

Interpretable Vertebral Fracture Quantification via Anchor-Free Landmarks Localization

Alexey Zakharov^{a,b,1}, Maxim Pisov^{a,c,1}, Alim Bukharaev^{a,1}, Alexey Petraikin^d, Sergey Morozov^e, Victor Gombolevskiy^f, Mikhail Belyaev^{a,b,*}

^a*IRA Labs Ltd, Moscow, Russia*

^b*Skolkovo Institute of Science and Technology, Moscow, Russia*

^c*Kharkevich Institute for Information Transmission Problems, Moscow, Russia*

^d*Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department*

^e*Osimis SA, Liège, Belgium*

^f*Artificial Intelligence Research Institute, Moscow, Russia*

Abstract

Vertebral body compression fractures are early signs of osteoporosis. Though these fractures are visible on Computed Tomography (CT) images, they are frequently missed by radiologists in clinical settings. Prior research on automatic methods of vertebral fracture classification proves its reliable quality; however, existing methods provide hard-to-interpret outputs and sometimes fail to process cases with severe abnormalities such as highly pathological vertebrae or scoliosis. We propose a new two-step algorithm to localize the vertebral column in 3D CT images and then detect individual vertebrae and quantify fractures in 2D simultaneously. We train neural networks for both steps using a simple 6-keypoints based annotation scheme, which corresponds precisely to the current clinical recommendation. Our algorithm has no exclusion criteria, processes 3D CT in 2 seconds on a single GPU, and provides an interpretable and verifiable output. The method approaches expert-level performance and demonstrates state-of-the-art results in vertebrae 3D localization (the average error is 1 mm), vertebrae 2D detection (precision and recall are 0.99), and fracture identification (ROC AUC at the patient level is up to 0.96). Our anchor-free vertebra detection network shows excellent generalizability on a new domain by achieving ROC AUC 0.95, sensitivity 0.85, specificity 0.9 on a challenging VerSe dataset with many unseen vertebra types.

Keywords: Vertebral Fractures, Object Detection, Keypoints Localization, Convolutional Neural Network, Chest Computed Tomography

*Corresponding author

Email address: m.belyaev@skoltech.ru (Mikhail Belyaev)

¹Equal contribution

1. Introduction

Osteoporosis is a systemic skeletal disease manifested by low bone mass and deterioration of bone microarchitecture followed by increased bone fragility. The clinical manifestation of osteoporosis is the occurrence of bone fractures (Kanis et al., 2019) which are common in older adults and resulted in more than two million Disability Adjusted Life Years in Europe (Johnell and Kanis, 2006). Typically, osteoporotic fractures are localized in the spine, hip, distal forearm, and proximal humerus.

Osteoporotic fracture risk models are becoming increasingly popular, while bone mineral density (BMD) is a major contributing factor. However, the prevalence and severity of vertebral compression fractures (VCFs) are predictive for the risk of new osteoporotic fractures independently of bone mineral density (BMD) measurements (Malgo et al., 2017). In particular, vertebrae fractures usually occur before hip fractures (Riggs and Melton Iii, 1995) and dramatically increase the probability of the subsequent fractures (Klotzbuecher et al., 2000); thus can be used as an early marker of osteoporosis.

Medical imaging, such as Computed Tomography (CT), is a useful tool to identify VCFs (Lenchik et al., 2004), especially as an incidental finding. However, radiologists usually analyze CT by navigating through axial slices as, first, computed tomography produces axial slices by design, and second, this view is sufficient to analyze the majority of pathological conditions, e.g., lung diseases. In contrast, vertebral fractures identification is an exception and must be analyzed in the sagittal plane. Multiplanar reconstructions are not generated automatically in many hospitals and require some additional manual steps from radiologists (Gossner, 2010). As a result, radiologists frequently miss fractures, especially if they are not specializing in musculoskeletal imaging, with the average error rate being higher than 50% (Mitchell et al., 2017). At the same time, rapidly evolving lung cancer screening programs or active usage of CT for COVID-19 diagnosis and management provide a solid basis for opportunistic screening of vertebral fractures.

The medical image computing community thoroughly investigated fractures detection and/or classification on vertebrae-level (Roth et al., 2016; Valentinitisch et al., 2019; Burns et al., 2017; Antonio et al., 2018; Husseini et al., 2020), whole study-level (Tomita et al., 2018; Bar et al., 2017; Chetrit et al., 2020), or jointly on both levels (Nicolaes et al., 2019; Yilmaz et al., 2021), see Section 2 for more details. Many of these approaches require prior vertebrae detection (Antonio et al., 2018; Valentinitisch et al., 2019; Nicolaes et al., 2019; Husseini et al., 2020), or spine segmentation (Burns et al., 2017; Roth et al., 2016; Bar et al., 2017). Though both problems are active areas of research with prominent results, fractured vertebrae are the most complex cases for these algorithms (Sekuboyina et al., 2017), and even good average detection/segmentation accuracy may not be sufficient for accurate fracture estimation. As a result, researchers had to exclude some studies from the subsequent fracture classification due to errors in prior segmentation (Valentinitisch et al., 2019), or due to scoliosis (Tomita et al., 2018).

The second important issue is the mismatch between computer science problem statements and the radiological way to define fractures. The Genant scale ([Genant et al., 1993](#)) is a widely used medical criterion recommended by the International Osteoporosis Foundation ([Genant and Bouxsein, 2011](#)). It relies on the measurements of h_a, h_m, h_p - the anterior, middle and posterior heights of vertebral bodies (Fig. 1d, 1e):

$$G = \frac{\min\{h_a, h_m, h_p\}}{\max\{h_a, h_m, h_p\}}, \quad (1)$$

G values provide an easy to interpret continuous index, whereas existing methods are usually trained to predict a binary label extracted from radiological reports ([Tomita et al., 2018](#); [Bar et al., 2017](#)) or multiclass labels based on threshold levels for G ([Valentinitisch et al., 2019](#); [Burns et al., 2017](#)). A related problem is the interpretability of the methods' outputs. The only available information is the network's attention ([Tomita et al., 2018](#)) or a similar score ([Nicolaes et al., 2019](#)) somehow related to the probability of fracture presence. At the same time, the medical community is not satisfied with the level of interpretability of such approaches ([Ghassemi et al., 2021](#)).

Our contribution is twofold. First, our method estimates six keypoints to detect each vertebra and estimate its heights h_* simultaneously (Fig. 1c-e), which results in excellent fracture classification quality with the area under ROC curve equal to 0.95 or higher. The predictions are highly interpretable as they can be validated by a doctor using a simple ruler. Second, we demonstrate the generalizability of our approach by evaluating the vertebrae detection model on the VerSe ([Löffler et al., 2020](#)) dataset *without any training/fine-tuning*, which results in only a minor drop of fracture severity classification quality.

This work extends our previous conference paper ([Pisov et al., 2020](#)) published at MICCAI-2020. The additional primary contributions are

- We propose a new anchor-free approach to detect vertebrae. We compare this idea with our previous anchor-based method and show that it offers not only a more elegant but also more accurate algorithm, see Section 6.2.1.
- We tested the generalizability of the proposed fracture quantification method in two ways. First, we transfer it from LungCancer-500 to VerSe with a negligible drop in quality. Second, we tested the proposed approach within the mosmed.ai initiative, an experiment for systematic comparison of various AI solutions for medical imaging ([Pavlov et al., 2021](#)).
- We extended the previous public release of vertebra annotation on LungCancer-500 dataset by adding missing annotations for lumbar and cervical vertebrae.

2. Previous Work

We split previous methods into three major groups:

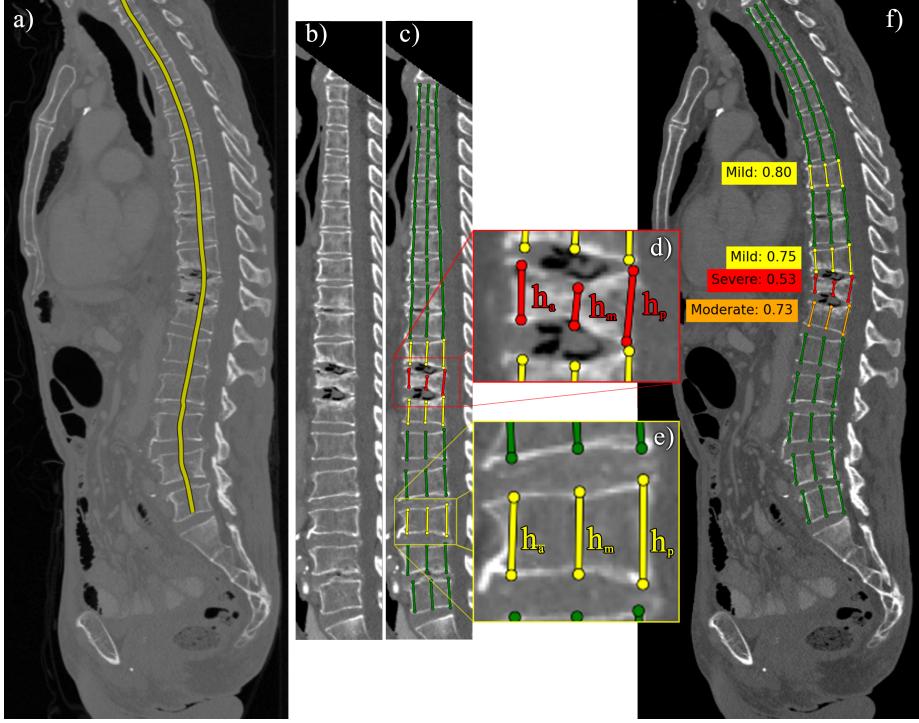


Figure 1: Overview of the proposed model. **Step 1:** a) localizing vertebrae centers in 3D CT (a sagittal projection is shown); b) generating a new 2D image via spine ‘straightening’. **Step 2:** c) identifying key-points and the corresponding heights; d-e) a closer look at some vertebrae (colors denote the fracture severity). **Finally:** f) the original image with vertebrae types (right side) and their estimated fracture severities (left side).

- *Vertebra-level analysis.* The goal of these methods is to estimate severity for every vertebra. Usually, these methods rely on an external vertebra detection method to provide a small 3D path for each vertebra.
- *Patient-level analysis.* These methods operate with the whole 3D image and assign a label to the whole CT series. Though vertebra-level predictions can be aggregated, e.g., by max pooling, some of the methods work with a single classification label for the whole 3D image, so it cannot be decomposed into a series of vertebra-level predictions. Patient-level methods usually provide the full pipeline and don’t use external vertebra localization as an input.
- *Other spine-related methods.* Some interesting ideas are proposed in other spine imaging analysis problems, which aren’t directly related to fracture detection.

2.1. Vertebra-level fracture classification

The automatic classification of vertebral fractures has received much attention from the medical image analysis community. A quantitative image analysis method was proposed in ([Burns et al., 2017](#)) to classify individual vertebrae. First, the spinal column is segmented by an external method detecting intervertebral discs. Then each vertebra is split into 17 sections to extract a set of simple features such as mean density from the segmentation mask. Finally, a support vector machine classifies vertebrae based on the obtained 51 features. The system provides excellent sensitivity (98.7%) but quite low specificity (77.3%). A similar approach was used in ([Valentinitisch et al., 2019](#)) where authors calculated computer vision features such as histograms of oriented gradients from vertebra masks and achieved ROC AUC 0.88. A plain deep learning-based version of this two-step approach was proposed in ([Antonio et al., 2018](#)), where classical ResNet was trained on 3-channel 2D images obtained from the prior segmentation mask by taking central sagittal, axial and coronal slices for each vertebra.

Finally, in ([Husseini et al., 2020](#)) a severity-aware training procedure is proposed. The authors use a tripled-loss-inspired loss function which motivates the network to cluster the representations of each image/patch with similar severities according to their Genant index. This significantly improves the learnt representations, which is empirically demonstrated by training an SVM classifier for the healthy/fractured task. However, because the solved task is classification at either vertebra or image level, the method still lacks interpretability and ease of clinical validation as other classification-based approaches.

It is important to note that all the methods above rely on prior segmentation, which may result in removing some cases with severe abnormalities. Indeed, the authors of ([Valentinitisch et al., 2019](#)) reported that 11 cases out of 154 were excluded from the analysis due to incorrect prior spine segmentation largely caused by high-grade fractures.

2.2. Patient-level fracture classification

This requirement was relaxed in several papers. In ([Nicolaes et al., 2019](#)) the authors proposed a two-step pipeline for vertebrae detection: first, a segmentation neural network is used to generate pixel-level predictions (background, normal, fracture), then the predicted maps are aggregated. Instead of the whole spine mask, the authors used the ground-truth coordinates of vertebrae centroids to produce vertebrae-level predictions and achieved ROC AUC 0.93. A simple idea was used in ([Tomita et al., 2018](#)), where the authors selected the central sagittal slices as the spine is usually located in the middle of the image. In particular, they processed only 6.9 central slices per study (on average). As a result, this approach fails to identify fractures in patients with at least moderate scoliosis, and they had to exclude 156 out of 869 subjects from the analysis, primarily due to scoliosis. Though the average prevalence of scoliosis is 8.85%, it positively correlated with age and increases from 10.95% in 60-69 to 50% in 90+ age groups ([Kebaish et al., 2011](#)), so this cohort can not be ignored in

vertebral fractures screening. The classification method from (Tomita et al., 2018) consists of a ResNet34 which processes each of the central sagittal slices separately; then the obtained scores are aggregated by a simple LSTM network.

An original approach was proposed in (Bar et al., 2017). Though the method also relies on external spine segmentation, the mask is used to extract the spinal cord and create a new virtual sagittal slice. Next, small patches are extracted from this slice and classified by a convolutional network; finally, a recurrent neural network (RNN) is used to aggregate the predictions from each patch. Although the training database is the largest among the reviewed works (consisting of 1673 cases), the model achieves 83.9% sensitivity (with 93.8% specificity), likely due to poor study-level binary annotation extracted from the radiological reports.

This approach was further extended in (Chetrit et al., 2020). First, a YOLO-like object detection network is used to localize the vertebrae on axial slices; the resulting locations are then linearly interpolated to obtain the spinal cord location. The localized spinal cord is split into multiple volumetric patches to tile the vertebrae with minimal overlap. Next, a patch-wise network is used to obtain a fixed-shape representation for each patch. Finally, an aggregation network maps these representations to a final label (healthy/fractured). The patch-based approach gives the possibility to combat various exclusion criteria such as scoliosis. On the other hand, patch-level predictions cannot be accurately mapped to individual vertebrae, giving only rough localization; thus only a single label per image can be predicted, which greatly reduces interpretability and potential for clinical validation.

Another interesting patient-level method was proposed in (Yilmaz et al., 2021). It follows a similar high-level scheme and employs a hierarchical convolutional network (Buerger et al., 2020) to localize vertebrae. Then a simple 4-layer CNN analyzes patches to estimate vertebrae deformity and fracture grades. Finally, the obtained vertebra-level scores are aggregated via maximum function.

2.3. Other spine-related methods

An interesting vertebrae detection and labeling idea is proposed in (Windsor et al., 2020). First, a 2D network is separately applied to sagittal slices in order to (1) detect vertebrae corners and centroids and (2) assign corners to their respective centroids: each "corner" pixel is treated as a potential corner of a vertebra's bounding quadrilateral and 4 displacement fields (for each corner type) are predicted. The displacement fields are then used to group the corners pointing to the same centroid. Finally, the obtained vertebrae are labeled using 2 additional convolutional neural networks combined with a language model, which, by design, guarantees the monotonicity of the resulting labels. However, in addition to its complexity, the approach doesn't always yield geometrically valid predictions, because the corners are not originally tied to centroids. For this reason the authors applied additional post-processing in order to alleviate the problem of too many or too few corners per quadrilateral.

Another iterative approach for vertebrae segmentation is discussed in (Ma-suzawa et al., 2020): first a 3D segmentation network is used to roughly localize the thoracic, lumbar and cervical vertebrae. Next, an iterative convolutional network is used to segment individual vertebrae in each region as well as predict the rough location of the next vertebra.

A carefully designed pipeline for vertebra identification and labelling was proposed within VerSe-2020 competition (Sekuboyina et al., 2021) by Christian Payer who won the challenge. First, the spine is localized by a spinal center-line heatmap regression predicted by U-net. Second, SpatialConfiguration-Net (Payer et al., 2019) is employed to detect centres of the vertebral bodies as landmarks by combining its local appearance with global joint configuration of all vertebrae.

3. Method

The high-level structure of the method follows our previous work (Pisov et al., 2020) as we use a 2-stage pipeline.

1. Spine localization. We propose a new soft-argmax based approach to identify the vertebral column in 3D CT and, as a consequence, reducing the problem to 2D by producing the corresponding mid-sagittal slice (Buckens et al., 2013) to measure h_a, h_m, h_p for each vertebra (Fig. 1a,b). Our method is trained to directly solve the localization problem rather than spine segmentation and demonstrates excellent localization quality with the average error less than 1 mm. Also, it allows us to process all studies with no exceptions, including cases with severe scoliosis.

This step is similar to other recent pipelines: the VerSe-2020 (Sekuboyina et al., 2021) winning solution by C. Payer employs heatmap regression to detect spinal center-line; (Chetrit et al., 2020) detects the center-line points by a 2D YOLO. In fact, both soft-argmax and heatmap regression are actively used in landmark detection while the former approach shows better results in some pose estimation tasks (Luvizon et al., 2018). Detection methods like YOLO seems to be less appropriate as we need to find just one point for every slice, not an arbitrary number of objects.

2. Vertebra detection and fracture quantification. The second task at hand is very similar to object detection: for a given image it is required to localize all objects of a given class, the number of objects may vary, however it is limited by a constant.

The main difference is the encoding of sought objects: in classical object detection axis-aligned bounding boxes (AABBS) are used, while in our case each vertebra is represented by 6 coplanar points in an image.

We propose a modified 2D object detection network to predict the Genant segments directly.

Our contribution is the following:

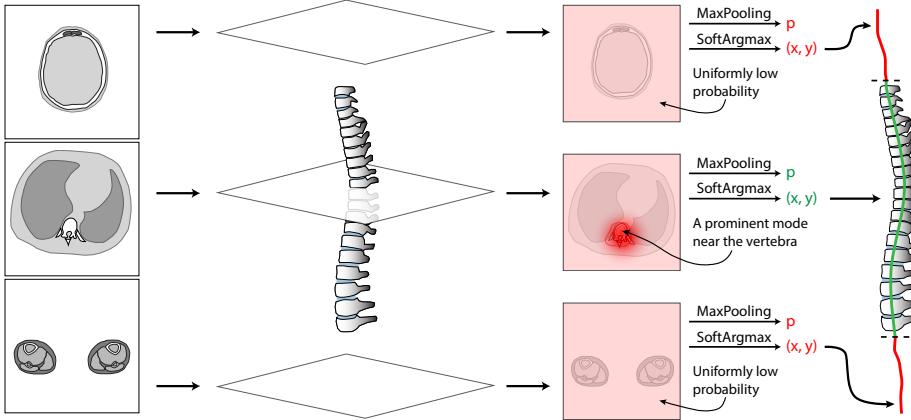


Figure 2: The spine localization pipeline: a) three axial slices from different body regions: head, thorax, legs; b) the slices’ spatial position relative to the spine; c) the predicted probability maps for each slice, the color intensity denotes the probability magnitude; d) the final curve passing through the vertebral column (green), as well as the redundant parts (red) delimited by the spine limits (black, dashed).

- We lift the previous limitation to chest CT images by incorporating the information regarding the *vertebrae limits* into the first network, thus making it applicable to any input.
- We simplify the second network by *removing the anchor boxes*.

3.1. Spine localization

We start our pipeline by localizing the spine. For this purpose we use a 3D UNet-like (Milletari et al., 2016) fully convolutional neural network. For each voxel we interpret the network’s output as the probability of being situated near the vertebral column. Next, we process the prediction in two ways (Fig. 2):

1. We take the 2D soft-argmax (Luvizon et al., 2019) operation along the xOy axes in order to obtain the spine coordinates in axial planes.
2. We take the global max-pooling operation along the same axes in order to obtain the probability of containing the spine at a given z coordinate.

We train the model by optimizing a sum of two loss functions:

1. *Mean absolute error* between the predicted coordinates and the ones smoothly interpolated between vertebrae centers, calculated from annotation (Fig. 4a).
2. *Binary cross-entropy* between the predicted slice-level probabilities and the binary limits, also extracted from the vertebrae annotation.

At inference time we threshold the slice-level probabilities by 0.5 and take the convex hull in order to obtain the curve limits. The predicted curve is then cropped according to these limits. No additional post-processing is used.

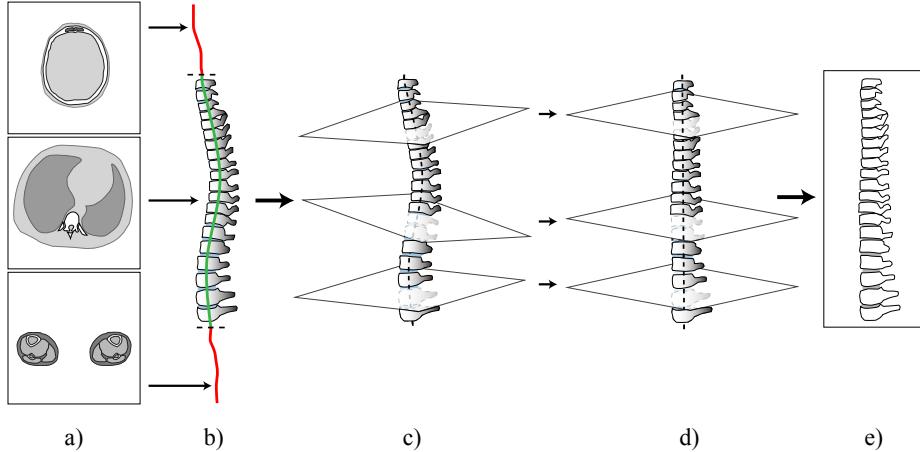


Figure 3: The spine straightening pipeline: a) three axial slices from different body regions: head, thorax, legs; b) the combined points from each slice result in a 3D curve, colors denote the probability of being inside the vertebral column limits: green - high probability, red - low; c) planes orthogonal to the curve (for better visualization most planes are omitted); d) a straightened vertebral column (the planes become parallel); e) the new central sagittal plane.

3.2. Spine straightening

Using the predictions from the first network, we obtain a 3D curve passing through the vertebral column, as well as the limits in which it is defined. We then crop the image accordingly and interpolate it onto a new 3D grid on which the obtained curve becomes a straight vertical line.

In order to find such a grid, we select a number of equidistant points on the curve and construct corresponding orthogonal planes. Then, we generate a grid for which all the planes become parallel, which effectively straightens the curve, because the plane normals are tangent to the curve². Finally we select a new sagittal plane where all vertebrae are visible, namely the one that contains the entire curve. Fig. 3 shows a detailed illustration.

3.3. Vertebrae localization

In classical object detection axis-aligned bounding boxes (AABBS) are used as a relatively adequate description of both localization and shape of a given object. In our case the Genant segments play the same role while also containing more task-specific information, namely the level of deformation. This fact suggests that AABBS can be completely removed from our training pipeline.

During **target generation** we simplify the idea from (Ren et al., 2015) and use anchor-based translation-invariant encoding (Fig. 4b-c): each pixel is

²See <https://github.com/neuro-ml/straighten> for full code for interpolation along curves.

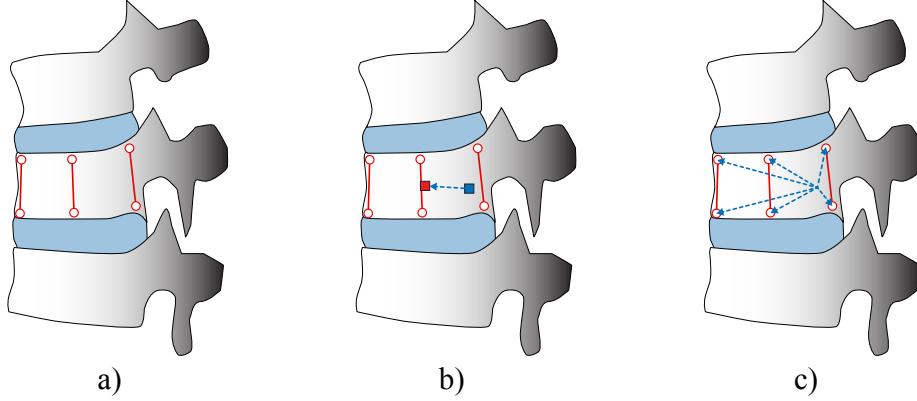


Figure 4: Target generation steps: a) example of an annotated vertebra; b) the distance between the anchor pixel (blue square) and the vertebra centroid (red square); the **objectness** O is 1, if the distance is smaller than a given threshold; c) the **keypoints' coordinates** relative to the anchor pixel.

treated as an anchor relative to which the 6 points' coordinates are calculated:

$$e^x = g^x - a^x; \quad e^y = g^y - a^y, \quad (2)$$

where $(g^x, g^y), (e^x, e^y)$ are the global and encoded coordinates of a given point respectively and (a^x, a^y) - are the coordinates of an anchor pixel.

Additionally, as in standard object detection, each location requires an *objectness* label, which is used to filter out the pixels not related to any vertebrae. The objectness is 1 if the given pixel is closer than a fixed threshold to any vertebra centroid (Fig. 4b), and 0 otherwise.

Finally, we use the same **loss function** as in (Pisov et al., 2020) to train our second network:

$$L = BCE(\hat{o}, o) + \frac{1}{\sum I[o_i = 0]} \sum_{i=1}^N \frac{I[o_i > 0]}{G_i} \cdot MAE(\hat{e}_i, e_i), \quad (3)$$

where BCE is the *log-loss* between the real (o) and predicted (\hat{o}) objectness, MAE is the *mean absolute error* between real (e_i) and predicted (\hat{e}_i) encoded keypoints' coordinates (2) for the i -th vertebra and G_i is the respective Genant score (1) used for loss reweighting. We found such a reweighting of regression loss to be very effective in balancing the network's performance across vertebrae with different fracture severities.

3.4. Non-maximum suppression

Non-maximum suppression (NMS) is an essential component of the majority of current object detection pipelines. Its main purpose is to reduce the number of predictions referring to the same object, and ideally leaving exactly one final

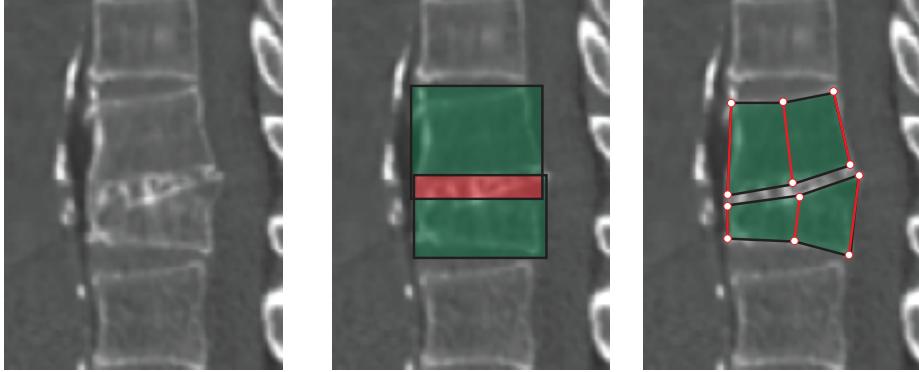


Figure 5: A comparison of IoU and BIoU for two deformed vertebrae: a) a portion of the vertebral column on a straightened image; b) vertebrae bounding boxes, the red region denotes their intersection; c) vertebrae "butterfly" hexagons, built using the Genant segments (red) note the empty intersection and thus zero BIoU.

prediction per object. NMS requires a measure of *closeness* between two given predictions, i.e. a measure of overlap between bounding boxes, as well as an *objectness* score.

In our case, the objectness score is directly predicted by the model.

As for closeness, we replace the intersection-over-union (IoU) between axis-aligned bounding boxes from (Pisov et al., 2020) by new a measure, named as *Butterfly-IoU* (BIoU). In general, we compute BIoU in the same way as standard IoU, but instead of axis-aligned bounding boxes (Fig. 5b), we use hexagons built on six vertebra key points. First, we build these hexagons, as shown on Fig. 5c, for two sets of key points. Second, we calculate areas of intersection and union of two hexagons. Finally, we divide the intersection area by the union area to calculate BIoU. During non-maximum suppression, we use BIoU values to remove similar predictions as usual.

BIoU is specifically designed to better capture the shape of vertebrae and yields significantly better results in cases of severely damaged vertebrae (Fig. 5).

4. Data

4.1. LungCancer500

Our main dataset consists of 100 randomly selected images from the *Moscow Radiology CTLungCa-500* dataset³ (Morozov et al., 2021). The cohort includes studies from a lung cancer screening program, so osteoporotic fractures are common incidental findings as patients aged from 50 to 75.

³https://mosmed.ai/en/data-sets/ct_lungcancer_500/

The images have various voxel spacing ranging from $.5mm \times .5mm \times .8mm$ to $1mm \times 1mm \times .8mm$ and different numbers of visible vertebrae: from 10 to 15.

In (Pisov et al., 2020) an annotation of genant segments for this dataset was published. However, the work was focused solely on detection of thoracic vertebrae. We extended the annotation by adding vertebrae from the remaining regions⁴. The re-annotation was performed by 3 experts with 1 to 5 years of experience in radiology and a board-certified radiologist with 12 years of experience in the field. In total the dataset contains 3565 vertebra annotations (2-3 per single vertebra). To annotate vertebra heights, radiologists (1) look for a sagittal slice passing through the middle of the vertebra and (2) marks six keypoints on this slice. Then the next vertebra is annotated in the same manner, but the selected slice can be different if a patient has scoliosis or patient positioning is wrong.

The distribution of vertebral fractures is the following: 440 mild, 250 moderate, 54 severe deformations and 2821 normal vertebrae. Patient-wise we have a somewhat balanced distribution with 11, 23, 44 and 22 patients with none, mild, moderate and severe deformations respectively.

4.2. VerSe-2020

For additional validation on external data we use the *VerSe-2020* dataset (Löffler et al., 2020; Sekuboyina et al., 2020b,a; Glockner et al., 2013, 2012). The dataset consists of over 300 multidetector CT images of various regions of the spine.

For each image, every vertebra has an associated segmentation mask as well as centroid coordinates. Additionally, a subset of VerSe-2020 contains vertebrae deformation labels calculated using the same Genant scale. However, the vertebrae near image borders sometimes lack annotation, which mostly impacts the estimated precision of our method (see Section 6 for details). Additionally, because during training we rely on interpolation *between* vertebrae, the image containing a single annotated vertebra (verse116) was excluded from the training set.

4.3. Private dataset

In addition to public datasets, we also tested the pipeline trained on a large private dataset which includes

- 402 chest and abdominal CT studies with annotated Genant segments using the same protocol as LungCancer500. The distribution of vertebral fractures is the following: 667 mild, 364 moderate, 153 severe deformations and 4364 normal vertebrae. The annotation was done by three experienced radiologists with at least 10 years of experience.

⁴<https://github.com/neuro-ml/anchor-free-genant/>

- 191 additional studies of chest, abdominal and brain CT with annotated limits to improve the spine localization network performance.

4.4. Mosmed.ai test dataset

In addition to testing the models on public datasets, we also evaluated them on an independent test dataset collected by the Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department. The dataset was prepared to test various algorithms using the principles described in (Pavlov et al., 2021) and consists of 50 studies with vertebral fractures and 50 age-matched healthy studies (Genant score is 0.75 or higher). Cases were prepared carefully to test the algorithms under various conditions, such as vertebral ankylosis, vertebroplasty and osteoblastic metastases, among others.

Only patient-level metrics are available for this dataset. The data is hidden from the developers; the test was conducted in real time with 60 seconds response time requirement.

5. Experimental setup

We trained our **spine localization** network with Adam optimizer (Kingma and Ba, 2014) with standard parameters for 10^5 iterations with batches of size 3 and a learning rate of $1 \cdot 10^{-4}$ ($3 \cdot 10^{-5}$ for VerSe). As a preprocessing step we normalize the voxel intensities to the interval $[0, 1]$ as well as resample the images to a spatial size of $2mm \times 2mm \times 4mm$. Our motivation behind this is to standardize the spatial features, because CNNs are not invariant to scaling, as well as reduce memory consumption for images with too high resolution. No additional postprocessing is applied to final predictions, aside from the probability maps binarization described in Section 3.1.

The **vertebrae detection** network was trained for 80k iterations with batches of size 8. We used Adam optimizer (Kingma and Ba, 2014) with standard parameters and a gradually decreasing learning rate, which enabled the models to reach better optima. The initial learning rate was set to $3 \cdot 10^{-4}$ and decreased by a factor of 2 after 6k, 10k, 16k, 28k, 40k and 56k iterations. As a preprocessing step we normalize the voxel intensities to zero mean and unit variance. As described in Section 3.3, we generate a grid onto which the new image is interpolated. The grid is constructed in such a way, so that the spatial size of each voxel becomes $1mm \times 1mm \times 1mm$, which can be regarded as another implicit preprocessing step. Finally, during non-maximum suppression we leave the predictions with an *objectness* greater than 0.7 and use a threshold of 0.1 for the closeness function (3.4).

6. Results

We report results obtained using 5-fold cross-validation. For every network we trained 5 experiments with different cross-validation splits in order to estimate mean and standard deviation for each score. To obtain patient-level

Table 1: Spinal line localization metrics for Lung-Cancer-500 and VerSe-2020 datasets.

Data		Points	Limits
Train	Test	Mean l2, mm	MAE, mm
Cancer500 Private	Cancer500	.92 (.07)	.03 (.04)
		.74 (.06)	.07 (.08)
VerSe Private	VerSe val	1.81 (.20)	19.23 (3.42)
		1.30 (.15)	18.37 (3.82)
VerSe Private	VerSe test	1.72 (.17)	18.76 (2.59)
		1.19 (.16)	18.07 (3.10)

predictions, we use the most severe fracture among all the vertebrae, which is equivalent to taking the minimal Genant score. As we have multiple annotations per study, we also report the inter-expert variability.

Unlike Lung-Cancer-500, VerSe has a publicly available division on train, validation and test. For this reason, we used the last two subsets for testing.

6.1. Spine localization

We trained models on different datasets with various characteristics:

1. Lung-Cancer-500, which consists of pretty standard Chest CT scans primarily with thoracic vertebrae.
2. More informative Private and VerSe datasets. As Lung-Cancer-500 contains a very limited number of lumbar and cervical vertebrae, we didn't transfer models trained on Lung-Cancer-500 to VerSe.

We report the localization quality of the first step of our method in Table 1. Because most of the images from Lung-Cancer-500 (Section 4.1) are fully covered by vertebrae, this dataset is not very challenging for the limits classification head. Thus, in order to thoroughly evaluate the localization network, we use an additional dataset with annotated vertebrae centroids - VerSe-2020 (Section 4.2).

The difference in the Mean l2 metric between the two datasets can be explained by the fact that the images of VerSe are much more diverse than those of Lung-Cancer-500. Nevertheless, given the input voxel size of $2mm \times 2mm \times 4mm$, the results on both the datasets suggest that the model's performance is close to maximal.

Moreover, as mentioned in Section 4.2, the vertebrae near image limits are sometimes lacking annotation in VerSe, see several examples on Fig. 6. We argue that this is the main cause of such a large *Limits MAE* in Table 1. For this reason, we additionally extrapolate the predictions of our spine localization network trained on VerSe by 2cm at both limits, leaving the second network with

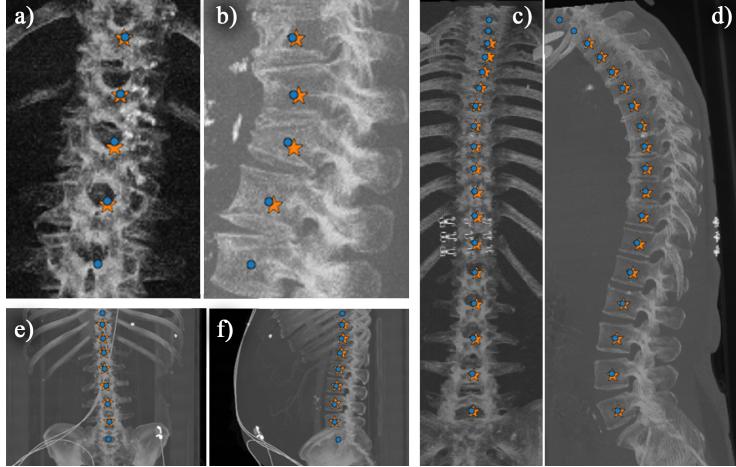


Figure 6: Several examples of "false positives" on the VerSe dataset in sagittal and coronal projections. The annotated vertebrae are marked with orange stars, the predicted ones - with blue circles. Note the missing annotations for "border" vertebrae: near image borders (a-b), vertebral column borders (c-f).

more potential vertebrae to detect. Surprisingly, this simple post-processing increased the overall recall of the pipeline, but doesn't deteriorate the precision significantly (Table 2).

6.2. Vertebrae detection and severity classification

6.2.1. Cross-validation on Cancer500

We first compare three modifications of the pipeline

1. *Anchor-Boxes*. The approach from our previous work (Pisov et al., 2020). Both spine localization and vertebra analysis networks were trained on Lung-Cancer-500.
2. *Ours*. The proposed anchor-free approach; both networks were trained on Lung-Cancer-500.
3. *Ours Private*. The proposed approach; both networks were trained on the Private dataset.

In order to calculate vertebrae-level metrics a matching procedure of predicted and real vertebrae is required. For Cancer500 we use the Butterfly-IoU (namely BIoU) between each (*prediction*, *target*) pair, see Fig. 5 above. All pairs with a closeness smaller than a given threshold are discarded. For the remaining pairs, for each *target* we select the closest corresponding *prediction*. This way all unmatched targets are considered false negatives (FN) and all unmatched predictions - false positives (FP). Table 2 shows vertebrae detection metrics averaged by patients as well as by vertebrae.

We report two types of metrics: precision and recall for vertebrae identification and binary classification metrics for vertebra fracture severity classification;

Table 2: Vertebrae detection and severity classification metrics. *Anchor-Boxes* and *Ours* metrics are based on 5-fold cross-validation. Severity classification metrics are reported on the vertebra level for identifying Moderate and Severe vertebrae ($G \leq 0.74$). We report sensitivity for a specificity fixed at 90% to make models directly comparable.

		Vertebra Detection		Severity Classification	
		Precision	Recall	ROC AUC	Sens. at spec.=0.9
Cancer500	Anchor-Boxes	.994 (.001)	.953 (.001)	.955 (.004)	.863 (.030)
	Ours	.991 (.002)	.990 (.002)	.959 (.002)	.885 (.002)
	Experts	.999 (.001)	.994 (.001)	.971 (.005)	.936 (.018)
	Ours Private	.993	.991	.981	.950
VerSe	Ours	.947	.886	.951	.848
	Ours ext.	.935	.951		
	Ours Private	.896	.973	.970	.908

see the first four lines of Table 2. Anchor-free model benefits from higher detection recall and slightly better classification metrics. The same model trained on the Private dataset shows perfect severity classification metrics, even outperforming expert agreement level. Figure 7 provides some insights about models’ performance. Ours-Private models achieves more symmetrical distribution of errors and a generally narrower interval. Figure 8 shows some qualitative analysis of predictions for the Cancer500 dataset.

6.2.2. External tests

For external test on VerSe and mosmed.ai we use two setups:

1. *Ours*. Spine localization network is now trained on VerSe as LungCancer-

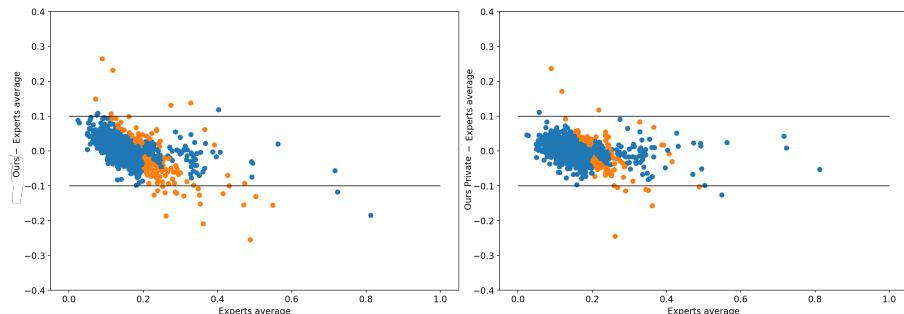


Figure 7: Bland-Altman plots for Ours and Ours-Private models. Blue points denotes vertebra with correct Genant grade classification; orange ones shows wrong classification.

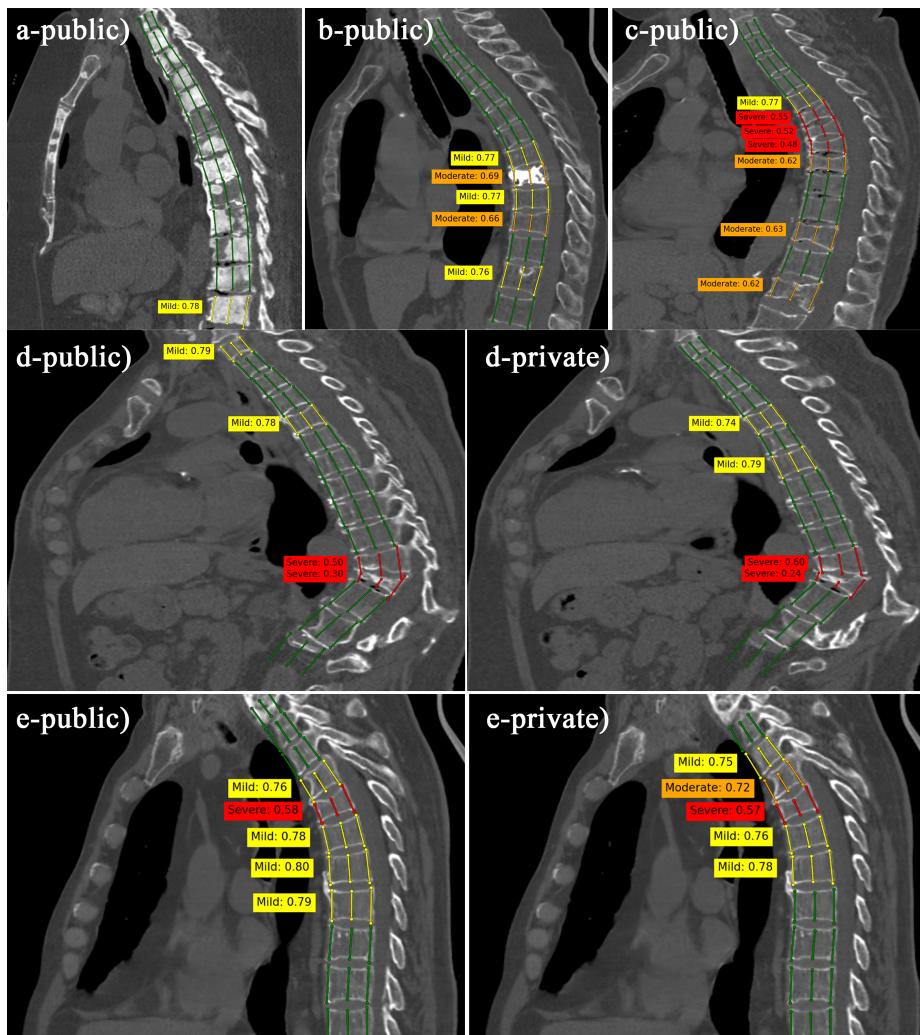


Figure 8: Qualitative analysis of predictions for the Cancer500 dataset for both our public and private networks.

Table 3: Binary classification metrics on VerSe for various grades of fractures: at least *Mild* ($G \leq 0.8$) and at least *Moderate* ($G \leq 0.74$). All numbers are given as mean (std).

Task	Model	Vertebra-level		Patient-level	
		ROC AUC	Sens. at spec.=0.9	ROC AUC	Sens. at spec.=0.9
G0 vs G1, G2, G3	Public	.877	.713	.882	.703
	Private	.906	.777	.911	.807
G0 vs G2, G3	Public	.963	.889	.953	.856
	Private	.979	.952	.962	.919
G0, G1 vs G2, G3	Public	.951	.848	.936	.806
	Private	.970	.908	.960	.910
mosmed.ai	Private	N/A	N/A	.99	1.0

500 contains primarily thoracic vertebrae. The second network is simply reused without any adaptation to VerSe.

2. *Ours-Private* is used without any modifications as its training data set is diverse enough.

For VerSe another matching procedure is required, given the fact that the annotation represents vertebrae centers. Similarly to (Löffler et al., 2020) we use the Euclidean distance between centers as a closeness function with a threshold of 20 mm.

In addition to vertebra-level, we also report patient level by using maximal deformation $1 - G$ to determine the whole image label. To analyze the performance of vertebrae fracture severity classification, we report metrics for various threshold values of G following the radiological definition of severity (Genant et al., 1993), see Table 3. We assume that the most relevant problem for chest CT is the identification of at least Moderate fractures ($G \leq 0.74$) as healthy vertebrae in the thoracic spine are wedged, so normal variation can be misclassified as a Mild fracture ($0.74 < G \leq 0.8$) (Lenchik et al., 2004). To enable direct comparison with (Husseini et al., 2020) we also add *G0 vs G2-G3* where these Mild fractures are removed from the data.

Moreover, we analyze the generalizability of our approach by evaluating the model on the VerSe dataset, which contains additional annotations regarding the fracture severity of each vertebra. In this experiment we only trained the spine localization network on VerSe and reused the vertebrae detection network (trained on Cancer500) *as is*, without any additional tuning. The results are shown in the last rows of Tables 2 and 3. Note the minimal drop in classification performance.

It is important to note that the results for the VerSe dataset from Table 2 are underestimated, especially the precision. This significant drop in quality is due

to lacking annotations for vertebrae near images’ edges. Fig. 6 shows several examples with such “false positives”. According to our calculations about 95% of FPs are due to such partially annotated cases.

Similar values of ROC AUC were obtained at vertebra (0.88 (Valentinitisch et al., 2019), 0.93 (Nicolaes et al., 2019)) and patient levels (0.92 (Tomita et al., 2018)).

Finally, Fig. 9 gives a qualitative analysis of our method’s performance on VerSe by showing several interesting (both bad and good) predictions. The rest of predictions on the entire VerSe test subset can be found in a separate repository⁵.

We can see that the network is robust to multiple severe deformations (Fig. 9a), extreme noise (Fig. 9b), generalizes well to whole body scans, various regions of the vertebral column and presence of contrast (Fig. 9c).

On the other hand, some predictions suffer from severely misplaced keypoints, mainly in the cervical and lumbar regions. An interesting example is Fig. 9g, which shows a faulty prediction for the C2 vertebra, which has an atypical shape as compared to other vertebrae. Another good example is Fig. 9e, which has multiple false-negatives due to a noticeable amount of artifacts. Also, Fig. 9f shows a case with multiple problems, all of which were caused by the spine localization network.

Overall, our analysis shows that, at a great extent, most of the causes of bad performance can be alleviated by training the network on a larger and more challenging dataset. See Fig. 9d-g for a comparison between our public and private networks.

7. Discussion

We proposed an interpretable method for vertebral compression fractures quantification. It’s based on clinical recommendations and provides easy-to-verify outputs.

We extended and simplified the method for automatic identification of vertebrae-level fractures classification using the Genant score proposed in (Pisov et al., 2020). First of all we demonstrated, for the task at hand, the redundancy of bounding-boxes, and showed that Genant segments represent a more suitable description of vertebrae shape and localization. Furthermore, we extended the spine localization model by adding limits prediction, which makes it more suitable for clinical applications.

Finally, we demonstrated, the generalizability of our approach by evaluating our vertebrae detection network on the publicly available dataset VerSe-2020.

Acknowledgments. Alexey Zakharov, Maxim Pisov, Victor Gombolevskiy and Mikhail Belyaev were supported by the Russian Science Foundation grant 20-71-10134.

⁵<https://github.com/neuro-ml/anchor-free-genant/>

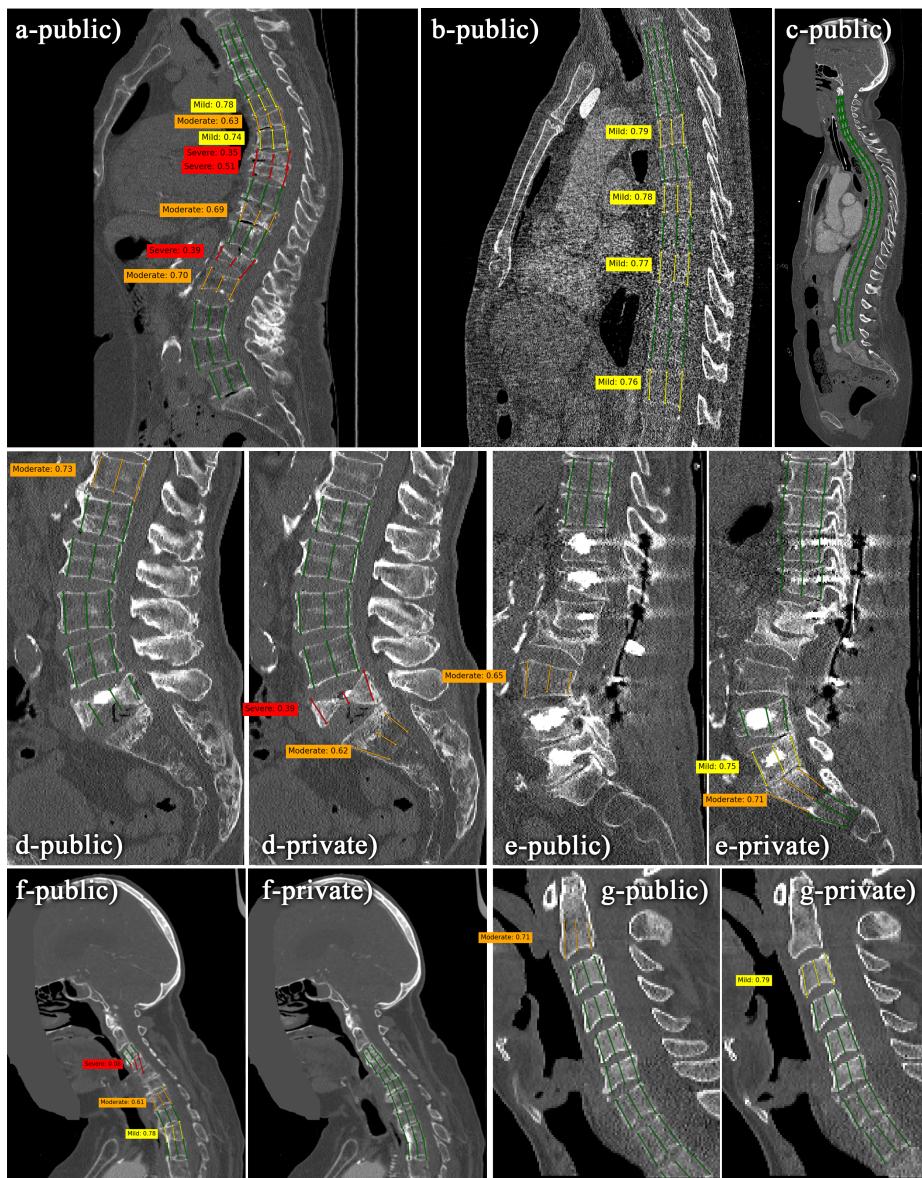


Figure 9: Qualitative analysis of predictions for the VerSe dataset for both our public and private networks.

References

- Antonio, C.B., Bautista, L.G.C., Labao, A.B., Naval, P.C., 2018. Vertebra fracture classification from 3D CT lumbar spine segmentation masks using a convolutional neural network, in: Asian Conference on Intelligent Information and Database Systems, Springer. pp. 449–458.
- Bar, A., Wolf, L., Amitai, O.B., Toledano, E., Elnekave, E., 2017. Compression fractures detection on CT, in: Medical Imaging 2017: Computer-Aided Diagnosis, International Society for Optics and Photonics. p. 1013440.
- Buckens, C.F., de Jong, P.A., Mol, C., Bakker, E., Stallman, H.P., Mali, W.P., van der Graaf, Y., Verkooijen, H.M., 2013. Intra and interobserver reliability and agreement of semiquantitative vertebral fracture assessment on chest computed tomography. *PloS one* 8.
- Buerger, C., von Berg, J., Franz, A., Klinder, T., Lorenz, C., Lenga, M., 2020. Combining deep learning and model-based segmentation for labeled spine CT segmentation, in: Medical Imaging 2020: Image Processing, International Society for Optics and Photonics. p. 113131C.
- Burns, J.E., Yao, J., Summers, R.M., 2017. Vertebral body compression fractures and bone density: automated detection and classification on CT images. *Radiology* 284, 788–797.
- Chetrit, D., Meir, T., Lebel, H., Orlovsky, M., Gordon, R., Akselrod-Ballin, A., Bar, A., 2020. 3D convolutional sequence to sequence model for vertebral compression fractures identification in CT, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, Springer International Publishing, Cham. pp. 743–752.
- Genant, H.K., Bouxsein, M.L., 2011. Vertebral Fracture Initiative: Executive summary. https://www.iofbonehealth.org/sites/default/files/PDFs/IOF_VFI-Executive_Summary-English.pdf.
- Genant, H.K., Wu, C.Y., Van Kuijk, C., Nevitt, M.C., 1993. Vertebral fracture assessment using a semiquantitative technique. *Journal of bone and mineral research* 8, 1137–1148.
- Ghassemi, M., Oakden-Rayner, L., Beam, A.L., 2021. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health* 3, e745–e750.
- Glocker, B., Feulner, J., Criminisi, A., Haynor, D.R., Konukoglu, E., 2012. Automatic localization and identification of vertebrae in arbitrary field-of-view CT scans, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 590–598.

- Glocker, B., Zikic, D., Konukoglu, E., Haynor, D.R., Criminisi, A., 2013. Vertebrae localization in pathological spine CT via dense classification from sparse annotations, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 262–270.
- Gossner, J., 2010. Missed incidental vertebral compression fractures on computed tomography imaging: More optimism justified. *World Journal of Radiology* 2, 472.
- Husseini, M., Sekuboyina, A., Loeffler, M., Navarro, F., Menze, B.H., Kirschke, J.S., 2020. Grading loss: A fracture grade-based metric loss for vertebral fracture detection, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, pp. 733–742.
- Johnell, O., Kanis, J., 2006. An estimate of the worldwide prevalence and disability associated with osteoporotic fractures. *Osteoporosis international* 17, 1726–1733.
- Kanis, J.A., Cooper, C., Rizzoli, R., Reginster, J.Y., 2019. European guidance for the diagnosis and management of osteoporosis in postmenopausal women. *Osteoporosis international* 30, 3–44.
- Kebaish, K.M., Neubauer, P.R., Voros, G.D., Khoshnevisan, M.A., Skolasky, R.L., 2011. Scoliosis in adults aged forty years and older: prevalence and relationship to age, race, and gender. *Spine* 36, 731–736.
- Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .
- Klotzbuecher, C.M., Ross, P.D., Landsman, P.B., Abbott III, T.A., Berger, M., 2000. Patients with prior fractures have an increased risk of future fractures: a summary of the literature and statistical synthesis. *Journal of bone and mineral research* 15, 721–739.
- Lenchik, L., Rogers, L.F., Delmas, P.D., Genant, H.K., 2004. Diagnosis of osteoporotic vertebral fractures: importance of recognition and description by radiologists. *American Journal of Roentgenology* 183, 949–958.
- Luvizon, D.C., Picard, D., Tabia, H., 2018. 2d/3d pose estimation and action recognition using multitask deep learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5137–5146.
- Luvizon, D.C., Tabia, H., Picard, D., 2019. Human pose regression by combining indirect part detection and contextual information. *Computers & Graphics* 85, 15–22.
- Löffler, M.T., Sekuboyina, A., Jacob, A., Grau, A.L., Scharr, A., El Husseini, M., Kallweit, M., Zimmer, C., Baum, T., Kirschke, J.S., 2020. A vertebral segmentation dataset with fracture grading. *Radiology: Artificial Intelligence* 2, e190138.

- Malgo, F., Hamdy, N., Ticheler, C., Smit, F., Kroon, H., Rabelink, T., Dekkers, O., Appelman-Dijkstra, N., 2017. Value and potential limitations of vertebral fracture assessment (VFA) compared to conventional spine radiography: experience from a fracture liaison service (FLS) and a meta-analysis. *Osteoporosis International* 28, 2955–2965.
- Masuzawa, N., Kitamura, Y., Nakamura, K., Iizuka, S., Simo-Serra, E., 2020. Automatic segmentation, localization, and identification of vertebrae in 3D CT images using cascaded convolutional neural networks, in: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, Springer International Publishing, Cham. pp. 681–690.
- Milletari, F., Navab, N., Ahmadi, S.A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: *3D Vision (3D V)*, 2016 Fourth International Conference on, IEEE. pp. 565–571.
- Mitchell, R., Jewell, P., Javaid, M., McKean, D., Ostlere, S., 2017. Reporting of vertebral fragility fractures: can radiologists help reduce the number of hip fractures? *Archives of osteoporosis* 12, 71.
- Morozov, S., Gombolevskiy, V., Elizarov, A., Gusev, M., Novik, V., Prokudaylo, S., Bardin, A., Popov, E., Ledikhova, N., Chernina, V., et al., 2021. A simplified cluster model and a tool adapted for collaborative labeling of lung cancer CT scans. *Computer Methods and Programs in Biomedicine* 206, 106111.
- Nicolaes, J., Raeymaeckers, S., Robben, D., Wilms, G., Vandermeulen, D., Libanati, C., Debois, M., 2019. Detection of vertebral fractures in CT using 3D convolutional neural networks. arXiv preprint arXiv:1911.01816 .
- Pavlov, N.A., Andreychenko, A.E., Vladzimyrskyy, A.V., Revazyan, A.A., Kirpichev, Y.S., Morozov, S.P., 2021. Reference medical datasets (MosMedData) for independent external evaluation of algorithms based on artificial intelligence in diagnostics. *Digital Diagnostics* 2, 49–66.
- Payer, C., Štern, D., Bischof, H., Urschler, M., 2019. Integrating spatial configuration into heatmap regression based CNNs for landmark localization. *Medical image analysis* 54, 207–219.
- Pisov, M., Kondratenko, V., Zakharov, A., Petraikin, A., Gombolevskiy, V., Morozov, S., Belyaev, M., 2020. Keypoints localization for joint vertebra detection and fracture severity quantification, in: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, Springer International Publishing, Cham. pp. 723–732.
- Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks, in: *Advances in neural information processing systems*, pp. 91–99.

- Riggs, B.L., Melton Iii, L., 1995. The worldwide problem of osteoporosis: insights afforded by epidemiology. *Bone* 17, S505–S511.
- Roth, H.R., Wang, Y., Yao, J., Lu, L., Burns, J.E., Summers, R.M., 2016. Deep convolutional networks for automated detection of posterior-element fractures on spine CT, in: Medical Imaging 2016: Computer-Aided Diagnosis, International Society for Optics and Photonics. p. 97850P.
- Sekuboyina, A., Bayat, A., Husseini, M.E., Löffler, M., Li, H., Tetteh, G., Kukačka, J., Payer, C., Štern, D., Urschler, M., Chen, M., Cheng, D., Lessmann, N., Hu, Y., Wang, T., Yang, D., Xu, D., Ambellan, F., Amiranashvili, T., Ehlke, M., Lamecker, H., Lehnert, S., Lirio, M., de Olaguer, N.P., Ramm, H., Sahu, M., Tack, A., Zachow, S., Jiang, T., Ma, X., Angerman, C., Wang, X., Wei, Q., Brown, K., Wolf, M., Kirszenberg, A., Élodie Puybareauq, Valentinitisch, A., Rempfler, M., Menze, B.H., Kirschke, J.S., 2020a. VerSe: A vertebrae labelling and segmentation benchmark for multi-detector CT images. [arXiv:2001.09193](https://arxiv.org/abs/2001.09193).
- Sekuboyina, A., Husseini, M.E., Bayat, A., Löffler, M., Liebl, H., Li, H., Tetteh, G., Kukačka, J., Payer, C., Štern, D., et al., 2021. VerSe: a vertebrae labelling and segmentation benchmark for multi-detector CT images. *Medical image analysis* 73, 102166.
- Sekuboyina, A., Kukačka, J., Kirschke, J.S., Menze, B.H., Valentinitisch, A., 2017. Attention-driven deep learning for pathological spine segmentation, in: International Workshop and Challenge on Computational Methods and Clinical Applications in Musculoskeletal Imaging, Springer. pp. 108–119.
- Sekuboyina, A., Rempfler, M., Valentinitisch, A., Menze, B.H., Kirschke, J.S., 2020b. Labeling vertebrae with two-dimensional reformations of multidetector CT images: An adversarial approach for incorporating prior knowledge of spine anatomy. *Radiology: Artificial Intelligence* 2, e190074.
- Tomita, N., Cheung, Y.Y., Hassanpour, S., 2018. Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans. *Computers in biology and medicine* 98, 8–15.
- Valentinitisch, A., Trebeschi, S., Kaesmacher, J., Lorenz, C., Löffler, M., Zimmer, C., Baum, T., Kirschke, J., 2019. Opportunistic osteoporosis screening in multi-detector CT images via local classification of textures. *Osteoporosis international* 30, 1275–1285.
- Windsor, R., Jamaludin, A., Kadir, T., Zisserman, A., 2020. A convolutional approach to vertebrae detection and labelling in whole spine MRI, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, pp. 712–722.
- Yilmaz, E.B., Buerger, C., Fricke, T., Sagar, M.M.R., Peña, J., Lorenz, C., Glüer, C.C., Meyer, C., 2021. Automated deep learning-based detection of

osteoporotic fractures in CT images, in: International Workshop on Machine Learning in Medical Imaging, Springer. pp. 376–385.