Context-Aware Transformers For Spinal Cancer Detection and Radiological Grading

Rhydian Windsor[1], Amir Jamaludin[1], Timor Kadir[1,2], and Andrew Zisserman[1]

[1] Visual Geometry Group, Department of Engineering Science, University of Oxford
[2] Plexalis Ltd, Oxford
rhydian@robots.ox.ac.uk

Abstract. This paper proposes a novel transformer-based model architecture for medical imaging problems involving analysis of vertebrae. It considers two applications of such models in MR images: (a) detection of spinal metastases and the related conditions of vertebral fractures and metastatic cord compression, (b) radiological grading of common degenerative changes in intervertebral discs. Our contributions are as follows: (i) We propose a Spinal Context Transformer (SCT), a deep-learning architecture suited for the analysis of repeated anatomical structures in medical imaging such as vertebral bodies (VBs). Unlike previous related methods, SCT considers all VBs as viewed in all available image modalities together, making predictions for each based on context from the rest of the spinal column and all available imaging modalities. (ii) We apply the architecture to a novel and important task – detecting spinal metastases and the related conditions of cord compression and vertebral fractures/collapse from multi-series spinal MR scans. This is done using annotations extracted from free-text radiological reports as opposed to bespoke annotation. However, the resulting model shows strong agreement with vertebral-level bespoke radiologist annotations on the test set. (iii) We also apply SCT to an existing problem – radiological grading of inter-vertebral discs (IVDs) in lumbar MR scans for common degenerative changes. We show that by considering the context of vertebral bodies in the image, SCT improves the accuracy for several gradings compared to previously published models.

Keywords: Metastasis · Vertebral Fracture · Metastatic Cord Compression· Radiological Reports · Radiological Grading · Transformers

1 Introduction

When a radiologist is reporting on a patient's imaging study, they will typically make assessments based on multiple scans of the same patient. For example, in a MRI study they will often consider multiple pulse sequences showing the same region and note differences between them. Related conditions may appear similar in one modality and only be distinguishable when more context is given via an additional imaging modality. Similarly, when reporting on spinal images,

radiologists can learn a lot about a specific vertebra based on its appearance *relative to the vertebrae neighbouring it*. For example, a region of hyperintensity may represent a lesion or may just be an artifact of the scanning protocol used; this often can be elucidated by looking at the rest of the spine.

Accordingly, automated models for spinal imaging tasks should also be able to make vertebra-level predictions with context from multiple sequences *and from neighbouring vertebrae*. In this work, we propose the *Spinal Context Transformer* (SCT), a model which aims to do exactly this. Crucially, SCT leverages lightweight transformer-based models which allow for variable numbers of inputs (both during training and at test-time) in terms of both: (a) the number of vertebral levels, and (b) the number of input modalities depicting each vertebra.

As well as introducing SCT, this paper also explores a new application of deep learning in medical imaging: automated diagnosis of spinal metastases and related conditions. This is an important task for several reasons. Firstly, metastases are very common; overall 2 in 3 cancer cases will metastasise [25] (rising to nearly 100% incidence in patients who die of cancer [20]). The spine is one of the most common places for these metastases to occur [23]. Secondly, early detection is vital; metastases generally indicate advanced cancer and, if left untreated, cause conditions such as vertebral fractures and metastatic cord compression, both of which can lead to significant pain and disability for the patient and are difficult and expensive to treat [28, 29]. Current clinical practice relies on heuristic-based scoring systems to quantify the development of metastases, such as SINS [7] and the Tokuhashi score [27]. We aim to develop models to aid in the rapid detection and consistent quantification of this serious condition. To further validate our model, we also test it on a previously published dataset for grading common intervertebral disc (IVD) degenerative changes in lumbar MR scans.

Another theme explored in this work is the ability of free-text radiological reports to generate supervisory signals for medical imaging tasks. The common praxis when training models on radiological images is to extract a relevant dataset from an imaging centre, pseudo-anonymise it and then annotate the dataset for a condition of interest. While this method is simple, effective and very popular, the additional step of annotating the dataset can be a significant bottleneck in the size of datasets used. This is a more serious problem than in conventional computer vision datasets as medical image annotation generally requires a specialist, whose time is limited and expensive. A much more scalable method of curating annotated datasets is to use existing hospital records for generating supervisory signals. However, in radiology, these records are usually in the form of free-text reports. Due to variation and lack of structure in reporting, extracting useful information from these reports has been considered a difficult task [5], and, as such, most studies annotate their own data retrospectively. In this paper, we explore how useful the information in free-text reports can be for training models for our specific task of detecting spinal cancer, and propose methods for dealing with ambiguities and omissions from the text.

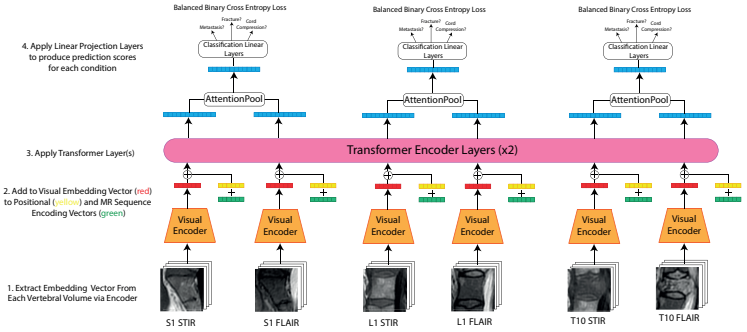
2 Related Work

The SCT, and its training and application, builds on three related areas of work: deep learning for analysis of spinal scans, obtaining supervision from free-text radiology reports and transformers for aggregating multiple sources of information. Automated detection and labelling of vertebrae is a long standing task in spinal MRI [18] and CT [10] analysis, with deep learning now the standard approach [3, 8, 26, 31, 32]. One study particularly relevant to this work is that of Tao *et al.* [26] who use a transformer architecture for this task. Beyond vertebra detection and labelling, deep learning has also been used to assess spinal MRIs for degenerative changes. For example, SpineNet of Jamaludin *et al.* [15] and Windsor *et al.* [33] is a multi-task classification model acting on intervertebral discs which grades for multiple common conditions. We compare to this approach here. DeepSpine of Lu *et al.* [19] and also Lewandrowski *et al.* [17] use ground truth derived from radiological reports to train models for grading stenosis and disc herniation respectively, a theme we also explore in this work. In terms of spinal cancer, [34] and [30] both train models to detect metastases based on 2D images extracted from MRI studies; and [22] proposes a model to automatically detect metastatic cord compression, although this is restricted to axial T2 scans or the cervical region. Several other works have considered vertebra and metastases detection in other modalities such as CT (e.g. [2, 10, 11, 12]), although MR remains the clinical gold-standard in early spinal cancer detection. Finally, our approach also has parallels with works in video sequence analysis that combine a 2D CNN backbone with a temporal transformer for tasks such as representation learning, tracking, object detection and segmentation [1, 4, 9, 21]; though in our case, visual features are extracted from multiple vertebrae shown in multiple MR sequences rather than consecutive frames.

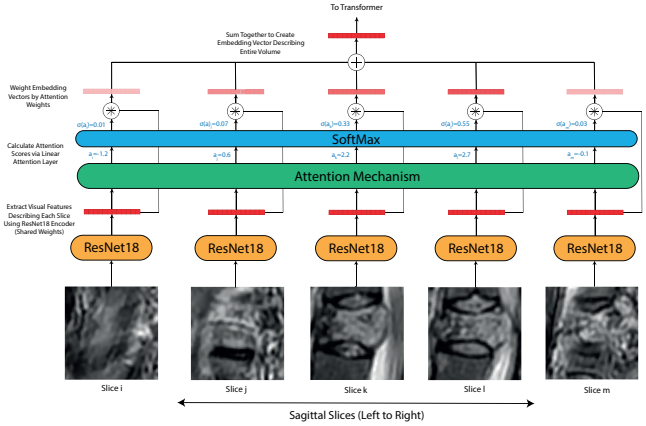
3 Spinal Context Transformer

Like other models used to assess vertebral disorders such as SpineNet [15,33] and DeepSpine [19], the Spinal Context Transformer (SCT) exploits the repetitive nature of the spine by using the same CNN model to extract visual features from each vertebra given as input. However, there are two key conceptual differences between how SCT and these other models operate which are described in this section.

Firstly, instead of using 3D convolutions and pooling layers to extract features from multislice MR images, SCT instead extracts features from each slice independently using a standard 2D ResNet18 [13] architecture adapted for single-channel images. The extracted feature vectors for each slice are then aggregated using an attention mechanism. This is shown in Figure 1b. This design choice is made for several reasons: (i) 3D convolutions assume all images have the same slice thickness. In real-world clinical practice a variety of scanning protocols are used. This means slice thickness can vary significantly from sample to sample. Of course, resampling can be employed in pre-processing to ensure a consistent slice



(a) The SCT architecture illustrated for the spine cancer prediction task operating on multiple MRI sequences of three vertebrae. In the actual cancer detection task, the whole spinal column is used, and SCT is able to handle arbitrary variations in the number and type of input sequences and vertebrae without adaptation.



(b) The visual encoder network used to extract features from each vertebral volume, This model can be trained by itself or in conjunction with the transformer architecture as shown in (a). The same model weights are used for all MR sequences.

Fig. 1: Architecture diagrams for SCT: (a) the full SCT architecture; (b) the encoder used to extract visual features from each vertebra. STIR and FLAIR sequences are shown here, however the radiological grading task uses strictly T1 and T2 sequences, whereas the spinal cancer task uses a range of common sequences (T1, T2, STIR, FLAIR etc).

thickness for input images, however this introduces visual artifacts which may degrade performance. (ii) Trained encoder models can operate on 2D or 3D scans without adaptation. This also means image CNNs trained on large 2D datasets such as ImageNet [6] can be used to initialize the encoder backbone. (iii) The attention mechanism forces the model to explicitly determine the importance of each slice for its prediction. This allows peripheral slices with partial-volume effects to be ignored and can be useful when it comes to interpreting the output of the model. An example of this effect is shown in the appendix. Secondly, as well

as using attention to aggregate feature vectors from each slice, SCT also uses attention to collate information from each MR sequence and vertebra in the spinal column. This is done by feeding the vertebra's visual embedding vectors from each MR sequence to a lightweight 2-layer transformer encoding model, with additional embedding vectors describing the level of each input vertebra and the imaging modality used. Both these additional embeddings are calculated by linear layers operating on a one-hot encoding of the vertebra's name and sequence. The output feature vectors for the same vertebra shown in different sequences are then pooled together using a final attention mechanism identical to that used to pool features across slices. This creates a single output vector for each vertebra. Linear classification layers convert this output vector into predictions for each classification task. The full process is shown in Figure 1a. Annotated PyTorch-style pseudo-code is given in the appendix.

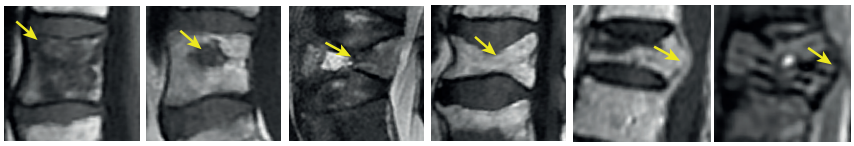
4 Detecting Spine Cancer Using Radiological Reports

This section describes applying SCT to a novel task: detecting the presence of spinal metastases in whole spine clinical MR scans, as well as the related conditions of vertebral fractures/collapse and spinal cord compression, employing information from free-text radiological reports for supervision. We use a dataset of anonymised clinical MRI scans and associated reports extracted from a local hospital trust PACS (Oxford University Hospitals Trust) with appropriate ethical clearance. Inclusion criteria were that the patient was over 18 years old, had at least one whole spine study, and was referred from a cancer-related specialty. In order to ensure a mix of positive and negative cases, all patients that matched this criteria from April 2015 until April 2021 were included in the dataset, regardless of whether or not they had metastases.

To avoid having to annotate each scan independently we instead rely on the radiological reports written at the time the scan was taken to provide supervision for training our models. A major advantage of this approach is that it is much faster and thus readily scalable to larger datasets – since no radiologist is required to review each image independently, an annotation set can be obtained much quicker than would otherwise be possible. Furthermore, significantly less clinical expertise is required to generate these annotations – only a basic understanding of vocabulary related to the conditions of interest is needed.

A particular challenge of using annotations derived from reports for training is dealing with ambiguities in the text. Firstly, reports are often inconclusive; e.g. "The nature of this lesion is indeterminate" or "Metastases are a possibility but further investigation is required". In these cases we label vertebrae as 'unknown' for a specific condition and do not apply a loss to the model's corresponding predictions. Note if one vertebra is marked as positive in a spine and the other levels are ambiguous, then the others are marked as 'unknown'. This accounts for cases where a specific metastasis/fracture is being commented on due to a change from a previous scan but other unchanged metastases/fractures are not fully described. Cord compression, on the other hand, is marked as negative unless

explicitly stated otherwise, as it is a severe condition which is highly unlikely to go unmentioned in a report. Secondly, reports often state that "metastatic disease is widespread" throughout the spine but do not indicate at which specific levels this disease occurs. To derive annotations from these ambiguous reports, we introduce an additional, global label for each study, indicating if the given condition is present in the scan at any vertebral level or not. We then aggregate each vertebra-level prediction to produce a spinal-column level prediction of whether the condition is present. This setting is an example of multiple-instance learning (MIL), whereby instead of each training instance being individually labelled, sets of instances (known as 'bags') are assigned a single label, indicating if at least one instance in the bag is positive. The challenge is then to determine which instances in the bag are positives. In this case each bag represents an entire spinal column and each instance is a single vertebra. We can then train by a hybrid single-instance and multiple-instance learning approach; vertebra-level annotations are used for supervision where given by the report in addition to the scan-level annotations given in each case. A breakdown of the labels extracted by this method is given in Table 1. Examples of vertebral bodies from each class are shown in Figure 2. Further explanation and example annotations are given in the appendix. To ensure our model produces results similar to those given by bespoke annotations, our test dataset is labelled by an expert spinal surgeon.



(a) Metastases (b) Collapsed Vertebrae (c) Cord Compression

Fig. 2: Example vertebrae with the three spinal-cancer related conditions of interest. Note that in many cases vertebrae have a combination of these conditions.

Split	Patients	Studies	Total	Vertebral Bodies								
				Metastases			Fractures			Compression		
				+	−	?	+	−	?	+	−	?
Training	258	558	12459	829	6099	5531	216	12124	119	90	12369	0
Validation	33	63	1391	81	829	482	25	1360	7	15	1377	0
Test (Reports)	32	66	1430	92	646	692	22	1408	0	11	1419	0
Test (Expert)	32	66	1430	374	1056	0	46	1384	0	31	1399	0

Table 1: The dataset used to train SCT to detect metastases, fractures and compression. The first-three rows indicate the number of labels extracted from free-text reports for the training, validation and test splits. The bottom row indicates the same test dataset with each vertebra independently annotated by an expert. Vertebra are labelled as positive (+), negative (−), or unknown (?). Note the expert annotator was required to make their best guess in cases of uncertainty.

5 Radiological Grading Of Spinal Degenerative Changes

To validate SCT on an existing vertebra analysis problem, we also measure its performance at grading common degenerative changes around the intervertebral discs. Specifically, we grade the following common conditions: Pfirrmann Grading, Disc Narrowing, Endplate Defects (Upper/Lower), Marrow Changes (Upper/Lower), Spondylolisthesis and Central Canal Stenosis. We do this on Genodisc, a dataset of clinical lumbar MRIs of 2295 patients from 6 different clinical imaging sites. This problem setting closely follows that discussed in [15] and [33], which give the existing state-of-the-art methods for many of these grading tasks. Each vertebral disc from L5/S1 to T12/L1 in each scan is graded by an expert radiologist for the aforementioned gradings. In this setting, SCT takes as input volumes surrounding intervertebral discs (IVD) and outputs predictions for each grading. The study protocol for scans vary, though the vast majority of scans have at least a T1-weighted and T2-weighted sagittal scan. In this work we only consider sagittally sliced scans. However, there is no reason why SCT could not also operate on axial scans in conjunction. A more complete breakdown of the Genodisc dataset is given in [14]. We compare to SpineNet V1 [15] and SpineNet V2 [33], existing models trained on the same dataset.

6 Experimental Results

This section describes experiments to evaluate the performance of SCT at the tasks of: (a) detecting spinal cancer; (b) grading common degenerative changes.

Preprocessing, Implementation & Training Details: For both tasks, vertebrae are detected and labelled in the scans using the automated method described in [33]. This is then used to extract volumes surrounding VBs (for the cancer task) and IVDs (for the grading task). Volumes are resampled to $S \times 112 \times 112$ and $S \times 112 \times 224$ respectively, where S is the number of sagittal slices. Additional training details are given in the appendix.

Baselines: For each task, we compare SCT to baseline models operating on a single vertebra/IVD at a time. For the cancer task we compare to the visual encoder shown in Figure 1b operating alone on a single sequence (since this cannot be trivially extended to multiple sequences). For the radiological grading task we compare with SpineNet V1 [15], SpineNet V2 [33] and the visual encoder alone. We also train SCT on T1 and T2 scans independently.

Results: Table 2 shows the performance of all models at the cancer task. The SCT model performs well across all tasks. In particular, it performs much better at detecting subtle metastases (AUC: **0.80→0.931** for the expert labels). This effect can be seen clearly in Figure 3. This makes sense as patients will often present with multiple metastases and thus obvious metastases in one area of the spine will inform predictions on marginal cases elsewhere in the spine. We note

slightly depreciated performance at the compression task. We believe this is due to overfitting as there are relatively few compression cases in the training set.

Table 3 shows the performance of all models at the radiological grading tasks, and compares to the state-of-the-art SpineNet model [14]. SCT outperforms SpineNet V1 & V2 on the same dataset for all tasks except central canal stenosis where the difference in performance is minimal . In total average performance increases from **85.9%**→**87.4**%. As expected, the multiple sequence model exceeds the performance of single sequence models in most tasks. The largest improvements can be seen in the endplate defect task (**82.9**/**87.8%** →**87.2**/**90.7%**), disc narrowing (**76.1%**→**77.4%**) and Pfirrmann grading (**71.0%** →**73.0%**). The Pfirrmann grading score of the T1-only model is far worse than the models with T2 sequences (64.5%). This matches expectations as one of the criteria of Pfirrmann grading is the intensity of the intervertebral discs in T2 sequences [24]. Overall, the improved performance is a clear indication of the benefit of considering context from multiple sequences and IVDs together.

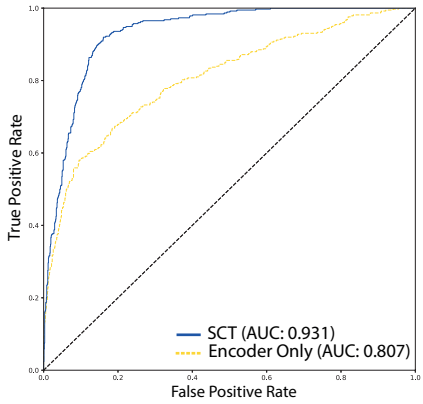


Fig. 3: ROC curves for the metastases task for SCT and the encoder baseline. Curves for the other tasks are given in the appendix.

Table 2: AUC scores for the three tasks with various models. We compare to the report-extracted annotations and also expert annotations of each image.

Expert-Labelled Test Set Annotations

	ROC AUC		
	Mets	Frac.	Cmprs.
Baseline	0.80	0.975	**0.930**
SCT	**0.931**	**0.980**	0.868

Report-Extracted Test Set Annotations

	ROC AUC		
	Mets	Frac.	Cmprs.
Baseline	0.934	0.902	**0.955**
SCT	**0.944**	**0.901**	0.918

Model	# Classes:	Pfirrmann 5	Disc Narrowing 4	C.C.S. 2	Spondylolisthesis 2
SpineNet V1 (T2) [16]		71.0	76.1	**95.8**	95.4
SpineNet V2 (T2) [33]		70.9	76.3	93.2†	95.0
Baseline: SCT Encoder (T2)		70.8	74.6	94.3	96.2
SCT (T1)		64.5	73.9	93.6	**96.6**
SCT (T2)		71.7	76.9	93.6	96.2
SCT (T1,T2)		**73.0**	**77.4**	94.9	95.5

		Endplate Defect		Marrow Change		**Average**
		Upper	Lower	Upper	Lower	
Model	# Classes:	2	2	2	2	
SpineNet V1 (T2) [16]		82.9	87.8	89.2	88.4	85.8
SpineNet V2 (T2) [33]		84.9	89.8	88.9	88.2	85.9
Baseline: SCT Encoder (T2)		87.2	88.2	90.0	89.2	86.3
SCT (T1)		84.7	87.0	88.4	88.7	84.7
SCT (T2)		85.3	89.3	**91.0**	**90.2**	86.8
SCT (T1,T2)		**87.2**	**90.7**	90.1	89.9	**87.4**

Table 3: Results of the IVD grading task. The balanced accuracy for each subtask is shown. CCS represents central canal stenosis. †For SpineNetV2, CCS is originally graded with 4 degrees of severity. We combine the severity classes 2-4 (mild, moderete & severe CCS) into a single class compared to class 1 (no CCS).

7 Conclusion

In this paper we present SpinalContextTransformer (SCT) for the analysis of multiple vertebrae and multiple MRI sequences. We demonstrate SCT being applied to a new task, detecting conditions related to spinal cancer including metastases, vertebral collapses and metastatic cord compression. We also show that the model improves on existing models for radiological grading of common spinal conditions and can be used flexibly with varying input imaging modalities. Finally, it is worth noting that the SCT architecture is applicable to the analysis of other modalities (e.g. CT, X-Ray), fields-of-view (e.g. coronal, axial) and to other repeated anatomical sequential structures, such as teeth or ribs.

Acknowledgements and Ethics: Ethics for the spinal cancer dataset extraction are provided by OSCLMRIC (IRAS project ID: 207857). We are grateful to Dr. Sarim Ather, Dr. Jill Urban and Prof. Jeremy Fairbank for insightful conversations on the clinical aspect of this work as well as Prof. Ian McCall for annotating the data. Finally, we thank to our funders: Cancer Research UK via the EPSRC AIMS CDT and EPSRC Programme Grant Visual AI (EP/T025872/1).

References

1. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: IEEE International Conference on Computer Vision (2021)

2. Burns, J.E., Yao, J., Wiese, T.S., Muñoz, H.E., Jones, E.C., Summers, R.M.: Automated detection of sclerotic metastases in the thoracolumbar spine at CT. Radiology **268**(1), 69–78 (2013)
3. Cai, Y., Osman, S., Sharma, M., Landis, M., Li, S.: Multi-Modality Vertebra Recognition in Arbitrary Views Using 3D Deformable Hierarchical Model. IEEE Transactions on Medical Imaging **34**(8), 1676–1693 (2015)
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Proc. ECCV. pp. 213–229 (2020)
5. Casey, A., Davidson, E., Poon, M., Dong, H., Duma, D., Grivas, A., Grover, C., Suárez-Paniagua, V., Tobin, R., Whiteley, W., Wu, H., Alex, B.: A systematic review of natural language processing applied to radiology reports. BMC Medical Informatics and Decision Making **21**(1) (2021)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proc. CVPR (2009)
7. Fisher, C.G., DiPaola, C.P., Ryken, T.C., Bilsky, M.H., Shaffrey, C.I., Berven, S.H., Harrop, J.S., Fehlings, M.G., Boriani, S., Chou, D., Schmidt, M.H., Polly, D.W., Biagini, R., Burch, S., Dekutoski, M.B., Ganju, A., Gerszten, P.C., Gokaslan, Z.L., Groff, M.W., Liebsch, N.J., Mendel, E., Okuno, S.H., Patel, S., Rhines, L.D., Rose, P.S., Sciubba, D.M., Sundaresan, N., Tomita, K., Varga, P.P., Vialle, L.R., Vrionis, F.D., Yamada, Y., Fourney, D.R.: A Novel Classification System for Spinal Instability in Neoplastic Disease: An Evidence-Based Approach and Expert Consensus From the Spine Oncology Study Group. Spine **35**(22), E1221–E1229 (Oct 2010)
8. Forsberg, D., Sjöblom, E., Sunshine, J.L.: Detection and Labeling of Vertebrae in MR Images Using Deep Learning with Clinical Annotations as Training Data. Journal of Digital Imaging **30**(4), 406–412 (2017)
9. Gabeur, V., Sun, C., Alahari, K., Schmid, C.: Multi-modal Transformer for Video Retrieval. In: European Conference on Computer Vision (ECCV) (2020)
10. Glocker, B., Feulner, J., Criminisi, A., Haynor, D.R., Konukoglu, E.: Automatic localization and identification of vertebrae in arbitrary field-of-view CT scans. In: Proc. MICCAI. pp. 590–598 (2012)
11. Glocker, B., Zikic, D., Konukoglu, E., Haynor, D.R., Criminisi, A.: Vertebrae localization in pathological spine CT via dense classification from sparse annotations. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013. pp. 262–270 (2013)
12. Hammon, M., Dankerl, P., Tsymbal, A., Wels, M., Kelm, M., May, M., Suehling, M., Uder, M., Cavallaro, A.: Automatic detection of lytic and blastic thoracolumbar spine metastases on computed tomography. European Radiology **23**(7), 1862–1870 (2013)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. Proc. CVPR (2016)
14. Jamaludin, A., Kadir, T., Zisserman, A.: Self-Supervised Learning for Spinal MRIs. In: MICCAI Workshop on Deep Learning in Medical Image Analysis (2017)
15. Jamaludin, A., Kadir, T., Zisserman, A.: SpineNet: Automated classification and evidence visualization in spinal MRIs. Medical Image Analysis **41**, 63–73 (2017)
16. Jamaludin, A., Lootus, M., Kadir, T., Zisserman, A., Urban, J., Battié, M.C., Fairbank, J., McCall, I.: Automation of reading of radiological features from magnetic resonance images (mris) of the lumbar spine without human intervention is comparable with an expert radiologist. European Spine Journal (2017)
17. Lewandrowskl, K.U., Muraleedharan, N., Eddy, S.A., Sobti, V., Reece, B.D., ramírez León, J.F., Shah, S.: Feasibility of deep learning algorithms for report-

ing in routine spine magnetic resonance imaging. International Journal of Spine Surgery **14** (2022)
18. Lootus, M., Kadir, T., Zisserman, A.: Vertebrae detection and labelling in lumbar mr images. In: MICCAI Workshop: Computational Methods and Clinical Applications for Spine Imaging (2013)
19. Lu, J.T., Pedemonte, S., Bizzo, B., Doyle, S., Andriole, K.P., Michalski, M.H., Gonzalez, R.G., Pomerantz, S.R.: Deep spine: Automated lumbar vertebral segmentation, disc-level designation, and spinal stenosis grading using deep learning. In: Machine Learning for Healthcare (2018)
20. Maccauro, G., Spinelli, M.S., Mauro, S., Perisano, C., Graci, C., Rosa, M.A.: Physiopathology of Spine Metastasis. International Journal of Surgical Oncology **2011** (2011)
21. Meinhardt, T., Kirillov, A., Leal-Taixe, L., Feichtenhofer, C.: Trackformer: Multi-object tracking with transformers. In: Proc. CVPR (2022)
22. Merali, Z., Wang, J.Z., Badhiwala, J.H., Witiw, C.D., Wilson, J.R., Fehlings, M.G.: A deep learning model for detection of cervical spinal cord compression in MRI scans. Scientific Reports **11**(1) (2021)
23. Ortiz Gomez, J.: The incidence of vertebral body metastases. International Orthopaedics **19**(5) (1995)
24. Pfirrmann, C.W.A., Metzdorf, A., Zanetti, M., Hodler, J., Boos, N.: Magnetic resonance classification of lumbar intervertebral disc degeneration:. Spine **26**(17) (2001)
25. Shaw, B., Mansfield, F.L., Borges, L.: One-stage posterolateral decompression and stabilization for primary and metastatic vertebral tumors in the thoracic and lumbar spine. Journal of Neurosurgery **70**(3), 405–410 (1989)
26. Tao, R., Zheng, G.: Spine-transformers: Vertebra detection and localization in arbitrary field-of-view spine CT with transformers. In: Proc. MICCAI (2021)
27. Tokuhashi, Y., Uei, H., Oshima, M., Ajiro, Y.: Scoring system for prediction of metastatic spine tumor prognosis. World Journal of Orthopedics **5**(3), 262–271 (2014)
28. van Tol, F.R., Massier, J.R.A., Frederix, G.W.J., Öner, F.C., Verkooijen, H.M., Verlaan, J.J.: Costs Associated With Timely and Delayed Surgical Treatment of Spinal Metastases. Global Spine Journal (2021)
29. van Tol, F.R., Versteeg, A.L., Verkooijen, H.M., Öner, F.C., Verlaan, J.J.: Time to Surgical Treatment for Metastatic Spinal Disease: Identification of Delay Intervals. Global Spine Journal (2021)
30. Wang, J., Fang, Z., Lang, N., Yuan, H., Su, M.Y., Baldi, P.: A multi-resolution approach for spinal metastasis detection using deep Siamese neural networks. Computers in biology and medicine **84**, 137–146 (2017)
31. Windsor, R., Jamaludin, A.: The ladder algorithm: Finding repetitive structures in medical images by induction. In: IEEE ISBI (2020)
32. Windsor, R., Jamaludin, A., Kadir, T., Zisserman, A.: A convolutional approach to vertebrae detection and labelling in whole spine MRI. In: Proc. MICCAI (2020)
33. Windsor, R., Jamaludin, A., Kadir, T., Zisserman, A.: SpineNetV2: Automated detection, labelling and radiological grading of clinical MR scans. In: Technical Report, arXiv:2205.01683 (2022)
34. Zhao, S., Chen, B., Chang, H., Wu, X., Li, S.: Discriminative dictionary-embedded network for comprehensive vertebrae tumor diagnosis. In: Proc. MICCAI (2020)

8 Appendix

8.1 Pytorch-Style Pseudocode for SCT

```
class SpinalContextTransformer():
    def initialize(self, n_sequences, n_vertebrae):
        self.visual_encoder = VisualEncoderModel() # extracts features from each vertebra
        self.transformer    = TransfomerEncoderLayers(n_layers=2) # transforms features using context
        self.E              = 128 # embedding/token size for transformer
        # embedders convert one-hot encoded vert level/positions and type MR sequences
        # into something that can be summed with the features extracted from each vertebra
        self.positional_embedder = LinearLayer(in_dim=n_sequences, out_dim=self.E)
        self.sequence_embedder   = LinearLayer(in_dim=n_vertebrae, out_dim=self.E)

        self.sequence_attention = LinearLayer(in_dim=self.E, out_dim=self.E)
        # output classification layers
        self.met_classifier         = LinearLayer(in_dim=self.E, out_dim=1)
        self.fracture_classifier    = LinearLayer(in_dim=self.E, out_dim=1)
        self.compression_classifier = LinearLayer(in_dim=self.E, out_dim=1)

    def forward(self, vert_vols, one_hot_position_encoding, one_hot_sequence_encoding):
        B, N, C, S, H, W = vert_vols.shape
        B, N, C, n_sequences = one_hot_sequence_encoding.shape
        B, N, C, n_vertebrae = one_hot_position_encoding.shape

        # B=Batch size, N=Num Vertebrae, C=Num Channels/MR Sequences,
        # S=Number of Slices, H=Height, W=Width
        vert_vols = vert_vols.flatten(start_dim=0, end_dim=2)
        # [B,N,C,S,H,W] -> [B*N*C,S,H,W]
        vert_features = self.visual_encoder(vert_vols)
        # vert_features has dimension [B*N*C,E] where E is the embedding size (E=128 here)
        vert_features        = vert_features.reshape(B,N,C,S,self.E)
        positional_embedding = self.positional_embedder(one_hot_position_encoding)
        sequence_embedding   = self.sequence_embedder(one_hot_sequence_encoding)
        # add in positional and sequence embedding
        input_tokens  = vert_features + positional_embedding + sequence_embedding
        # used transformer to consider context between all vertebrae and sequences
        output_tokens = self.transformer(input_tokens)
        output_tokens = output_tokens.reshape(B,N,C,self.E)
        # now pool along MR sequence dimension (C)
        sequence_attention_scores = self.sequence_attention(output_tokens)
        # softmax across sequence dimension
        sequence_attention_weights = SoftMax(sequence_attention_scores,dim=2)
        pooled_output_tokens = (output_tokens*sequence_attention_weights).sum(dim=2)
        # pooled_output_tokens has dimension [B,N,E].
        # Now apply lineaer classifiers to get output of shape [B,N,1] for each task
        met_score      = self.met_classifier(pooled_output_tokens)
        fracture_score = self.fracture_classifier(pooled_output_tokens)
        mscc_score     = self.compression_classifier(pooled_output_tokens)
        return met_score, fracture_score, mscc_score
```

8.2 Additional Model Training Details

All models are trained until 10 successive validation epochs do not result in a decreased loss. An Adam optimizer is used with a learning rate of 10^{-4} and $\beta = (0.9, 0.999)$. A batch size of 20 is used for the lumbar-only degenerative changes task and 6 for the whole-spine cancer task (each sample consists of each VB in each MR sequence the scan). In total training for both tasks consumes around 40GB of GPU memory and takes approximately 6 hours using 2 Tesla P40s. The classification linear layers are then finetuned for each task for 10 epochs with the encoder and transformer weights frozen. Standard augmentations for each vertebra are used including rotation ($\pm 15°$), translation ($\pm$32px), scaling ($\pm 10\%$) and intensity augmentation ($\pm 10\%$). For the multi-sequence models trained on T1 and T2 sequences, a single sequence or both sequences are dropped during training with probabilities of 0.4 and 0.1 respectively, leaving the model make predictions for a volume it has not seen based on context alone. 50% dropout is used for the transformer layers. Pytorch 1.10 was used to implement all models.

8.3 Slicewise Predictions

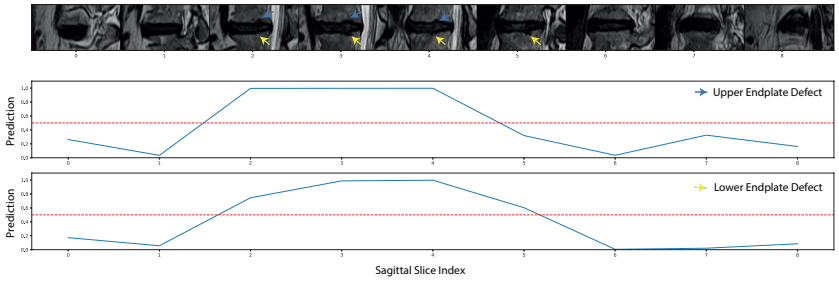


Fig. 4: Slicewise outputs for SCT operating on an IVD (top row) with upper (blue arrows) and lower endplate defects (yellow arrows). The second and third rows show the predictions for upper and lower endplate defects respectively when each sagittal slice of the scan is fed into the model independently. SCT is highly accurate at detecting pathology even at the single slice level, detecting the upper endplate defects in slices 2,3 & 4 and the lower endplate is slices 2,3,4 & 5. This can considered a form of attribution. The dotted red line shows $p(\text{endplate defect}) = 50\%$ for both slicewise predictions.

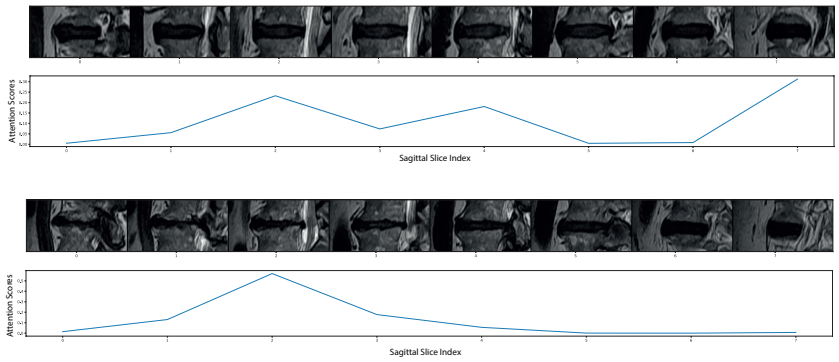


Fig. 5: Slicewise attention values for a healthy (top) and an unhealthy (bottom) IVD. In the healthy IVD, attention is spread across all mid-sagital slices whereas in the unhealthy vertebra, the attention is concentrated on slices which clearly show pathology (endplate defects, narrowing and slight stenosis in slices 1,2 & 3).

8.4 Example Annotation

Report:
Clinical Information: kidney cancer on follow up. CT scan shows stable disease. MRI spine to stage spine ? further progression

MRI Spine Whole : comparison is made with the MR scan dated [date redacted]
1. There has been a decrease in the size of the bone metastasis in L4. However, there has been further loss of vertebral body height at this level and there is protrusion of the posterior bony cortex into the spinal canal, with a slight increase in canal stenosis. There remains narrowing of the left exit foramen.
2. Abnormal bone marrow signal in L5 does not appear significantly changed.
3. The metastasis in T12 extending into the left posterior elements appears unchanged. There is no significant encroachment into the spinal canal at this level.
4. There is fatty replacement of marrow from T11-L5 in keeping with previous radiotherapy.
5. Right upper pole renal mass noted.
6. There is a large metastasis in the left side of the sacrum. This was incompletely imaged on the previous MR and cannot be compared.
7. No metastases are identified in the upper spine from C1 to T 11. No cord compression. There is a small metastasis in the clivus, unchanged.

Conclusion: Minor improvement in L4 metastasis but increase in spinal canal stenosis at this level due to a malignant compression fracture rather than tumour extension.

Scan Mid Sagittal Slices + Detected Vertebrae:

STIR Lumbar | STIR Cervical | T1 Lumbar

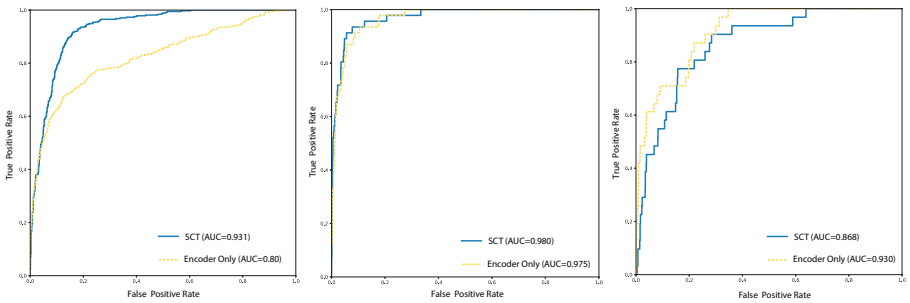
Extracted Annotations:

Condition	Widespread?	S1	L5	L4	L3	L2	L1	T12	T11	T10	T9
Metastasis?	✓			✓				✓			
Fracture?				✓							
Compression?				✓							

Condition	T8	T7	T6	T5	T4	T3	T2	T1	C7	C6	C5	C4	C3	C2
Metastasis?														
Fracture?														
Compression?														

Fig. 6: An example of annotations extracted from a free-text report. There are metastases at T12 and L4, as well as cord compression and fractures ("malignant compression fracture") at L4. The report also suggests there maybe other metastases, e.g. in L5, however these are not explicitly stated to be present. For this reason the 'widespread' column is ticked for metastases. As a results all vertebrae other than T12 and L4 (i.e. levels where metastases are explicitly mentioned) are treated as 'unknown'.

8.5 ROC Curves for All Cancer Tasks



(a) Metastasis (b) Vertebral Fracture (c) Cord Compression

Fig. 7: ROC Curves for the individual spine cancer tasks. These curves show the agreement of the model's predictions with the annotations produced by an expert spinal surgeon.