

IMAGING INFORMATICS AND ARTIFICIAL INTELLIGENCE Open Access



VERIDAH: solving enumeration anomaly–aware vertebra labeling across imaging sequences

Hendrik Möller^{1,2*} , Hanna Schön^{1,3}, Robert Graf^{1,2}, Matan Atad^{1,2}, Nathan Molinier⁴, Anjany Sekuboyina⁵, Bettina K. Budai⁶, Fabian Bamberg⁷, Steffen Ringhof⁷, Christopher Schlett⁷, Tobias Pischon⁸, Thoralf Niendorf⁹, Josua A. Decker¹⁰, Marc-André Weber³, Bjoern Menze⁵, Daniel Rueckert^{2,11} and Jan S. Kirschke¹

Abstract

Objectives Although the identification of enumeration anomalies has potential clinical implications for chronic back pain and operation planning, the thoracolumbar junction is poorly assessed and rarely described in clinical reports. Additionally, even though multiple deep-learning-based vertebra labeling algorithms exist, there is a lack of methods to automatically label enumeration anomalies.

Materials and methods We introduce “Vertebra Identification with Anomaly Handling” (VERIDAH), a novel vertebra labeling algorithm based on multiple classification heads combined with a weighted vertebra sequence prediction algorithm. We use the NAKO dataset for MRI and a CT cohort to retrospectively evaluate the labeling performance by computing how often the model is correct in predicting the vertebra labels, including enumeration anomalies. Statistical significance was assessed using the Wilcoxon signed-rank test.

Results On 1000 T2w sagittal (mean age 52 ± 11 , 52% female) and 365 CT (mean age 74 ± 21 , 44% female) images, we show that our approach surpasses existing models on T2w TSE sagittal (98.30% vs. 94.24% of subjects with all vertebrae correctly labeled, $p < 0.01$) and CT imaging (99.18% vs. 77.26% of subjects with all vertebrae correctly labeled, $p < 0.01$) and works in arbitrary field-of-view images. VERIDAH correctly labeled the presence of thoracic enumeration anomalies in 87.80% and 96.30% of T2w and CT images, respectively, and lumbar enumeration anomalies in 94.48% and 97.22% for T2w and CT, respectively.

Conclusion VERIDAH offers enumeration anomaly robust vertebra labeling in CT and T2w sagittal imaging and is publicly released along with a spine segmentation model.

Key Points

Question Automatically identifying enumeration anomalies in the spine is important in determining the correct vertebra labels, especially in limited field-of-view imaging.

Findings A carefully designed classifier with multiple classification heads and a sequence predictor provides superior results on large benchmarks, both in CT and MRI.

Clinical relevance The proposed method identifies vertebra levels even in very limited field-of-view imaging, and correctly predicts enumeration anomalies, making downstream tasks such as surgery planning easier. Its public availability enables studies and applications to take these spinal anomalies into account.

*Correspondence:

Hendrik Möller

hendrik.moeller@tum.de

Full list of author information is available at the end of the article

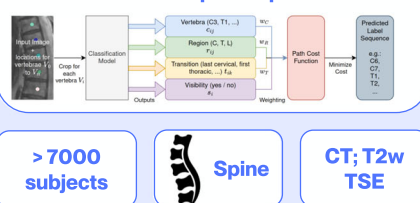
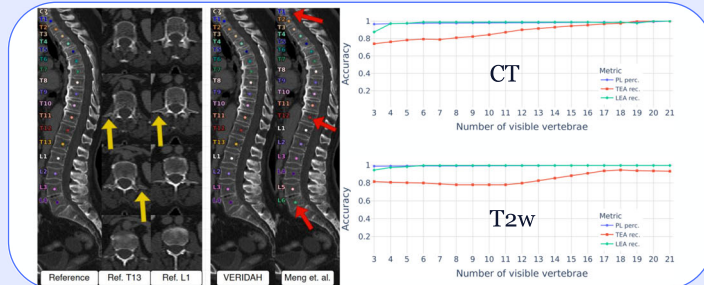
Keywords Computed tomography (CT), Enumeration anomaly, Machine learning, Magnetic resonance imaging (MRI), Spine

Graphical Abstract

VERIDAH: solving enumeration anomaly-aware vertebra labeling across imaging sequences

Can we predict vertebra labels in arbitrary field-of-view CT/MRI, even in the presence of enumeration anomalies?

- One CT and one MRI cohort
- Multi-head classifier
- Constrained sequence predictor



Achieves new state-of-the-art vertebra labeling performance, enabling enumeration anomaly-aware labeling even in limited field-of-view imaging.

Eur Radiol (2026) Möller H, Schön H, Graf R, et al;
DOI: 10.1007/s00330-026-12966-x

EUROPEAN SOCIETY OF RADIOLOGY
European Radiology

Introduction

Automated medical image analysis has repeatedly shown benefits for clinical assessment [1–3]. For spinal analysis, accurate vertebra identification is essential, for example, in fracture prediction [4]. The human spine typically consists of seven cervical, twelve rib-bearing thoracic, and five lumbar vertebrae [5]. However, there can be enumeration anomalies, causing a deviation of the number of thoracic and/or lumbar vertebrae by one [6, 7]. These thoracic (TEA) and lumbar (LEA) enumeration anomalies often co-occur with transitional anomalies such as thoracolumbar transitional vertebrae (TLTV), which exhibit mixed thoracic and lumbar morphology [8]. Studies have shown that the presence of enumeration anomalies is strongly correlated with transitional ones and has potential clinical implications for chronic back pain and degenerative changes [9–11]. Ignoring enumeration anomalies can also lead to vertebrae being labeled incorrectly. This can result in wrong-site surgery, where the intended vertebra is missed and adjacent vertebrae are operated on by mistake [6, 12].

There are a multitude of common methods for vertebra identification. One can count the vertebrae in

cranial-to-caudal order, assuming seven cervical and twelve thoracic vertebrae. Alternatively, the last lumbar vertebra can be identified and counted upward, assuming five lumbar vertebrae [13]. However, this approach fails in the presence of enumeration anomalies or limited fields of view (FOVs). A rib-based classification identifies the last thoracic vertebra by the presence of a rib longer than 38 mm and counts the vertebrae in both directions from there [7]. However, when the entire thorax is not visible, the number of thoracic vertebrae present becomes unclear. Moreover, measuring 38 mm in edge cases requires significant manual effort. The L5 can also be identified by the presence of the iliolumbar ligament connecting the L5 transverse process to the ilium. Lastly, nerve root morphology can be used to identify L5, as it has no proximal division [14]. However, both methods fail in a limited FOV where the L5 is not visible, and neither can distinguish thoracolumbar enumeration anomalies. In summary, because CT and MRI scans often have limited FOVs, all conventional methods fail, leading to inconsistent manual counting. Enumeration anomalies further increase the risk of labeling errors [15].

Table 1 Study cohort demographics

German National Cohort	Summary	Train	Test
Subjects (<i>n</i>)	6291	5291	1000
Age range (years)	20–72	20–72	22–72
M. age (years) $\pm$ SD	52 $\pm$ 11	52 $\pm$ 12	52 $\pm$ 11
M. weight (kg) $\pm$ SD	80 $\pm$ 16	80 $\pm$ 16	80 $\pm$ 16
M. height (m) $\pm$ SD	1.7 $\pm$ 0.1	1.7 $\pm$ 0.1	1.7 $\pm$ 0.1
Sex (% female)	52.4	52.4	52.1
TEA rate (%)	5.8	5.4	8.2
LEA rate (%)	9.7	8.8	14.6
M. visible vertebrae $\pm$ SD	23.1 $\pm$ 0.3	23.1 $\pm$ 0.3	23.1 $\pm$ 0.4
CT cohort	Summary	Train	Test
Subjects (<i>n</i>)	1536	1171	365
Age range (years)	18–100	18–100	20–100
M. age (years) $\pm$ SD	74 $\pm$ 21	74 $\pm$ 21	74 $\pm$ 21
M. weight (kg) $\pm$ SD	76 $\pm$ 16	75 $\pm$ 14	81 $\pm$ 21
M. height (m) $\pm$ SD	N/A	N/A	N/A
Sex (% female)	49.1	50.9	44.3
TEA rate (%)	5.5	5.1	6.8
LEA rate (%)	7.3	7.1	7.9
M. visible vertebrae $\pm$ SD	16.7 $\pm$ 3.6	16.7 $\pm$ 3.5	16.8 $\pm$ 4.0

The table shows the demographics for the whole datasets and individual train and test splits

yrs years, *kg* kilogram, *m* meters, *M.* mean, *SD* standard deviation, *TEA* thoracolumbar enumeration anomaly, *LEA* lumbar enumeration anomaly

The automatic identification of vertebrae has been widely studied [16–22]. There have been many development efforts, especially in CT [17, 18, 22]. However, some approaches are only tested on small cohorts [23], and almost none of them consider TEA or LEA, as the existing systems can only predict a maximum of 12 thoracic and five lumbar vertebrae [24]. We found three automated approaches that theoretically take this into account. Sekuboyina et al [25] developed a method for localization and identification of vertebrae in CT, using a conditional random field and user input to refine errors. Meng et al [26] utilized segmentation masks of vertebrae and a graph optimization problem for vertebra identification. Finally, Warszawer et al [27] locate specific intervertebral discs during the task of segmentation in order to identify the sequence of vertebrae visible in the image by counting from those discs in both directions. However, these methods either fail to label TEA and LEA cases reliably and robustly in arbitrary FOVs or have never been validated on a large dataset.

We address these limitations using large CT and MRI cohorts and propose a robust vertebra labeling method that handles enumeration anomalies and arbitrary FOVs. Our contributions are: (1) a multi-classification head training strategy for vertebra labeling, (2) a label sequence predictor based on classifier outputs, (3) demonstration of superior performance over state-of-the-art methods on CT and T2w MRI, and (4) public release of our approach and model weights.

Materials and methods

Data

We utilized the external German National Cohort (NAKO) (prospectively collected, participants gave informed consent) [28, 29], and an in-house CT dataset (waived informed consent by the local ethics committee, 593/21 S-NP).

For the T2w sagittal MRI data, we utilized a random subset of 6291 subjects of the German National Cohort (NAKO) and split the data roughly into 70/15/15% train/validation/test (see Table 1). We utilized SPINEPS [30], a segmentation network, to derive instance segmentation annotations for all vertebrae in our T2w sagittal MRI data. We excluded 33 subjects (0.52%) where this segmentation failed. We did not observe obvious reasons for these failures, such as hyperintense spots or pathologies. The segmentations yielded localizations for each vertebra.

The T2w sagittal images of the NAKO always show the whole spine, containing up to 26 fully visible vertebrae (see Fig. 1). They were produced with a consistent SIEMENS scanner and field strength of 3 Tesla. For CT, we utilized an in-house cohort consisting of images from various scanners (SIEMENS, Philips, GE, and TOSHIBA) and both bone and soft-tissue kernels. Some CT images are contrast-enhanced. The data includes segmentation annotations for the vertebrae, supervised by an expert with 24 years of experience. In contrast to the T2w MRI of the NAKO, the FOV of the CTs varied heavily, containing a range of 8–25 fully visible vertebrae (average 16.7 ± 3.6 vertebrae). We utilized 1536 images in which the first thoracic vertebra is present (see Fig. 1). This is necessary to manually assess TEAs and create the reference labels. In both MRI and CT, we consciously split in a way so that the test set contains more TEA and LEA to better evaluate the models.

When manually rating the thoracolumbar junction, we adapted the categorization of Park et al [8]. We defined a thoracic vertebra as one having a true rib, regardless of the rib's length. If there is only a unilateral or bilateral accessory ossification center (comparable to Type 4 as defined in [8]), we determined this to be a lumbar vertebra. In ambiguous cases, we judged by morphology (e.g., facet orientation) whether a vertebra is thoracic, based on

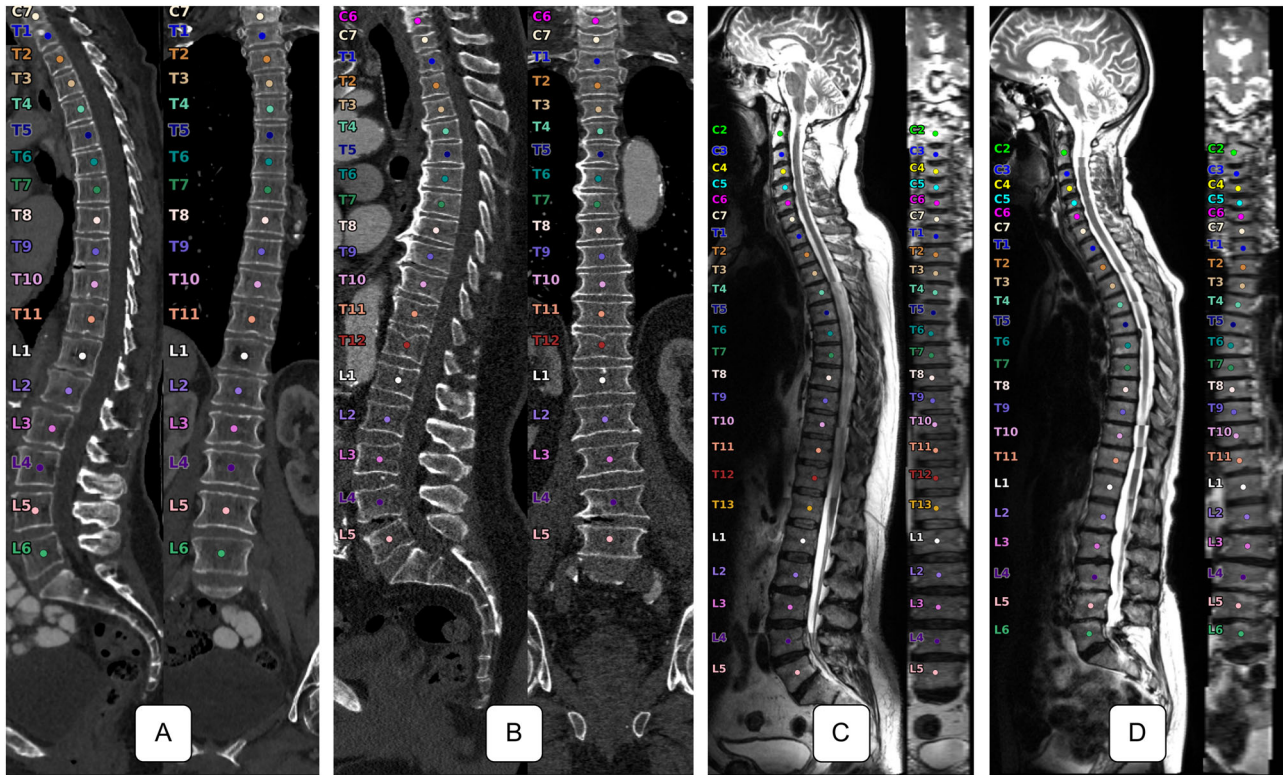


Fig. 1 Example data samples from our CT (**A, B**) and MRI (**C, D**) datasets. For each sample, it shows a sagittal and coronal projection along the spine as well as the corpus centroids with vertebra labels associated with each. While the CT images contain various field-of-views, the NAKO T2w sagittal MRI images always cover the whole spine. **A** Spine with eleven thoracic vertebrae, **B** normal spine with twelve thoracic and five lumbar vertebrae, **C** has an additional thoracic vertebra, and **D** a subject with only eleven thoracic vertebrae but six lumbar ones

Schinz et al [15]. They report an excellent overall inter-rater agreement ($\kappa > 0.9$). For the lumbosacral transition, we adapted the analysis of Konin et al [31]. In the presence of a lumbosacral transitional Castellvi anomaly, we defined the vertebra with this anomaly as a (partially) sacralized lumbar vertebra, regardless of the Castellvi type, i.e., still labeled this vertebra as the last lumbar vertebra. The segmentation model SPINEPS we used on the NAKO data was already trained to make that distinction, which we only had to manually correct in 16 cases (0.25%). Similarly, the annotations of the CT data already distinguished between the lumbar and sacral vertebrae. Apart from this, while manually rating each case, we did not find other segmentation-based localization errors that required manual intervention.

One expert with 3 years of experience (H.M.) performed the manual annotations for all data, supervised by an expert with 24 years of experience (J.S.K.). The supervision included an independent rating of 100 CT and MRI cases for the lumbosacral junction, followed by a joint discussion about discrepancies. We compute the inter-rater agreement over the initial, isolated created ratings.

Afterward, the supervisor confirmed all cases with stump ribs and anomalous numbers of vertebrae (1297 cases for MRI, 117 in CT).

VERIDAH

Our proposed approach consists of two main parts, a classification model and a sequence predictor (see Fig. 2). The sequence predictor is a constraint optimization algorithm that guarantees a logical sequence of vertebra labels. The input variables for this algorithm are derived per individual vertebra, which are obtained using a classification model.

For training and inference, our approach requires the location of each vertebra in each image. Since we already have instance segmentation annotations for our data, we utilize those to compute the center of mass locations for each vertebra. With those locations, we calculate crops for each vertebra, where the center of mass location becomes the center. We computed a crop over the whole training dataset that guarantees fully visible neighboring vertebrae to give the classification model context, rounding up to the next value that is divisible by eight for proper

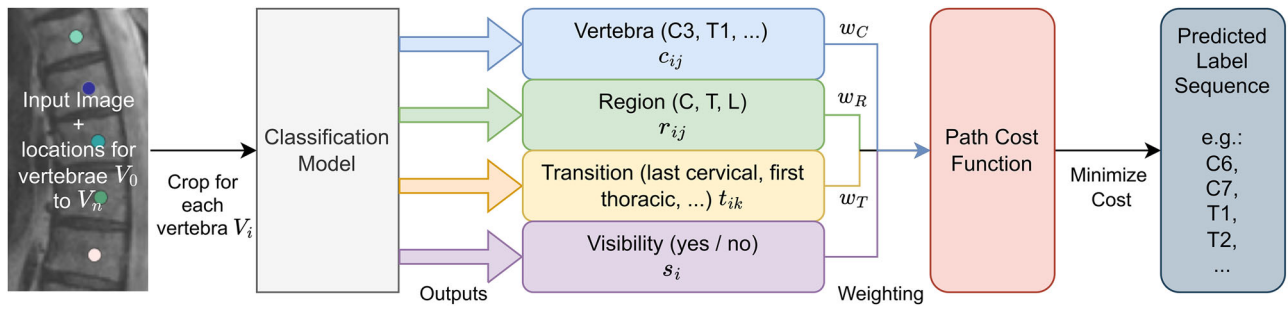


Fig. 2 Overview of the VERIDAH approach. We take an input image and the vertebra localizations to crop around each vertebra. We then input these 3D cropped images separately into our classification model. The classification model has four output heads. We weigh their predictions based on our calibration and combine them into a cost matrix. Finally, an algorithm finds the optimal label sequence path that minimizes this cost, making our final sequence prediction

convolution training. We reorient these cutouts consistently and align the resolution depending on whether it is T2w MRI or CT (see Supplemental Table 1).

We trained a 3D DenseNet169 [32] on individual vertebra cutouts. With a typical one-classification head approach, one would train it to output the vertebra label (i.e., C3, T5, ...). In contrast, VERIDAH utilizes four different classification heads (see Fig. 2): (1) Target vertebra label (such as C3, T5, ...) with 23 classes; (2) target region with the three classes cervical, thoracic, and lumbar; (3) and target vertebra transition with six classes (is the vertebra the last cervical, first thoracic, last thoracic, first lumbar, last lumbar, or none above); (4) target visibility (if the vertebra is fully visible or not) as one class output. We replace the original classification head with the four prediction heads. Each has an additional fully connected linear layer and a ReLU activation function, followed by the output linear layer and the softmax function. Since T13 cases are especially rare, training directly on class labels like T13 would create a huge class imbalance problem for the classifier (initial observations showed bad recall results of around 60). To circumvent this, we replace the T13 and L6 labels in the reference annotation with T12 and L5, respectively. Thus, the existence of a T13 and L6 is indicated by two consecutive T12 or L5 labels. This means that during training, the label difference between a T12 and a T13 is only present to the model through the vertebra transition output head. The T13 is labeled as the last thoracic vertebra; the T12 in a case with 13 thoracic vertebrae is not. All the vertebrae in our datasets are sufficiently visible to train on. Thus, we introduce the visibility label during training as data augmentation (denoted RandCut). We randomly crop a part of the image, and if the vertebra segmentation is not fully visible anymore, we set the visibility reference label to zero, one otherwise. We used Pytorch v2.04.1 [33] with Pytorch Lightning v2.0.8 [34] and the MONAI libraries v1.3.0 [35]. To account for the rare

occurrence of these TEA and LEA anomalies, we sample each anomalous case in the training set four times during each epoch instead of only once. For the hyperparameters employed, see Supplemental Table 1.

To get the final vertebra labeling sequence for any field of view, we incorporate global constraints similar to Meng et al [26]. For that, we combine the outputs of our classification model into a cost function. The intuition is to find the sequence with the maximum agreement between the different classification outputs across the vertebrae. For instance, one classification head might indicate a T12, while the region output predicted the vertebra belonging to the lumbar region. The classifier predicts the visibility, which we utilize to weigh each vertebra. The less visible the model predicts the vertebra to be, the less it impacts the sequence predictor. Additionally, by applying a Gaussian filter, we automatically weigh the boundary vertebrae less. We do this since the classifier has less surrounding information when looking at a boundary vertebra or when the vertebra itself is not fully visible, and we observed less accurate individual predictions there. Thus, we can combine the visibility output with the other per-vertebra outputs and assign a probability to each possible path of vertebra labels. In order to find the optimal solution, we utilize dynamic programming and recursively run over all possible path options and assign a cost (for details, see Appendix A). However, we can enforce constraints by setting them at an infinite cost. Our constraints, regardless of FOV, are as follows:

- The labels must be consecutive (e.g. T5 must be followed by T6). To account for T11 cases, we allow an L1 directly after a T11.
- The labels must be in spatial order (not T6 and then T5).
- There can only be a maximum of seven cervical vertebrae, 11 to 13 thoracic vertebrae, and only four to six lumbar vertebrae.

- We allow a T12 followed by a T12 exactly once to account for T13 and do the same with two consecutive L5s for L6 cases.

In contrast to Meng et al [26], we do not add artificial costs to the paths containing a TEA or LEA. Once we have computed the optimal path, all we need to do is relabel occurrences of two consecutive T12s to [T12, T13] and two consecutive L5s to [L5, L6] to receive the final result.

Experiments

For both CT and MRI, we individually train a classification model. We train the T2w sagittal MRI model on the native resolution of the NAKO images, which is approximately (0.85, 0.85, 3.3) mm, with the orientation being Posterior, Inferior, Right consistently. We resampled the CT images consistently to an isotropic (1.0, 1.0, 1.0) mm resolution.

For benchmarking, we evaluate VERIDAH against the approach from Meng et al [26] in CT, while in T2w sagittal MRI, we use SpineNetV2 [24] and Warszawski et al [27]. Notably, these baselines detect the vertebra locations in the image in addition to labeling. To mitigate propagation errors originating from merged or missed detections and thus perform a fairer comparison to VERIDAH, we replaced the detection process in the Meng et al algorithm with the ground-truth locations. We removed any wrong locations that were predicted by any of those methods. Thus, we only compare the vertebra labeling algorithms of those methods. Meng et al [26] introduce an artificial cost to predict TEA and LEA, which may cause under-prediction of TEA and LEA; thus, we additionally compare against Meng et al, where these penalty costs are set to zero. We denote this as Meng et al ZP. We train a baseline model for ablation studies using only the vertebra label classification output, representing the naive approach. We then evaluate several configurations on each image sequence independently. We abbreviate the vertebra label, region, transition, and visibility heads as V, R, T, and S, respectively. We compare (1) a model trained solely on vertebra-level output and predicting labels directly without a sequence predictor (VH-V), (2) a model trained on vertebra-level output but using the sequence predictor constrained to this single output (VH-P), (3) models trained with each individual head in combination with the normal vertebra-level output (VRH-P, VTH-P, VSH-P, respectively), and (4) the complete VERIDAH approach, trained with four classification outputs and utilizing the complete sequence predictor (4H-P). Additionally, to test a naive integration strategy, we construct a combined cost matrix from all classification outputs and assign labels by taking the per-vertebra argmax, bypassing

the sequence predictor. We denote this variant as 4H-AM. The final ablation is about FOV consistency. We analyzed this by cropping the test images to all permutations of different numbers of visible vertebrae and then comparing the predictions over these cut images against the cut reference labels. For more ablations, including stratified evaluation on scoliosis cases, fractures, and hardware and runtimes, see Appendix B and C.

Evaluation

We compare the labeling sequences of predictions with the references. We define the perfect label percentage (PL perc.) as the percentage of subjects for which the labeling was correct for all vertebrae in that subject. We denote the subject correctness (sub. corr.) as the average number of correct vertebrae in a subject. To focus on the performance of enumeration anomalies, we also define the TEA recall (TEA rec.) and LEA recall (LEA rec.) as the percentage of correctly identified TEA and LEA, respectively. Specifically, we define a correct T11 case if the model predicts the sequence [T11, L1] and the reference has the same labels at the same positions. A correctly labeled T13 case is one where the model predicts two consecutive T12s at matching positions. Correct L4 and L6 cases are identified by both sequences ending on L4 or two consecutive L5, respectively. Similarly, we denote TEA precision (TEA prec.) and LEA precision (LEA prec.) as the ratio of correctly predicted anomalies in all anomalous predictions. Notably, other labeling errors apart from the enumeration anomalies are not reflected in these scores. We employ the Wilcoxon signed-rank test for statistical significance, with a p -value < 0.05 indicating significance.

Results

Inter-rater analysis

Two raters independently rated a total of 200 subjects regarding the lumbosacral junction (100 MRI, 100 CT, randomly chosen). The task was to determine the level of the last lumbar vertebrae based on the image. The Cohen's kappa is 0.95. This highlights great consistency regarding the manual ratings.

CT

VERIDAH achieves significantly better metric values than the labeling from Meng et al (see Table 2, all $p < 0.01$). While Meng et al provide decent results for LEA recall, they fail to correctly annotate TEA cases, regardless of whether we set their anomaly penalty cost to zero. Notably, VERIDAH achieves better performance on non-anomalous cases, as indicated by the superior subject correctness with much higher meaning and lower standard deviation. When we set the punishment cost for TEA

Table 2 Labeling performance on CT test data

Metric	VERIDAH	Meng et al [23]	Meng et al ZP [23]
PL perc.	99.18	77.81	77.26
Sub. Corr.	99.63 ± 4.37	89.04 ± 27.73	88.48 ± 28.28
TEA rec.	96.3	11.11	14.81
LEA rec.	97.22	61.11	61.11
TEA prec.	92.86	50	36.36
LEA prec.	97.22	64.71	66.67

Meng et al ZP is the approach from Meng et al, but setting the TEA and LEA penalty costs to zero. Bold values indicate best performance.

rec recall, *prec* precision, *TEA* thoracolumbar enumeration anomaly, *LEA* lumbar enumeration anomaly

to zero, their perfect label percentage drops slightly while the TEA recall significantly increases. However, with 14.81% compared to VERIDAH's 96.30%, this is still inferior. For a qualitative example, see Fig. 3. Notably, on our test set, in contrast to Meng et al, the predicted labels of VERIDAH were never shifted by more than one (see Fig. 4, for example).

T2w MRI

As shown in Table 3, VERIDAH outperforms the other baselines labeling methods in every metric on the NAKO test set (all $p < 0.01$). Notably, while both baselines have decent LEA accuracies around 83%, they fail to accurately detect TEAs. With VERIDAH, we achieved a perfect labeling sequence in 98.3% of subjects of the test set, while correctly predicting 87.8% and 94.48% of

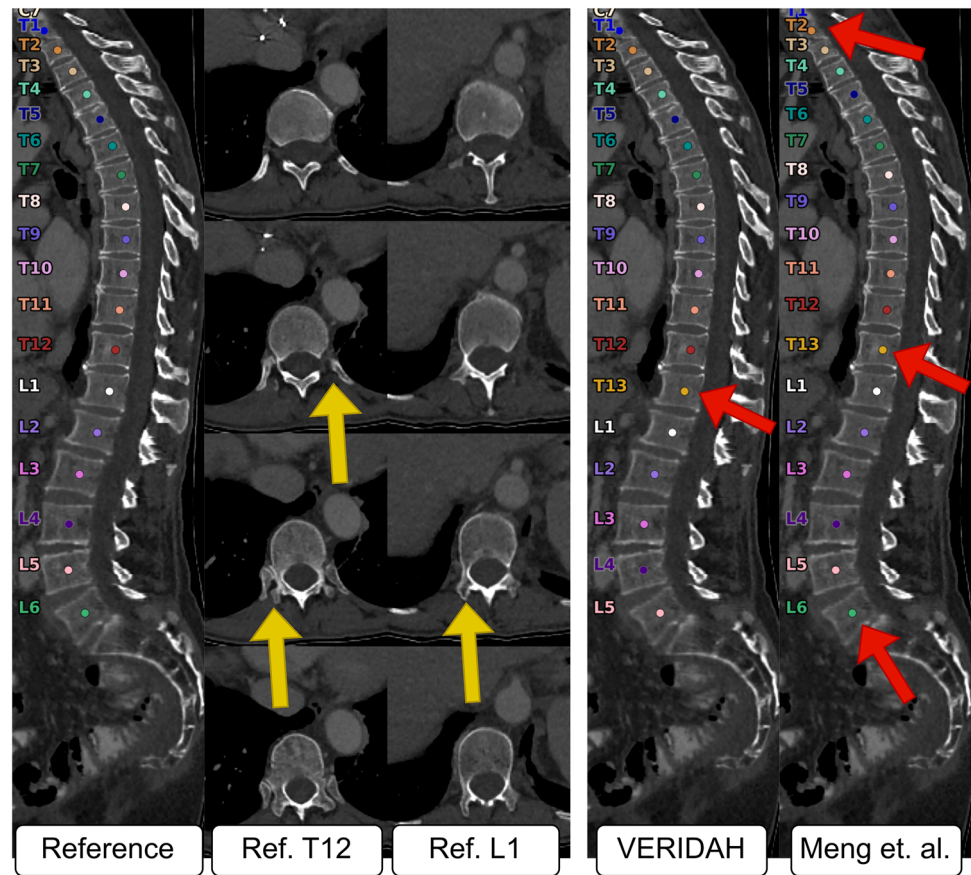


Fig. 3 Example of a CT test subject where all models failed. The reference data (left) shows an L6 LEA case with an ordinary number of twelve thoracic vertebrae. VERIDAH (right) mistakes the L1 for a thoracic vertebra, resulting in thirteen thoracic and five lumbar vertebrae (red arrows). The prediction of Meng et al has a shift error by one, but correctly labels the L1. The center shows the reference T12 and L1 in multiple axial slices (top to bottom). While the T12 has clear ribs on both sides (yellow arrows), the L1 has thoracic-looking facets while having no ribs, making this a TLTV

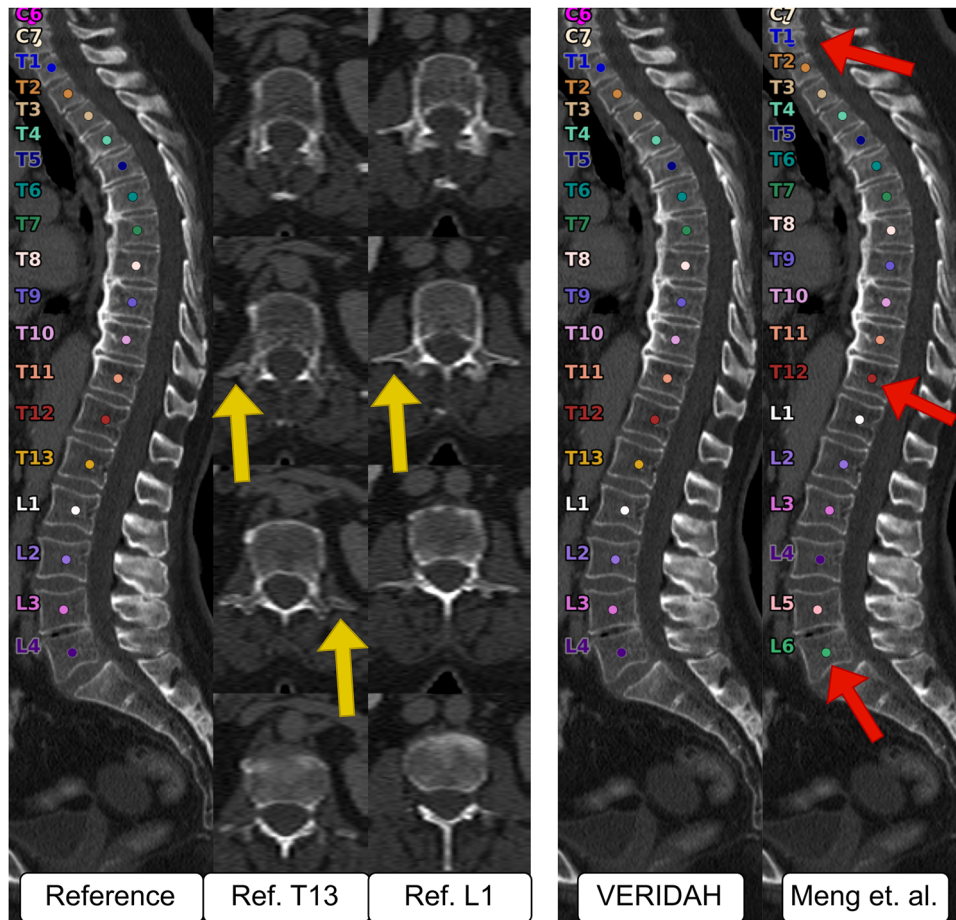


Fig. 4 Example of a CT test subject with a T13 and only four lumbar vertebrae. The axial slices of the reference data (left) show that the T13 has a rib on either side (yellow arrows). While VERIDAH (right) correctly classifies the whole sequence of labels, the prediction of Meng et al made two major errors, leading to six lumbar vertebrae instead of four (red arrows)

Table 3 Labeling performance on NAKO test data

Metric	VERIDAH	SpineNetv2 [21]	TotalSpineSeg [24]
PL perc.	98.3	91.49	94.24
Sub. Corr.	99.58 ± 3.19	97.89 ± 9.32	98.98 ± 4.13
TEA rec.	87.8	0	27.63
LEA rec.	94.48	83.72	83.04
TEA prec.	89.87	0	95.45
LEA prec.	92.57	69.68	80.17

rec recall, prec precision, TEA thoracolumbar enumeration anomaly, LEA lumbar enumeration anomaly. Bold values indicate best performance.

the TEA and LEA cases, respectively. The subject correctness is consistently high across the board with relatively low standard deviation, indicating that erroneous labels mostly involve one or two vertebrae within one subject.

Ablations

Both the multi-classification head approach and our proposed sequence predictor provide significant performance boosts (see Table 4, all $p < 0.01$). Notably, without the sequence predictor, the added benefit of four classification outputs compared to one (4H-AM vs. VH-V) is barely noticeable. The four classification outputs become more meaningful when combined with the sequence predictor (4H-P vs. VH-P). These behaviors are very similar both in CT and MRI.

Figure 5 highlights the robustness of VERIDAH on arbitrary FOVs. The perf. Label perc. metric is consistently high, while only the TEA recall suffers from small FOVs. We also observed that mostly a couple of samples were misclassified as soon as an individual vertebra was present in the FOV. This leads to the reason that individual vertebrae are challenging for VERIDAH, not specific FOVs. When we relate how consistent VERIDAH's

Table 4 Ablation labeling performance

Metric	VH-V	VH-P	4H-AM	VRH-P	VTH-P	VSH-P	4H-P
T2w sagittal							
PL perc.	82	95.6	82.1	97.67	97.9	97.6	98.3
Sub. Corr.	97 $\pm$ 6.75	98.92 $\pm$ 5.05	97.45 $\pm$ 6.69	99.43 $\pm$ 3.7	99.48 $\pm$ 3.54	99.41 $\pm$ 3.79	99.58 $\pm$ 3.19
TEA rec.	56.1	80.49	57.32	86.59	85.37	86.59	87.8
LEA rec.	65.52	80.69	67.59	93.79	93.79	93.79	94.48
TEA prec.	66.67	70.21	67.14	85.54	88.61	84.52	89.87
LEA prec.	74.22	87.97	73.68	90.67	91.89	90.07	92.57
CT							
PL perc.	73.42	95.62	73.42	98.36	96.99	96.44	99.18
Sub. Corr.	95.95 $\pm$ 8.66	98.75 $\pm$ 5.89	95.92 $\pm$ 8.74	99.5 $\pm$ 3.94	99.13 $\pm$ 4.97	99 $\pm$ 5.29	99.63 $\pm$ 4.37
TEA rec.	74.07	62.96	74.07	96.3	77.78	74.07	96.3
LEA rec.	30.56	63.89	30.56	88.89	75	69.44	97.22
TEA prec.	83.33	73.91	80	83.87	80.77	76.92	92.86
LEA prec.	68.75	88.46	68.75	94.12	93.1	92.59	97.22

rec recall, prec precision, TEA thoracolumbar enumeration anomaly, LEA lumbar enumeration anomaly, VH-V taking only the vertebra labeling without sequence predictor, VH-P inputting only the vertebra output to the sequence predictor, 4H-AM taking the Argmax of the four outputs instead of the sequence predictor, VRH-P only using the vertebra and region head with the sequence predictor, VTH-P only using the vertebra and vertebra relation head with the sequence predictor, VSH-P only using the vertebra and visibility head with the sequence predictor, 4H-P the proposed approach of utilizing the four classification outputs with the sequence predictor. Bold values indicate best performance.

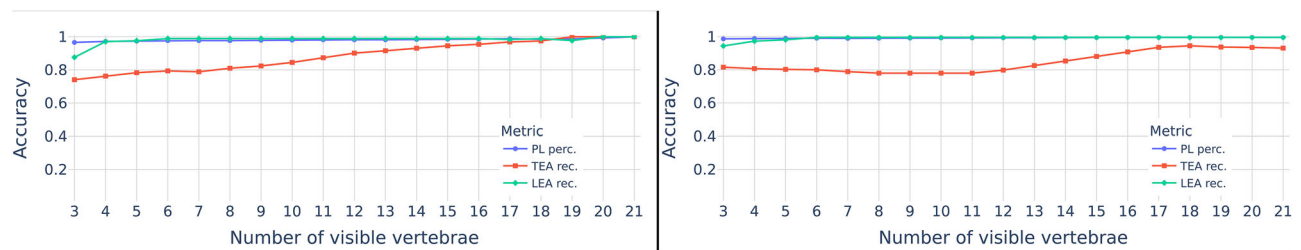


Fig. 5 The behavior of three metrics over different FOVs of the test data. The x-axis shows the number of visible vertebrae fed into VERIDAH, while the y-axis showcases the different metric values

prediction is in any FOV compared to VERIDAH's prediction on the whole spine, there are virtually no inconsistencies except for a handful of permutations of samples out of the test samples. We observed that these are mostly the same samples and FOVs that are falsely predicted from VERIDAH in those limited FOVs.

Discussion

We present VERIDAH, a 3D vertebra labeling approach that combines multiple classification heads with a constraint optimization solver. The method outperforms state-of-the-art approaches for T2w MRI and CT, works on arbitrary fields of view, and consistently handles

vertebral enumeration anomalies. Previous methods were unable to reliably label these anomalies, which can inherently contradict non-anomalous labeling schemes and limit achievable performance.

Automatic vertebra labeling in CT and MRI is clinically relevant for consistent assessments and reducing radiologists' workload. Especially in smaller FOV images, manual counting approaches can be more error-prone and inconsistent [15].

Nevertheless, our study has its limitations. The most important one to address is the requirement for vertebral localizations. VERIDAH does not perform a localization or detection step prior to labeling but instead requires the

centroids of vertebrae beforehand. However, the method SPINEPS [30], which we used for our MRI data, had a localization error in 0.53% of subjects. Public datasets such as VerSe [20] exist with similar automated detection approaches for CT. Therefore, as publicly available models can perform localization through segmentation, this requirement should not hinder fellow researchers and clinicians. Nevertheless, we do acknowledge the increased effort required to effectively utilize VERIDAH.

Since the VERIDAH model trained for T2w MRI was only trained on a healthy cohort (NAKO), we cannot guarantee the same performance on other datasets from clinical routine. Since the CT data used is very diverse, VERIDAH should perform more robustly in that modality.

Overall, we believe that VERIDAH provides a foundation that bridges the gap to enumeration anomaly-aware labeling. Since many studies analyze different aspects of the spine in relation to vertebra labels, accurate delineation of vertebra levels is crucial. While previous tools never really addressed enumeration anomalies, our approach is publicly available and offers more accurate labels, opening up more detailed research opportunities for the future.

Abbreviations

FOV	Field of view
LEA	Lumbar enumeration anomaly
TEA	Thoracic enumeration anomaly
TLTV	Thoracolumbar transitional vertebrae
VERIDAH	Vertebra Identification with Anomaly Handling

Supplementary information

The online version contains supplementary material available at <https://doi.org/10.1007/s00330-026-12966-x>.

Funding

This study has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (101045128-iBack-epic-ERC2021-COG). This project was conducted with data (Application No. NAKO-732/733) from the German National Cohort (NAKO) (www.nako.de). The NAKO is funded by the Federal Ministry of Education and Research (BMBF) (project funding reference numbers: 01ER1301A/B/C, 01ER1511D, and 01ER1801A/B/C/D), federal states of Germany, the Helmholtz Association, the participating universities, and the institutes of the Leibniz Association. We thank all participants who took part in the NAKO study and the staff of this research initiative. Open Access funding enabled and organized by Projekt DEAL.

Compliance with ethical standards

Guarantor

The scientific guarantor of this publication is Hendrik Möller.

Conflict of interest

The authors of this manuscript declare relationships with the following companies: J.S.K. and A.S. are co-founders and shareholders of Bonescreen GmbH (see <https://bonescreen.de/>). During this study, the authors who analyzed and controlled the data are not employees, co-founders, or

shareholders of Bonescreen GmbH. All remaining authors declare to have no competing interests.

Statistics and biometry

No complex statistical methods were necessary for this paper.

Informed consent

Written informed consent was waived by the Institutional Review Board.

Ethical approval

Institutional Review Board approval was obtained (593/21 S-NP).

Study subjects or cohorts overlap

There is no subject overlap with internal data. We utilize a subset of the German National Cohort (NAKO). This article has a preprint: Möller H, Schoen H, Graf R et al (2026) "VERIDAH: solving enumeration anomaly aware vertebra labeling across imaging sequences". <https://doi.org/10.48550/arXiv.2601.14066> (arXiv).

Methodology

- Retrospective
- Cross-sectional study
- Performed at one institution

Author details

¹Department of Diagnostic and Interventional Neuroradiology, TUM University Hospital, Munich, Germany. ²Chair for AI in Healthcare and Medicine, Technical University of Munich (TUM) and TUM University Hospital, Munich, Germany. ³Department of Diagnostic and Interventional Radiology, Pediatric Radiology and Neuroradiology, Rostock University Medical Center, Rostock, Germany. ⁴NeuroPoly Lab, Institute of Biomedical Engineering, Polytechnique Montreal and Mila (Quebec AI Institute), Quebec, Canada. ⁵Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland. ⁶Department of Diagnostic and Interventional Radiology, University Hospital Heidelberg, Heidelberg, Germany. ⁷Department of Diagnostic and Interventional Radiology, Medical Center, University of Freiburg, Freiburg, Germany. ⁸Molecular Epidemiology Research Group, Max-Delbrueck-Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin, Germany. ⁹Berlin Ultrahigh Field Facility (B.U.F.F.), Max Delbrueck Center for Molecular Medicine, Berlin, Germany. ¹⁰Department of Diagnostic and Interventional Radiology, University Hospital Augsburg, Augsburg, Germany. ¹¹Department of Computing, Imperial College London, London, UK.

Received: 11 March 2026 Revised: 11 August 2026 Accepted: 17 September 2026

Published online: 16 October 2026

References

1. Sim Y, Chung MJ, Kotter E et al (2020) Deep convolutional neural network-based software improves radiologist detection of malignant lung nodules on chest radiographs. *Radiology* 294:199–209
2. Kickingeder P, Isensee F, Tursunova I et al (2019) Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. *Lancet Oncol* 20:728–740
3. De Fauw J, Ledsam JR, Romera-Paredes B et al (2018) Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med* 24:1342–1350
4. Löffler MT, Sekuboyina A, Jacob A et al (2020) A vertebral segmentation dataset with fracture grading. *Radiol Artif Intell* 2:e190138
5. Thawait GK, Chhabra A, Carrino JA (2012) Spine segmentation and enumeration and normal variants. *Radiol Clin North Am* 50:587–598
6. Tatara Y, Niimura T, Sekiya T, Mihara H (2021) Changes in lumbosacral anatomy and vertebral numbering in patients with thoracolumbar and/or lumbosacral transitional vertebrae. *JBJS Open Access* 6:e20
7. Wigh RE (1980) The thoracolumbar and lumbosacral transitional junctions. *Spine* 5:215–222

8. Park SK, Park JG, Kim BS, Huh JD, Kang H (2016) Thoracolumbar junction: morphologic characteristics, various variants and significance. *Br J Radiol* 89:20150784
9. Nardo L, Alizai H, Virayavanich W et al (2012) Lumbosacral transitional vertebrae: association with low back pain. *Radiology* 265:497–503
10. Apaydin M, Uluc ME, Sezgin G (2019) Lumbosacral transitional vertebra in the young men population with low back pain: anatomical considerations and degenerations (transitional vertebra types in the young men population with low back pain). *Radiol Med* 124:375–381
11. Sugiura K, Morimoto M, Higashino K et al (2021) Transitional vertebrae and numerical variants of the spine: prevalence and relationship to low back pain or degenerative spondylolisthesis. *Bone Joint J* 103-B:1301–1308. <https://doi.org/10.1302/0301-620X.103B7.BJJ-2020-1760.R1>
12. Jhawar BS, Mitsis D, Duggal N (2007) Wrong-sided and wrong-level neurosurgery: a national survey. *J Neurosurg Spine* 7:467–472. <https://doi.org/10.3171/SPI-07/11/467>
13. Hahn PY, Strobel JJ, Hahn FJ (1992) Verification of lumbosacral segments on MR images: identification of transitional vertebrae. *Radiology* 182:580–581
14. Peckham ME, Hutchins TA, Stilwell SE et al (2017) Localizing the L5 vertebra using nerve morphology on MRI: an accurate and reliable technique. *AJNR Am J Neuroradiol* 38:2008–2014
15. Schinz D, Möller H, Ramschütz C et al (2026) A vertebral shape-based classification for labeling the thoracolumbar junction on CT imaging. *Eur Spine J* 35:2195–2203. <https://doi.org/10.1007/s00586-025-09364-0>
16. Cai Y, Osman S, Sharma M, Landis M, Li S (2015) Multi-modality vertebra recognition in arbitrary views using 3D deformable hierarchical model. *IEEE Trans Med Imaging* 34:1676–1693
17. Liao H, Mesfin A, Luo J (2018) Joint vertebrae identification and localization in spinal CT images by combining short-and long-range contextual information. *IEEE Trans Med Imaging* 37:1266–1275
18. Payer C, Stern D, Bischof H, Urschler M (2019) Integrating spatial configuration into heatmap regression based CNNs for landmark localization. *Med Image Anal* 54:207–219
19. Alomari RS, Corso JJ, Chaudhary V (2010) Labeling of lumbar discs using both pixel-and object-level features with a two-level probabilistic model. *IEEE Trans Med Imaging* 30:1–10
20. Sekuboyina A, Husseini ME, Bayat A et al (2021) VerSe: a vertebrae labelling and segmentation benchmark for multi-detector CT images. *Med Image Anal* 73:102166
21. Tao R, Liu W, Zheng G (2022) Spine-transformers: vertebra labeling and segmentation in arbitrary field-of-view spine CTs via 3D transformers. *Med Image Anal* 75:102258
22. Huang Y, Uneri A, Jones CK et al (2021) 3D vertebrae labeling in spine CT: an accurate, memory-efficient (Ortho2D) framework. *Phys Med Biol* 66:125020
23. Kawathekar ID, Areeckal AS, Aparna V (2024) A novel deep learning pipeline for vertebra labeling and segmentation of spinal computed tomography images. *IEEE Access Pract Innov Open Solut* 12:15330–15347
24. Windsor R, Jamaludin A, Kadir T, Zisserman A (2024) Automated detection, labelling and radiological grading of clinical spinal MRIs. *Sci Rep* 14:14993
25. Sekuboyina A, Irmay J, Shit S, Kirschke JS, Andres B, Menze B (2023) Pushing the limits of an FCN and a CRF towards near-ideal vertebrae labelling. In: *Proceedings of the 2023 IEEE 20th international symposium on biomedical imaging*. IEEE, Cartagena, pp 1–5
26. Meng D, Mohammed E, Boyer E, Pujades S (2022) Vertebrae localization, segmentation and identification using a graph optimization and an anatomic consistency cycle. In: *International workshop on machine learning in medical imaging*. Springer, Cham, pp 307–317
27. Warszawer Y, Molinier N, Valošek J et al (2024) Fully automatic vertebrae and spinal cord segmentation using a hybrid approach combining nnU-Net and iterative algorithm. In: *Proceedings of the International Society for Magnetic Resonance in Medicine, scientific meeting and exhibition*. <https://doi.org/10.58530/2024/4847>
28. Bamberg F, Schlett CL, Caspers S et al (2024) Baseline MRI examination in the NAKO health study. *Dtsch Arztebl Int* 121:587–593. <https://doi.org/10.3238/arztebl.m2024.0151>
29. Bamberg F, Kauczor H-U, Weckbach S et al (2015) Whole-body MR imaging in the German National Cohort: rationale, design, and technical background. *Radiology* 277:206–220
30. Möller H, Graf R, Schmitt J et al (2024) SPINEPS—automatic whole spine segmentation of T2-weighted MR images using a two-phase approach to multi-class semantic and instance segmentation. *Eur Radiol*. <https://doi.org/10.1007/s00330-024-11155-y>
31. Konin GP, Walz DM (2010) Lumbosacral transitional vertebrae: classification, imaging findings, and clinical relevance. *AJNR Am J Neuroradiol* 31:1778–1786
32. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: *Proceedings of the 2017 IEEE conference on computer vision and pattern recognition (CVPR)*. IEEE, Honolulu, pp 2261–2269
33. Imambi S, Prakash KB, Kanagachidambaresan GR (2021) PyTorch. Programming with TensorFlow. In: *Solution for Edge Computing Applications*. Springer, pp 87–104
34. Falcon W, The PyTorch Lightning team (2019) PyTorch Lightning. Zenodo. <https://doi.org/10.5281/zenodo.22687155>
35. Cardoso MJ, Li W, Brown R et al (2022) Monai: an open-source framework for deep learning in healthcare. Preprint at <https://doi.org/10.48550/arXiv.2211.02701>

Publisher's Note

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