Interpretable Vertebral Fracture Diagnosis

Paul Engstler[1⋆], Matthias Keicher[1∗], David Schinz[2], Kristina Mach[1], Alexandra S. Gersing[2,3], Sarah C. Foreman[2], Sophia S. Goller[3], Juergen Weissinger[3], Jon Rischewski[3], Anna-Sophia Dietrich[2], Benedikt Wiestler[2], Jan S. Kirschke[2], Ashkan Khakzar[1], and Nassir Navab[1,4]

[1] Technical University of Munich
[2] Klinikum Rechts der Isar (Technical University of Munich)
[3] Klinikum der Universität München (University of Munich)
[4] Johns Hopkins University

Abstract. Do black-box neural network models learn clinically relevant features for fracture diagnosis? The answer not only establishes reliability, quenches scientific curiosity, but also leads to explainable and verbose findings that can assist the radiologists in the final and increase trust. This work identifies the concepts networks use for vertebral fracture diagnosis in CT images. This is achieved by associating concepts to neurons highly correlated with a specific diagnosis in the dataset. The concepts are either associated with neurons by radiologists pre-hoc or are visualized during a specific prediction and left for the user's interpretation. We evaluate which concepts lead to correct diagnosis and which concepts lead to false positives. The proposed frameworks and analysis pave the way for reliable and explainable vertebral fracture diagnosis. The code is publicly available.[5]

Keywords: Vertebral Fracture Diagnosis · Interpretability

1 Introduction

Osteoporosis is regarded as one of the most relevant diseases of the elderly, with 22 million women and 5.5 million men affected in the EU alone [5, 14]. Early detection of incidental osteoporotic fractures in routinely-acquired computed tomography (CT) scans is important, as these often remain clinically silent for a long time [12]. Furthermore, osteoporotic fractures are an independent predictor of further fractures with an approx. 12-fold increased risk and are associated with an 8-fold increased mortality [6, 24]. The sequelae include major socioeconomic consequences and an individual reduction in quality of life [4, 16, 13, 7]. Despite the clinical significance, around 85% of osteoporotic fractures are not adequately described in the radiological reports of routinely acquired CT scans, possibly as a result of a disproportionate increase in radiologists' workload [31, 2].

[⋆] P. Engstler and M. Keicher - Both authors share first authorship.

[5] https://github.com/CAMP-eXplain-AI/Interpretable-Vertebral-Fracture-Diagnosis

Automatic detection of vertebral body fractures with deep learning models can remedy this and increase incidental findings. However, most of these methods are black-box models that do not give insights into the decision-making process. Revealing the inside of these models can allow for investigation of failure cases and, when addressed, increase robustness and trust in the system.

Thus far, interpretable diagnosis is mostly investigated via feature attribution (saliency) approaches [19] such as class activation maps [36]. These interpretations reveal where important features for the prediction are located. Although being a valuable tool for running a sanity check on the network inference mechanism, feature attribution does not disclose further information regarding prediction. Moreover, only knowing about the location of important features is not useful information for fracture diagnosis as it is easy to see where the fracture is located, and it is of interest to know "what" features are important.

To this end, we leverage the network dissection [3] approach and analyze the internal units of the neural network and their associated clinical concepts, inspired by its applications in chest radiography [19] and mammography [32]. Subsequently, we ask the clinicians to identify the concepts associated with highly correlated activations by inspecting the inputs that activate those neurons the highest. We investigate what concepts the network has learned and whether they are aligned with what clinicians use. Moreover, we visualize the concepts used for prediction on a single input to get a conceptual understanding of the decision-making mechanism of the model. We perform the analysis for on the open-source VerSe [29] dataset and a larger private dataset procured at our hospitals. The objective of this work is to investigate what features the network uses for fracture diagnosis, whether they overlap with clinical knowledge, and how they can be used for more verbose and explainable fracture diagnosis.

1.1 Related Work

Vertebral Fracture Detection Most approaches use Convolutional Neural Networks (CNN) on Computer Tomography (CT) spine images. CNN-based methods can be categorized into 2D and 3D convolutions. 2D methods usually rely on a feature aggregation with Recurrent Neural Networks to model inter-slice dependencies [1, 30]. Husseini et al. [15] reformat the image to use the most informative mid-sagittal slice of each vertebra and, in addition to fracture detection, grade fractures using an ordinal regression loss for representation learning. Pisov et al. [27] also reformat the 3D volume to retrieve a spine-centered 2D image and detect key points for measuring the compression of each vertebra, detecting and grading fractures.

Detecting fractures on a voxel-level and then post-processing, Nicolaes et al. [26] for the first time used 3D convolutions for the detection of vertebral fractures. More recent works using 3D convolutions include modeling the dependency between the 3D volumes of each vertebra with a sequence-to-sequence model [8] and detecting osteoporotic fractures on a patient-level [33]. Related to the task of fracture detection and grading, recently Li et al. [22], and Feng et al. [10] explored the distinction between benign and malign vertebral fractures.

Interpretability of models is narrowly explored in the domain of vertebral fracture diagnosis and [34] interprets the models by feature attribution (saliency) approaches to identify which regions in the input contributed to the prediction. In fact, in most medical image analysis applications, feature attribution is the dominant approach [19]. However, attribution methods are limited in the information they can disclose regarding the decision-making mechanism of the model. Moreover, the feature attribution problem remains largely unsolved, and although there are many attribution approaches (CAM [36], LRP [25], DeepSHAP [23], IBA [28, 35, 20]. . .), the methods disagree with the identified important features [18, 35, 17]. This disagreement problem is a caveat for domain experts while utilizing these attribution methods. Thus there is a need for interpretation approaches that are reliable and reveal more information than "which region is important." Network Dissection [3, 32, 19] allows to identify the concepts encoded by internal units (neurons) of the network. Methodologically, our work differs from [19, 3] in that we do not use an annotation dataset and instead identify the highly correlated neurons with the output. Furthermore, the main contribution of this work is establishing trust by investigating the alignment between the learned concepts and vertebrae fracture analysis domain knowledge.

2 Methodology

2.1 Vertebral Fracture Detection

We model the vertebral fracture detection task as a binary classification problem, where the positive class indicates a fracture. The network function is defined as $f_{\Theta}(x) : \mathbb{R}^{H \times W \times D} \rightarrow \mathbb{R}$. The predicted probability is $\hat{y} = sigmoid(f_{\Theta}(x))$. We use a 3D U-Net [9] for the vertebral fracture classification task, replacing its upsampling path with a classification head.

2.2 Semantic Concept Extraction (Correlation)

In neural networks, each neuron is activated by a specific input pattern. The corresponding pattern of each neuron can be equivalently deemed as its associated *concept*. In convolutional neural networks, each neuron can be considered either as an activation map or an activation unit within the map. As the activation units within an activation map all represent the same function (only for different spatial locations), they represent the same concept [3]. For our purposes, we refer to the output of a convolutional filter after the activation function as a unit. We denote the output activations of the final convolutional layer of the network by the tensor $A \in \mathbb{R}^{H^{'} \times W^{'} \times K}$ where K represents the number of channels in that layer. After computing the distribution of individual unit activations a_k, we determine the top quantile level $\mathcal{T}_k$ for each unit k such that $P(a_k > \mathcal{T}_k) = 0.005$ [3]. We then derive the binary segmentation mask $M_k(\boldsymbol{x}) := A_k(\boldsymbol{x}) > \mathcal{T}_k$ and denote the set of enabled units for an input $\boldsymbol{x}$ as $E_x := \{k \mid \sum M_k(\boldsymbol{x}) > 0\}$.

Positive Prediction Correlation Some units might capture concepts that are highly useful to determine whether a sample is fractured, establishing a stronger correlation with a true positive prediction than other units. To find these units, we compute:

$$c_k := \frac{\sum_{x \in P} \mathbf{1}_{E_x}(k)}{|P|} \tag{1}$$

where P is the set of positive samples and $\mathbf{1}$ is the indicator function. With $c_{k_1} > c_{k_2} > ...$, k_1 is the unit most strongly correlated with a true positive prediction, followed by k_2.

2.3 Visualization of Highly Correlating Concepts at Inference

Due to the variability of observed defects in fractured vertebrae, different concepts are relevant during the inference of a sample. We compute the relevance of a unit k during inference of input $\boldsymbol{x}$ as follows:

$$r_k := \sum M_k(\boldsymbol{x}) \odot A_k(\boldsymbol{x}) \tag{2}$$

For units k_1, k_2 with $r_{k_1} > r_{k_2}$, k_1 is more relevant for the inference of $\boldsymbol{x}$ than k_2. Now, when visualizing highly correlated concepts for a sample $\boldsymbol{x}$, we compute the inference relevance of each detector unit and display the activation maps $A_{k_1}(x)$, $A_{k_2}(x)$, ... with $r_{k_1} > r_{k_2} > ...$, showing the corresponding responses for the input sample $\boldsymbol{x}$.

3 Experimental Setup

Data Preparation The network is trained on the VerSe dataset [29] as well as an in-house dataset acquired at Hospital A and Hospital B. The latter includes 465 patients with a median age of $\sim 69(\pm 12)$ years, containing a heterogeneous collection of field of views, scanner settings, and healthy and fractured vertebra, including metallic implants and foreign materials. This combined dataset contains CT scans of patients with healthy and fractured vertebrae of osteoporotic or malignant nature from a heterogeneous collection of CT scanners. To address the inherent class imbalance in the data, negative samples are undersampled and positive (fractured) samples are oversampled in training to achieve a perfect class balance each epoch. As osteoporotic and malignant fractures rarely occur in cervical vertebrae (C1-C7), they are excluded from the dataset. We extract $96 \times 96 \times 96$ sized 3D patches for each vertebrae with a 1mm resolution. These patches are centered on the vertebral body and oriented along the spine by aligning the vertical axis with a spline constructed with the vertebral centroids provided by the dataset similar to [15]. The intensity values of the resulting crops are cropped to a Hounsfield Unit range of $[-1000, 1000]$ and then scaled to $[0, 1]$. During training, intensity (Gaussian noise, smoothing, and contrast) and heavy spatial data augmentations (similarity transformation and elastic deformation) are applied. For these tasks, NiBabel 3.2.1 and MONAI 0.8.0 are used.

Implementation Details The 3D U-Net is implemented in PyTorch Lightning 1.5.10 on top of PyTorch 1.10.2, and trained using the Adam [21] optimizer (learning rate 0.001) without weight decay. Training is concluded if the validation F1 score has not improved for 50 epochs. Dropout with probability 0.3 is applied.

4 Results and Discussion

In the following, we first evaluate the performance of our vertebral fracture detection neural network before dissecting it into its individual detector units. We then validate detector units highly correlated with a true positive prediction by showing that they represent clinically meaningful concepts. Lastly, we present a system to display the units most relevant to a single inference.

Vertebral Fracture Detection We consider the threshold-based evaluation metrics F1-score and accuracy, evaluated at the vertebra level. To remove the dependence on a manually chosen threshold whose optimum might vary between trained networks, the area under curve (AUC) and average precision (AP) metrics are also evaluated. We report the mean and standard deviation of these metrics from five separate training trials for each model.

Training	Testing	F1 (%)	Acc. (%)	AUC (%)	AP (%)
VerSe	VerSe	71.2 ± 10.8	78.2 ± 12.0	84.5 ± 9.1	76.4 ± 14.5
VerSe, in-house	VerSe	86.1 ± 2.6	**90.9 ± 1.6**	**96.2 ± 0.9**	94.1 ± 1.6
VerSe, in-house	VerSe, in-house	**88.0 ± 0.7**	88.0 ± 0.4	94.7 ± 0.5	**95.0 ± 0.4**

Table 1. Performance of the trained neural networks on the test holdout of the smaller VerSe dataset as well as the combined dataset, comprised of VerSe and non-public data acquired from Hospital A and Hospital B. In total, the VerSe dataset contains 3,920 non-cervical vertebrae (254 of which are fractured), whereas the combined dataset comprises 10,675 T1-L5 vertebrae (1,246 fractured).

For networks trained on the smaller VerSe dataset, we observe performance akin to "naive" two-dimensional vertebral fracture detection approaches on the same dataset [15], and a high dependence on a beneficial random seed. These networks, however, do not yield detector units that exhibit any discerning patterns. This is achieved by training a network with the larger dataset, combining VerSe and in-house data collected at Hospital A and Hospital B, that is reliably superior in performance. Its detector units exhibit a variety of patterns that are investigated in the subsequent sections.

4.1 Clinical Meaningfulness of Extracted Semantic Concepts

Given the network trained on the larger dataset, we extract its semantic concepts with Network Dissection [3], which we extended to the three-dimensional space.

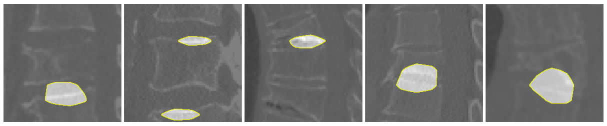
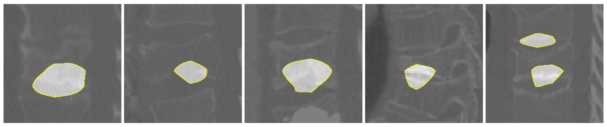
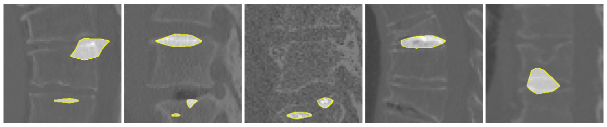
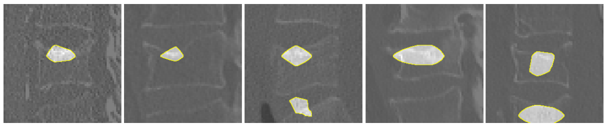
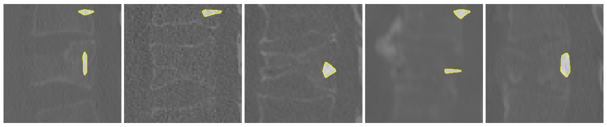
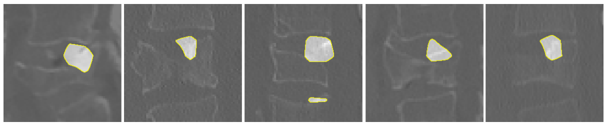
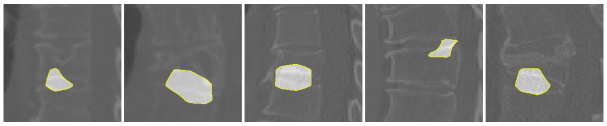
Rank	Sample Activations	Clinical Explanation
1		Abnormal endplate and intervertebral disc shapes
2		Primarily defects of the inferior endplate, associated with severe fractures
5		Abnormal endplate shapes with partial observation of adjacent inferior vertebrae
7		Central defect of the superior endplate, commonly observed in compression fractures, with partial observation of adjacent inferior vertebrae
8		Observation of the spongiosa in the primary vertebrae as well as the adjacent superior one
9		Injury to the middle column of the vertebral bodies, associated with clinically significant myelon compression and consecutive paresis
10		Abnormal endplate and intervertebral disc shapes

Table 2. Visualization of the detector units most strongly correlated with a true positive prediction along with an interpretation of their activations by clinical experts. All displayed samples are fractured and represented by a slice with high activation after thresholding.

To reduce the 512 detector units of the 3D U-Net to a tractable number, we determine the top ten units highly correlated with a true positive prediction as detailed in Section 2.2. For these units, we exported a single-slice collage of 25 strongly activating fractured samples serving as an overview of the units' activations. For the five samples that activated the unit most strongly, all two-dimensional slices as well as three-dimensional NIfTI files are exported, allowing for a detailed inspection.

Based on these exports, we consulted two clinical experts with a combined experience of 22 years in spine imaging about the clinical meaningfulness of these detector units. Omitting three units where no immediate association was possible, we show the detector units identified by their correlation rank with their corresponding clinical explanation in Table 2. The provided samples show a diverse collection of detector unit activations, with each unit exhibiting consistent patterns across multiple samples. We also observe that these units' main focus is the primary vertebra, even if there is some activation in the surroundings. It is noteworthy that the patterns align with the bone anatomy and present themselves in clinically significant locations. As severe fractures are associated with changes in the superior and inferior vertebral endplates, we find the majority of activations in these regions. Although multiple detector units target these areas, they focus on different locations and exhibit varying sizes of regions of interest, with some integrating further information from the intervertebral discs as well as the adjacent vertebra. These insights are clinically meaningful to detect moderate and severe vertebral deformations (Genant grade 1 or higher [11]), and thus show that our network learned concepts that have a clinical correspondence.

For the omitted cases, we observed either no statistically significant activations, i.e. $M_k(\boldsymbol{x}) = \mathbf{0}$, or sporadic activations that do not present any clear patterns, even though they are highly correlated with a true positive prediction. Overall, such detector units represent a minority and can therefore be disregarded in light of those that exhibit tangible patterns.

4.2 Single-Inference Concept Visualization

Having shown that the network learns clinically relevant concepts, we have validated its ability to make use of conducive features. We further seek to illuminate the black box decision-making process of the network by providing the user with a visual explanation for a single inference. To this end, we propose a system that visualizes the concepts considered most important by the network during inference.

Using the method described in Section 2.3 to identify the units representing the most relevant concepts, we retrieve their respective top activating images from our combined dataset. We then display two visualizations for each unit: (i) the activations of those units for the input sample, and (ii) the activations for their corresponding top images. This provides the user with a detector unit's particular response for the given input sample as well as a larger context to understand its general concept. For both visualizations, a single slice with high activation (after thresholding) is shown. An example of (i) is given with Table 3, which gives evidence of the network corroborating its prediction with a diverse set of concepts. These concepts illustrate the network accurately identifying relevant indications for the wedge-shaped deformity and incorporating information from an adjacent vertebra.

This system enables users to comprehend the network's decision making, increasing trust in the system and allowing them to identify failure cases more easily. Furthermore, this approach does not require any prior concept matching by experts, as the user is able to interpret the general concept of a detector unit and make informed judgements about its importance for a particular sample.

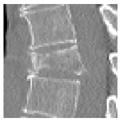
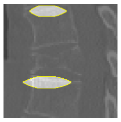
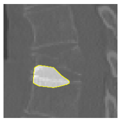
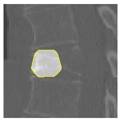
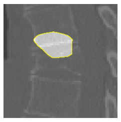
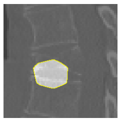
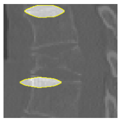
							
Unit	22	111	301	122	277		197
Relevance	1	2	3	4	5	. . .	10

Table 3. Visualization of the most relevant detector units during class prediction of the sample shown on the left, which the network correctly predicted as fractured. Each detector unit is represented by a single slice activation for that particular sample. We also show its ranking in units highly correlated with a true positive prediction. We observe that the network uses concepts associated with wedge-shaped deformity and incorporates information from an adjacent vertebra

5 Conclusion

We show that a 3D U-Net learns a diverse set of concepts to tackle the task of detecting vertebral fractures. To gauge their meaningfulness, we first proposed a method to identify units highly correlated with a fracture detection. Then, we showed the overlap of these units with clinical concepts as validated by experts. Finally, we introduced a system to visually explain a single inference by showing the concepts most relevant for the classification of the sample, giving users insight into the network's decision making process. Further extensions of this system are conceivable, such as pre-filling a radiology report based on activations in a group of semantically similar detector units.

Acknowledgements

The authors acknowledge the financial support by the Federal Ministry of Education and Research of Germany (BMBF) under project DIVA (FKZ 13GW0469C). Ashkan Khakzar was partially supported by the Munich Center for Machine Learning (MCML) with funding from the BMBF under the project 01IS18036B. Kristina Mach was partially supported by the Linde & Munich Data Science Institute, Technical University of Munich PhD Fellowship.

References

1. Bar, A., Wolf, L., Amitai, O.B., Toledano, E., Elnekave, E.: Compression fractures detection on ct. In: Medical imaging 2017: computer-aided diagnosis. vol. 10134, p. 1013440. International Society for Optics and Photonics (2017)
2. Bartalena, T., Giannelli, G., Rinaldi, M.F., Rimondi, E., Rinaldi, G., Sverzellati, N., Gavelli, G.: Prevalence of thoracolumbar vertebral fractures on multidetector CT. European Journal of Radiology **69**(3), 555–559 (Mar 2009)

3. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: Quantifying interpretability of deep visual representations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6541–6549 (2017)
4. Bliuc, D.: Mortality Risk Associated With Low-Trauma Osteoporotic Fracture and Subsequent Fracture in Men and Women. JAMA **301**(5), 513 (Feb 2009). https://doi.org/10.1001/jama.2009.50
5. Cauley, J.A.: Public Health Impact of Osteoporosis. The Journals of Gerontology Series A: Biological Sciences and Medical Sciences **68**(10), 1243–1251 (Oct 2013)
6. Cauley, J., Thompson, D., Ensrud, K., Scott, J., Black, D.: Risk of mortality following clinical fractures. Osteoporosis international **11**(7), 556–561 (2000)
7. Center, J.R., Nguyen, T.V., Schneider, D., Sambrook, P.N., Eisman, J.A.: Mortality after all major types of osteoporotic fracture in men and women: an observational study. The Lancet **353**(9156), 878–882 (Mar 1999)
8. Chettrit, D., Meir, T., Lebel, H., Orlovsky, M., Gordon, R., Akselrod-Ballin, A., Bar, A.: 3d convolutional sequence to sequence model for vertebral compression fractures identification in ct. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 743–752. Springer (2020)
9. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: MICCAI. pp. 424–432. Springer (2016)
10. Feng, S., Liu, B., Zhang, Y., Zhang, X., Li, Y.: Two-stream compare and contrast network for vertebral compression fracture diagnosis. IEEE Transactions on Medical Imaging **40**(9), 2496–2506 (2021)
11. Genant, H.K., Wu, C.Y., Van Kuijk, C., Nevitt, M.C.: Vertebral fracture assessment using a semiquantitative technique. Journal of bone and mineral research **8**(9), 1137–1148 (1993)
12. Haczynski, J., Jakimiuk, A.: Vertebral fractures: a hidden problem of osteoporosis. Medical Science Monitor: International Medical Journal of Experimental and Clinical Research **7**(5), 1108–1117 (Oct 2001)
13. Hallberg, I., Bachrach-Lindström, M., Hammerby, S., Toss, G., Ek, A.C.: Health-related quality of life after vertebral or hip fracture: a seven-year follow-up study. BMC Musculoskeletal Disorders **10**(1), 135 (Dec 2009). https://doi.org/10.1186/1471-2474-10-135
14. Hernlund, E., Svedbom, A., Ivergård, M., Compston, J., Cooper, C., Stenmark, J., McCloskey, E.V., Jönsson, B., Kanis, J.A.: Osteoporosis in the European Union: medical management, epidemiology and economic burden: A report prepared in collaboration with the International Osteoporosis Foundation (IOF) and the European Federation of Pharmaceutical Industry Associations (EFPIA). Archives of Osteoporosis **8**(1-2), 136 (Dec 2013). https://doi.org/10.1007/s11657-013-0136-1
15. Husseini, M., Sekuboyina, A., Loeffler, M., Navarro, F., Menze, B.H., Kirschke, J.S.: Grading loss: a fracture grade-based metric loss for vertebral fracture detection. In: MICCAI. Springer (2020)
16. Jalava, T., Sarna, S., Pylkkänen, L., Mawer, B., Kanis, J.A., Selby, P., Davies, M., Adams, J., Francis, R.M., Robinson, J., McCloskey, E.: Association Between Vertebral Fracture and Increased Mortality in Osteoporotic Patients. Journal of Bone and Mineral Research **18**(7), 1254–1260 (Jul 2003). https://doi.org/10.1359/jbmr.2003.18.7.1254
17. Khakzar, A., Baselizadeh, S., Navab, N.: Rethinking positive aggregation and propagation of gradients in gradient-based saliency methods. arXiv preprint arXiv:2012.00362 (2020)

18. Khakzar, A., Khorsandi, P., Nobahari, R., Navab, N.: Do explanations explain? model knows best. arXiv preprint arXiv:2203.02269 (2022)
19. Khakzar, A., Musatian, S., Buchberger, J., Valeriano Quiroz, I., Pinger, N., Baselizadeh, S., Kim, S.T., Navab, N.: Towards semantic interpretation of thoracic disease and covid-19 diagnosis models. In: MICCAI. Springer (2021)
20. Khakzar, A., Zhang, Y., Mansour, W., Cai, Y., Li, Y., Zhang, Y., Kim, S.T., Navab, N.: Explaining covid-19 and thoracic pathology model predictions by identifying informative input features. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 391–401. Springer (2021)
21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
22. Li, Y., Zhang, Y., Zhang, E., Chen, Y., Wang, Q., Liu, K., Yu, H.J., Yuan, H., Lang, N., Su, M.Y.: Differential diagnosis of benign and malignant vertebral fracture on ct using deep learning. European Radiology **31**(12), 9612–9619 (2021)
23. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems (2017)
24. Melton III, L.J., Atkinson, E.J., Cooper, C., O'Fallon, W.M., Riggs, B.L.: Vertebral Fractures Predict Subsequent Fractures. Osteoporosis International **10**(3), 214–221 (Sep 1999). https://doi.org/10.1007/s001980050218
25. Montavon, G., Lapuschkin, S., Binder, A., Samek, W., Müller, K.R.: Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recognition (2017). https://doi.org/10.1016/j.patcog.2016.11.008
26. Nicolaes, J., Raeymaeckers, S., Robben, D., Wilms, G., Vandermeulen, D., Libanati, C., Debois, M.: Detection of vertebral fractures in ct using 3d convolutional neural networks. In: International Workshop and Challenge on Computational Methods and Clinical Applications for Spine Imaging. pp. 3–14. Springer (2019)
27. Pisov, M., Kondratenko, V., Zakharov, A., Petraikin, A., Gombolevskiy, V., Morozov, S., Belyaev, M.: Keypoints localization for joint vertebra detection and fracture severity quantification. In: MICCAI. pp. 723–732. Springer (2020)
28. Schulz, K., Sixt, L., Tombari, F., Landgraf, T.: Restricting the flow: Information bottlenecks for attribution. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id=S1xWh1rYwB
29. Sekuboyina, A., Husseini, M.E., Bayat, A., Löffler, M., Liebl, H., Li, H., Tetteh, G., Kukačka, J., Payer, C., Štern, D., et al.: Verse: a vertebrae labelling and segmentation benchmark for multi-detector ct images. Medical image analysis (2021)
30. Tomita, N., Cheung, Y.Y., Hassanpour, S.: Deep neural networks for automatic detection of osteoporotic vertebral fractures on ct scans. Computers in biology and medicine **98**, 8–15 (2018)
31. Williams, A.L., Al-Busaidi, A., Sparrow, P.J., Adams, J.E., Whitehouse, R.W.: Under-reporting of osteoporotic vertebral fractures on computed tomography. European Journal of Radiology **69**(1), 179–183 (Jan 2009)
32. Wu, J., Zhou, B., Peck, D., Hsieh, S., Dialani, V., Mackey, L., Patterson, G.: Deepminer: Discovering interpretable representations for mammogram classification and explanation. arXiv preprint arXiv:1805.12323 (2018)
33. Yilmaz, E.B., Buerger, C., Fricke, T., Sagar, M.M.R., Peña, J., Lorenz, C., Glüer, C.C., Meyer, C.: Automated deep learning-based detection of osteoporotic fractures in ct images. In: International Workshop on Machine Learning in Medical Imaging. pp. 376–385. Springer (2021)
34. Yilmaz, E.B., Mader, A.O., Fricke, T., Peña, J., Glüer, C.C., Meyer, C.: Assessing attribution maps for explaining cnn-based vertebral fracture classifiers. In: Inter-

pretable and Annotation-Efficient Learning for Medical Image Computing, pp. 3–12. Springer (2020)

35. Zhang, Y., Khakzar, A., Li, Y., Farshad, A., Kim, S.T., Navab, N.: Fine-grained neural network explanation by identifying input features with predictive information. Advances in Neural Information Processing Systems **34** (2021)
36. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)