

Probabilistic Point Cloud Reconstructions for Vertebral Shape Analysis

Anjany Sekuboyina, Markus Rempfler, Alexander Valentinitich,
Maximilian Loeffler, Jan S. Kirschke*, and Bjoern H. Menze*

Technische Universität München, Munich, Germany

Abstract

We propose an auto-encoding network architecture for point clouds (PC) capable of extracting shape signatures without supervision. Building on this, we (i) design a loss function capable of modelling data variance on PCs which are unstructured, and (ii) regularise the latent space as in a variational auto-encoder, both of which increase the auto-encoders' descriptive capacity while making them probabilistic. Evaluating the reconstruction quality of our architectures, we employ them for detecting vertebral fractures without any supervision. By learning to efficiently reconstruct only healthy vertebrae, fractures are detected as anomalous reconstructions. Evaluating on a dataset containing ~ 1500 vertebrae, we achieve area-under-ROC curve of $>75\%$, without using intensity-based features.

1 Introduction

One of the consequences of the numerous algorithms proposed for segmenting organs, tissues, the spine etc. involves analysing their anatomical shapes, eventually contributing towards population studies [6], disease characterisation [10], survival analysis [7], etc. Employing convolutional neural networks (CNN) for this task involves processing voxelised data due to its Euclidean nature. Such voluminous representation, however, is inefficient, especially when the masks are binary and the *shape information* corresponds to its surface profile. Alternatively, surface meshes (a collection of vertices, edges, and faces) or active contours could be used. Since the data is no longer Euclidean, a conventional CNN is unusable. Graph convolutional networks (GCN) [3] were thus developed by redefining the notion of 'neighbourhood' and 'convolution' for meshes and graphs. However, if the number of nodes is high, GCNs (esp. spectral) become

*Joint supervising authors.

bulky. Moreover, each mesh is treated as a domain, making mesh registration a requisite.

An alternative surface representation is a set of 3D points in space, referred to as the **point clouds** (PC). A PC represents the surface just with a set of N vertices, thus avoiding both the cubic-complexity of voxel-based representations and the $N \times N$ dimensional, sparse, adjacency matrix of meshes. However, despite their representational effectiveness, PCs are permutation invariant and do not describe data on a structured grid, preventing the usage of standard convolution. To this end, we work with an architecture capable of processing PCs (*point-net*, [9]), and design a network capable of reconstructing PCs thereby extracting shape signatures in an unsupervised manner.

1.1 Uncertainty and latent space modelling

Unlike supervised learning on PCs [5], we set out to obtain shape signatures from PCs without supervision, building towards a relatively less explored topic of *auto-encoding* point clouds. This involves mapping the PC to a latent vector and reconstructing it back. Since the PCs are unordered, PC-specific reconstruction losses replace traditional ones [4]. Extending an auto-encoder (AE) based on such a loss, we propose to improve its representational capacity by regularising the latent space to make it compact and by modelling the variance that exists in a PC population. We claim that this results in learning improved shape signatures, validating the claim by employing the extracted features for unsupervised vertebral fracture detection.

1.2 Vertebral fracture detection

There exists an inherent shape variation in vertebral shapes within the spine of a single patient (e.g. cervical–thoracic–lumbar) along with a natural variation in a vertebra’s shape in a population (e.g. L1 across patients, cf. Fig. 1). Additionally, osteoporotic fractures start without significant shape change and progress into a vertebral collapse. Hence, fracture detection in vertebrae is non-trivial. Added to this, limited availability of fractured vertebrae makes the learning of supervised classifiers non-trivial. In literature, several classification systems exist mainly based on vertebral height measurement [2] or analysing sub-regions of the spine in sagittal slices [11]. However, an explicit shape-based approach seems absent. Evaluating the representational ability of the proposed AE architectures, we seek to analyse vertebral shapes and eventually detect vertebral fractures using the extracted latent shape features.

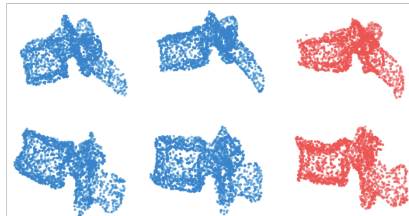


Figure 1: **Variation among vertebral shapes:** Compare the higher variation between healthy (blue) vertebrae of different classes (T3, top and L1, bottom) w.r.t the relatively lower variation within-class between fractured (red) and healthy vertebrae.

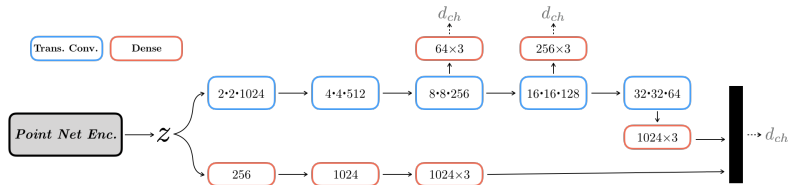


Figure 2: **Point Cloud Auto-encoder (pAE)**: Architectural details of decoding path constructing a point cloud from a latent vector. Top arm is convolutional while bottom arm is fully-connected. Transposed convolution ($\dashv\dashv\dashv$ channels) have a stride of 2. Since encoder is an adapted *point-net*[9], we detail its architecture in the supplement.

1.3 Our contribution

Summarising the contributions of this work: (1) We build on existing point-net-based architectures to propose a point-cloud auto encoder (*pAE*). (2) Reinforcing this architecture, we incorporate latent space modelling and a more challenging uncertainty quantification. (3) We present a comprehensive analysis of the reconstruction capabilities of our *pAE*s by investigating their utility in detecting vertebral fractures. We work with an in-house, clinical dataset (~ 1500 vertebrae) achieving an area-under-curve (AUC) of $>75\%$ in detecting fractures, even without employing texture or intensity-based features.

2 Methodology

We present this section in two stages: First, we introduce the notation used in this work and describe a point-net-based architecture capable of efficiently auto-encoding point clouds. Second, we build on this architecture to model the natural variance in vertebrae while regularising the latent space.

2.1 Auto-encoding point clouds

Given accurate voxel-wise segmentation of a vertebra, a *point cloud* (PC) can be extracted as a set of N points represented by $X = \{p_i\}_{i=0}^N$, where p_i represents a point by its 3D coordinate (x_i, y_i, z_i) . Additionally, p_i could also represent other point specific features such as normal, radius of curvature etc. So, each vertebra is represented by a PC of dimension $N \times m$ (in this work, $N = 2048$ vertices and $m = 3$ coordinates, with the vertices randomly subsampled from a higher resolution mesh). Recall the lack of a regular coordinate space associated with the point cloud and that any permutation of these N points represents the same point cloud. This requires incorporation of a unique variant of deep networks for processing PCs.

2.1.1 Architecture.

An AE consists of an encoder mapping the PC to the latent vector and a decoder reconstructing the PC back from this latent vector, i.e $X \mapsto z \mapsto X$. As

the encoder, we employ a variant of the point-net architecture [9]. The latent vector, z , respects the permutation invariance of the PC and represents its shape signature. As a decoder, taking cues from [4], we construct a combination of an up-convolutional and dense branches taking z as input and predicting \hat{X} , the reconstructed X . The convolutional path, owing to its neighbourhood processing, models the ‘average’ regions, while the dense path reconstructs the finer structures. This combination of the point-net and the decoder forms our point cloud auto-encoding (p AE, or interchangeably AE) architecture as illustrated in Fig. 2.

2.1.2 Loss.

Reconstructing point clouds requires comparing the predicted PC with the actual PC to back-propagate the loss during training. However, owing to the unordered nature of PCs, usual regression losses cannot be employed. Two prominent candidates for such a task are the Chamfer distance and the Earth Mover (EM) distance [4]. We observed that minimising EM distance ignores the natural variation in shapes (e.g. the processes of the vertebrae) and reconstructs only a mean representation (e.g. the vertebral body), as validated in [4]. Since we intend to model the natural variance in the data, using EM distance is undesirable in our case. We thus employ the Chamfer distance computed as:

$$d_{ch}(X, \hat{X}) = \mathcal{L}_{ae} = \sum_{p \in X} \min_{\hat{p} \in \hat{X}} \|p - \hat{p}\|_2^2 + \sum_{\hat{p} \in \hat{X}} \min_{p \in X} \|p - \hat{p}\|_2^2. \quad (1)$$

In essence, d_{ch} is the distance between a point in X and its nearest neighbour in \hat{X} and vice versa.

2.2 Probabilistic reconstruction

From a generative modelling perspective, an AE can be seen to predict the parameters of Gaussian distribution imposed on X , i.e. $p_{\Theta}(X) = \mathcal{N}(X|\hat{X}, \hat{\Sigma})$, parameterised by the weights of the AE denoted by Θ . Determining the distribution parameters, viz. optimising for the AE weights, now involves maximising the log-likelihood of X , resulting in:

$$\Theta^* = \arg \max_{\Theta} \log p_{\Theta}(X) = \arg \min_{\Theta} \frac{1}{2}(X - \hat{X})^T \hat{\Sigma}^{-1}(X - \hat{X}) + \frac{1}{2} \log |\hat{\Sigma}|. \quad (2)$$

This perspective towards auto-encoding enables us to extend the p AE to encompass the data variance ($\hat{\Sigma}$) while modelling the latent space, as described in following sections. It is important to note that the difference $X - \hat{X}$ is not well defined for point clouds, requiring us to opt for alternatives.

Assuming $\Sigma = \mathbb{I}$, implying an independence among the elements of X and an element-wise unit variance, results in the familiar mean squared error (MSE), $\mathcal{L} = \|X - \hat{X}\|^2$. Based on the parallels between MSE and the Chamfer distance (Eq. 1), we design σ -AE and σ -VAE, as illustrated in Fig 3.

2.2.1 σ -AE.

The assumption of unit covariance, as in AE, is inherently restrictive. However, modelling an unconstrained covariance matrix is infeasible due to quadratic complexity. A practical compromise is the independence assumption. Thus, representing covariance as, $\Sigma = \text{diag}\{\hat{\sigma}_{p_1}^2, \dots, \hat{\sigma}_{p_i}^2, \dots, \hat{\sigma}_{p_N}^2\}$, where $\hat{\sigma}_{p_i}^2$ denotes the variance corresponding to p_i , eq. (2) morphs to a loss function as:

$$\mathcal{L} = \sum_{\hat{p} \in \hat{X}} \sigma_{\hat{p}}^{-2} \|p_i - \hat{p}_i\|^2 + \log \sigma_{\hat{p}}^2 \quad (3)$$

This optimisation models the aleoteric uncertainty [8]. Eq. 3 is an attenuated MSE, where a high variance associated to a point down-weighs its contribution to the loss. However, due to the lack of a reference grid in the point cloud space, the notion of uncertainty being associated to a data point (eg. pixel, spatial location etc.) is absent. We propose to associate the notion of variance to every point, \hat{p}_i . This results in the variance-modelling Chamfer distance:

$$\mathcal{L}_{\sigma ae} = \sum_{p \in X} \min_{\hat{p} \in \hat{X}} \sigma_{\hat{p}}^{-2} \|p - \hat{p}\|_2^2 + \sum_{\hat{p} \in \hat{X}} \sigma_{\hat{p}}^{-2} \min_{p \in X} \|p - \hat{p}\|_2^2 + \log \sigma_{\hat{p}}^2 \quad (4)$$

Observe the slight abuse of notation in Eq. 4, wherein the variance at a predicted point, $\sigma_{\hat{p}}$, actually represents the variance of the coordinate elements of p , i.e $\{\sigma_{\hat{x}}, \sigma_{\hat{y}}, \sigma_{\hat{z}}\}$. Current notation is chosen to avoid clutter.

2.2.2 Variational and σ -Variational AE.

An alternative approach for modelling $p(X)$ involves modelling its dependency over a latent variable z , which is distributed according to a known prior $p(z)$. A variational auto-encoder (VAE) operates on these principles and involves maximising a lower bound on the log-evidence (referred to as ELBO) of the data described as below:

$$\log p(X) \geq \mathbb{E}_{z \sim q_\phi(z|X)} [\log p_\theta(X|z)] - \text{KL}[q_\phi(z|X) || p_\theta(z)], \quad (5)$$

where $q_\phi(z|x)$ is the approximate posterior of z learnt by the encoder and parameterised by ϕ . $p_\theta(X|z)$ is the data likelihood modelled by the decoder and parameterised by θ . $p_\theta(z)$ is the prior on z .

Maximising ELBO is equivalent to maximising the log-likelihood of X while minimising the Kullback-Leibler divergence between the approximate and true prior. Representing the combination as $\mathcal{L}_{rec} + \beta \mathcal{L}_{KL}$, where \mathcal{L}_{rec} is the reconstruction loss seen in earlier sections. β is a scaling factor weighing the

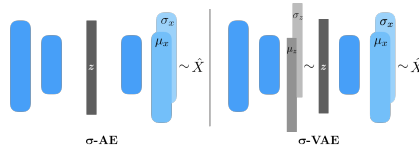


Figure 3: **Probabilistic reconstruction architectures:** \sim indicates a sampling operation. Since a point’s variance has a smaller scale compared to its mean, the variance is predicted using a softplus activation (added with $\epsilon = 10^{-6}$ for stabilising divisions) and uses a layer parallel to the one predicting the mean.

contribution of the two losses appropriately. Standard practice assigns Gaussian distributions for $q_\phi(z|x) \sim \mathcal{N}(z|\boldsymbol{\mu}_z, \boldsymbol{\sigma}_z)$ and $p(z) \sim \mathcal{N}(z|\mathbf{0}, \mathbf{1})$ (cf. Fig. 3). Thus, \mathcal{L}_{KL} models the latent space to follow a Gaussian distribution inline with the prior. Incorporating this into the point cloud domain, results in an objective function for a PC-based VAE (or σ -VAE) as $\mathcal{L}_{vae} = \mathcal{L}_{ae/\sigma ae} + \beta\mathcal{L}_{KL}$. Thus, σ -VAE acts as a AE capable of modelling the data variance while regularising the latent space. The prior on the latent space also imparts point cloud generation capabilities to σ -VAE.

2.3 Detecting fractures as anomalies

Examining the descriptive ability of our p AE architectures in auto-encoding PCs, we utilise them for an application of vertebral shape analysis: fracture detection in vertebrae. Assuming the AE is trained only on ‘normal’ patterns, a fracture can be detected as an ‘anomaly’ based on its ‘position’ in latent space. We inspect two measures for this purpose:

1. Reconstruction error or Chamfer distance: AEs trained on healthy samples fail to accurately reconstruct anomalous ones, resulting in a high d_{ch} .
2. Reconstruction probability or likelihood [1]: Expected likelihood $\mathbb{E}[p_\Theta(X)]$ of an input can be computed for σ - architectures (cf. Eq. 2). For any input PC, X_{in} , the likelihood is computed by $\mathcal{N}(X_{in}|\mu_\Theta, \Sigma_\Theta)$ with the predicted mean and variances. We expect fractured vertebrae to be less *likely* than healthy ones.

Intuitively, relying on the reconstruction error or likelihood for detecting anomalies requires the learnt ‘healthy’ latent space to be representative. Both σ -AE and the VAE work towards this objective. In σ -AE, predictive variance down-weighs the loss due to highly uncertain points in the PC. This suppresses the interference due to natural variation in the vertebral PCs. On the other hand, VAE acts directly on the latent space by modelling the encoding uncertainty ($X \mapsto z$). The σ -VAE encompasses both these features.

2.3.1 Inference.

During inference, the given vertebral PC is reconstructed and the reconstruction error and (or) likelihood are computed. This vertebra is said to be fractured if the reconstruction error is greater than a threshold, T_{rec} , or its likelihood is lesser than a threshold, T_l . The optimal values for these thresholds are chosen on a validation set.

3 Experiments & Discussion

We present this section in two parts: first, we explore the auto-encoding, variance modelling, and generative capabilities of our AE networks. Second, we deploy these architecture to detect vertebral fractures without supervision.

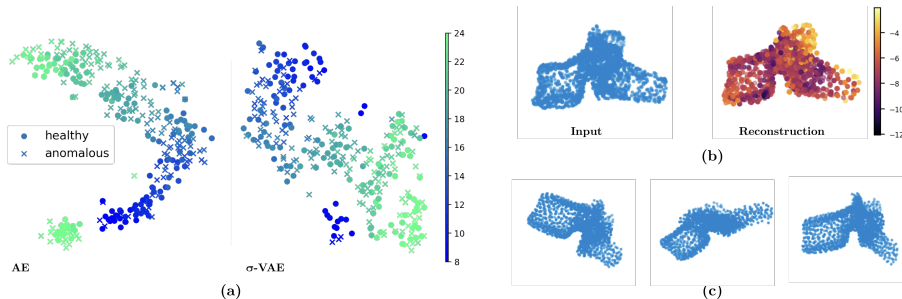


Figure 4: **Characteristics of σ -VAE:** (a) Comparison of TSNE embeddings of simple p AE with σ -VAE. Observe transition in clusters being inline with vertebral indices. Note that embedding becomes compact for the VAE. (b) A PC and its reconstruction coloured with $\log(\sigma^2)$ of every point. Observe high variance in vertebral processes. (c) Example generations from decoder with $z \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$.

Data preparation: We evaluate our architecture on an in-house dataset with accurate voxel-level segmentations converted into PCs. The dataset consists of 1525 healthy (validation & test subset: 50+100) and 155 fractured vertebrae (55+100 subset). For supervised learning, the data split changes to add fractured vertebrae to the train set. Thus, healthy and fractured subsets are altered to (50+55) and (15+55).

Training: The architecture of the encoder and the decoder is similar across all architectures (cf. Fig 3) except for the layers predicting variance. PCs are augmented online by perturbing the points with Gaussian noise and random rotations (± 15 deg). Finally, the PCs are median-centred to origin and normalised to have the same surface area. The networks are trained until convergence using an Adam optimiser with an initial learning rate of 5×10^{-4} . Specific to the VAE, we use KL-annealing by increasing β from 0 to 0.1.

3.1 Qualitative evaluation of AE architectures.

We investigate if meaningful shape features can be learnt without supervision. Validating this, in Fig. 4a, we plot a TSNE embedding of the test set latent vectors learnt by a naive p AE and σ -VAE trained only on healthy vertebrae. Observe the clusters formed based on the vertebral index and the transition between the indices. This corresponds to the natural variation of vertebral shapes in a human spine. Indicating the fractured vertebrae in the embedding, we highlight their degree of similarity with the healthy counterparts. Also, observe that embedding is more regularised representing a Gaussian in case of σ -VAE, indicating the continuity of the learnt latent space. Fig. 4b shows the predictive variance modelled by the σ -VAE. Posterior elements of a vertebrae are the most varying among population. Observe this being captured by the variance in the vertebral process regions. Lastly, illustrating σ -VAE’s generative capabilities, Fig. 4c shows vertebral PC samples generated by sampling the latent vector, $z \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$.

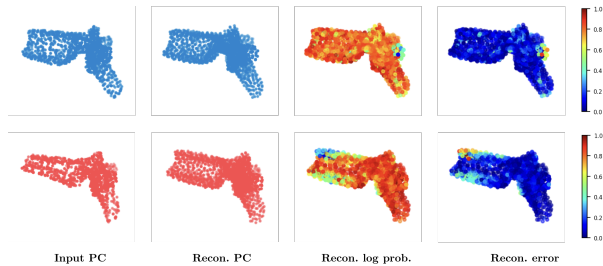


Figure 5: Probabilistic reconstruction of a healthy (top) and a fractured (bottom) vertebral point cloud. Observe p AE’s ‘healthy’ reconstruction of the fractured vertebra (normalised $[0,1]$ within PC).

Table 1: Performance comparison of unsupervised and supervised fracture detection approaches. Measures: Precision (P), Recall (R), F1-score, and Area-Under-ROC Curve (AUC) computed by varying thresholds on recon. error and recon. log-probabilities. Since supervised models have no threshold selection, AUC is not reported.

Measures	<i>recon. error</i>				<i>recon. log-likelihood</i>			
	PN	PN _{bal}	AE	VAE	σ -AE	σ -VAE	σ -AE	σ -VAE
<i>P</i>	100±0.0	68.6±3.4	57.6±4.1	61.1±1.9	67.1±6.5	68.4±3.3	62.3±4.3	61.6±1.4
<i>R</i>	13.9±3.1	57.6±7.5	85.0±9.8	79.0±3.6	74.3±4.0	71.7±4.1	72.7±6.1	79.7±2.5
<i>F1</i>	24.7±4.7	62.5±5.8	68.0±0.9	68.5±1.7	67.5±5.1	69.6±1.2	66.7±1.3	69.5±0.6
AUC	n.a	n.a	70.8±2.2	74.8±3.0	75.9±2.0	75.9±1.5	70.2±2.2	73.8±2.0

3.2 Vertebral fracture detection.

Evaluating the reconstruction quality of our p AE architectures, we employ them to detect fractures as anomalies. As baselines, we choose two supervised approaches: (1) point-net (PN), the encoding part in our p AE architectures, cast as a binary classifier and (2) the same point-net trained with median frequency balancing the classes (ref. as PN_{bal}) so as to accentuate the loss from minority fractured class. We report their performance in Table 1, over 3-fold cross-validation while retaining the ratio of healthy to fractured vertebrae in the data splits. Observe that frequency balancing improves the F1 score significantly, albeit not at the level of the proposed anomaly detection schemes.

Reconstruction for fracture detection: When detecting fractures based on reconstruction error (d_{ch}), we observe that a naive p AE already out-performs the supervised classifiers (cf. Table 1). On top of this, we see that latent space modelling and variance modelling individually offer an improvement in F1-scores while increasing the AUC, indicating a stable detection of fractures. The performance of both σ -AE and σ -VAE is similar indicating the role of loss attenuation. However, the advantage of explicitly regularising the latent space for σ -VAE can be seen in likelihood-based anomaly detection, where σ -VAE outperforms σ -AE. Fig. 8 compares a reconstruction of a healthy and fractured vertebrae of the same vertebral level. Note the high reconstruction error and a low log-likelihood spatially corresponding to the deformity due to fracture.

4 Conclusions

We presented point-cloud-based auto-encoding architectures for extracting descriptive shape features. Improving their description, we incorporated variance and latent space-modelling capability using specially defined PC specific losses. The former captures the natural variance in the data while the latter regularises the latent space to be continuous. Deploying these networks for the task of unsupervised fracture detection, we achieved an AUC of 76% without using any intensity or textural features. Future work will combine the extracted shape signatures with textural features e.g. bone density and trabecular texture of vertebrae to perform fracture-grade classification. Furthermore, the ability of our σ -(V)AE models in capturing other sources of variability, e.g inter-rater or scan resolution, is to be investigated.

References

- [1] An, J., Cho, S.: Variational autoencoder based anomaly detection using reconstruction probability. Tech. rep., SNU Data Mining Center (2015)
- [2] Baum, T., et al.: Automatic detection of osteoporotic vertebral fractures in routine thoracic and abdominal mdct automatic detection of osteoporotic vertebral fractures in routine thoracic and abdominal mdct. *Eur Radiology* (2014)
- [3] Bronstein, M.M., et al.: Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine* (2017)
- [4] Fan, H., et al.: A point set generation network for 3d object reconstruction from a single image. In: *CVPR* (2017)
- [5] Gutiérrez-Becker, B., et al.: Deep multi-structural shape analysis: Application to neuroanatomy. In: *MICCAI*. Springer (2018)
- [6] Ingalhalikar, M., et al.: Sex differences in the structural connectome of the human brain. *Proceedings of the National Academy of Sciences* (2014)
- [7] Isensee, F., et al.: Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge. arXiv e-prints arXiv:1802.10508 (2018)
- [8] Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: *NIPS* (2017)
- [9] Qi, C.R., et al.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: *CVPR* (2017)
- [10] Shakeri, M., et al.: Deep spectral-based shape features for alzheimer’s disease classification. In: *SeSAMI@MICCAI* (2016)
- [11] Tomita, N., et al.: Deep neural networks for automatic detection of osteoporotic vertebral fractures on ct scans. *Computers in Biology and Medicine* (2018)

Supplement

As supplementary content, we present: (1) A detailed description of the complete point-cloud auto-encoder, including the encoder architecture adapted from *point net* [1] (cf. Fig. 6), (2) additional illustrations of point-wise data uncertainty modelled by the proposed σ -VAE (cf. Fig. 7), and (3) Further qualitative results comparing probabilistic reconstructions of healthy and anomalous or fractured vertebrae, along with point-wise Chamfer distance and log-probability between the input and its reconstruction (cf. Fig. 8).

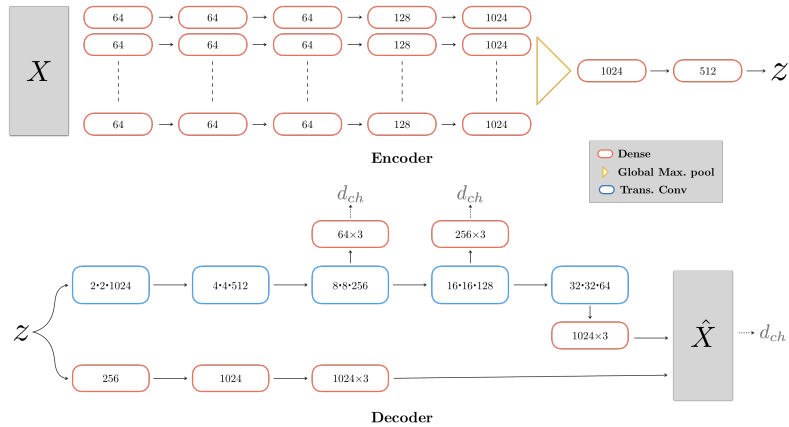


Figure 6: **Architecture for pAE:** Architectural details of encoding and decoding paths of the pAE. Note that every layer (except the last, in encoder and in the decoder) is followed by batch normalisation and leaky ReLU. The values in the dense layers indicate the number of nodes while the values in the transposed convolution layers ($\cdot \cdot \cdot$ channels) indicate the size of the resulting feature map. For example, the first transposed convolution layer, z is reshaped to $1 \cdot 1 \cdot 64$ and up-convolved to $2 \cdot 2 \cdot 1024$.

References

- [1] Qi, C.R., et al.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: CVPR (2017)
- [2] Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: NIPS (2017)

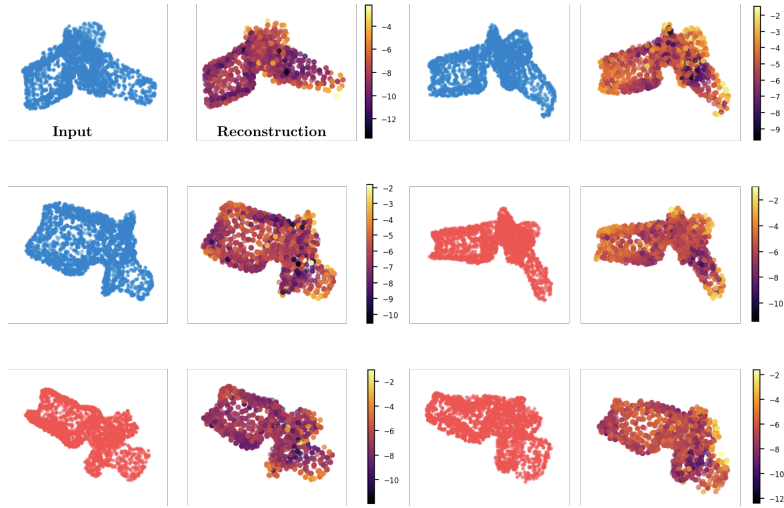


Figure 7: **Variance modelling** by the proposed σ -VAE for healthy (blue) and fractured (red) cases. Observe a higher variance in the vertebral processes representing the naturally occurring shape variance in a population. Note the lack of high uncertainty values for fractured vertebrae, in accordance with aleatoric uncertainty’s property of capturing only data variance [2]. Hence, predictive variance cannot be used as a means to detect fractures. However, it can reliably be employed as an attenuation factor for improving reconstruction or for computing the reconstruction probability (cf. Fig. 8), thereby enabling fracture detection.

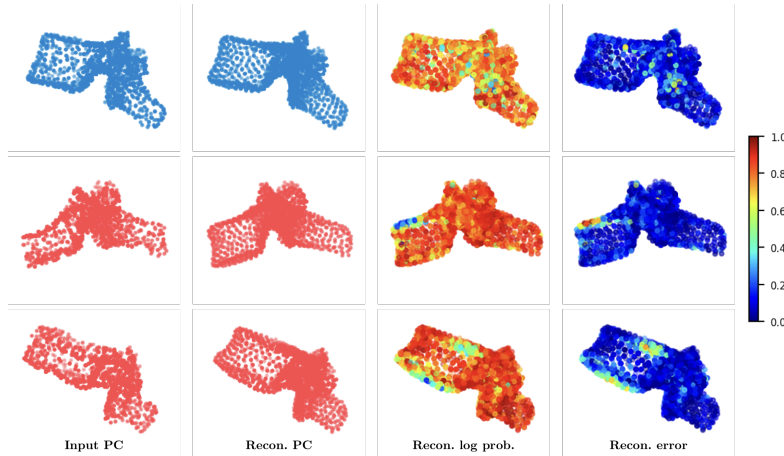


Figure 8: **Probabilistic reconstructions** of healthy and fractured vertebrae. Alongside spatial localisation of fractures, compare the dynamic range of the reconstruction log probability between healthy and fractured cases when normalised to $[0,1]$. Reconstruction probabilities are relatively uniformly-spread in the healthy case and are pushed to the extremes for a fractured one. This indicates a higher dynamic range in the unnormalised values while reconstructing fractured vertebrae. Thus, a low reconstruction probability (or a high reconstruction error) does indicate an outlier or a fracture.