A semi-data-driven approach for automating the assessment of the Spinal Instability Neoplastic Score

Naomi Vermeulen[a,b], Wietse S.C. Eppinga MD[a], Can Ozan Tan PhD[c], Mark H.F. Savenije PhD[a,b], Jorrit-Jan Verlaan MD, PhD[d], Antonetta C. Houweling PhD[a], Cornelis H. Slump PhD[c], Frank F.J Simonis PhD[e], Cornelis A.T van den Berg PhD[a,b]

Background: The Spinal Instability Neoplastic Score (SINS) is the most widely accepted classification system for spinal neoplastic disease. It is utilized as a referral tool for radiologists and oncologists, that supports them in making the decision whether a patient with spinal metastases should be referred to the orthopedic surgeon for stabilization of the spine. If the final score is equal to or higher than 7, a consultation with a spine surgeon is recommend. However, the radiographic assessment is time-consuming and its reliability as a referral tool can be questioned since the intra- and interobserver reliability among spine surgeons is significantly better than across other observers. Therefore, there is a need for a data science solution that automates the radiographic assessment of the SINS.

Purpose: This study describes the development and evaluation of a semi-data-driven approach for automating the radiographic assessment of the SINS based on PET/CT scans.

Methods: 87 PET/CT scans of patients with spinal metastases are included and split into a training (70%), validation (10%), and test set (20%). In this study, a semi-data-driven workflow is developed and assessed consisting of three sequential steps:

1) Three parallel pathways of convolutional neural networks (CNNs) for segmenting/labeling vertebrae, vertebral bodies and metastases respectively in the spine from PET/CT scans.
2) A post-processing step for the automated extraction of radiomic features.
3) The training of a machine learning model for SINS prediction utilizing the features.

For the training of the CNNs (step 1), the ground truths are manually segmented/labeled. The resulting segmentations are evaluated by calculating the Dice Similarity Coefficient (DSC) and Hausdorff Distance (HD). The labels are evaluated by calculating the percentage of correctly classified voxels per vertebra. The outcomes of the final model are assessed both as continuous (0-18) by calculating the R^2, and binary classes (do or do not refer) by calculating the sensitivity and specificity.

Results: After post-processing, the segmentations of the vertebrae resulted in a DSC and HD of 0.984 and 9mm respectively. For the segmentations of the vertebral bodies, these scores were 0.976 and 10mm and for the metastases, these were 0.720 and 12mm respectively. The percentages of voxels that were correctly classified by the model for labeling the vertebrae, varied from 42% for vertebral level Th7 to 96% for level L5. For the final prediction of the SINS, linear regression models showed the best performance (R^2 =0.56). As a binary referral tool (do/ do not refer patient), a threshold of 4 (without the pain component) resulted in the highest sensitivity of 0.93 with a corresponding specificity of 0.76.

Conclusions: To our knowledge this semi-data-driven approach for SINS classification is the first that shows promising results when it is utilized as a referral tool. However, further optimization and external validation of the model is needed before a reliable conclusion can be drawn.

Keywords: Spinal Instability Neoplastic Score, SINS, spinal instability, spinal metastases, segmentation spine, annotation spine

[a] Department of Radiotherapy, University Medical Centre Utrecht, Utrecht University, Utrecht, the Netherlands

[b] Computational Imaging Group for MRI Diagnostics and Therapy, Centre for Image Sciences, University Medical Center Utrecht, Utrecht, The Netherlands

[c] Robotics and Mechatronics Group, Technical Medical Centre, University of Twente, Enschede, Netherlands

[d] Departments Orthopedic Surgery, University Medical Centre Utrecht, Utrecht University, Utrecht, the Netherlands

[e] Magnetic Detection & Imaging, Technical Medical Centre, University of Twente, Enschede, Netherlands

1. Introduction

The Spinal Instability Neoplastic Score (SINS) is the most widely accepted classification system for spinal neoplastic disease (Table 1[1]). It assesses and scores six components based on radiographic and clinical components: 1) location of the affected segment within the spine, 2) presence of mechanical pain (i.e., with recumbency, movement or loading of the spine) , 3) bone lesion quality, 4) presence of deformity or malalignment at the affected segment, 5) extent of vertebral body involvement/destruction and 6) posterior element involvement (Table 1). The SINS score is the sum of the scores of each component calculated per vertebra. The final scores can be divided into three discrete categories: stable (0–6), potential unstable (7–12), and unstable (13–18). If the final score is equal to or higher than 7, a consultation with a spine surgeon is recommended.[1] However, the assessment of radiographic components can be time-consuming, even for experienced observers. Another potential downside is that both intra- and interobserver reliability among spine surgeons is significantly better than across other observers (0.919 *vs.* 0.673; P<0.0001).[2,3] Since spine surgeons are usually not involved in the routine care of cancer patients, the SINS' reliability as a referral tool can be questioned. Another limitation is that there are still various components that are not accounted for in this classification system, but that might also affect the stability of the spine, such as the volume[4] and specific location of the metastasis within the vertebral body[5] and destruction of cortical versus cancellous bone[6].

Table 1: Spinal Instability Neoplastic Score[1]

SINS component	Score
Location	
Junctional (occiput–C2, C7–T2, T11–L1, L5–S1)	3
Mobile spine (C3–6, L2–4)	2
Semirigid (T3–10)	1
Rigid (S2–5)	0
Pain*	
Yes	3
Occasional pain but not mechanical	2
Pain-free lesion	0
Bone lesion	
Lytic	2
Mixed (lytic/blastic)	1
Blastic	0
Spinal alignment	
Subluxation/translation present	4
De novo deformity (kyphosis/scoliosis)	2
Normal alignment	0
Vertebral body collapse	
>50% collapse	3
<50% collapse	2
No collapse with >50% body involved	1
None of the above	0
Posterolateral involvement of the spinal elementsy†	
Bilateral	3
Unilateral	1
None of the above	0

Criteria of instability. Total score (TS) 0–6 : stable spine, TS 7–12 : potential unstable spine, TS 13–18 : unstable spine. Recommendation : TS ≥7, consider surgical intervention. *Pain improvement with recumbency and/or pain with movement/loading of the spine, †Facet, pedicle, or costovertebral joint fracture or replacement with tumor

To overcome these limitations, there is a need for a solution that automates the assessment of the SINS and analyzes more components. The biggest benefit of this solution is to facilitate better and faster decision-making. It will lead to cost savings, higher reliability and a lower risk of underdiagnoses. The challenge is to investigate whether such a data science solution can achieve sufficient accuracy and robustness in a representative patient cohort. Previous research performed at the UMC Utrecht showed that automating every individual component of SINS based on post-processing (i.e., rule-based image processing) is prone to errors. To be able to assess the stability of the spine in an automated fashion, there is a need for a less rule-based (i.e., data-driven) approach. However, to train such a data-driven deep learning approach, a large amount of data should be available. In this study, a new approach is proposed that combines the robustness of a data-driven approach with the advantages of post-processing by adding prior knowledge of experts. More specifically, radiomic features, that are expected to correlate with the stability of the spine, are automatically extracted from the data and used for the training of a machine learning model for SINS prediction.

It is expected that the assessment of spinal stability based on radiomic features is a more standardized way of acquiring and presenting data, that can be utilized for clinical research. Hopefully, this will increase the insight into spinal neoplastic diseases and will enable the measurements of minor changes in vertebrae

appearance over time. When it becomes possible to detect minor changes early on, clinical deterioration can still be largely prevented with less invasive treatment strategies (e.g., medication, radiotherapy) as opposed to surgery. Nowadays, the SINS is often not used by radiologists or oncologists due to its limitations, leading to underdiagnoses or diagnostic delay. An automated assessment has the potential to accelerate the start of therapy for this palliative patient group that often has severe pain complaints, thereby potentially increasing the patients' quality of life.[7]

This study aims to investigate how sufficient radiomic image features can be extracted from PET/CT scans and if a data-driven machine learning technique can reliably (sensitivity≥0.9 and specificity≥0.8) predict whether a patient should be referred to a spine surgeon.

2. Methods

A deep learning workflow is developed consisting of three parallel image processing paths relying on convolution neural networks (CNNs) to obtain segmented/labeled vertebrae and vertebral bodies from CT scans and segmented metastatic volumes from PET scans. Radiomic features, that are expected to correlate with the stability of the spine, are extracted from these segmentations/labels and utilized for the training of a model for SINS prediction (Figure 1).

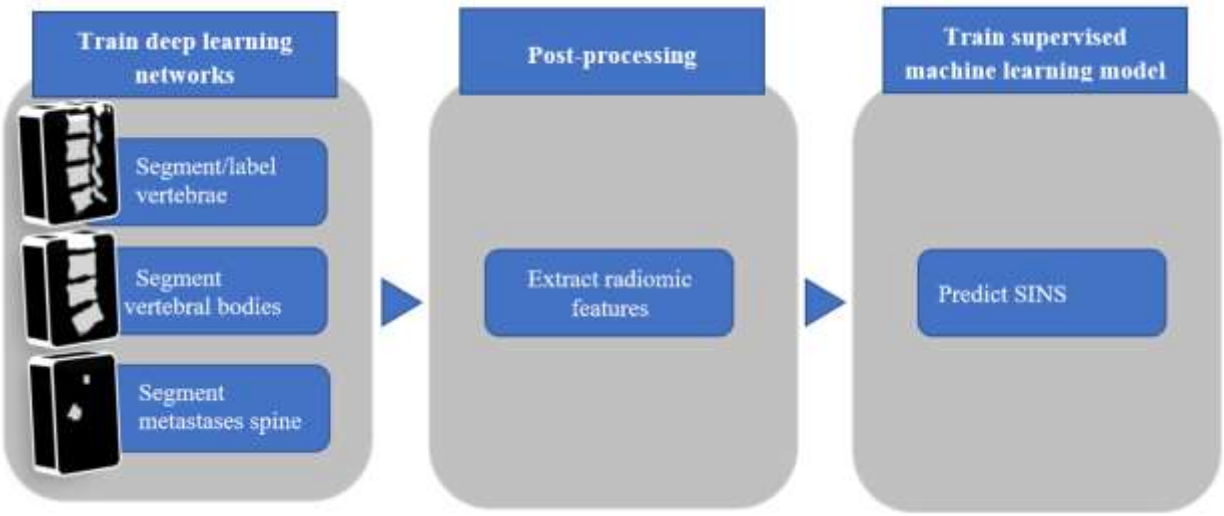

Figure 1: Workflow for automation assessment SINS

2.1 Data collection

To represent the patient population of SINS, there is a need for a data set with a wide variety of PET/CT scans (SIEMENS Biograph 40 mCT) ranging from oligo to diffuse metastases. Therefore, this study acquired scans from the PRESENT[8] data set (n=1904) retrospectively. This data set collected scans from three different hospitals in the Netherlands. These are all scans of patients diagnosed with bone metastases who received radiotherapy treatment at the UMC Utrecht. Scans are included according to the in- and exclusion criteria stated in Table 2.

Scans are acquired with a low-dose CT scan (100-120kV, 20-110 mAs). The CT scan is reconstructed in a field of view of 50 cm using filtered back projection, slice thickness of 2-3 mm, and spacing of 1-4 mm. Since bone metastases can result from multiple primary tumor types, the used radiotracers (FDG, PSMA, etc.) differed between the scans.

Table 2: In- and exclusion criteria

Inclusion criteria	Exclusion criteria
PET/CT scans	Spine only partially visualized
Metastases in the spine	Patients with a numeric variation (more/ less vertebrae) in the vertebral column
	Voxel size larger than 1×1×3mm
	Large artifacts

2.2 Convolutional neural networks

In total four different 3D CNNs are trained (Figure 2). The first network is trained to segment the vertebrae in a CT scan (CNN1a) and a second network (CNN1b) is trained to convert the resulting segmentations into labeled vertebrae. The last two networks are trained for segmenting the vertebral bodies (CNN2) and metastases (CNN3) respectively. For the CNN that segments the metastases in the spine, two approaches are assessed: One in which both the CT and its corresponding PET scan are provided and one in which only the PET scan is provided. In the following sections, more information will be given about the preliminary work, pre-processing, training (i.e., training configuration and network architecture), and post-processing.

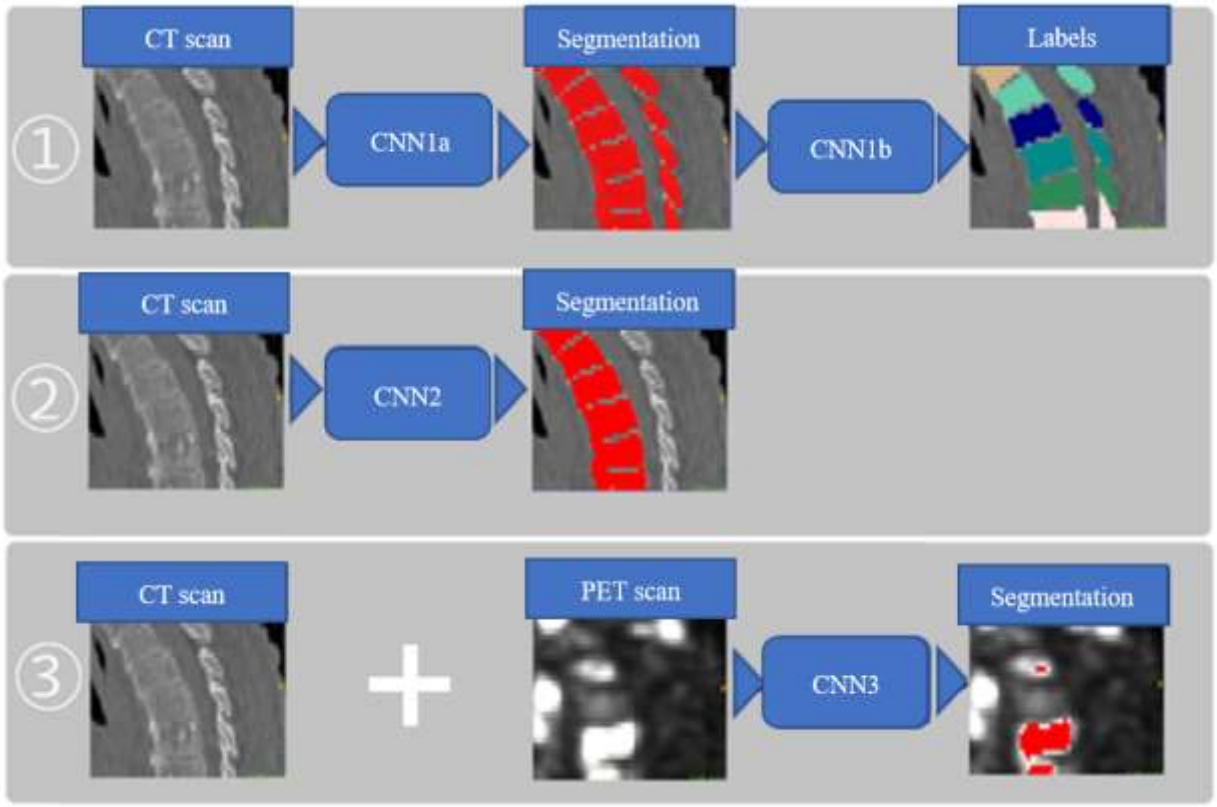

Figure 2: Four different CNNs for: 1) segmenting vertebrae, 2) labeling vertebrae, 3) segmenting vertebral bodies and 4) segmenting metastases.

2.2.1. Preliminary work

In a preliminary study performed at the UMC Utrecht a deep learning based workflow is developed consisting of two CNNs; one for segmenting the vertebrae (C1-L5) of the spine followed by a network for the labeling of these vertebrae. An overview of the training details can be found in Table 3 and more information about the training configuration in Appendix A.1 Table 3 and 4.

Table 3: Training details in-house developed network for segmenting/labeling vertebrae

	Training details
Data set	VerSe2019[9]
Training and validation set	CT scans (n=33)
Test set	CT scans (n=8)
Evaluation using test set of Verse2019	Mean Dice Similarity Coefficients (DSC) [min-max]: 0.740 [0.576-0.920] Mean Hausdorff Distances (HD) [min-max]: 60 [25-152] mm
Network architecture	DeepMedic[10]

The VerSe2019[9] data set, that was utilized for training, did not contain scans with bone metastases (resulting in intravertebral contrast differences or differences in shape due to collapse). Moreover, the scans contained a lot of noise and a limited field-of-view. Hence, it did not represent the patient population of the SINS and it failed to correctly segment/label collapsed vertebrae or vertebrae with diffuse metastases. Testing the network for segmenting the vertebrae with the PRESENT[8] data set, resulted in a mean DSC of 0.592 (±0.054) and a mean HD of 363 mm (±36) without post-processing. After the utilization of the second network for labeling the vertebrae, the mean DSC of the total spine was 0.973 (±0.008) and the mean HD was 9 mm (±1). On average, 62% of the voxels of each vertebra were correctly labeled by the network. To improve the performance, both networks are retrained with the PET/CT scans from the PRESENT[8] data set. In addition to these networks, two new networks are trained with this data set; one for the segmentation of the vertebral bodies and one network for the segmentation of metastases from the PET scans.

2.2.2. Pre-processing

For the creation of the ground truths of the labeled vertebrae, an existing network[11] is tested with the PRESENT[8] data set. Vertebral levels C1-L5 are decoded by the networks as masks with values ranging from 1 to 24. After the inference of the network on the CT scans, the incorrect segmented and labeled vertebrae (14.8%) are manually corrected using ITKSNAP 3.8.0[12]. The network[11] itself could not be utilized for this study, since it failed to segment the vertebrae properly when they were collapsed or contained metastases. Moreover, since it is not an in-house developed network, the medical device regulations would make it difficult to get the model clinically implemented.

All labels are converted to binary segmentations for the training of the network to segment the vertebrae. To obtain the segmentations of the vertebral bodies, the posterior elements are manually removed from the segmentations of the vertebrae by utilizing the 'scalpel mode' in ITKSNAP[12]. In addition, C1 is removed from these segmentations, due to the absence of a vertebral body.

To generate the ground truths of the network for segmenting the metastases in the spine, manual thresholding of PET sans is applied. Per scan, the threshold is chosen, in such a way that all spinal metastases, that are visible on the fused PET/CT scan, are segmented. Before the CT and PET scans are provided to the network, they are multiplied with the binary segmentation of the vertebrae and cropped with 50 voxels at the outer borders of the anterior-posterior and lateral directions of the spine. These borders are defined using the binary segmentation of the vertebrae.

Subsequently, the image intensity of the PET is clipped at its respective 99 percentile per patient volume and the intensity of the CT scans is clipped between -1000 and 1500. All scans and ground truths are resampled to a voxel size of 1×1×1mm and normalized to a zero-mean, unit-variance space**.** The reference segmentations/labels are created by one observer and randomly (20% of the scans) evaluated by a radiation oncologist.

2.2.3. Training

The previously described networks are (re-)trained with the included PET/CT scans, divided into a training (70%) validation (10%), and test set (20%). The pre-trained network, that segments/labels the vertebrae of the spine, was initially trained with 33 scans of the VerSe2019[9] data set and is retrained with 70% of the included scans from the PRESENT[8] data set.

For the creation of all CNNs, an 11-layers deep 3D CNN is employed: DeepMedic[10]. Its three-pathway architecture for multi-resolution processing of 3D patches enables the assessment of both the local and larger contextual information. Because it processes the input at multiple scales it achieves a larger receptive field for the final classification, while keeping the computational cost low. For training, DeepMedic made use of about 9 GB of GPU memory. The training configuration files of all networks are reported in Appendix A.1. The training configuration is kept as the original, except for a few parameters. The number of epochs is set to 25 and 45 for training the networks for segmenting and labeling the vertebrae respectively. In addition, for these networks, the proportion of samples extracted per category is set to 50/50% between the fore- and background. For the network that segments the metastases, this proportion is set to 25/75% and cross-entropy is adopted as an additional loss function. Especially for the metastatic segmentations (small volumes), an uniform amount of samples over the scan would result in a small amount of samples within the target volume. Therefore, it is hypostasized that the performance of the model will improve by increasing the amount of samples within the foreground. Cross-entropy is adopted as an additional loss function, because it is more suitable for small volumes such as spinal metastases. DSC is not an ideal metric to evaluate metastases, since it would be dominant by the larger target volumes. Moreover, the learning rate of the labeling network is decreased to 0.0005. Decreasing the learning rate, prevents the network from overwriting the parameters obtained during the previous training.

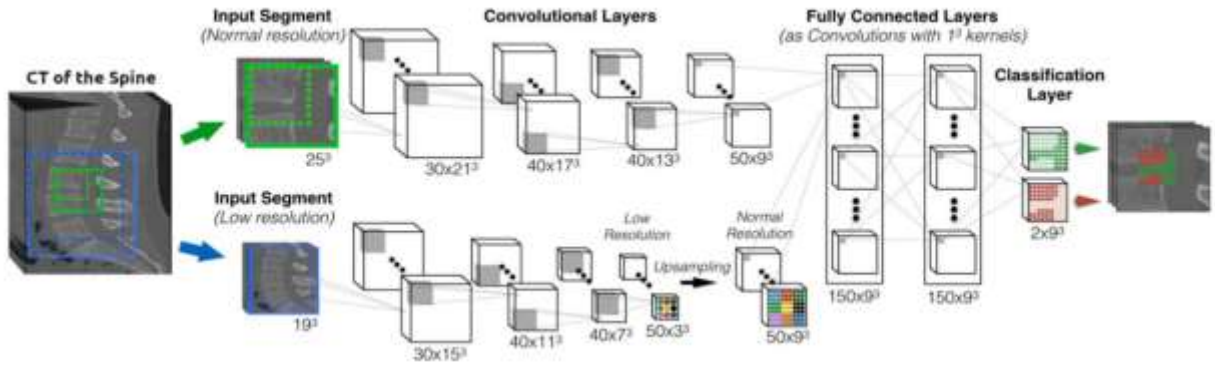

Figure 3: Architecture of DeepMedic (adapted from Kamnitsas et al. (2017)[10])

2.2.4. Post-processing and evaluation

After the training and inference of the networks, the segmentations are post-processed using Scikit-image 0.18.0. This is a Python (version 3.7.0; Python Software Foundation) package dedicated to image processing. During the post-processing of the segmentations, morphological operations of filling the small holes are applied and small segmented areas are removed.

Performances of the different models before and after the post-processing are evaluated in terms of DSC and HD. The metastatic segmentations are also evaluated by calculating sensitivity (Equation 1) and specificity (Equation 2) for detecting a metastasis. These segmentations of the metastases are evaluated per vertebra.

$$Sensitivity = \frac{True\ positives}{True\ positives + False\ negatives} \quad (1)$$

$$Specificity = \frac{True\ negatives}{True\ negatives + False\ positives} \quad (2)$$

Moreover, the results of the network for segmenting the spinal metastases, are compared with automatic threshold techniques such as Otsu[13] and Yen[14]. The results are also compared to an automated segmentation technique based on gradient edge detection. Before the edge detection is applied, a bilateral filter[15] is used to denoise the scan while preserving the edges. A watershed algorithm[16] is employed for the final segmentation of the gradient image. The processing of these segmentation methods is performed in Python, utilizing the packages Scikit-image 0.18.0, SimpleITK 2.0.2, and opencv-python 4.6.0.66.

2.3 Post-processing - Extraction of radiomic features

After the training of the networks, radiomic features are extracted from the segmentations/labels. Based on an expert meeting (consisting of radiation oncologists, spine surgeons, radiologists, and data scientists) radiomic features are defined that fulfill two important requirements: 1) are expected to influence the decision whether a patient should be referred to the spine surgeon or not 2) can be automatically extracted based on the PET/CT segmentations. Table 4 shows the 16 radiomic features that are extracted from the data.

Table 4: Names and descriptions of radiomic features

Name of feature	**Explanation**
HU vertebral body	Mean Hounsfield Unit (HU) within the vertebral body
HU tumor	Mean HU within the tumor
Location spine	The score for the location of the tumor within the spine (score is identical to the first component of Table 1)
Involvement cortex	The percentage of the vertebral cortex of the vertebral body that is affected by a tumor
Involvement body	The percentage of the vertebral body that is affected by a tumor
Involvement middle body	The percentage of the middle one-third of the axial plane of the vertebral body that is affected by a tumor
Involvement ventral body	The percentage of the ventral one-third of the vertebral body that is affected by a tumor
Subluxation	Dislocation relative to adjacent vertebrae (mm)
Collapse	The difference (mm) in height between the middle of the vertebra and the ventral/dorsal parts
Relative volume vertebra	The difference (mm^3) between the volume of each vertebra and its expected volume based on surrounding vertebra
Involvement spinal canal	The percentage of the spinal canal that is affected by tumor
Alignment (3 options):	
Alignment1	Absolute difference (mm) between the ventral and dorsal height of vertebral body
Alignment2	For all vertebrae at the anterior side of the SVA-line (the line between C7 and L5) the dorsal height (mm) is subtracted from the ventral height (mm) and vice versa for the vertebrae at the dorsal side of the line. The result is multiplied by a weight (0-1) depending on the distance to the SVA-line.

Alignment3	The same as 'Alignment2'. However, for all vertebrae at the anterior side of the SVA-line, the ventral height (mm) is subtracted from the dorsal height (mm) and vice versa for the vertebra at the dorsal side of the line.
Location vertebral arch	Location of a tumor within the vertebral arch (score is identical to the last component of Table 1)
Involvement vertebral arch	The percentage of the vertebral arch that is affected by a tumor

In this study there will be referrals to these features by using the names stated in Table 4. An in-depth description of each radiomic feature can be found in Appendix A.2. Due to the deviant shape of the first two cervical vertebrae, features are only extracted from cervical level C3 until L5. For the extraction of all previously described features, Python programming language is utilized. Min-max normalization is applied to scale all values between 0 and 100. These minimum and maximum values are calculated by taking the 95%-confidence interval around the mean of the resulting features of all scans. Soft clipping is applied for all values below and above these minima and maxima, utilizing the mathematical expression for a sigmoid.

2.4 Train model for SINS prediction

The final step is to automatically convert these features into a prediction of the SINS. This step includes the training and comparison of five different supervised training methods: A linear mixed effect model (LMEM), Lasso regression, Ridge regression, SoftMax regression, and Support Vector Machine (SVM). The linear regression models (i.e., LMEM, Lasso regression, Ridge regression) consider the dependent variable as continuous (0-18), while the non-linear multinomial logistic regression model (i.e., SoftMax regression) and the SVM are used to classify the outcome variable in different categories; group 1) SINS 1-3, group 2) SINS 4-6, group 3) SINS 7-12 and group 4) SINS 13-18).

The SINS is manually assessed for all vertebrae that contain metastases by one observer and randomly (20% of the scans) verified by a radiation oncologist. The scores do not include the pain component of the SINS, because this information cannot be automatically extracted from the scans. The SINS calculations are utilized as a ground truth for the SINS predictions of the model. If the model predicts the SINS for a particular vertebra, while there is no reference score available (i.e., model incorrectly indicates vertebra has tumor), it is scored with a SINS of 0.

Model development is undertaken using five-fold cross-validation, splitting the scores of all vertebrae with metastases into a training (70%) and validation (10%) set. The assessment is performed using an independent test set (20%). For the optimization of the model, back- and forward feature selection is applied to eliminate the features that do not improve the models' performance. Multicollinearity between the independent features of the model is prevented by calculating the variance inflation factor (VIF) and removing the features that are highly correlated to each other (i.e., VIF>5). In addition, the correlations between the features themselves and the individual components of the SINS are analyzed with Pearson's correlation coefficient. Finally, the models are evaluated and compared by calculating the proportion of the variation in the resulting SINS that is predictable from the model (R^2) and relative contribution of each feature (i.e., proportional reduction of the R^2). The equations for calculating R^2, VIF, and the proportional reduction of R^2 are stated below in equations 3,4, and 5.

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_i (SINSactual - SINSpredicted)^2}{\sum_i (SINSpredicted - SINSpredicted\ mean)^2} \quad (3)$$

$$VIF_i = \frac{1}{1 - R_i^2} \quad (4)$$

$$Proportional\ reduction\ R^2 = \frac{R^2\ model\ including\ all\ features - R^2 model\ without\ 1}{R^2\ model\ including\ all\ features} \quad (5)$$

All models are trained and evaluated using Scikit-learn 1.0.2 and Statsmodels 0.13.2, both machine learning libraries in the Python programming language. GridsearchCV, a tool of Scikit-learn, is utilized for an automated optimization of the models' hyperparameters. It evaluates all different combinations of the specified hyperparameters and selects the best combination based on the resulting R^2. Based on these results, the best-performing model is chosen for further evaluation. The results of this model are split into two groups by thresholding: 1) do not refer and 2) do refer. The sensitivity (Equation 1) and specificity (Equation 2) are calculated for all thresholds between 4 and 7. Finally, the results of this indirect binary approach are compared with a machine learning model that directly classifies the vertebrae into the two groups. Depending on the best-performing model, it is compared with a non-linear (e.g., non-linear support vector classification) or linear (e.g., stochastic gradient descent (SGD) classifier) binary classification model.

3. Results

The results are separated into four sections: the patient characteristics, the assessment of the networks for segmenting/labeling the vertebrae and metastases, the optimization of a model for SINS prediction, and the assessment of the best performing model.

3.1 Patient characteristics

90 consecutive patients that received a PET/CT scan in the period between January 2014 and December 2020 were selected for the study. Three (3.3%) of them are excluded for having a numeric variation in the vertebral column. The exclusions resulted in a final data set of 87 PET/CT scans of unique patients and 551 included vertebrae with metastases. Table 5 shows that the patient characteristics are well balanced over the training and test set.

Table 5: Rates of different characteristics between training/validation and test set

	N.o.(%)	
	Training and validation sets (n=68)	**Test set (n=19)**
Age median(minimum-maximum)	60 (40-84)	58 (42-82)
Male	29 (43%)	10 (50%)
Female	39 (57%)	10 (50%)
Spinal location[a]		
Cervical (C3-C7)	45 (10%)	12 (11%)
Thoracic (Th1-Th12)	261 (59%)	60 (57%)
Lumbar (L1-L5)	139 (31%)	34 (32%)
Multilevel (>1 tumor)	55 (81%)	16 (84%)
Histology		
Breast	14 (21%)	5 (1%)
Prostate	21 (31%)	0 (0%)
Renal cell carcinoma	2 (3%)	1 (5%)
Lung	24 (35%)	1 (5%)
Other	5 (7%)	0 (0%)
Unknown	0 (0%)	16 (84%)
SINS score[a,b]		
Median(minimum-maximum)	4 (0-12)	4 (0-10)
≤3	142 (31%)	37 (35%)
≥4	254 (57%)	61 (58%)
≥7	49 (11%)	8 (7%)

[a] *Is calculated over the total amount of included vertebrae. A total of 445 vertebrae are included in the training set and 106 in the test set.*

[b] *SINS without the SINS component for pain*

Figure 4 shows the distribution of SINS scores within both the training and test set. 95% of the scores are between 0 and 8 (SINS minus pain component).

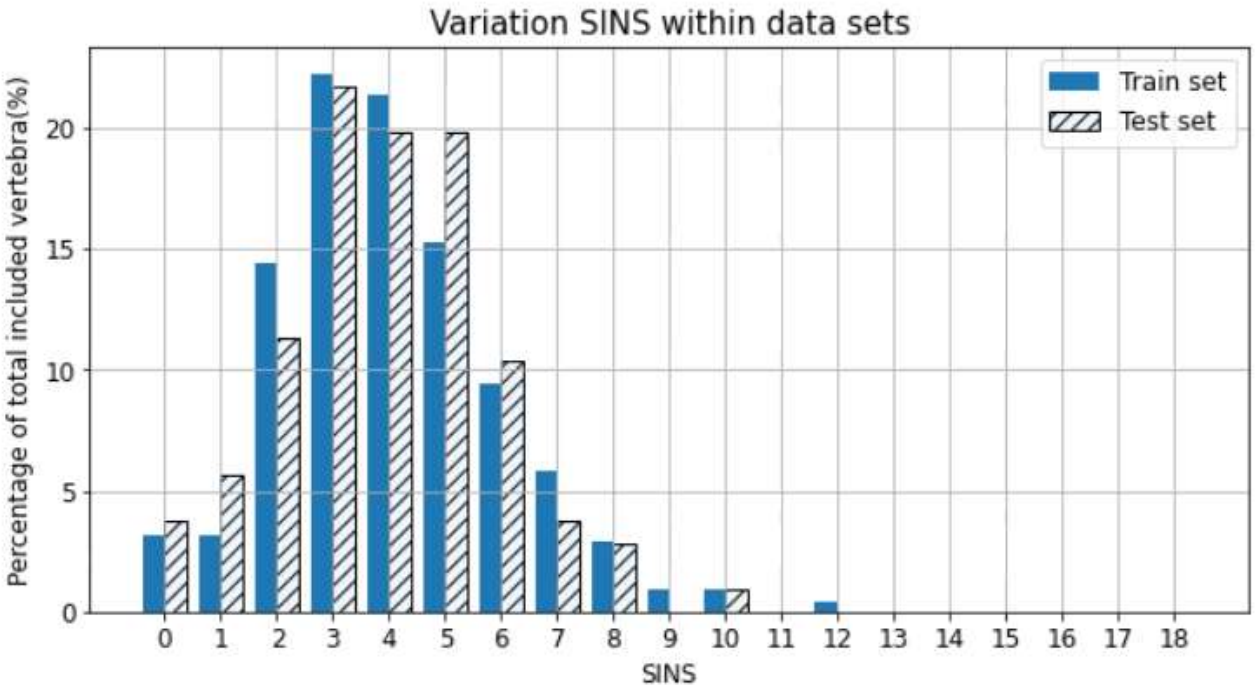

Figure 4: Percentage of vertebra with specific SINS score

3.2 Assessment segmentations/labels

Figure 5 shows a two-dimensional representation of the results from the segmentation/label networks.

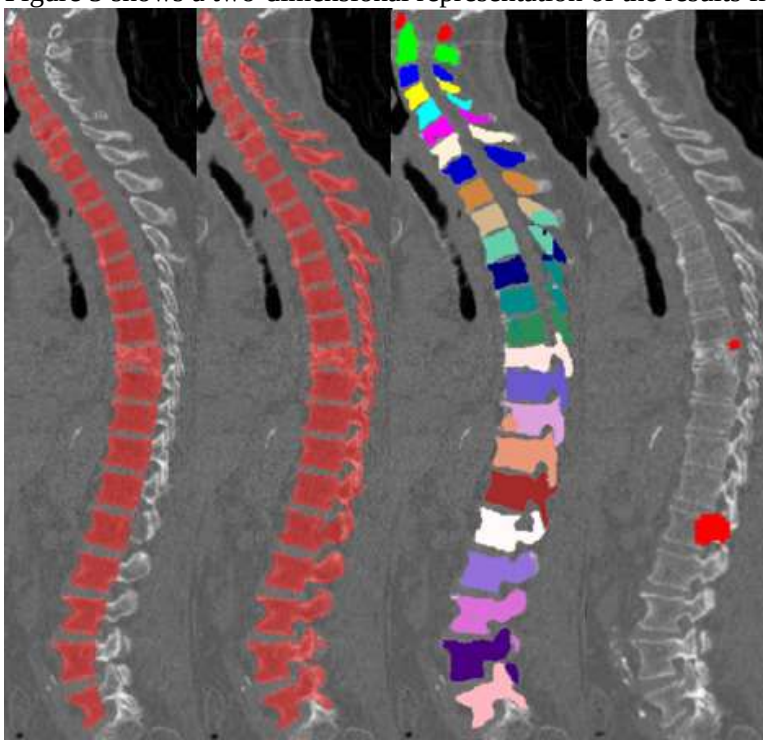

Figure 5: Result of the model for segmenting the vertebral bodies, segmenting the vertebrae, labeling the vertebrae, and segmenting the metastases respectively (test set).

Table 6 demonstrates the HD and DSC of all trained networks, evaluated using the test set. After the post-processing of the segmentations of the vertebrae and vertebral bodies, the HDs are significantly reduced. The CNN for segmenting the metastases shows better results compared to automatic threshold techniques such as Otsu (HD: 15±3, DSC: 0.622±0.0494) and Yen (HD: 21±3, DSC: 0.493±0.0465). Moreover, it performed better than segmentation based on gradient edge detection (HD:19±3, DSC: 0.513±0.0712).

Table 6: Mean HD and DSC of segmentations (test set)

	HD (95% CI) in mm	**DSC (95% CI)**
Segmentations vertebrae		
Without post-processing	288 (246-332)	0.978 (0.975-0.982)
With post-processing		
Cervical (C3-C7)	2 (1-3)	0.988 (0.986-0.990)
Thoracal (Th1-Th12)	2 (1-3)	0.991 (0.990-0.993)
Lumbar (L1-L5)	2 (1-3)	0.993 (0.992-0.995)
All	9 (6-13)	0.981 (0.977-0.984)
Segmentations vertebral bodies		
Without post-processing	260 (215-306)	0.965 (0.955-0.971)
With post-processing		
Cervical (C3-C7)	3 (2-4)	0.980 (0.976-0.985)
Thoracal (Th1-Th12)	3 (2-4)	0.988 (0.985-0.991)
Lumbar (L1-L5)	3 (2-4)	0.993 (0.991-0.995)
All	10 (6-14)	0.976 (0.973-0.979)
Segmentations metastases		
PET scan is provided	13 (11-15)	0.698 (0.655-0.742)
PET and CT scans are provided	12 (9-14)	0.720 (0.681-0.760)

The segmentations of the best-performing models are analyzed in more depth. In 10% of the resulting segmentations of the vertebral bodies and in 5% of the segmentations of the vertebrae, a part of the sacrum is segmented (Figure 6). In 5% of the segmentations of both the vertebral bodies and vertebrae, segmentation mistakes occurred due to contrast differences (Figure 7). In an additional 10% of the segmentations of the vertebral bodies, the second cervical vertebra (C2) is only partly segmented (±50% of its actual volume).

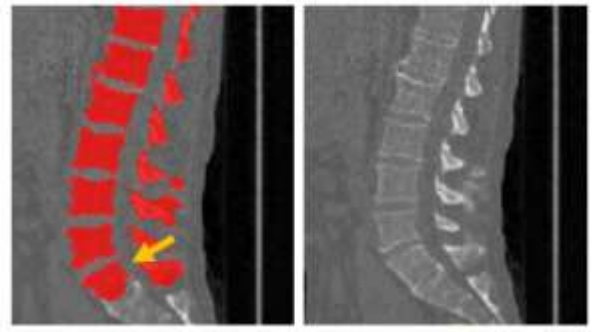

Figure 6: Part of sacrum is segmented

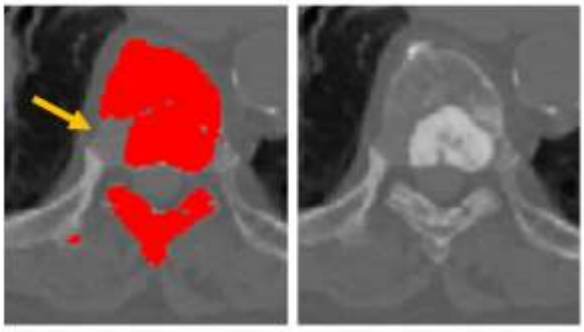

Figure 7: Incorrectly segmented vertebra

Some ground truths that are created by thresholding of the PET scans, showed segmented voxels within a vertebra, while that particular vertebra did not contain a metastasis based on the visual assessment of the scans. An example of such incorrect segmented voxels can be found near the arrow in Figure 8. Based on

the visual examination, the metastasis is only located in the vertebrae above it, but the segmentation of the PET also shows red voxels within the vertebra below it.

Figure 8: Example of incorrect segmented voxels

Another example that resulted in such errors, is that the SINS is not manually calculated for small metastases consisting of only a few voxels. The model calculates the SINS for all vertebrae with a segmented metastasis, even when just one voxel is segmented. Since these errors can also occur in the output segmentations of the network, it is decided to filter out these vertebra by the model instead of correcting the ground truths. Therefore, an optimal volumetric threshold is chosen that keeps the amount of vertebra that are missed by the model at 0.0%. All metastases with a volume lower than 0.25% of the surrounding vertebrae are for this reason not involved in the SINS prediction (Table 7). It shall be noted that the results of Table 7 are based on the ground truths. Testing the network for segmenting metastases from PET scans, resulted in a sensitivity of 94% and specificity of 97% after this thresholding.

Table 7: The number of errors after thresholding: All metastases with volume ≤ 0.25% of volume vertebrae are removed (ground truths)..

	Before thresholding	**After thresholding**
Training set		
Model incorrectly indicates that vertebra has a tumor	20.5 %	3.3 %
Model incorrectly indicates that vertebra has no tumor	0.0 %	0.0 %
Test set		
Model incorrectly indicates that vertebra has a tumor	20.8 %	3.8 %
Model incorrectly indicates that vertebra has no tumor	0.0 %	0.0 %

The network for labeling the vertebrae is evaluated by calculating the proportion of each vertebra that is correctly labeled. Figure 9 shows these percentages and their confidence intervals for both the pre-trained and retrained network.

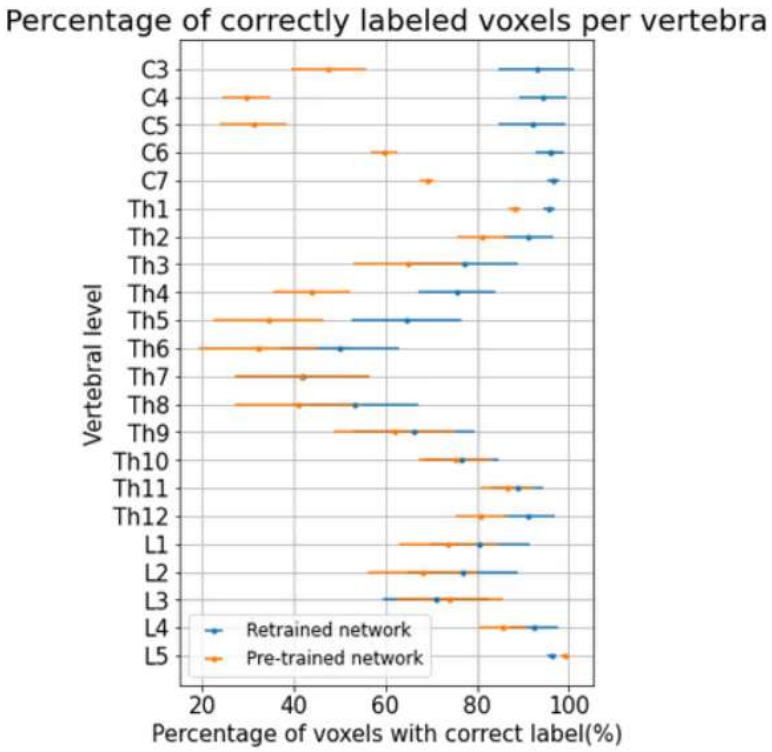

Figure 9: Percentage of correctly labeled voxels

Figure 10 shows an example of labeling errors that occurred. Multiple labels are assigned to the same vertebra.

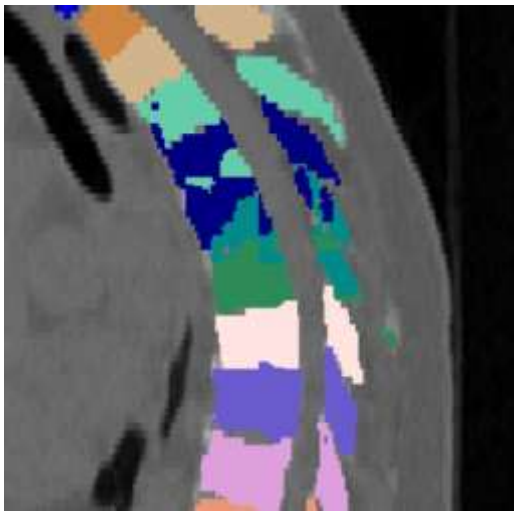

Figure 10: Labeling errors

Due to the limited performance of the models for segmenting the metastases and labeling the vertebrae, this study continued with the ground truths.

3.3 Optimizing a model for SINS prediction

Five machine learning models are optimized by applying feature selection and five-fold cross-validated hyperparameter tuning. Based on the resulting R^2-values that are stated in Table 8, the linear models (i.e.,

Lasso regression, Ridge regression, and LMEM) have the best performance. There is no significant difference between the values of these three models.

Table 8: R^2 of different models

	R^2 validation set	**R^2 test set**
Support Vector Machine	0.28 (0.28-0.29)	0.32
SoftMax regression	0.31 (0.30-0.32)	0.31
Lasso regression	0.53 (0.53-0.55)	0.62
Ridge regression	0.56 (0.55-0.56)	0.63
Linear Mixed Effect Model	0.56 (0.55-0.56)	0.60

This study continued with the LMEM for further assessment. For- and backward feature selection resulted in the combination of features with the best performance of the model. This combination corresponds to the features in Figure 11, which all show a positive contribution to the performance of the model based on the validation set (i.e., an increase in the R^2).

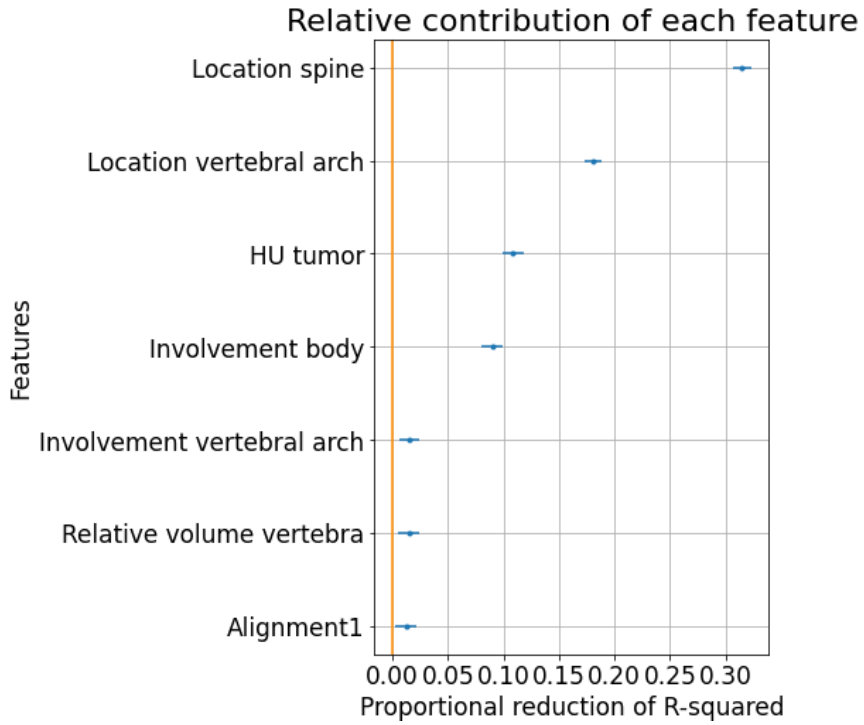

Figure 11: Relative contribution of each feature

In addition, the correlations between the features themselves and the subcomponents of the SINS are assessed. As expected, Appendix A.3 Figure 1 shows a high correlation (0.80) between the features 'Involvement cortex, 'Involvement body', 'Involvement middle body', and 'Involvement ventral body'. Of these features, feature 'Involvement body' resulted in the highest R^2. Adding one of the other three features, resulted in a VIF of ≥5. Therefore, these features are excluded from the model. Appendix A.3 Figure 2 shows a high correlation (0.87) between feature 'Location spine' and the SINS component regarding the location in the spine. It also shows a high (0.93) correlation between feature 'Location vertebral arch' and the SINS component regarding the location of the tumor within the vertebral arch. The highest (-0.53) correlation with the SINS component regarding the type of bone lesion is with feature 'HU tumor'. Appendix A.3 shows two confusion matrices (Figure 3 and 4) and a boxplot (Figure 5) plot of the previously

described correlations. The SINS component regarding the alignment had the highest correlation with features 'Involvement middle body', 'Involvement ventral body', and 'Alignment1' (0.17). The component regarding the vertebral collapse had the highest (0.50) correlation with feature 'Involvement body'.

The correlations between the features and the components of the SINS are also assessed by calculating the mean differences between the actual and predicted SINS scores (Appendix A.3 Table 1) for different groups. Each group consists of vertebrae with a specific score for an individual SINS component. The highest errors occurred in the groups with a score ≥2 for the alignment. The errors that occurred within the group with a score of 4 for the alignment are evaluated in more detail. Table 2 Appendix A.3 shows that there are four vertebrae with a score of 4 for this component. Three of these vertebrae are located at L5, having an error of 4, and one is located at C6, having an error of 0.

3.4 Assessment of the best performing model

Figure 12 shows a scatter plot of the actual versus the predicted SINS scores. The correctly predicted SINS scores are located on the line. As described in section 2.4, predicted SINS scores of 0 represent the vertebrae without an actual SINS score. If these zeros are excluded from the test set the resulting R^2 value is 0.64 instead of 0.60 (Table 8).

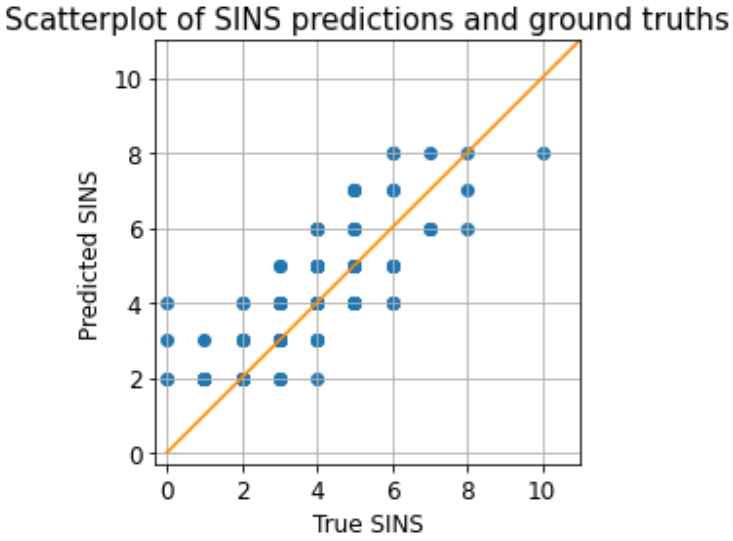

Figure 12: Scatterplot of true versus predicted SINS scores (test set)

Figure 13 shows a bar plot of the percentage of vertebrae with a specific error between actual and predicted SINS score. The mean error of all vertebra is -0.30 (95% CI: -0.53-0.07) The cumulative percentages of vertebra with an absolute error of zero, ≤1, and ≤2 are 40.4%, 82.0%, and 93.0% respectively.

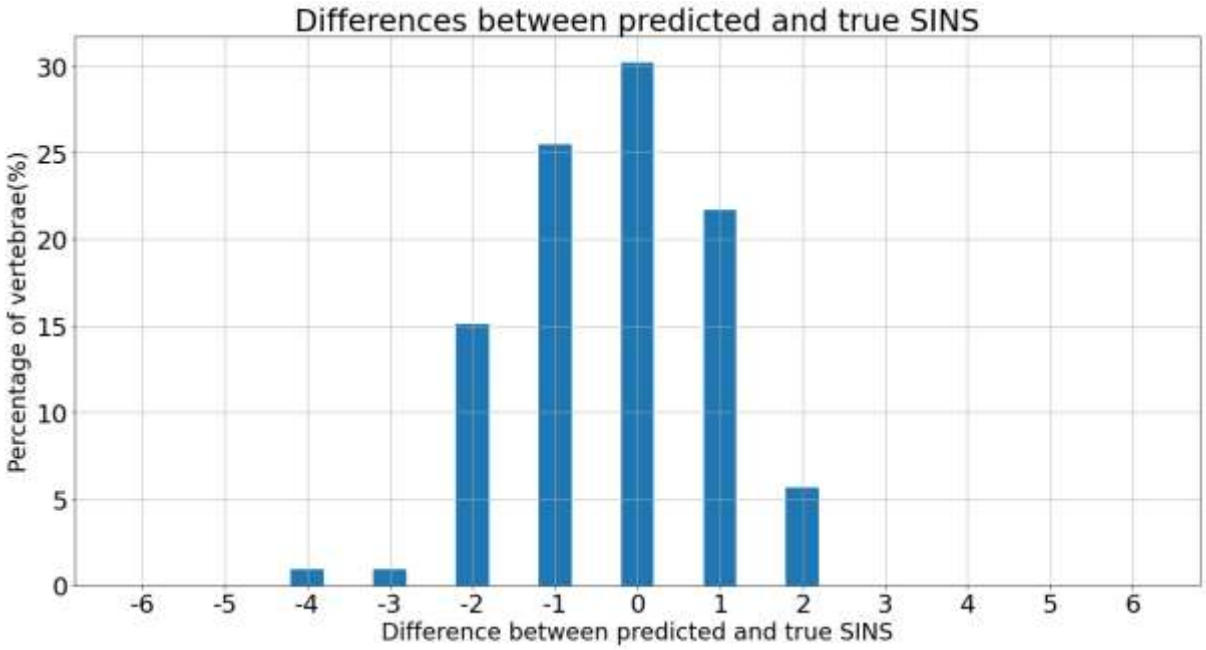

Figure 13: Percentages of vertebra with a specific error between actual and predicted SINS score. The predicted score is subtracted from the actual score (test set).

The performance of the LMEM model is also evaluated by splitting the scores (0-18) into two groups; do and do not refer a patient. Table 9 shows the sensitivity and specificity for different thresholds. A threshold of 4 resulted in the highest sensitivity but lowest specificity. By increasing the threshold, the sensitivity decreased, while the specificity increased.

Table 9: Sensitivity and specificity for referring patients at different thresholds (test set)

	Sensitivity	**Specificity**
Threshold = 4	0.93	0.76
Threshold = 5	0.70	0.83
Threshold = 6	0.63	0.87
Threshold = 7	0.50	0.93

The confusion matrix in Figure 14 shows the number of true negatives, true positives, false negatives and false positives if a threshold of 4 is chosen. The left upper quadrant includes 34(31%) patients with an actual score <4 that should not be referred (i.e., true negatives) and the right upper quadrant includes 11(11%) patients with a score <4 that should be referred according to the model (i.e., false positives). In addition, the lower left quadrant shows that 5(5%) patients with an actual score ≥4 should not be referred (i.e., false negatives) and the lower right quadrant shows that 57(53%) patients with a score ≥4 should be referred according to the model (i.e., true positives). A more detailed confusion matrix can be found in Appendix A.3 Figure 6.

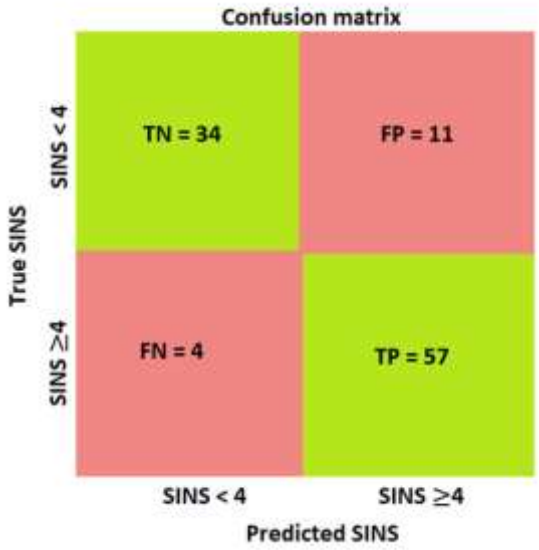

Figure 14: Confusion matrix of true versus predicted SINS score (test set), including the true negative (TN), false positive (FP), false negative (FN) and true positive (TP) values.

The results of this indirect approach are compared with the results of a SGD classifier model that directly classifies the vertebra into two groups (i.e., do and do not refer) using a threshold of 4. This direct binary approach resulted in a sensitivity of 0.89 and specificity of 0.73. This is lower than the sensitivity and specificity of the indirect approach (Table 9).

4. Discussion

To our knowledge, this is the first study that describes the development and assessment of a tool for automated SINS prediction. The first step is the training of CNNs that segment/label the vertebrae, segment the vertebral bodies and segment the metastases of the spine. Existing literature mainly focused on methods for segmenting healthy vertebrae[17] or did not include the cervical vertebrae[18,19]. Therefore, this is the first study that proposed an automated 3D deep learning workflow suitable for SINS prediction. Comparing the DSC and HD scores described in section 2.2.1 with Table 6, shows that retraining on a new data set improved the performance of the network for segmenting the vertebrae. The HD and DSC scores (Table 6) of the segmentations of both the vertebrae and vertebral bodies are sufficient for a reliable SINS prediction. A segmentation error, such as segmenting a part of the sacrum, does not affect the performance of the model, since it is easily removed in the post-processing phase where it is combined with the labeled vertebrae. The same applies when a part of C2 is not correctly segmented because this vertebral level is not included in the model. The errors that do affect the performance, are the ones that occur due to intravertebral contrast differences (5% of all vertebrae). To make the networks less vulnerable to these differences, more samples should be taken within these vertebrae and more of these vertebrae should be included in the training set. Another thing to take into consideration is that, when it was not possible to discriminate between the vertebra and its surrounding tissue based on differences in contrast (e.g., due to lytic metastasis within in the cortex), the ground truths were created by following its expected contour. Hence, the contour can slightly differ from its original contour, and influence further assessment based on shape.

Further, tracer-independent approaches for segmenting the metastases from PET/CT scans are investigated. To our knowledge, this is the first study that proposed a tracer-independent approach instead of tracer-dependent approach for segmenting spinal metastases. The results show that a CNN trained with both PET and CT scans performs the best. However, based on the results in Table 6, the segmentations of the

metastases lack robustness for a reliable SINS prediction. An explanation could be that different radiopharmaceuticals all have different physiological behavior, which means that the Standardized Uptake Values (SUV) of the PET scans can be very different for all scans. If it would be possible to automatically extract information about the utilized tracer type, this would enable a tracer-dependent approach that might show better results[20,21].

Also, the network for labeling the vertebrae needs some improvements. Figure 9 shows an improvement of the performance of the labeling network after retraining it with the PRESENT data set, but its performance is still not good enough for a reliable SINS prediction. Interestingly, most labeling errors occur in the mid-thoracic vertebrae. An anatomical explanation for this can be the location of these vertebrae within the curved part of the spine, showing smaller intervertebral discs. This in combination with the similarity in shape between these vertebrae might make it more difficult for the network to separate individual vertebral bodies. Preferably the whole segmented spine should be provided to the network at once instead of providing patches because it is easier to discriminate between the vertebra based on larger contextual information (e.g., location). However, this would require more computational power. Therefore, it is preferred to utilize a larger patch size or to add an additional down-sampled pathway to increase its receptive field. Another option could be to increase the number of samples taken within the mid-thoracal vertebrae. When providing a CT scan as a secondary input channel, the additional information could also make it easier for the network to discriminate between the vertebrae. In addition, cropping the scans around the spine would reduce the computational workload without the loss of important information. Moreover, post-processing could improve the labeling, such as; performing a watershed to separate the vertebrae, removing labeling mistakes by majority voting per vertebra, and by verifying whether the vertebrae are labeled in chronological order from top to bottom. Combining the outcomes of a new network, trained for labeling the vertebral bodies, with the labels of the vertebrae might improve the results even further.

While optimizing a model for SINS prediction, seven key predictive features are identified and included: Features ‘HU tumor’, ‘Location spine’, ‘Involvement body’, ‘Relative volume vertebra’, ‘Alignment1’, ‘Location vertebral arch’, and ‘Involvement vertebral arch’. In addition, the correlation between the features and the individual SINS components is analyzed. This demonstrated a very weak correlation between the features and the component regarding the alignment and a moderate correlation with both the components about collapse and the type of bone lesion. However, the correlation between the component regarding the type of bone lesion and feature ‘HU tumor’ might be underestimated, because the score for this individual component is determined per patient instead of per vertebra. Appendix A.3 Table 1 confirms the weak correlations between the features and the components regarding the alignment and collapse, by showing the highest error between the actual and predicted SINS in the groups with a score >0 for these individual components. A suggestion to improve the assessment of the components regarding the collapse and alignment would be to train a network that scores the vertebrae based on their shape[22].

A more in-depth analysis of the component regarding the alignment (Appendix A.3 Table 2) suggests that subluxation cannot be measured by the model when it is located at L5. An explanation could be that sacrum is not segmented, meaning that the degree of subluxation can only be assessed relative to L4. A recommendation for improvement would therefore be to train a network that also segments (a reference point within) the sacrum. Another solution might be to add weights to the vertebra including metastases. Currently, the degree of subluxation is automatically assessed by calculating the distance between the midpoints of each vertebra and a polynomial fit through the midpoints of all vertebrae (Appendix A.2.7). By adding weights, the polynomial function can be fitted more towards the midpoints of the vertebrae without metastases. Another difficulty that is noticed during the assessment of the subluxation, is when two adjacent vertebrae are subluxated relative to each other. If both vertebrae contain a metastasis, it is difficult to decide which one caused the dislocation and should therefore obtain a score of 4 for this individual

component. Further research is essential to get more insight into how this problem is addressed within the clinical workflow before it can be assessed in a standardized and automated fashion.

In addition to the assessment of the individual features, the performance of the model is evaluated. The resulted R^2-values in Table 8 show a poor to moderate correlation with the SINS, suggesting that the predictive power of the model is not reliable enough for clinical implementation. However, when using it as a referral tool, there is no need for knowing the exact SINS score. More important is to know whether a patient should be referred or not. For this reason, the number of output classes can be reduced to two. Two approaches are assessed for the classification of the vertebrae in these two groups (i.e., do and do not refer): a direct and an indirect approach. The best performing approach, the indirect approach, resulted in a high sensitivity (i.e., 93%) and specificity (i.e.,76%) when utilizing an optimal threshold of 4 (7 minus the score for the pain component). Especially, when comparing it with the literature. Fourney et al.(2011) analyzed the SINS predictions among international spine oncology surgeons using a threshold of 7 (including the pain component), resulting in a sensitivity and specificity of 96% and 80% respectively.[23] Thus a comparable performance can be achieved between the model and specialized human observers assuming information on mechanical pain is present. An overview of the SINS predictions can be found in Figure 14, showing four false negatives with an actual SINS of 4. However, it is debatable whether these are actual false negatives because the SINS score does not include the component regarding pain. Only when these patients had a score of 3 for this individual component, they should have been referred. Therefore, there might be an underestimation of the sensitivity. The specificity, on the other hand, might be overestimated for this reason. A critical side note is that Table 5 and Figure 4 notably show a relative underreported number of vertebrae with a SINS≥7. This, in combination with the limited representation of the components regarding the collapse and alignment in the model, might indicate that the performance would be worse if more vertebrae with a score ≥7 are assessed. This also applies to the results of Figure 13, since larger errors between the actual and predictive SINS occur more easily at higher SINS scores. Another thing to consider is that the results are based on the ground truths and not on the resulting segmentations/ labels of the networks. Segmentation/labeling mistakes might influence the reliability and performance of the final model.

Furthermore, there are some potential limitations regarding the generalizability of the outcomes of this study. Scans with large artifacts, a numeric variation in the number of vertebrae, and a partially visualized spine are excluded from the data set. Therefore, the SINS cannot be calculated for these scans. Moreover, a PET/CT scan must be available and because the vertebra located at C1 and C2 are excluded from the data set, the SINS score cannot be calculated for these vertebral levels. A limitation regarding the validation of this study is that all ground truths and manual SINS calculations are randomly validated by one radiation oncologist. Hence, the reliability of the ground truths can be questioned. It is also important to note that the manual assessment of the SINS, is carried out based on PET/CT scans, while in the clinical field mainly CT scans are utilized. For example, the percentage of the vertebral body that is involved, is assessed by looking at the ratio between the volume of the vertebral body and the regions with an increased SUV on the PET. This ratio might deviate from the ratio calculated based on CT scans only. For these reasons, an external validation with multiple observers is recommended.

Another recommendation is to look into a different radiation-free approach using MRI scans instead of PET/CT scans. Several studies[24,25] demonstrated the feasibility of generating 3D CT-like images for the visualization of bone structures derived from a MRI scans. With this solution, more information (i.e., features) can be obtained about the soft tissue, such as the tumor or spinal cord, without losing information about the bony structures.

A difficulty in assessing the stabilization of the spine in an automated fashion is that multiple features and vertebrae might influence each other. For example, depending on the location within the spine and the distance to the SVA-line, the misalignment of vertebrae has more or less influence on the spinal stability. Another example is that the presence of a metastasis in an adjacent vertebra increases the risk of mechanical complications[7]. To overcome these hurdles, a recommendation for improving the performance of the model is to add a feature that scores the vertebra based on the presence of tumors in the surrounding vertebrae. It would also be interesting to examine the correlations between the SINS and the factorial feature interactions and the features extracted by the open-source Python package; Pyradiomics[26]. Furthermore, it should be investigated if it is possible to extract clinical features such as the primary tumor type, experienced pain, age, and gender, from the patients' medical records automatically. Including these features might improve the models' performance[27]. Another way to increase its accuracy and robustness, is to combine the predictions of two or more models (i.e., ensemble learning). A whole different approach would be to directly train a deep learning model that labels each vertebra of the spine with a SINS score, instead of automating the assessment based on radiomic features. However, this approach requires sufficient data that was not available for this study.

Looking at the SINS from a different perspective, as a decision tool instead of a referral tool, will lead to even more interesting research opportunities. Currently, the prognostic value of scores between 7 and 12 is still controversial and therefore called the 'grey zone'. It is difficult to find a cutoff where the spine requires stabilization.[7] The features extracted in this study, can be utilized for the training of a new model with a decision-making purpose. Instead of using the SINS as a reference, the clinical outcomes (i.e., stabilized or not stabilized) need to be used as ground truths. Since many extracted features are not directly related to the SINS, but are important according to medical experts, it might be possible that there is a higher correlation with these clinical outcomes.

4.1 Summary of the future recommendations

- Recommendations for improvement of the segmentations of the vertebrae/vertebral bodies:
    - Include more vertebrae with intravertebral contrast differences and/or increase the number of samples within these vertebrae. Both PET/CT scans and CT scans can be utilized for the training of these networks.
- Recommendations for improvement of the PET segmentations:
    - Investigate the possibility of a tracer-dependent approach for segmenting metastases from PET scans.
- Recommendations for improvement of the labels:
    - Increase patch size and/ or add more down-sampled pathways.
    - Increase the number of patches within mid-thoracal vertebrae.
    - Provide a CT scan as a secondary input channel.
    - Crop scans around the spine.
    - Post-processing:
        - Perform a watershed to separate the vertebrae.
        - Remove labeling mistakes by majority voting between all labels within a vertebra.
        - Automated verification of whether the vertebrae are ordered chronologically (1-24).
        - Train a new network for labeling the vertebral bodies and combine these results with those of the vertebrae.
- Recommendations for optimization of the model for SINS prediction:
    - Improve features regarding collapse and alignment:
        - Train a network that scores the vertebrae based on their shape.
    - Improve the feature for scoring the degree of subluxation:

        - Train new networks that also segment/ label (a reference point within) the sacrum.
        - Add higher weights to vertebrae without a metastasis
        - Get more insight into how this component is assessed in the clinical workflow.
    - Ad a feature that scores the vertebra based on the presence of tumors in the adjacent vertebrae.
    - Investigate the correlations between the SINS and the factorial feature interactions.
    - Investigate the correlations between the SINS and the features extracted by Pyradiomics[26].
    - Investigate if it is possible to automatically extract clinical features from the patients' medical records and which features can predict the SINS.
    - Combine multiple models with ensemble learning.
    - Perform an external validation with multiple observers.
    - Train a new model with the extracted features to predict whether a vertebra needs to be stabilized or not (i.e., decision tool).
    - Investigate a different approach using MRI scans as input for the model.
    - Investigate a different approach: Train a deep learning model that directly labels the vertebra with a SINS score.

5. Conclusion

In conclusion, the results of this study suggest that the developed semi-data-driven approach is suitable for automating the assessment of the SINS when it is utilized as a referral tool (i.e., do/ do not refer) instead of a score prediction tool (i.e., 0-18). However, no reliable conclusion can be drawn until some (section 4.1) improvements are made and the results are externally validated with a data set that includes more vertebrae with high SINS scores.

References

1. Fisher CG, Dipaola CP, Ryken TC, et al. A novel classification system for spinal instability in neoplastic disease: An evidence-based approach and expert consensus from the spine oncology study group. *Spine (Phila Pa 1976)*. 2010;35(22). doi:10.1097/BRS.0B013E3181E16AE2
2. Pennington Z, Ahmed AK, Cottrill E, Westbroek EM, Goodwin ML, Sciubba DM. Intra- and interobserver reliability of the Spinal Instability Neoplastic Score system for instability in spine metastases: a systematic review and meta-analysis. *Ann Transl Med*. 2019;7(10):218-218. doi:10.21037/atm.2019.01.65
3. Arana E, Kovacs FM, Royuela A, et al. Spine Instability Neoplastic Score: agreement across different medical and surgical specialties. *Spine J*. 2016;16(5):591-599. doi:10.1016/J.SPINEE.2015.10.006
4. Salvatore G, Berton A, Giambini H, et al. Biomechanical effects of metastasis in the osteoporotic lumbar spine: A Finite Element Analysis. *BMC Musculoskelet Disord*. 2018;19(1). doi:10.1186/S12891-018-1953-6
5. Georgy BA. Metastatic Spinal Lesions: State-of-the-Art Treatment Options and Future Trends. *Am J Neuroradiol*. 2008;29(9):1605-1611. doi:10.3174/AJNR.A1137
6. Tschirhart CE, Finkelstein JA, Whyne CM. Biomechanics of vertebral level, geometry, and transcortical tumors in the metastatic spine. *J Biomech*. 2007;40(1):46-54. doi:10.1016/J.JBIOMECH.2005.11.014
7. Serratrice N, Faddoul J, Tarabay B, et al. Ten Years After SINS: Role of Surgery and Radiotherapy in the Management of Patients With Vertebral Metastases. *Front Oncol*. 2022;12:67. doi:10.3389/FONC.2022.802595/BIBTEX
8. Prospective Evaluation of Interventional Studies on Bone Metastases - the PRESENT Cohort - Full Text View - ClinicalTrials.gov. https://clinicaltrials.gov/ct2/show/NCT02356497. Accessed September 30, 2021.
9. Löffler MT, Sekuboyina A, Jacob A, et al. A Vertebral Segmentation Dataset with Fracture Grading. *https://doi.org/101148/ryai2020190138*. 2020;2(4):e190138. doi:10.1148/RYAI.2020190138
10. Kamnitsas K, Ledig C, Newcombe VFJ, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med Image Anal*. 2017;36:61-78. doi:10.1016/J.MEDIA.2016.10.004
11. Payer C, Štern D, Bischof H, Urschler M. Coarse to fine vertebrae localization and segmentation with spatialconfiguration-Net and U-Net. *VISIGRAPP 2020 - Proc 15th Int Jt Conf Comput Vision, Imaging Comput Graph Theory Appl*. 2020;5:124-133. doi:10.5220/0008975201240133
12. Yushkevich PA, Piven J. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. NeuroImage.
13. Xu X, Xu S, Jin L, Song E. Characteristic analysis of Otsu threshold and its applications. *PaReL*. 2011;32(7):956-961. doi:10.1016/J.PATREC.2011.01.021
14. Yen JC, Chang FJ, Chang S. A New Criterion for Automatic Multilevel Thresholding. *IEEE Trans Image Process*. 1995;4(3):370-378. doi:10.1109/83.366472
15. Tomasi C, Manduchi R. Bilateral filtering for gray and color images. *undefined*. 1998:839-846. doi:10.1109/ICCV.1998.710815
16. Jain AK, Murty MN, Flynn PJ. Data clustering. *ACM Comput Surv*. 1999;31(3):264-323. doi:10.1145/331499.331504
17. Qu B, Cao J, Qian C, et al. Current development and prospects of deep learning in spine image analysis: a literature review. *Quant Imaging Med Surg*. 2022;12(6):3454-3479. doi:10.21037/QIMS-21-939/COIF
18. Klein G, Martel A, Sahgal A, Whyne C, Hardisty M. Metastatic Vertebrae Segmentation for Use in a Clinical Pipeline. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)*. 2020;11963 LNCS:15-28. doi:10.1007/978-3-030-39752-4_2/FIGURES/4
19. Hardisty M, Gordon L, Agarwal P, Skrinskas T, Whyne C. Quantitative characterization of metastatic disease in the spine. Part I. Semiautomated segmentation using atlas-based deformable registration and the level set method. *Med Phys*. 2007;34(8):3127-3134. doi:10.1118/1.2746498
20. Yuan Y. Automatic Head and Neck Tumor Segmentation in PET/CT with Scale Attention Network. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)*. 2021;12603 LNCS:44-52. doi:10.1007/978-3-030-67194-5_5/TABLES/2
21. Kao YS, Yang J. Deep learning-based auto-segmentation of lung tumor PET/CT scans: a systematic review. *Clin Transl Imaging*. 2022;10(2):217-223. doi:10.1007/S40336-022-00482-Z/TABLES/4
22. Vertebral Abnormality Scoring - Grand Challenge. https://grand-challenge.org/algorithms/vertebral-abnormality-scoring/. Accessed July 3, 2022.
23. Fourney DR, Frangou EM, Ryken TC, et al. Spinal instability neoplastic score: An analysis of reliability and

validity from the spine Oncology Study Group. *J Clin Oncol*. 2011;29(22):3072-3077. doi:10.1200/JCO.2010.34.3897
24. van der Kolk BBYM, Slotman DJ, Nijholt IM, et al. Bone visualization of the cervical spine with deep learning-based synthetic CT compared to conventional CT: A single-center noninferiority study on image quality. *Eur J Radiol*. 2022;154:110414. doi:10.1016/J.EJRAD.2022.110414
25. Morbée L, Chen M, Herregods N, Pullens P, Jans LBO. MRI-based synthetic CT of the lumbar spine: Geometric measurements for surgery planning in comparison with CT. *Eur J Radiol*. 2021;144:109999. doi:10.1016/J.EJRAD.2021.109999
26. Van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res*. 2017;77(21):e104-e107. doi:10.1158/0008-5472.CAN-17-0339
27. Gui C, Chen X, Sheikh K, et al. Radiomic modeling to predict risk of vertebral compression fracture after stereotactic body radiation therapy for spinal metastases. *J Neurosurg Spine*. 2021;36(2):294-302. doi:10.3171/2021.3.SPINE201534

A. Appendix

A.1. Training configuration of networks

Table 1: Training configuration network for segmenting metastases

	Network segmenting metastases from PET/CT scans
Number of epochs	35
Batch size training	10
Optimizer	Root Mean Squared Propagation
Losses and their weights for total cost	{"xentr": 0.5, "dsc": 0.5}
Drop-out rate	[0.0, 0.5, 0.5]
Activation function	Prelu
Proportion of examples to extract per category	[0.25,0.75]
Dimensions of the kernel	[[3,3,3], [3,3,3], [3,3,3], [3,3,3], [3,3,3], [3,3,3], [3,3,3], [3,3,3]]
Size of image segments (training)	[37,37,37]
Initial learning Rate (LR)	0.001
Epochs to lower LR (divide by 2)	[17,22,27,30,33]

Table 2: Training configuration network for segmenting vertebral bodies

	Network labeling vertebral bodies CT scan
Number of epochs	*35*
Batch size training	*10*
Optimizer	*Root Mean Squared Propagation*
Losses and their weights for total cost	*"dsc"*
Drop-out rate	*[0.0, 0.5, 0.5]*
Activation function	*Prelu*
Proportion of examples to extract per category	*[0.5,0.5]*
Dimensions of the kernel	*[[3,3,3], [3,3,3], [3,3,3], [3,3,3], [3,3,3], [3,3,3], [3,3,3], [3,3,3]]*
Size of image segments (training)	*[37,37,37]*
Initial learning Rate (LR)	*0.001*
Epochs to lower LR (divide by 2)	*[17,22,27,30,33]*

Table 3: Training configuration network for segmenting vertebrae

	Network labeling vertebral bodies CT scan
Number of epochs	*25*
Batch size training	*10*
Optimizer	*Root Mean Squared Propagation*
Losses and their weights for total cost	*"dsc"*
Drop-out rate	*[0.0, 0.5, 0.5]*
Activation function	*Prelu*
Proportion of examples to extract per category	*[0.5,0.5]*
Dimensions of the kernel	*[[3,3,3], [3,3,3], [3,3,3], [3,3,3], [3,3,3], [3,3,3], [3,3,3], [3,3,3]]*
Size of image segments (training)	*[37,37,37]*
Initial learning Rate (LR)	*0.001*
Epochs to lower LR (divide by 2)	*[17,22,27,30,33]*

Table 4: Training configuration network for labeling vertebrae

	Network labeling vertebral bodies CT scan
Number of epochs	*45*
Batch size training	*10*
Optimizer	*Root Mean Squared Propagation*
Losses and their weights for total cost	*"dsc"*
Drop-out rate	*[0.0, 0.5, 0.5]*
Activation function	*Prelu*
Proportion of examples to extract per category	*[0.5,0.5]*
Dimensions of the kernel	*[[3,3,3], [3,3,3], [3,3,3], [3,3,3], [3,3,3], [3,3,3], [3,3,3], [3,3,3]]*
Size of image segments (training)	*[37,37,37]*
Initial learning Rate (LR)	*0.0005*
Epochs to lower LR (divide by 2)	*[17,22,27,30,33]*

A.2. Description of extraction radiomic features

First, a PET/CT scan should be provided to the four CNNs, resulting in segmentations/labels of the vertebrae, segmentations of the vertebral bodies, and segmentations of the metastases in the spine. To obtain labeled vertebral bodies, the segmentations of the vertebral bodies are multiplied by the labels of the vertebrae. In addition, the segmentations of the metastases can be labeled per vertebra, by multiplying the segmentations with the labels of the vertebrae. From all these segmentations/labels, radiomic features can be extracted. How these are extracted, is described in the next sections. Features are only extracted from vertebrae that contain a metastasis of which the volume is ≥0.25% of the total vertebral volume.

A.2.1 HU vertebral body

For the radiomic feature regarding the mean Hounsfield Unit (HU) of the vertebral body, a 3D mask is created for each vertebral body that contains a metastasis (Figure 1).

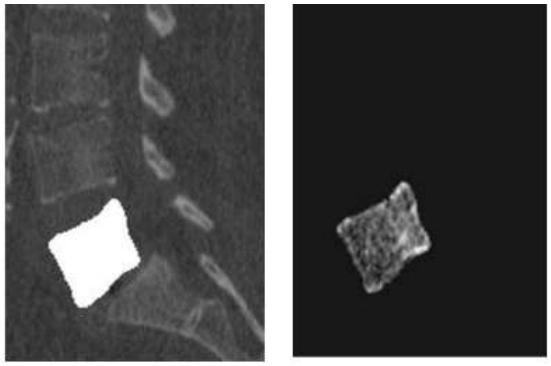

Figure 1: Left) CT scan with mask of vertebral body, right) image intensities (HU) within vertebral body

This mask is utilized to calculate the mean HU of the CT scan within the mask (Figure 2). This calculated HU is normalized between 0 and 100.

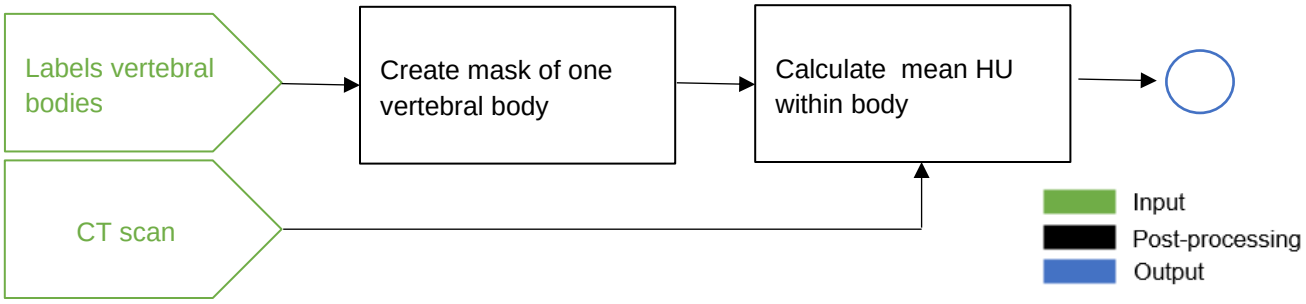

Figure 2: Post-processing workflow feature 'HU vertebral body'

It is expected that this feature has a negative correlation with the SINS, since a lower HU might indicate that the bone is destructed/softened and therefore it increases the likelihood of a vertebral collapse.[1]

A.2.2 HU tumor

For the radiomic feature regarding the mean Hounsfield Unit (HU) of the metastases within a vertebra, a 3D mask is created of all metastases within a particular vertebra. Similar to the previous section, this mask is utilized to calculate the mean HU of the CT scan within this mask and the outcome value is normalized between 0 and 100 (Figure 3).

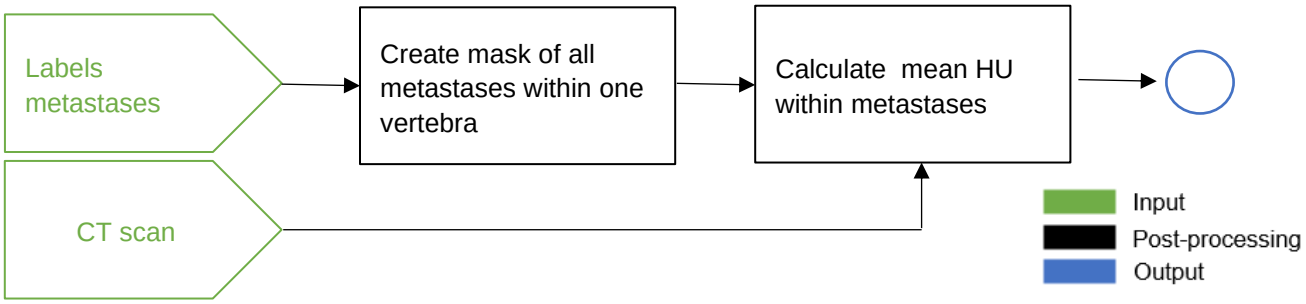

Figure 3: Post-processing workflow feature 'HU tumor'

Lytic lesions often appear darker on the CT scan, while blastic lesions appear lighter. It is therefore expected that this feature has a negative correlation with the SINS component regarding the type of bone lesion.[2]

A.2.3 Location spine

The feature regarding the location of the vertebra is based on the SINS component about the location. Depending on the label, a score will be assigned to the vertebra that is similar to the scores for this SINS component (Figure 4). The output is therefore categorical: 1, 2, or 3.

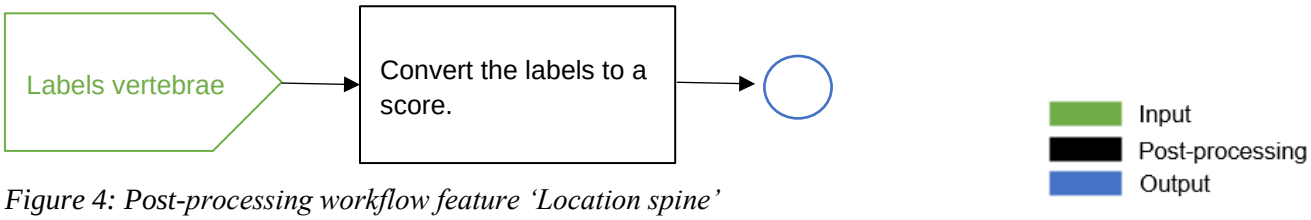

Figure 4: Post-processing workflow feature 'Location spine'

A.2.4 Involvement cortex

To calculate the percentage of tumor involvement within the vertebral cortex, both the labels of the vertebral bodies and metastases are utilized. First, a 3D mask is created from the vertebral body. These masks are eroded with 1 voxel around the borders. Subsequently, the eroded mask is subtracted from the original mask resulting in a segmentation of the vertbral cortex with a thickness of 1 voxel. This segmentation is multiplied with a mask of the metastases within that vertebra. Finally, to know the percentage of involvement within the cortex, all voxel values are summed up, divided by the total amount of voxels in the cortex, and multiplied with 100 (Figure 5).

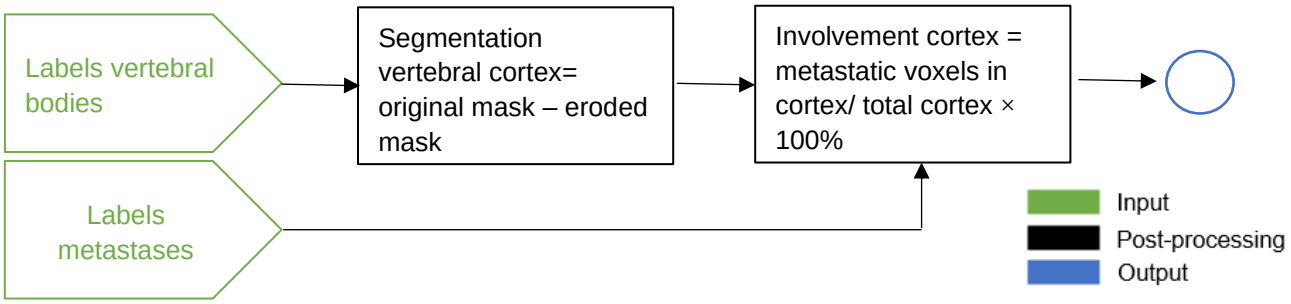

Figure 5: Post-processing workflow feature 'Involvement cortex'

It is expected that the involvement of the tumor in the vertebral cortex does affect the spinal stability more than when it is involved in the cancellous bone[3].

A.2.5 Involvement body

To calculate the percentage of tumor involvement within the vertebral body, both the labels of the vertebral bodies and metastases are utilized. The volume of the metastases is calculated per vertebral body, divided by the total volume of the vertebral body, and multiplied by 100. The result is the percentage of tumor involvement (Figure 6).

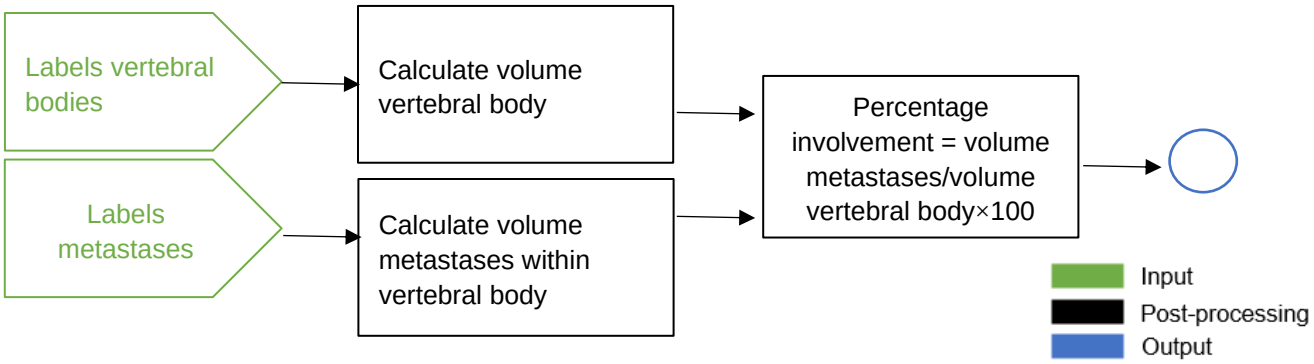

Figure 6: Post-processing workflow feature 'Involvement body'

It is hypothesized that larger tumors, affect the stability of the spine more than smaller tumors[4].

A.2.6 Involvement middle/ventral body

In addition to the involvement of the tumor within the vertebral body, the involvement of the tumor within both the middle axial plane and ventral plane of the body are analyzed. To be able to analyze the different parts of the vertebral body, it needs to be separated into smaller volumes. The vertebral body can be depicted as a cube composed of 27 smaller cubes (Figure 7).

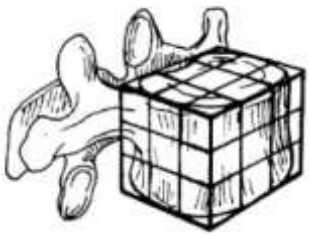

Figure 7: Comparison between a vertebral body and a cube[5]

To create such a vertebral cube, first, a sagittal slice should be taken through the center of the vertebral body (Figure 8). A bounding box can be created that surrounds the vertebra and the coordinates of the corners can be calculated. With these corner points, it becomes possible to determine the vertical and horizontal division lines of the cube (at 1/3th and 2/3th between the corners).

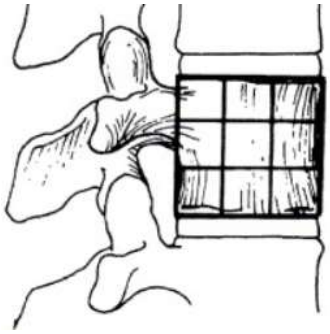

Figure 8: Sagittal plane of vertebral body (cube) [5]

If the same is applied to a slice in the coronal direction, a cube can be created of the vertebral body (Figure 9).

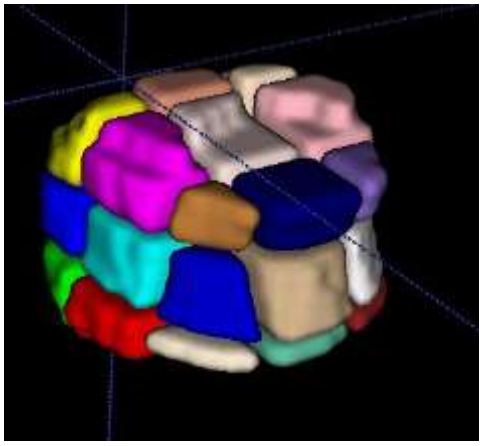

Figure 9: 3D vertebral body, divided in 27 segments

A.2.6.1 Involvement middle body

To be able to extract the percentage of involvement within the middle axial plane, the vertebral body needs to be divided into three horizontal parts (Figure 10).

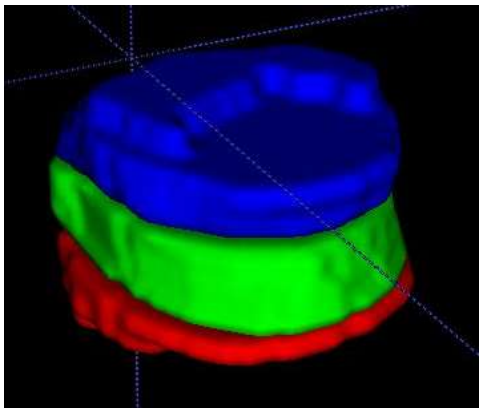

Figure 10: 3D vertebral body, divided in three horizontal planes

It is now possible to determine the total volume of the metastases within the middle axial plane. This volume will be divided by the total volume of this plane and multiplied by 100 (Figure 11).

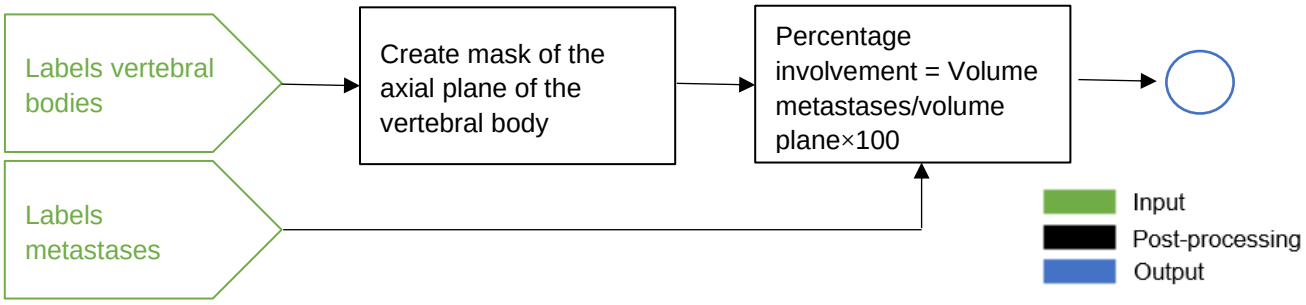

Figure 11: Post-processing workflow feature 'Involvement middle body'

It is investigated that destruction of the middle third in the axial plane results in gross instability, whereas destruction of the middle third in the sagittal plane may not be associated with significant destabilization[5].

A.2.6.2 Involvement ventral body

To be able to extract the percentage of involvement within the ventral plane, the vertebral body needs to be divided into two vertical planes: The ventral one-third and the dorsal side (Figure 12).

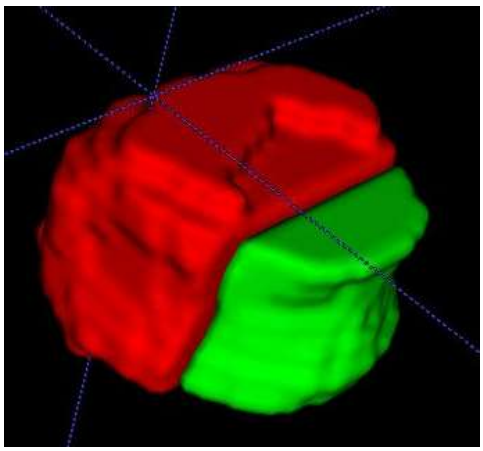

Figure 12: 3D vertebral body, divided in two vertical planes

The total volume of the metastases can be calculated within the ventral part. This volume will be divided by the total volume of this part and multiplied by 100.

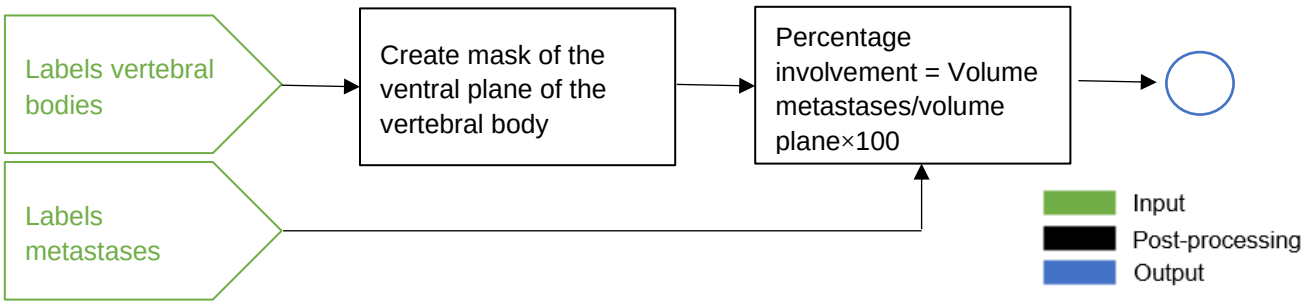

Figure 13: Post-processing workflow feature 'Involvement ventral body'

It is investigated that a lesion in the ventral portion of the vertebral body in the coronal plane affects stability more than a lesion in the middle or dorsal portions[5].

A.2.7 Subluxation

For the calculation of the degree of subluxation, the centers of gravity of all vertebral bodies are calculated and a polynomial function is fit through these points in the sagittal plane (Figure 14). Thereafter, the distances between the function and the coordinates of each vertebra can be calculated. For each vertebra, the difference between the distance for that vertebra and the mean distance of the 4 surrounding vertebrae (2 above and 2 below) is calculated. This makes it less vulnerable to errors when the polynomial function does not fit perfectly to a part of the spine.

This feature is based on the component of the SINS about the subluxation. However, instead of giving a subluxated vertebra a score of 4, the outcome is a continuous value (mm) that increases if the degree of subluxation increases (Figure 15). The final value is normalized between 0 and 100.

Figure 14: Polynomial function fitted through center points of vertebrae

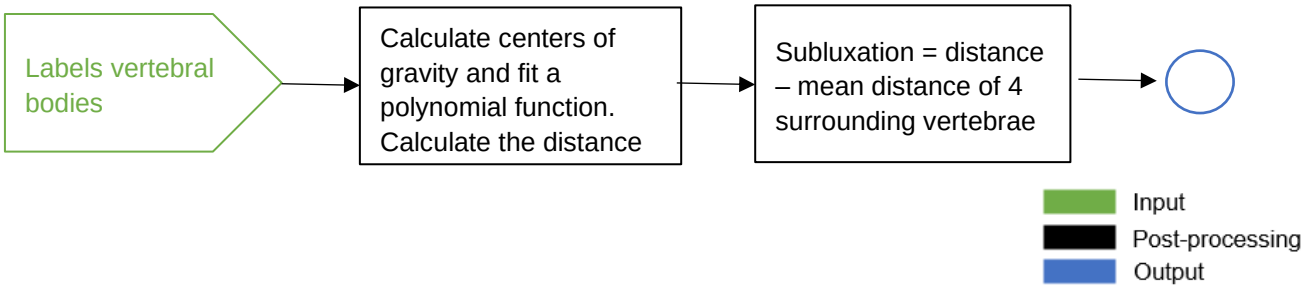

Figure 15: Post-processing workflow feature 'Subluxation'

A.2.8 Collapse

For the feature regarding the collapse, first, a bounding box is created around the vertebral body in the sagittal plane through the center of gravity. The vertebral height is calculated at three positions along the upper and lower sides of the bounding box: at 11/24th, 12/24th and 13/24th. At these positions, lines are drawn parallel to the left and right side of the box. These lines are used for the calculation of the points of intersection with the vertebral contour. When the intersection points at the contour are calculated, the height can be measured at these positions and the mean can be calculated. Thereafter, it will be compared with the height of the bounding box that surrounds that vertebra (Figure 16).

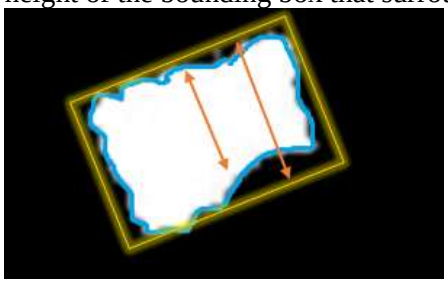

Figure 16: Vertebra with bounding box (yellow) and contour (blue)

The height in the middle of the vertebra is subtracted from the height of the bounding box. This resulting value represents the collapse (mm) within the middle of the vertebra (Figure 17). The final value is normalized between 0 and 100.

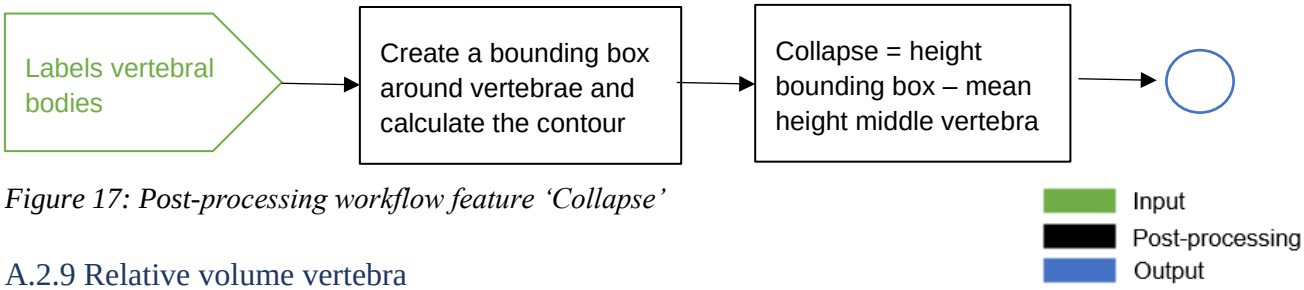

Figure 17: Post-processing workflow feature 'Collapse'

A.2.9 Relative volume vertebra

It is hypothesized that the volume of the vertebra, has a correlation with the SINS component about the collapse. However, a difficulty is that all vertebrae have different natural volumes. There is need for a reference volume for each vertebra, to be able to calculate the difference (mm^3). Therefore, the volumes of all vertebrae are calculated and plotted. Thereafter, a polynomial function is fit through these points (Figure 18).

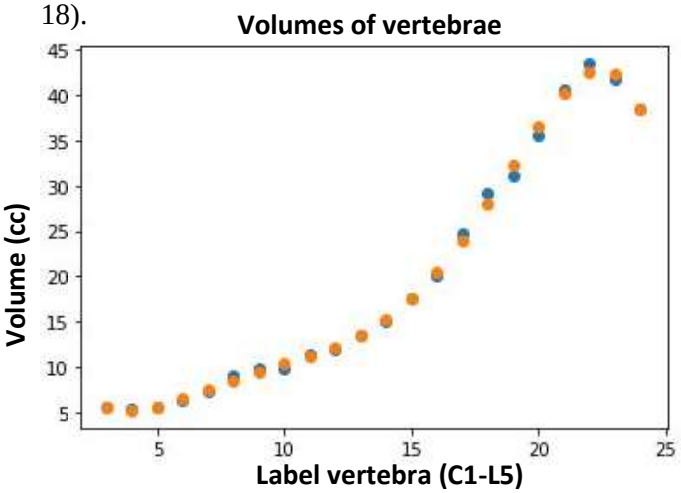

Figure 18: Volumes of vertebrae (blue) and expected volumes according to polynomial function (orange)

All negative values (when the vertebra is larger than expected) are set to zero, since these do not influence the stability in a negative way. The final value is normalized between 0 and 100. An overview of the workflow is showed in Figure 19.

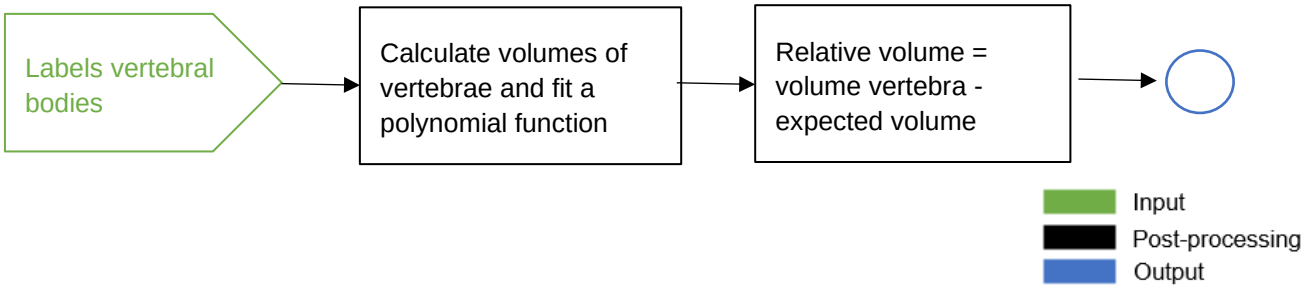

Figure 19: Post-processing workflow feature 'Relative volume vertebra'

A.2.10 Involvement spinal canal

Spinal cord compression is associated with neurological impairment, decreased mobility and worse quality of life[6]. Therefore, it is important to measure the involvement of the tumor in the spinal canal. To achieve this, a segmentation of the spinal canal needs to be obtained. From each label of the vertebrae, a vertebral mask is generated. These masks are dilated and eroded. Thereafter the original masks are subtracted. The result is a segmented volume within the spinal canal (Figure 20).

Figure 20: Coronal slices of the closed segmentation vertebrae (left), original segmentation (middle), and the spinal canal (right)

Of all these volumes, the centers of gravity are calculated and a polynomial function is fit through the points in both the coronal and sagittal planes. Subsequently, the distances from the centers of gravity to the function are calculated and when the distance lays outside the 95%-confidence interval, the points are removed. After the removal of the outliers, a line is fit through the resulting points and dilated to the size of the spinal canal. Finally, the original segmentation of the vertebra is subtracted from the result and small volumes are removed. The final result is shown in Figure 21.

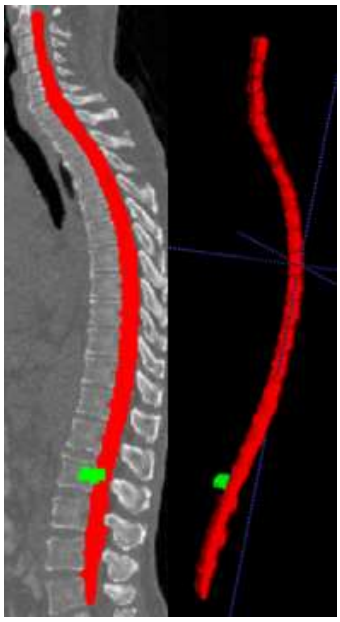

Figure 21: Left) 2D representation of the spinal canal (red) and tumor (green), right) 3D segmentation spinal canal and tumor

With the segmentation of the spinal canal, the involvement within the canal per vertebra can be calculated. If a vertebra contains a metastasis, a bounding box will be created around the vertebra in the sagittal plane through the center of gravity. The functions of the upper and lower sides of this box are utilized to create two cut planes through the spinal canal. The volumes outside the cut planes are removed, and the volume in-between the planes is multiplied with the labels of the metastases. The resulting amount of segmented voxels (mm^3), is the involvement of the tumor in this part of the spinal canal (Figure 22). The final value is normalized between 0 and 100.

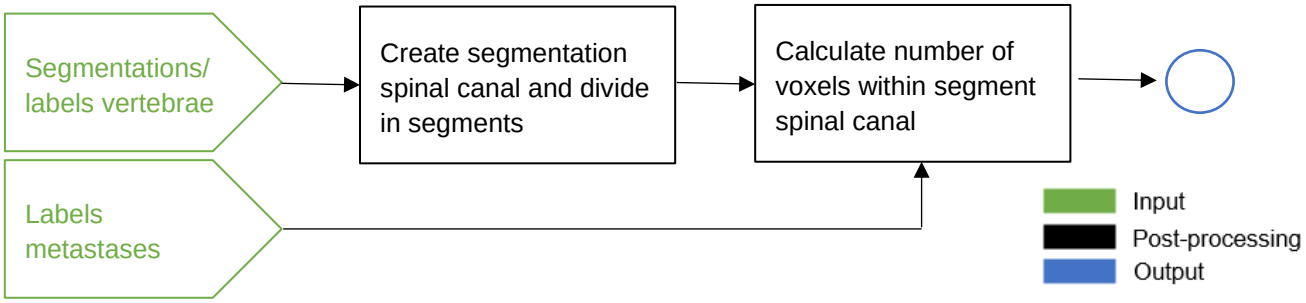

Figure 22: Post-processing workflow feature 'Involvement spinal canal'

A.2.11 Alignment

The feature 'alignment' is based on the SINS component regarding the alignment and more specifically on its subcomponent regarding the deformity of the spine (i.e., kyphosis/scoliosis). Therefore, it is important to calculate the height differences between the ventral and dorsal sides of the vertebral body. To be able to calculate these differences, a sagittal plane is taken through the midpoint of each vertebra. In this 2D sagittal image, the contour of the vertebra is determined and a bounding box is created around the contour (Figure 23).

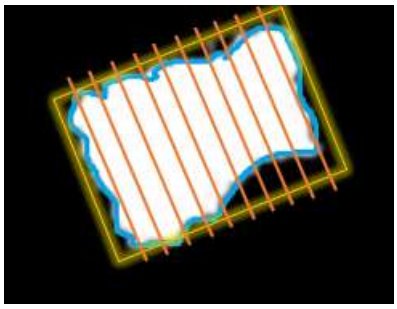

Figure 23: Sagittal slice of the vertebral body with a bounding box (yellow), a contour (blue) and 10 lines parallel to left and right side bounding box (orange)

At ten predefined locations along the bounding box, lines will be drawn parallel to the left and right sides of the box. By calculating the intersection points of these lines with the contour, ten points that represent the upper side and ten points that represent the lower side of the vertebra are generated. Subsequently, coordinates outside the 95%-confidence interval are removed and two lines are fitted through the resulting points. Finally, the ventral and dorsal height of the vertebra can be calculated (Figure 24).

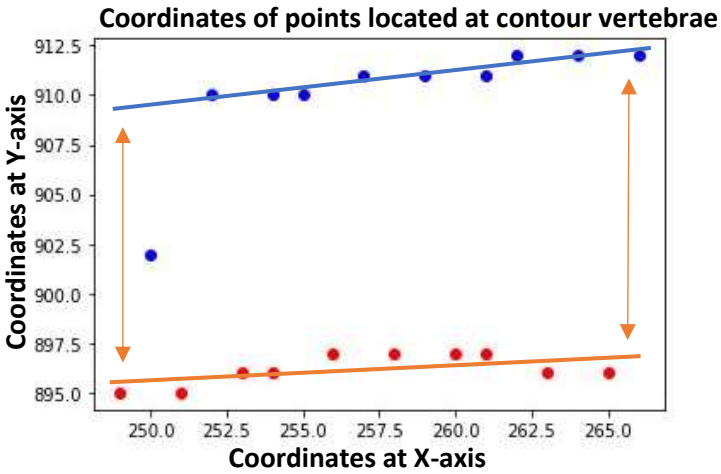

Figure 24: The blue and red dots represent the 10 points along the upper and lower side of the vertebra respectively. Two lines are fitted through the points within the 95%- confidence interval. The arrows show the height of the ventral (left) and dorsal (right) of the vertebra.

Three different approaches are made for this feature. These will be described in the following sections.

A.2.11.1 Alignment1

For the first approach, the absolute difference (mm) is calculated between the ventral and dorsal height of the vertebral body (Figure 25). The final value is normalized between 0 and 100.

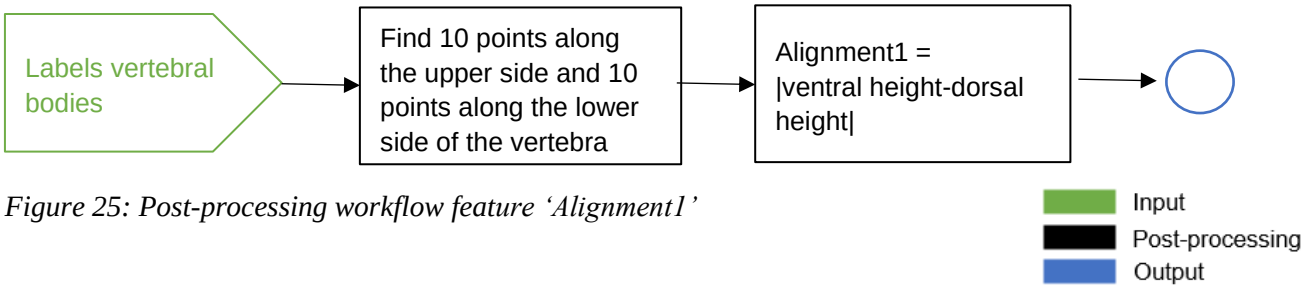

Figure 25: Post-processing workflow feature 'Alignment1'

A.2.11.2 Alignment2

For the second approach, the distance from each vertebra to the line between C7 and L5 (SVA-line) is taken into account together with its location relative to this line. Therefore, first, the line between C7 and L7 (Figure 26[7]) will be drawn, and the distances and relative locations of all vertebrae are defined. Normally, when a vertebra is located at the anterior side of line, the dorsal part of the vertebral body is lower than the ventral part. Therefore, the dorsal height will be subtracted from the ventral height for these vertebrae. When the resulting value is lower than zero, this means the vertebra is collapsed in such a way that the ventral part became lower than the dorsal part. Since it is excepted that these vertebra, with a value lower than zero, affect the spinal stability less, these will be evaluated in a separate feature 'Alignment3' (in section 7.2.11.3). For this feature, these negative values are set to zero. The same applies to the vertebrae located at the dorsal side of the line, however the other way around: The ventral part is subtracted from the dorsal side.

Figure 26: SVA-line

The greater the distance to the SVA-line, the higher the impact of the alignment is on the stability of the spine. Therefore the final result is multiplied by a weight (0-1) depending on the distance to this line (Figure 27). To calculate these weights, the distances from all included vertebra (n=551) to this line are calculated. Min-max normalization is applied to scale all values between 0 and 1. These minimum and maximum values are calculated within the 95%-confidence interval around the mean of the resulting features of all scans. Soft clipping is applied for all values below and above 0 and 1 respectively, utilizing the mathematical expression for a sigmoid. The final value is normalized between 0 and 100.

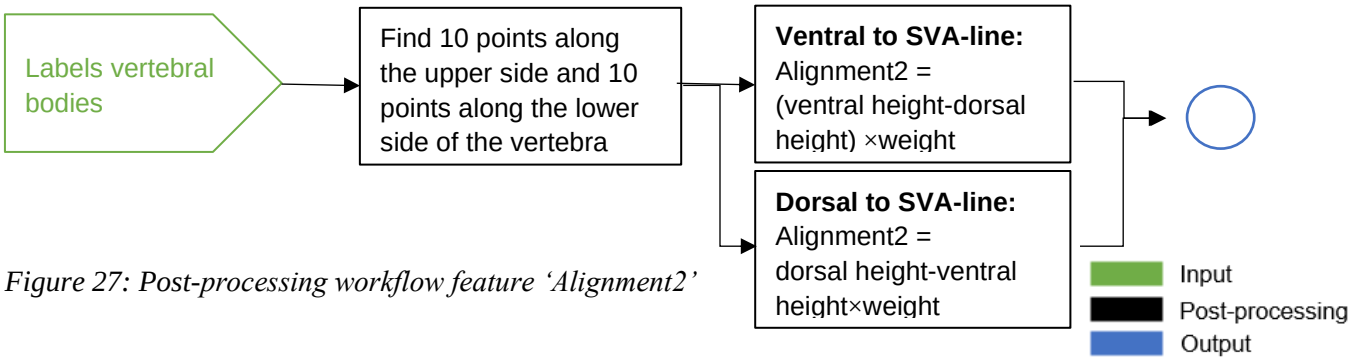

Figure 27: Post-processing workflow feature 'Alignment2'

A.2.11.3 Alignment3

For feature 'Alignment3', the opposite of 'Alignment2' is performed (Figure 28).

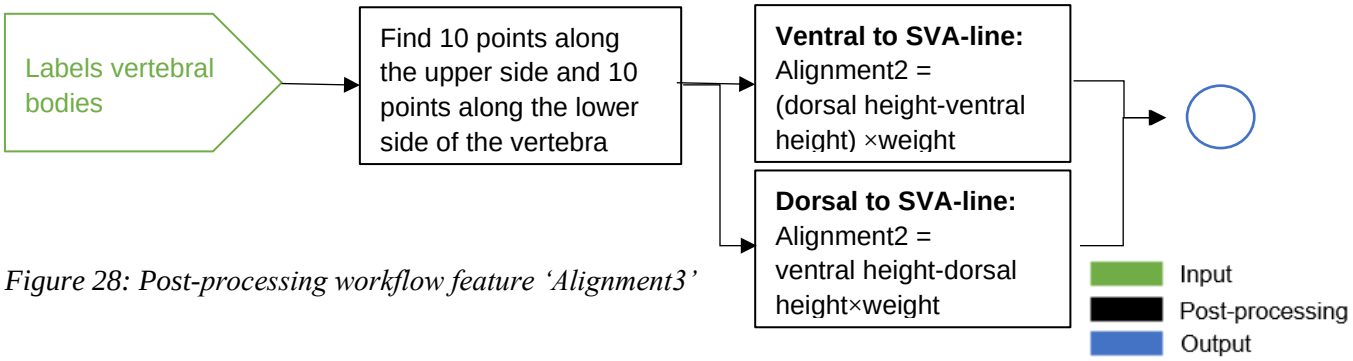

Figure 28: Post-processing workflow feature 'Alignment3'

A.2.12 Location/ Involvement vertebral arch

For the feature 'location vertebral arch' and 'Involvement vertebral arch', there is a need for a segmentation of the vertebral arch. For each vertebra, this segmentation is obtained by subtracting the segmentations of the vertebral bodies from the segmentations of the vertebrae (Figure 29).

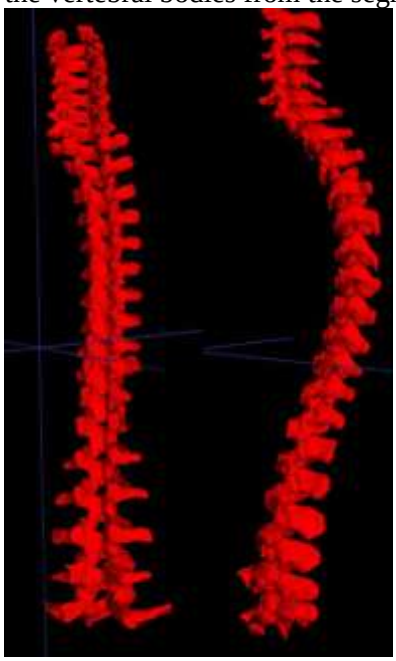

Figure 29: Two 3D segmentations of the vertebral arch

A.2.12.1 Location vertebral arch

To define the location of the metastasis within in the vertebral arch, the resulting segmentation should be divided into several parts (Figure 30). Therefore, first, the center of gravity is calculated twice; once for the original segmentation of the vertebral arch and once for the same segmentation but mirrored in the midsagittal plane. The mean of these values is considered the actual center of the vertebral arch since it is less vulnerable to segmentation mistakes or vertebrae that are not symmetrical. From this center point, one sagittal plane is shifted 4 voxels to the left and one is shifted 4 voxels to the right. These are the cut planes to divide the vertebral arch into three parts.

Figure 30: Coronal slice of the vertebral arch divided in three parts

Since the processes spinous can be wider than 8 voxels, especially for the cervical vertebrae, another method is employed to make sure that all its voxel are classified as the middle part of the vertebral arch. Therefore, for each vertebral arch, the coordinates of the most anterior and posterior voxel are determined. At 3/4th, for the cervical vertebrae, and at 2/3th , for the thoracal and lumbar vertebrae, a coronal cut plane is applied. All voxels at the posterior side of the plane are classified as the middle part of the vertebral arch. Finally, a score will be given, depending on the location of the metastases (Figure 31). If it is located in one of the lateral parts and/or in the middle part, a score of 1 is given. If it is located in both parts and/or the middle part, a score of 3 is given. If a metastases is only located in the middle part, a score of 1 is given. When it is not located in any of the parts, it is scored with 0.

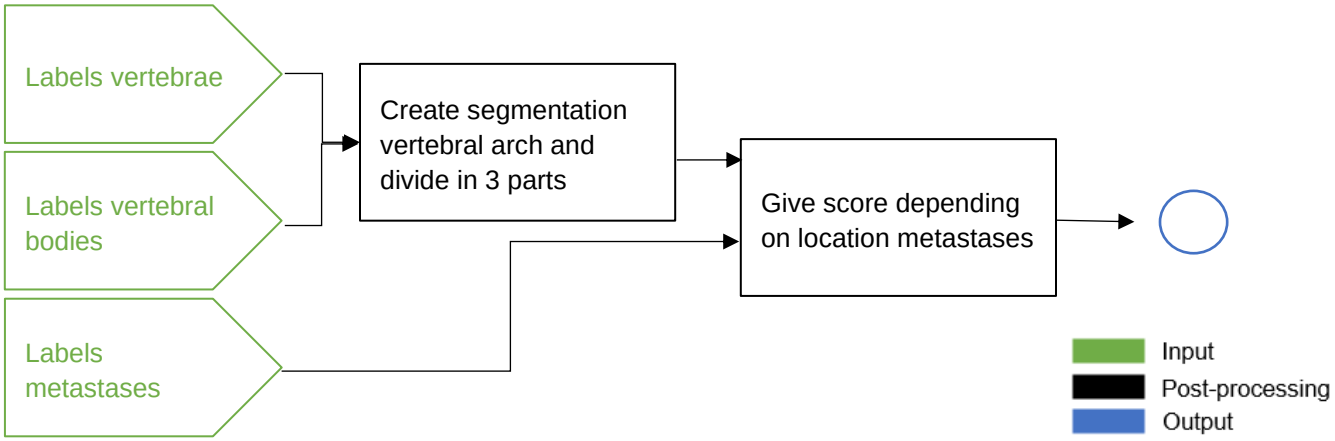

Figure 31: Post-processing workflow feature 'Location vertebral arch'

A.2.12.2 Involvement vertebral arch

For the feature 'Involvement vertebral arch', the segmentation of the vertebral arch is multiplied by the segmentation of the metastases. Subsequently, the volume of the result is calculated, divided by the total volume of the vertebral arch, and multiplied by 100 (Figure 32).

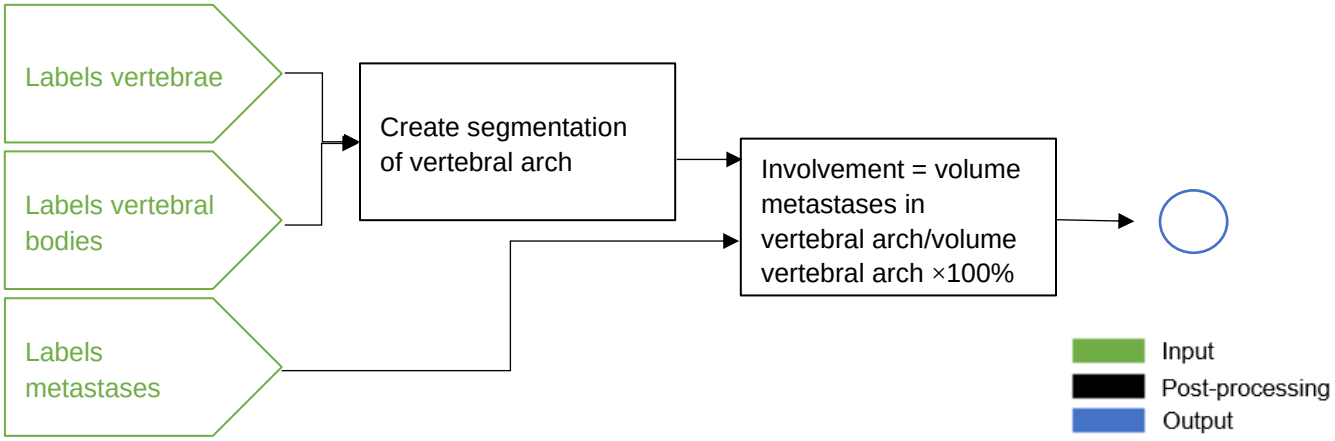

Figure 32: Post-processing workflow feature 'Involvement vertebral arch

A.2.13 References

1. Zijlstra H, Wolterbeek N, Drost RW, et al. Identifying predictive factors for vertebral collapse fractures in multiple myeloma patients. *Spine J*. 2020;20(11):1832-1839. doi:10.1016/J.SPINEE.2020.07.004
2. Vassiliou V, Kalogeropoulou C, Petsas T, Leotsinidis M, Kardamakis D. Clinical and radiological evaluation of patients with lytic, mixed and sclerotic bone metastases from solid tumors: Is there a correlation between clinical status of patients and type of bone metastases? *Clin Exp Metastasis*. 2007;24(1):49-56. doi:10.1007/S10585-007-9056-Z/TABLES/3
3. Tschirhart CE, Finkelstein JA, Whyne CM. Biomechanics of vertebral level, geometry, and transcortical tumors in the metastatic spine. *J Biomech*. 2007;40(1):46-54. doi:10.1016/J.JBIOMECH.2005.11.014
4. Salvatore G, Berton A, Giambini H, et al. Biomechanical effects of metastasis in the osteoporotic lumbar spine: A Finite Element Analysis. *BMC Musculoskelet Disord*. 2018;19(1). doi:10.1186/S12891-018-1953-6
5. Georgy BA. Metastatic Spinal Lesions: State-of-the-Art Treatment Options and Future Trends. *Am J Neuroradiol*. 2008;29(9):1605-1611. doi:10.3174/AJNR.A1137
6. Serratrice N, Faddoul J, Tarabay B, et al. Ten Years After SINS: Role of Surgery and Radiotherapy in the Management of Patients With Vertebral Metastases. *Front Oncol*. 2022;12:67. doi:10.3389/FONC.2022.802595/BIBTEX
7. De Rezende Pratali R, De Oliveira Luz C, Eduardo Gonçales Barsotti C, et al. Analysis of sagittal balance and spinopelvic parameters in a brazilian population sample. doi:10.1590/S1808-18512014130200399

A.3. Results

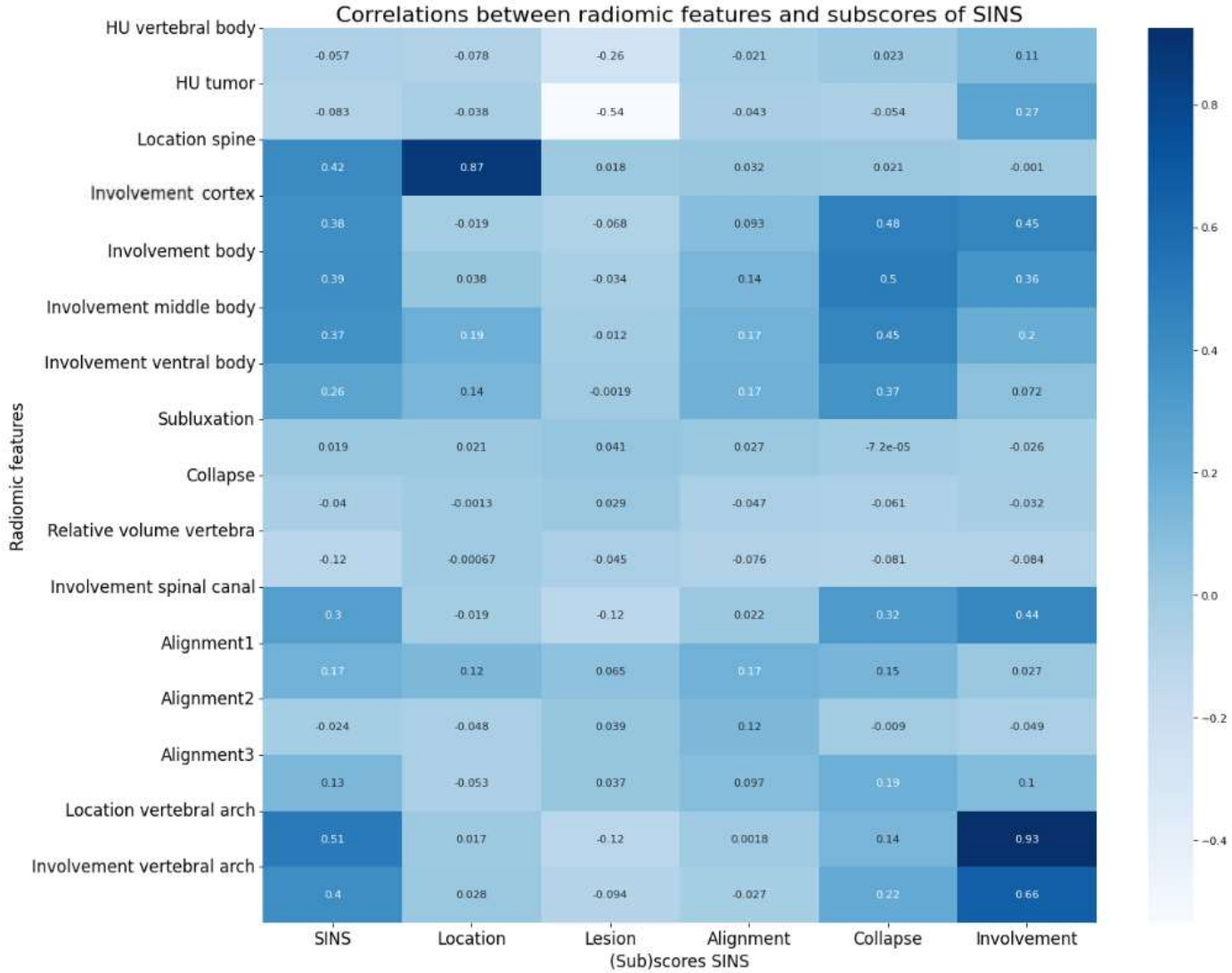

Figure 1: Correlations between features and individual components SINS

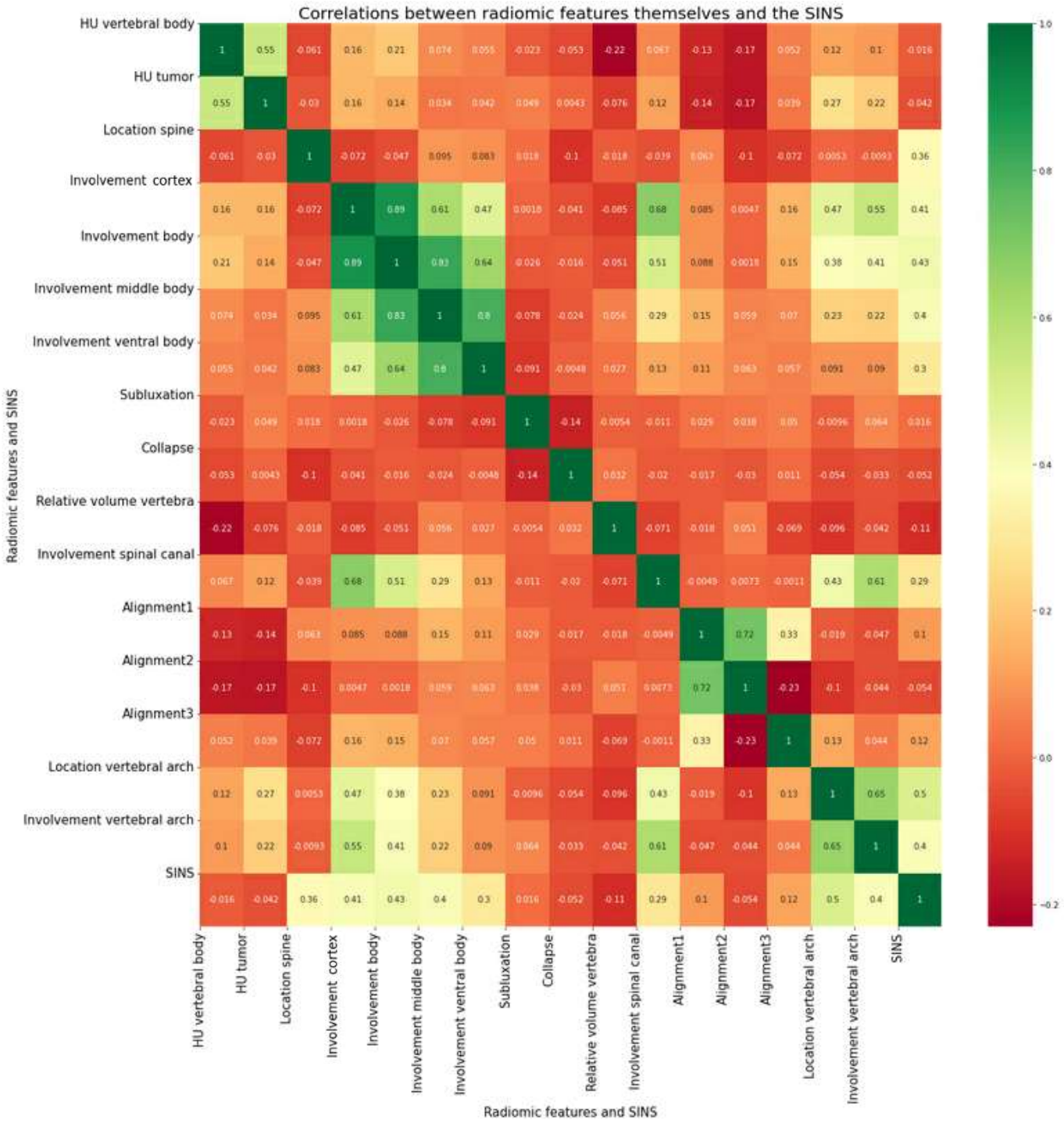

Figure 2: Correlations between features and SINS components

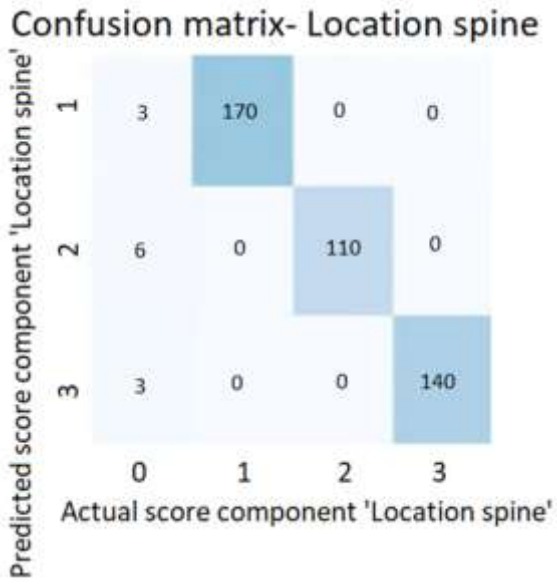

Figure 3: Confusion matrix feature and SINS component regarding the vertebral location

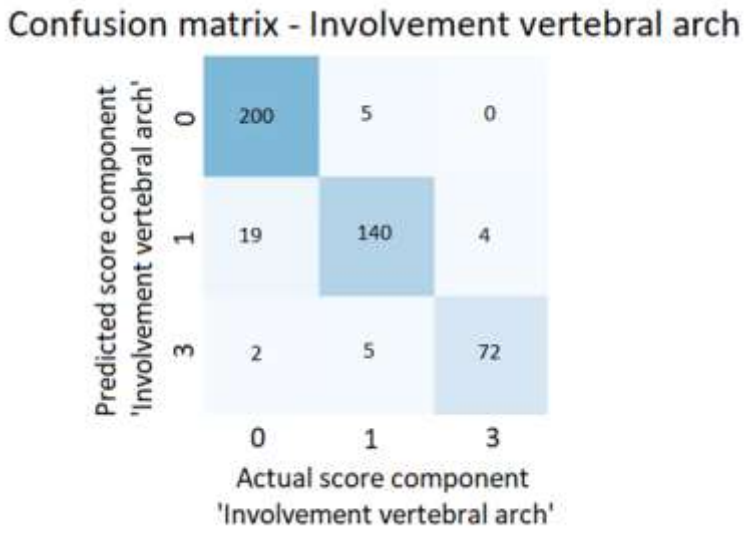

Figure 4: Confusion matrix feature and SINS component regarding the location tumor in vertebral arch

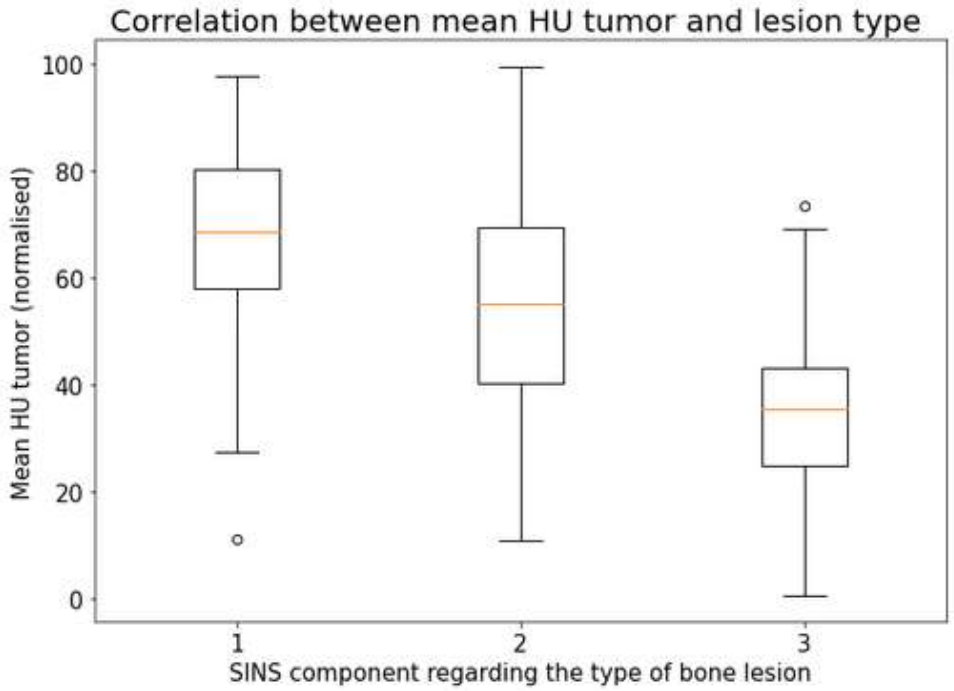

Figure 5: Correlation between mean HU and SINS component regarding the type of bone lesion

Table 1: Errors between actual and predicted SINS calculated for different groups (train set)

Actual scores for individual SINS component	***Mean absolute difference between actual and predicted SINS (95% CI)***
Alignment	
=2	*2.9 (1.7- 4.1)*
=4	*4.5 (3.6, 5.4)*
Collapse	
=1	*1.0 (0.7-1.2)*
=2	*2.1 (1.6- 2.5)*
=3	*3.5 (1.2- 5.8)*
Vertebral arch	
=1	*0.9 (0.7- 1.1)*
=3	*0.9 (0.6- 1.1)*
Bone lesion	
=1	*0.6 (0.5-0.7)*
=2	*1.1 (0.8-1.3)*
Location	
=1	*0.8 (0.7-0.9)*
=2	*0.7 (0.5-0.8)*
=3	*0.9 (0.8-1.1)*

Table 2: Errors between actual and predicted SINS calculated for different groups (train + test set)

Vertebral level	***Actual SINS***	***Predicted SINS***
C6	*7*	*7*
L5	*10*	*6*
L5	*12*	*8*
L5	*8*	*4*

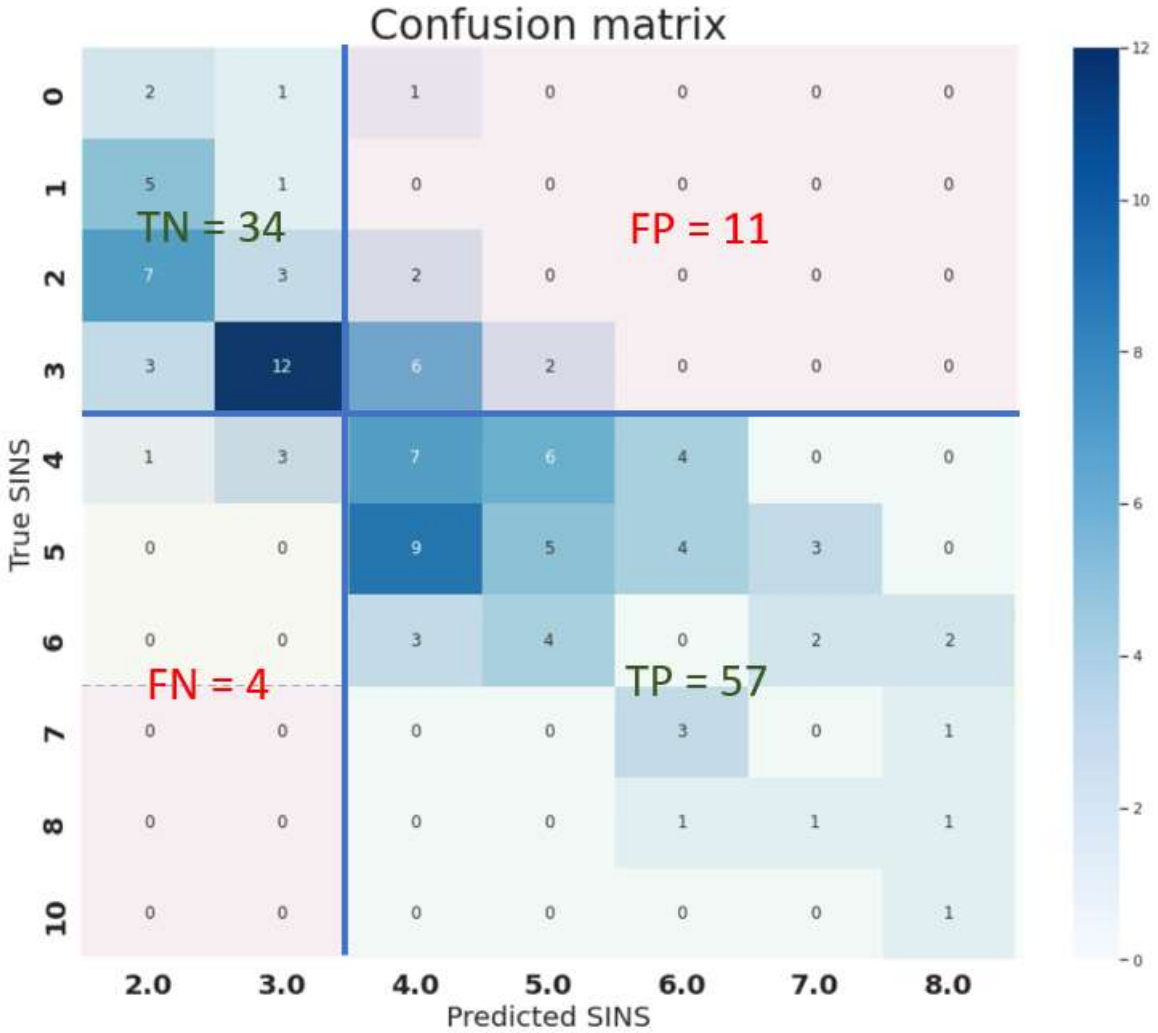

Figure 6: Percentages of vertebra with a specific error between actual and predicted SINS score. The predicted score is subtracted from the actual score (test set).