol 45(1) (2004) 53-8.
53. Y. Shin, K. Han, Y.H. Lee, Temporal Trends in Cervical Spine Curvature of South Korean Adults Assessed by Deep Learning System Segmentation, 2006-2018, JAMA Netw Open 3(10) (2020) e2020961.
54. P.L. Sinnott, S.K. Dally, J. Trafton, J.L. Goulet, T.H. Wagner, Trends in diagnosis of painful neck and back conditions, 2002 to 2011, Medicine 96(20) (2017).
55. S. Srinivasan, S.D. Kumar, S. R, D.D. Jebaseelan, N. Yoganandan, RajasekaranS, Effect of heterotopic ossification after bryan-cervical disc arthroplasty on adjacent level range of motion: A finite element study, Journal of Clinical Orthopaedics and Trauma. (2020).
56. A. Suzani, A. Seitel, Y. Liu, S. Fels, R.N. Rohling, P. Abolmaesumi, Fast Automatic Vertebrae Detection and Localization in Pathological CT Scans - A Deep Learning Approach, in: N. Navab, J. Hornegger, W.M. Wells, A.F. Frangi (Eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham, 2015, pp. 678-686.
57. S. Tabibu, P.K. Vinod, C.V. Jawahar, Pan-Renal Cell Carcinoma classification and survival prediction from histopathology images using deep learning, Sci Rep 9(1) (2019) 10509.
58. S. Tan, J. Yao, L. Yao, M.M. Ward, High precision semi-automated vertebral height measurement using computed tomography: A phantom study, Annu Int Conf IEEE Eng Med Biol Soc 2012 (2012) 1554-7.
59. A. Tiulpin, J. Thevenot, E. Rahtu, P. Lehenkari, S. Saarakkala, Automatic Knee Osteoarthritis Diagnosis from Plain Radiographs: A Deep Learning-Based Approach, Scientific Reports 8(1) (2018) 1727.
60. J. Urrutia, T. Zamora, R. Yurac, M. Campos, J. Palma, S. Mobarec, C. Prada, An Independent Inter- and Intraobserver Agreement Evaluation of the AOSpine Subaxial Cervical Spine Injury Classification System, Spine 42(5) (2017).
61. X. Wang, S. Zhai, Y. Niu, Automatic Vertebrae Localization and Identification by Combining Deep SSAE Contextual Features and Structured Regression Forest, J Digit Imaging 32(2) (2019) 336-348.
62. K.L. Weiss, J.M. Storrs, R.B. Banto, Automated spine survey iterative scan technique, Radiology 239(1) (2006) 255-62.
63. G. Wells, B. Shea, D. O'Connel, P.J.V. Welch, M. Losos, P. Tugwell, The NewcastleOttawa Scale (NOS) for assessing the quality of nonradomised studies in metaanalysis, http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp ((Accessed 25 December 2020)).
64. X. Xi, H. Hong-Wei, Y. Xu-Cheng, L. Ning, S.H. Shafin, Automatic segmentation of cervical vertebrae in X-ray images, The 2012 International Joint Conference on Neural Networks (IJCNN), 2012, pp. 1-8.
65. Y. Xu, A. Hosny, R. Zeleznik, C. Parmar, T. Coroller, I. Franco, R.H. Mak, H. Aerts, Deep Learning Predicts Lung Cancer Treatment Response from Serial Medical Imaging, Clin Cancer Res 25(11) (2019) 3266-3275.
66. G. Zamora, H. Sari-Sarraf, R.L. Long, Hierarchical segmentation of vertebrae from x-ray images, Proc.SPIE, 2003.
67. Y. Zhang, E.M. Lobo-Mueller, P. Karanicolas, S. Gallinger, M.A. Haider, F. Khalvati, CNN-based survival model for pancreatic ductal adenocarcinoma in medical imaging, BMC Medical Imaging 20(1) (2020) 11.

Appendix A

Complete overview of the Risk of Bias assessment for conventional Machine Learning segmentation articles.

Author (Year)	Search	1.1	1.2	1.3	1.4	2	3.1	3.2	3.3	Total
Zamora (2003)	Additional	1	1	1	1	1	1	0	0	6
Burnett (2004)	PubMed	1	1	1	1	1	1	0	0	6
Weiss (2006)	PubMed	1	1	1	1	0	1	0	0	5
Schmidt (2007)	PubMed	0	1	1	1	2	1	1	0	7
Klinder (2009)	PubMed	1	1	1	1	2	1	0	0	7
Huang (2009)	PubMed	1	1	1	1	2	0	0	0	6
Banik (2010)	PubMed	1	1	1	1	0	1	1	0	6
Giulietti (2011)	PubMed	0	1	1	1	1	1	1	0	6
Chen (2012)	PubMed	0	0	1	1	0	1	0	0	3
Glocker (2012)	PubMed	1	1	1	1	2	0	1	0	7
Xu (2012)	Additional	1	1	1	1	1	1	0	0	6
Glocker (2013)	PubMed	1	1	1	1	2	0	1	1	8
Mirzaalian (2013)	PubMed	1	1	1	1	1	0	1	0	6
Larhmam (2014)	PubMed	1	1	1	1	1	1	1	0	7
Daenzer (2014)	PubMed	1	1	1	1	2	1	1	1	9
De Leener (2015)	PubMed	1	1	1	1	2	1	1	0	8
Clogenson (2015)	Embase	0	1	1	1	0	0	1	1	5
Al Arif (2016)	Additional	1	1	1	1	2	1	1	0	8
Mehmood (2017)	Additional	1	1	1	1	1	0	0	0	5
Hanaoka (2017)	PubMed	0	1	0	1	2	1	0	0	5

Color coded with red (high risk of bias), orange (intermediate risk of bias) and green (low risk of bias).

Appendix B

Complete overview of the Risk of Bias assessment for Deep Learning segmentation articles.

Author (Year)	Search	1.1	1.2	1.3	1.4	2	3.1	3.2	3.3	Total
Suzani (2015)	Additional	1	1	1	0	2	1	1	0	7
Chen (2015)	Additional	1	1	1	0	2	1	1	0	7
Cai (2016)	PubMed	1	1	1	0	2	1	0	0	6
Forsberg (2017)	PubMed	1	1	1	1	1	1	1	0	7
Liu (2018)	PubMed	1	1	1	1	2	1	1	0	8
Al Arif (2018)	PubMed	1	1	1	0	2	1	1	0	7
Jakubicek (2019)	PubMed	1	1	0	1	1	1	1	0	6
Bae (2019)	Embase	1	1	1	1	1	1	1	0	7
Rak (2019)	PubMed	1	1	1	0	1	1	1	0	6
Wang (2019	PubMed	1	1	1	1	2	1	1	0	8
Chen (2020)	Embase	1	1	1	0	2	1	1	0	7
Jakubicek (2020)	PubMed	1	1	1	1	2	1	1	0	8

Color coded with red (high risk of bias), orange (intermediate risk of bias) and green (low risk of bias).

Appendix C

Complete overview of the Risk of Bias assessment for cervical spine analysis articles.

Author (Year)	Search	1.1	1.2	1.3	1.4	2	3.1	3.2	3.3	Total
Schmitz (2004)	PubMed	0	1	0	0	1	1	0	0	3
Pekar (2007)	PubMed	1	1	1	1	1	1	0	0	6
du Bois (2007)	Embase	0	1	1	1	0	1	1	0	5
Benjelloun (2009)	PubMed	1	1	1	1	1	0	0	0	5
Lecron (2012)	PubMed	1	1	1	1	2	1	0	0	7
Tan (2012)	PubMed	0	1	1	1	0	1	1	0	5
Dzyubachyk (2013)	PubMed	1	1	1	0	1	1	1	0	6
Balkovec (2017)	Embase	0	1	1	1	1	1	1	0	6
Nikkhoo (2019)	Additional	0	1	1	1	1	1	1	0	6
Makaremi (2019)	Web of Science	1	0	1	1	1	1	1	0	6
Hopkins (2019)	PubMed	0	1	0	1	1	1	1	0	5
Jin (2019)	PubMed	1	1	1	1	1	1	1	0	7
Kok (2019)	PubMed	0	0	1	1	1	1	0	1	5
Rashad (2019	PubMed	1	1	1	0	1	1	1	0	6
Srinivasan (2020)	Embase	0	1	1	1	1	1	0	0	5
Nikkhoo (2020)	Embase	0	1	1	1	1	1	1	0	6
Amasya (2020)	PubMed	1	1	1	1	1	1	1	0	7
Kage (2020)	PubMed	0	1	0	1	0	1	1	1	5
Shin (2020)	PubMed	1	1	1	1	2	1	1	0	8

Color coded with red (high risk of bias), orange (intermediate risk of bias) and green (low risk of bias).