Separable tissue representations for attributable risk prediction

Victor Wåhlstrand[1[0000−0001−6569−120X]], Jennifer Alvén[1[0000−0003−4195−9325]], Lisa Johansson[2[0009−0005−6005−3838]], Kristian Axelsson[2], Mattias Lorentzon[2[0000−0003−0749−1431]], and Ida Häggström[1,2[0000−0001−9178−6683]]

[1] Chalmers University of Technology, SE-41296 Göteborg, Sweden
[2] Sahlgrenska Academy at University of Gothenburg, SE-40530 Göteborg, Sweden
victor.wahlstrand@chalmers.se

Abstract. Attributing model predictions to a set of variables is a crucial part of methods in medicine and the sciences. However, in multimodal settings, ablating the contribution of a particular part of an image is often challenging. We present the STRAP framework (*separable tissue representations for attributable risk prediction*) using a novel masked autoencoder (MAE) enabling learning representations of a varying number of image patch tokens, enhancing memory efficiency and flexibility. We apply this framework on a fracture risk prediction task using clinical features and high-resolution peripheral quantitative computed tomography (HR-pQCT) images, to investigate the contribution of bone vs. muscle and fat tissues. Unlike previous work, we are able to selectively include specific tissues in risk prediction, and attribute their contribution to the risk using ablation and state-of-the-art interpretability methods. For the first time, we demonstrate that including soft-tissue from HR-pQCT increases prediction performance both in terms of C-index and overall AUC. Source-code is openly published online: https://github.com/waahlstrand/strap.

Keywords: risk prediction · representation learning · interpretability · attribution

1 Introduction

Osteoporosis-induced fractures are a serious complication from loss of bone mass, especially affecting elderly and women, and may severely increase mortality and reduce quality of life. Identifying individuals at risk is challenging but important to apply effective preventative measures [19]. Clinical assessment is commonly done by prediction models, such as Cox Proportional Hazards (CoxPH)[3] or the Fracture Risk Assessment Tool (FRAX®) [22], which typically estimate the fracture risk through a set of clinical risk factors like body mass index (BMI), age, history of fractures and bone mineral density (BMD) [12,11]. The lattermost is typically extracted from dual-energy X-ray (DXA) or computed tomography (CT) images, making imaging an important part of consultation, yet image data

is rarely included directly, and methods instead rely on similar pre-designed features. [21,4,23,10] Further, in traditional modelling, ablation of these variables is often used as a method of attributing a change in risk, and a tool to understand the factors of the underlying condition.

In the context of osteoporosis, the relation between the state of the bones, low BMD and corresponding fractures is well-established [12,9], and while the contribution of soft tissues (e.g. muscle, fat) may be modelled with extracted features such as BMI and DXA trochanteric soft tissue thickness [26,1,25], the authors are unaware of attempts to directly investigate the separate contribution of e.g. CT soft tissues to fracture risk prediction, without a priori designed features of interest.

1.1 Contributions

We propose the multimodal STRAP framework, *separable tissue representations for attributable risk prediction*, using novel modification on vision transformers to enable inclusion of certain ROIs in an images, increasing memory efficiency, improving performance and enabling attribution to specific areas in an image. We test the STRAP method on a survival analysis application, estimating the risk of fracture given tabular risk factors plus bone, muscle and fatty tissues in high-resolution peripheral quantitative computed tomography (HR-pQCT), where image size is a limiting factor. The main contributions of this work are:

a) a novel ViT architecture reducing memory use by means of a variable number of patch tokens from segmented regions,
b) flexible training and attribution of risk contributions from image ROIs,
c) and improved fracture risk prediction over baselines from HR-pQCT using only soft tissue representations.

1.2 Related works

Recent work on osteoporosis-induced fractures rely on extracting manual features or radiomics. Jaiswal et al. (2025) [10] show increased risk prediction performance across a number of machine learning algorithms, using finite element method measures from HR-pQCT bone tissues. Lu et al. (2023) [21] use texture metrics from HR-pQCT to improve fracture classification. Such methods are naturally interpretable, but limited, since the features must be known beforehand, as well as computable. Efforts to include full image data such as DXA or CT in risk prediction typically employ relatively simple methods, using convolutional neural networks (CNN). Kim et al. (2024) [15] use DenseNet modules along the imaging axes of a hip CT, aggregating the results by voting. Kong et al. (2020) [16] present a multimodal solution concatenating clinical risk factors and CNN features from spinal DXAs to predict vertebral risk fracture, and recently also applied a recurrent CNN to vertebral CT [17], visualizing network attribution using attention. However, these methods produce latent representations of the entire image, and are generally not able to separate attribution from specific

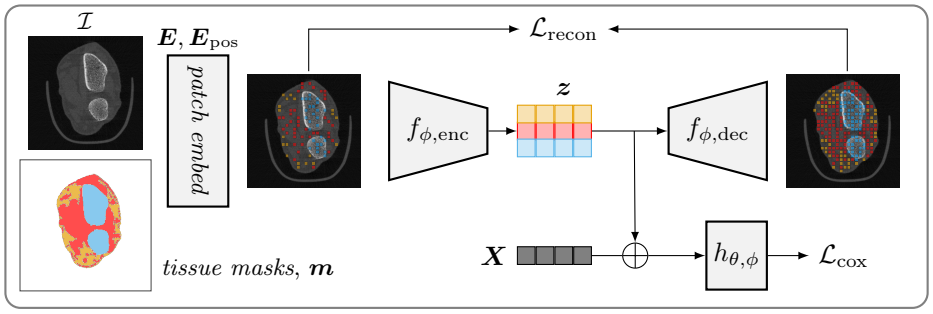

Fig. 1: *Overview of STRAP.* The STRAP approach, producing variable length sequences from ROI masks, using a custom ViT encoder and decoder to produce ROI-specific representations $\boldsymbol{z} = \{z_{\text{bone}}, z_{\text{muscle}}, z_{\text{fat}}\}$. These are combined with clinical risk factors $\boldsymbol{X}$ to predict survival.

parts of an image. Including parts of an image typically relies on cropping [35,6], which is not suitable when image features are embedded in each other. Methods such as Grad-CAM and occlusion [27,34] have made significant contributions to image attribution. Integrated gradients (IG) [28] enable a consistent framework for joint attribution of both image features and tabular data. Vision Transformers (ViT) [5] are compatible with many interpretability methods, natively using attention for attribution, but scale poorly with image size, especially noticeable in medical contexts, where images are typically very large.

Masked image modelling [32] and in particular masked autoencoders (MAEs) [8] are an adaptation of ViTs inspired by masked language modelling, using an encoder-decoder architecture to reconstruct image patch tokens. The input image is embedded as a fixed-length sequence, followed by discarding a high degree of masked patches. The decoder is tasked with reconstructing the discarded complement of the image given the remaining ones. However, MAE inherits the computational inefficiency of ViT, and requires full images. Attempts to accommodate varying size images or post-hoc token selection [30,24,33], still do not consider a varying number of input tokens (e.g. a ROI), even though this is the explicit objective of the original transformers [31].

2 Methodology

2.1 Preliminaries on risk prediction

Given a set of N patients with corresponding covariates $\boldsymbol{x}_i \in \mathbb{R}^D$ and fracture events $(\delta_i, t_i), i = 1, ..., N$ where $t_i \in]0, T]$ is the fracture event time until end of study T, and $\delta_i = \{0, 1\}$ is a binary indicator of an event or censoring, we wish to estimate a per-patient risk. Traditional methods for risk prediction estimate a log-hazard function $h_\theta : \mathbb{R}^D \to \mathbb{R}$ for ranking the patients at risk. For example, CoxPH assumes a linear, time-invariant hazard such that $h_\theta(\boldsymbol{x}) = \boldsymbol{x}^T\theta$. The

parameters θ are optimized by minimizing the negative partial log-likelihood,

$$\mathcal{L}_{\text{cox}} = -\sum_{i=1}^{N} \left(h_\theta(\boldsymbol{x}_i) - \log \sum_{j:\, t_i \geq t_j} \exp h_\theta(\boldsymbol{x}_j) \right). \tag{1}$$

Recent methods for risk prediction often use variations of DeepSurv [13], which models h_θ as a feed-forward neural network, or ConvDeepSurv, using a CNN as its hazard model. [36,14,2,6]

2.2 STRAP approach

An overview of our multimodal framework is seen in Fig. 1. Given an image $\mathcal{I} \in \mathbb{R}^{H \times W}$ and K ROI masks $M \in \{0,1\}^{K \times H \times W}$ with height H and width W, we construct two sets of patches with patch size P: image patches $\boldsymbol{x} : x_p \in \mathbb{R}^{K \times P^2}$ and mask patches $\boldsymbol{m} : m_p \in \{0,1\}^{K \times P^2}$, where $p = 1, ..., N$, $N = HW/P^2$. The patches x_p are selected if covered by the mask, such that the sequences are filtered based on a threshold hyperparameter τ, keeping only the patches x_p with τ ratio of mask pixels m_{ijp},

$$\boldsymbol{x}' = \{x_p : \frac{1}{P^2} \sum_{ij} m_{ijp} > \tau\} \tag{2}$$

resulting in filtered sequence lengths $N_k \leq N, k = 1, ..., K$. Like in ViT [5], the image patches are linearly transformed by a set of D-dimensional embedding layers $\boldsymbol{E} \in \mathbb{R}^{N_k \times D}$, followed by added positional embeddings $\boldsymbol{E}_{\text{pos}} \in \mathbb{R}^{N_k}$ corresponding to the coordinates of the patches in the original image, such that

$$\boldsymbol{z} = [x_{\text{cls}}, \boldsymbol{x}'\boldsymbol{E}] + \boldsymbol{E}_{\text{pos}} \tag{3}$$

is the initial set of tokens sent to the MAE, and x_{cls} is a specialized CLS token. Since ViTs have quadratic input complexity w.r.t. the number of tokens, using a smaller, variable number input tokens reduces the memory footprint significantly and enables e.g. larger batch sizes or higher resolution images - often not possible with standard ViTs. Like in MAE, these embeddings are passed through n transformer blocks, yielding an encoder embedding $f_{\phi,\text{enc}}(\boldsymbol{z}) = \boldsymbol{z}^{(n)}$. The representations $\boldsymbol{z}^{(n)}$ are subsequently passed to the decoder $f_{\phi,\text{dec}}$ for reconstruction using the mean squared error loss compared with ground truth patches, $\mathcal{L}_{recon}$.

STRAP variants. We consider three different variations on STRAP, and compare it with a simple analogue using a CNN.

Simultaneous training. Simultaneous training of a multilayer perceptron $h_{\theta,\phi}$ with Cox loss, Eq. (1), leads to the method we dub MAE-STRAP. We use the MAE approach to learn informative representations of each tissue ROI. This method is heavy computationally, since it trains both the encoder $f_{\phi,\text{enc}}$ and decoder $f_{\phi,\text{dec}}$, and thus limits the feasible batch size.

Self-supervision with probing. Training a MAE with both reconstruction loss and clinical risk factors with a supervised loss will inevitably lead to influence from the clinical risk factors on the tissue representations, which may be undesirable for interpretability. The conventional method of training MAE is with a self-supervised pretraining stage, followed by training the probing predictor neural network h_θ on top of the representations from the encoder. We will call this approach it MAE-STRAP-P.

The benefit of this approach is that it enables a greater batch size during the training of the predictor network, if that is desired, as well as representations uninfluenced by the clinical risk factors.

Only encoder. We also consider training the predictor $h_{\theta,\phi}$ using only our custom ViT encoder $f_{\phi,\text{enc}}$, without self-supervision to train the representations, increasing computational capacity by avoiding training the decoder at all. We simply call this method STRAP.

Masked CNN. Lastly, we construct an analogue for our task by simply masking images by each tissue, producing K images. Each masked image is passed through the CNN, yielding K separate sets of embeddings $\boldsymbol{z} \in \mathbb{R}^{K \times D}$, all of which are fed to a final predictor network. We call this approach CNN-STRAP.

3 Experiments and Results

3.1 Dataset

The Sahlgrenska University Hospital Prospective Evaluation of Risk of Bone Fractures (SUPERB) is a population-based study of 3028 older women (77.8 ± 1.6) from Gothenburg, Sweden [12,18,11,10] randomly selected from the national register, exploring the association of common risk factors and osteoporosis-induced fractures. HR-pQCT images were collected at up to two locations on the radius and the tibia, with a total of 11956 volumes. In this study, we consider the following clinical risk factors: age and BMI, parental history of hip fractures (FH), current smoker status, oral glucocorticoid use (OG), rheumatoid arthritis (RA), excessive alcohol consumption (EAC), secondary osteoporosis (SO) and femoral neck bone mineral density (BMD) t-scores. We evaluate the models on predicting any kind of fracture in terms of Harrell's and Uno's concordance (C) indexes [7,29], and cumulative time-dependent AUC, averaging model hazards over the sites, with a total of 36% incident events in the cohort. For computational reasons, we restrict ourselves to a slice in the middle of each image volume.

3.2 Model implementation and training

Fifty patients lacked tabular data or quality imaging, and were excluded from this study. We randomly held out 298 out of 2978 patients (10%) for final testing, and divided the remaining 2680 (90%) patients into 3 folds for cross-validation for model training and tuning. Age and BMI were standardized per fold.

Method	Attr	Input	Harrell's C	Uno's C	Mean AUC	3.5-year AUC
CNN-STRAP	✓	img.+clin.	0.629 ± 0.005	0.625 ± 0.006	0.652 ± 0.004	$\mathbf{0.657 \pm 0.009}$
MAE-STRAP	✓	img.+clin.	0.628 ± 0.002	0.623 ± 0.002	0.652 ± 0.012	0.656 ± 0.012
MAE-STRAP-P	✓	img.+clin.	0.626 ± 0.004	0.621 ± 0.003	0.648 ± 0.001	0.652 ± 0.001
STRAP	✓	img.+clin.	$\mathbf{0.632 \pm 0.004}$	$\mathbf{0.628 \pm 0.004}$	$\mathbf{0.655 \pm 0.004}$	$\mathbf{0.657 \pm 0.002}$
CoxPH [3]	✗	clin.	0.583 ± 0.003	0.588 ± 0.004	0.554 ± 0.002	0.563 ± 0.004
DeepSurv [13]	✗	clin.	0.587 ± 0.000	0.579 ± 0.001	0.561 ± 0.000	0.577 ± 0.000
ConvDeepSurv [36]	✗	img.+clin.	0.630 ± 0.004	0.624 ± 0.004	0.652 ± 0.009	0.654 ± 0.005

Table 1: *Baseline comparisons to our STRAP versions.* Columns indicate which models are tissue attributable and the model input (images and/or clinical variables). Results as mean±std over folds.

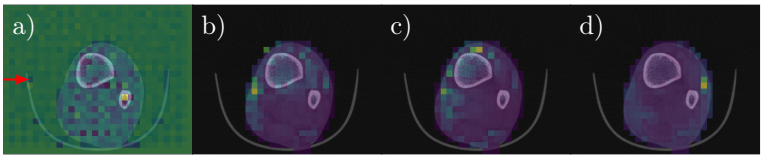

Fig. 2: *Visual attribution.* Absolute IG attribution of (a) ConvDeepSurv, showing tendency to focus on extraneous elements such as the leg/arm support (highlighted) and background , and (b) STRAP, (c) MAE-STRAP (d) MAE-STRAP-P constrained to the ROI.

All STRAP methods were trained with a patch size of 16, encoder embedding size 512 per tissue, decoder embedding size 1024, optimized with AdamW [20] (learning rate [LR] 5×10^{-4}, weight decay 5×10^{-3}) and batch size 8. MAE-STRAP and STRAP were trained for 200 epochs, and MAE-STRAP-P for 500 epochs, and patches were dropped randomly uniform with probability $p = 0.8$ and $p = 0.85$ respectively. During evaluation, the full set of tissue patches are encoded. Results are compared with CNN-STRAP trained for 150 epochs, using an embedding size of 512 per tissue (LR 1×10^{-4}, weight decay 5×10^{-3}) and batch size 8. We compare STRAP with the non-image-based CoxPH baseline ($\alpha = 0$, $N_{\text{iterations}} = 100$), using only clinical features, and a DeepSurv model (4 layers, sigmoid activation, LR 10^{-4}) for 50 epochs, as well as a baseline ConvDeepSurv with clinical variables but no masking (e.g. [36,16]), using a ResNet50 with 512 embedding size, (LR 5×10^{-4}) for 75 epochs.

3.3 Segmentation and preprocessing

The CT volumes were segmented using classical image analysis techniques. The images were resized from 1536×1536 to 512×512 for computational reasons, followed by initial segmentation of tissues and background using k-means at Hounsfield unit level. The leg/arm rest support was removed using successive erosion and dilation. Tissue segmentations were iteratively improved using dilation and hole-filling.

3.4 Model comparison

Table 1 shows the model ablations and baseline comparison results, indicating that the inclusion of HR-pQCT images increases predictive performance across all metrics. The STRAP approaches are all at a better or comparable level of performance compared to the baseline ConvDeepSurv, with added benefit of enabling a flexible number of tissues. Results also show that MAE-STRAP-P underperforms in terms of predictive capabilities compared to the best performing STRAP that uses only the novel encoder.

3.5 Interpretability and tissue attribution

Deep learning features have a distinct drawback, not being immediately interpretable, and encoding unknown image properties. We visualize attribution of the top-performing model STRAP and its self-supervised counterpart MAE-STRAP-P in image space and in representation space. Figure 2 shows a comparison of the absolute attribution from IG on ConvDeepSurv, STRAP, MAE-STRAP, and MAE-STRAP-P. While for ConvDeepSurv, attribution focuses on parts of the cortical bone, and outer parts of the tissues, it also yields undesired attributions on the leg/arms support (indicated with an arrow) and background. The STRAP methods are limited to the input patch tokens by design, but show different areas of attribution. The similarities between STRAP and MAE-STRAP vs. MAE-STRAP-P indicates that training with a decoder does not strongly influence the attributions.

At a representation level we aggregate the attributions of tissue representations. Figure 3a) shows the IG attribution for clinical variables and the sum of attributions from bone, muscle and fat, indicating agreement between models on clinical risk factors, and but not on the tissue representations, where fat have aggregated negative attribution to the log-hazard for MAE-STRAP-P, and positive for STRAP, and bone attribution is net zero. It is likely that different methods encode fundamentally different representations, requiring closer inspection. Figure 3b) shows the correlation between clinical features and tissues in MAE-STRAP-P, which shows that this method produces e.g. bone features with high correlation to BMD, in line with clinical intuition.

3.6 Tissue ablation and memory gains

We emphasize that the flexibility of our method allows us to train models informed by selected tissues. Using only bones leads to a 90% reduction in number of tokens. Since ViT has an input complexity of the input length squared, this yields an 80% reduction in memory use, by simply removing empty space. Since our framework enables an ablation of tissue variables, see Table 2, we can test the impact of soft tissues (muscle+fat) for prediction, especially in relation to the clinical BMD feature. By training models with only bone and soft tissues, but no BMD, we find that the signal of the BMD is likely strong enough to

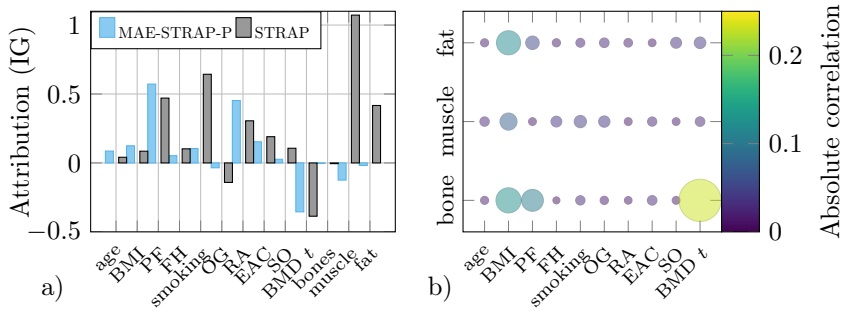

Fig. 3: *Feature contributions.* (a) Mean attribution, with methods agreeing on contribution from clinical features, but not from tissue representations. (b) Average abs. correlation between tissue representations and corresponding clinical features (MAE-STRAP-P), showing some clinically intuitive overlap.

bones	muscle+fat	BMD t	Harrell's C	Uno's C	Mean AUC	3.5-year AUC
✓	✓	✓	**0.632 ± 0.004**	**0.628 ± 0.004**	0.655 ± 0.004	0.657 ± 0.002
✗	✓	✓	0.631 ± 0.003	0.627 ± 0.001	**0.657 ± 0.009**	**0.660 ± 0.005**
✓	✗	✓	0.630 ± 0.003	0.625 ± 0.002	0.652 ± 0.007	0.658 ± 0.006
✓	✗	✗	0.604 ± 0.009	0.602 ± 0.006	0.608 ± 0.017	0.595 ± 0.013
✗	✓	✗	0.604 ± 0.008	0.604 ± 0.007	0.607 ± 0.017	0.593 ± 0.007
CoxPH		✗	0.563 ± 0.003	0.566 ± 0.001	0.548 ± 0.003	0.553 ± 0.003
		✓	0.583 ± 0.003	0.588 ± 0.004	0.554 ± 0.002	0.563 ± 0.004

Table 2: *Ablation of tissues.* Predictive performance of STRAP (mean± std over folds) from including soft tissues versus only bone features and BMD metrics.

supplant part of the contribution of the other variables, but Table 2 nonetheless shows that adding either tissue separately increases predictive power, and with the best model including all tissues and BMD. In conclusion, STRAP perform comparably or slightly better across metrics while providing meaningful attribution and greater flexibility.

4 Discussion and Conclusion

We present STRAP, a vision transformer modification using tissue masks of varying sizes for efficient computations and tissue differentiation. As demonstrated, STRAP enables calculation and analysis of separate tissue attributions to fracture prediction, improving interpretability over standard methods. Our results also show that multimodal inclusion of HR-pQCT and clinical risk variables yields better predictive performance of future fractures compared to standard methods, whilst also enabling the analysis of separate tissue contributions. Tissue ablations indicate that both bones and soft tissues carry predictive power. In addition, we found that STRAP features correlate to clinical variables as ex-

pected (e.g. bones representations to BMD and soft tissues to BMI), indicating that STRAP learns informative representations, encouraging future analysis.

One limitation of STRAP is its reliance on pre-segmented tissue masks. However, there are many classical approaches (like the one used here) and there exist many off-the-shelf segmentation models that can be integrated in a STRAP pipeline. Furthermore, we did not train a sophisticated aggregator over the imaging sites, but randomized site during training and averaged over all four during inference. Although we do not expect a learned aggregator to change the results significantly, this could be implemented in future work. Additionally, future work should adapt this work for full 3D volumes.

Finally, while this work focuses on bone, fat, and muscle tissue in CT for fracture risk prediction, our STRAP framework is general and broadly applicable for attributing multimodal survival to ROIs, e.g. cell types in histopathology, gene activation in spatial transcriptomics or areas of the brain in MRI.

5 Data Use Declaration and Acknowledgment

The collection of data for the SUPERB study was approved by the Ethical Review Authority Swedish Ethical Review Authority (DNR 929-12). The data is not publicly available as it contains sensitive health information. The source code and trained models can be downloaded for academic or non-commercial purposes.

Acknowledgments. This study was funded in part through the AIDA project grant DNR 2021-01420.

Disclosure of Interests. ML: lecture fees from Amgen, Astellas, Lilly, Meda, Renapharma, UCB Pharma; consulting fees: Amgen, Radius Health, UCB Pharma, Parexel International, Renapharma and Consilient Health. LJ: lecture fees from UCB Pharma. All other authors declare no conflicts of interest.

References

1. Bouxsein, M.L., Szulc, P., Munoz, F., Thrall, E., Sornay-Rendu, E., Delmas, P.D.: Contribution of trochanteric soft tissues to fall force estimates, the factor of risk, and prediction of hip fracture risk. J. Bone Miner. Res. **22**(6), 825–831 (2007)
2. Cheerla, A., Gevaert, O.: Deep learning with multimodal representation for pancancer prognosis prediction. Bioinformatics **35**(14), i446–i454 (2019)
3. Cox, D.R.: Regression models and life-tables. J. R. Stat. Soc. Ser. B Stat. Methodol. **34**(2), 187–220 (1972). https://doi.org/10.1111/j.2517-6161.1972.tb00899.x
4. De Vries, B.C.S., et al.: Comparing three machine learning approaches to design a risk assessment tool for future fractures: predicting a subsequent major osteoporotic fracture in fracture patients with osteopenia and osteoporosis. Osteoporos. Int. **32**(3), 437–449 (2021). https://doi.org/10.1007/s00198-020-05735-z
5. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv (2021), https://arxiv.org/abs/2010.11929

6. Hao, D., Li, Q., Feng, Q.X., Qi, L., Liu, X.S., Arefan, D., Zhang, Y.D., Wu, S.: Survivalcnn: A deep learning-based method for gastric cancer survival prediction using radiological imaging data and clinicopathological variables. Artif. Intell. Med. **134**, 102424 (2022)
7. Harrell Jr., F.E., Lee, K.L., Mark, D.B.: Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat. Med. **15**(4), 361–387 (1996). https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4
8. He, K., et al.: Masked autoencoders are scalable vision learners. arXiv (2021), https://arxiv.org/abs/2111.06377
9. Jaiswal, R., et al.: Increased bone material strength index is positively associated with the risk of incident osteoporotic fractures in older swedish women. J. Bone Miner. Res. **38**(6), 860–868 (2023). https://doi.org/10.1002/jbmr.4816
10. Jaiswal, R., et al.: Prediction of hip fracture by high-resolution peripheral quantitative computed tomography in older swedish women. J. Bone Miner. Res. p. zjaf020 (2025). https://doi.org/10.1093/jbmr/zjaf020
11. Johansson, L., et al.: Improved fracture risk prediction by adding VFA-identified vertebral fracture data to BMD by DXA and clinical risk factors used in FRAX. Osteoporos. Int. **33**(8), 1725–1738 (2022). https://doi.org/10.1007/s00198-022-06387-x
12. Johansson, L., et al.: Grade 1 vertebral fractures identified by densitometric lateral spine imaging predict incident major osteoporotic fracture independently of clinical risk factors and bone mineral density in older women. J. Bone Miner. Res. **35**(10), 1942–1951 (2020). https://doi.org/10.1002/JBMR.4108
13. Katzman, J.L., et al.: Deepsurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med. Res. Methodol. **18**(1) (2018). https://doi.org/10.1186/s12874-018-0482-1
14. Kim, D.W., Lee, S., Kwon, S., Nam, W., Cha, I.H., Kim, H.J.: Deep learning-based survival prediction of oral cancer patients. Scientific reports **9**(1), 6994 (2019)
15. Kim, Y., , et al.: A CT-based deep learning model for predicting subsequent fracture risk in patients with hip fracture. Radiology **310**(1), e230614 (2024). https://doi.org/10.1148/radiol.230614
16. Kong, S.H., et al.: Development of a spine X-ray-based fracture prediction model using a deep learning algorithm. Endocrinol. Metab. **37**(4), 674–683 (2022). https://doi.org/10.3803/EnM.2022.1461
17. Kong, S.H., et al.: A computed tomography–based fracture prediction model with images of vertebral bones and muscles by employing deep learning: Development and validation study. J. Med. Internet Res. **26**, e48535 (2024). https://doi.org/10.2196/48535
18. Lorentzon, M., et al.: Extensive undertreatment of osteoporosis in older Swedish women. Osteoporos. Int. **30**(6), 1297–1305 (2019). https://doi.org/10.1007/s00198-019-04872-4
19. Lorentzon, M.: Treating osteoporosis to prevent fractures: current concepts and future developments. J. Int. Med. **285**(4), 381–394 (2019). https://doi.org/https://doi.org/10.1111/joim.12873
20. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv (2019), https://arxiv.org/abs/1711.05101
21. Lu, S., et al.: Machine learning applied to HR-pQCT images improves fracture discrimination provided by DXA and clinical risk factors. Bone **168**, 116653 (2023). https://doi.org/10.1016/j.bone.2022.116653

22. McCloskey, E.V., et al.: Fracture risk assessment by the FRAX model. Climacteric **25**(1), 22–28 (2022). https://doi.org/10.1080/13697137.2021.1945027
23. Michalski, A., et al.: Opportunistic CT screening predicts individuals at risk of major osteoporotic fracture. Osteoporos. Int. **32**(8), 1639–1649 (2021). https://doi.org/10.1007/s00198-021-05863-0
24. Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., Hsieh, C.J.: Dynamicvit: Efficient vision transformers with dynamic token sparsification (2021), https://arxiv.org/abs/2106.02034
25. Robinovitch, S.N., McMahon, T.A., Hayes, W.C.: Force attenuation in trochanteric soft tissues during impact from a fall. J. Ortho. Res. **13**(6), 956–962 (1995)
26. Schacter, I., Leslie, W.D.: Estimation of trochanteric soft tissue thickness from dual-energy x-ray absorptiometry. J. Clin. Densitom. **17**(1), 54–59 (2014)
27. Selvaraju, R.R., et al.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. IJCV **128**(2), 336–359 (2019). https://doi.org/10.1007/s11263-019-01228-7
28. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. arXiv (2017), https://arxiv.org/abs/1703.01365
29. Uno, H., et al.: On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Statistics in Medicine **30**(10), 1105–1117 (2011). https://doi.org/10.1002/sim.4154
30. Varma, A., et al.: Varivit: A vision transformer for variable image sizes. In: Medical Imaging with Deep Learning (2024)
31. Vaswani, A., et al.: Attention is all you need. In: NeurIPS. p. 6000–6010. NeurIPS, Curran Associates, Inc., Red Hook, NY, USA (2017)
32. Xie, Z., et al.: SimMIM: A simple framework for masked image modeling. arXiv (2022), https://arxiv.org/abs/2111.09886
33. Yin, H., Vahdat, A., Alvarez, J., Mallya, A., Kautz, J., Molchanov, P.: A-ViT: Adaptive tokens for efficient vision transformer. In: Proc. CVPR (2022)
34. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. arXiv (2013), https://arxiv.org/abs/1311.2901
35. Zheng, Y., et al.: Preoperative CT-based deep learning model for predicting overall survival in patients with high-grade serous ovarian cancer. Front. Oncol. (2022). https://doi.org/10.3389/fonc.2022.986089
36. Zhu, X., Yao, J., Huang, J.: Deep convolutional neural network for survival analysis with pathological images. In: Proc. IEEE BIBM. pp. 544–547 (2016). https://doi.org/10.1109/BIBM.2016.7822579