Effects of Input Representation of Bone Shapes on Latent Space Organization

by

Luca Goemans

To obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on September 17th, 2025 at 12.30 PM.

Student number:	4954645
Thesis advisor:	Jesse Krijthe
Daily supervisor:	Gijs van Tulder
External committee member:	Thomas Höllt
Project duration:	January, 2025 – September, 2025

An electronic version of this thesis is available at `http://repository.tudelft.nl/`

Effects of Input Representation of Bone Shapes on Latent Space Organization

Luca Goemans

September 2025

Abstract

Osteoarthritis (OA) is a prevalent musculoskeletal disease, and radiographic assessment remains the standard for diagnosis and grading. However, expert grading is subjective and intensity-based automated methods are sensitive to imaging variability. As a potential solution to these problems, landmark-based approaches are worth exploring. Landmark-based representations of bone geometry offer an alternative to pixel-based inputs, reducing sensitivity to imaging artifacts and emphasizing structural variation. This thesis compares four landmark encodings (raw x,y coordinates, Procrustes-aligned points, pairwise distances, and polar coordinates) and evaluates them using both linear dimensionality reduction (PCA) and nonlinear generative modeling (VAEs) on hip radiographs from a publicly available dataset. We evaluate reconstruction fidelity, latent space traversal, correlation with clinical outcomes, and classification performance. Results show that raw point coordinates provide a strong baseline, often matching or outperforming more complex encodings in classification, while alternative representations improved interpretability but not discriminative power. PCA preserved clinically meaningful variability, whereas VAEs underperformed in this unsupervised setting. These findings suggest that landmark annotations already contain sufficient information for supervised OA tasks, while more advanced models may be needed for unsupervised or generative applications.

1 Introduction

Osteoarthritis (OA) is one of the most prevalent muscoskeletal disorders worldwide. Around 15% of the world's population is affected by the disease [1]. OA is mostly found in hand, knee and hip joints. It can cause pain and discomfort or even disability and eventually lead to patients having to undergo surgery for a replacement of the affected joint. The effects of OA can not be reversed through treatment, but they can be made less severe, for example through the use of physiotherapy. Early detection of the disease plays a crucial role in reducing the severity of symptoms.

There are multiple ways to detect OA, such as through radiography or clinical assessment. In the case of radiography the severity of OA is usually estimated from an X-ray image. The OA severity is commonly classified using the Kellgren-Lawrence (KL) grade [2]. This grading system assigns a score of 0-4 to an image, based on the presence of osteophytes and other factors such as joint space narrowing. Osteophytes are bony protrusions that develop at joint margins. Joint space narrowing refers to the reduction in distance between two bone surfaces, typically caused by cartilage loss. Both are considered main radiographic features of OA [3].

Traditionally the KL grade has been assigned by medical experts, by systematically reviewing X-ray images for presence of the aforementioned factors. Advancements in deep learning however have sparked growing interest in automated approaches to classification. Various models of convolutional neural networks have been used to classify X-ray images of OA patients [4], [5]. These networks learn from pixel intensities, allowing them to capture patterns such as joint space narrowing in the images. They do so by organizing these patterns in a latent space, an internal representation of the image features that the model uses to predict the KL grade. However, using intensity-based approaches also brings a set of limitations. Imaging conditions can vary, leading to differences in contrast, brightness and levels of noise in the images. The networks have to learn to ignore these differences, in order to focus on biologically or structurally relevant patterns, rather than variations introduced by the imaging process.

An alternative to pixel-based input is to represent the anatomy in a point-based manner. Instead of using raw X-ray intensities as input, we use a landmark-based representation derived from the X-ray: a set of coordinates at anatomically meaningful locations. These landmark points are typically obtained by manual annotation by experts or automated landmark detection algorithms. By using points instead of pixel values, the noise introduced by possible imaging artifacts and variations in brightness from the images can be omitted. Points are a more abstract way of representing the bone shapes needed for the classifications and allow a classification method to focus more on shape information such as placement and size. It is worth noting that this abstraction relies on the accurate placement of landmarks, which may be subject to annotation errors or limitations of detection algorithms.

A key consideration in moving from intensity-based input to point-based input is how these points are encoded as model input, since different representations can emphasize different structural properties. In this work we focus on three representations.

The first way to represent these points is using a coordinate-based input representation. We can use a set of x,y coordinates that outline the set of relevant bones. This is a straightforward way that is relatively easy to interpret and visualize.

However, using explicit x,y coordinates as input lacks a few properties that are important for distinguishing relevant anatomical variation from irrelevant differences. For example, they are not invariant to translation, rotation or scale. This means that a bone that is in different positions or orientations has different numerical values, even though it represents the same underlying anatomy. The model might interpret the differences as meaningful variation, rather than recognizing them as the same underlying shape. As a result, additional preprocessing or alignment steps are often required to ensure that the representation reflects structural differences rather than arbitrary imaging conditions.

In addition to their sensitivity to global transformations such as translation, rotation, and scale, coordinate-based inputs also make it difficult to capture local anatomical detail. Especially for OA classification intra-bone information is beneficial, as OA alters not only the overall alignment of the joint, but also the local bone shape and surface geometry. These local changes can be critical indicators of disease progression.

Although these local changes are clinically important, x,y coordinates make it difficult for a model to directly recognize how the points are related. It receives a set of coordinates as input, but it does not know which points are connected to form the bone outlines. Consider a set of bones represented by x,y coordinates. Although the points are consistently defined and ordered, a model does not automatically know how these coordinates relate to one another in terms of anatomical structure. A human might be able to connect the dots to form the set of bones, but a model would have to learn this relational information, which becomes especially important in settings where there are multiple bones.

This suggests that it is useful to carefully define what properties a representation should preserve or discard. Invariances are useful to remove things from the data that we do not want to keep, such as rotation, but we can also establish a set of requirements that we want to keep in the data rather than remove it. For example, one might want to keep the individual shapes that are captured in a dataset, or the information about the distances between shapes.

In this paper we study representations that offer some of these variances. The first is a distance-based representation. Instead of using x,y coordinates, points can be described using a set of pairwise distances, which would make the representation invariant to translations.

Another option to represent these points is an angle-based representation. Instead of describing points as a set of absolute points they are described as an angle with respect to a reference point. This provides translation invariance and highlights relative orientation of structures, allowing the model to focus on angular variation in bone geometry rather than absolute position.

Because each representation encodes the

anatomy in a different way, it is not immediately clear what information they preserve or discard, or how they might influence model behaviour. We therefore want to gain a better understanding of which information can be captured in the embeddings that machine learning models learn from these different representations. To this end, we examine the latent spaces produced by both classical statistical approaches and deep learning–based models. A latent space refers to a compressed internal representation of the data, where high-dimensional inputs such as coordinates are mapped into a lower-dimensional vector that retains the most relevant features. Studying this space allows us to see how different input formats emphasize or suppress certain information, for example by revealing which anatomical features vary smoothly along specific latent dimensions.

Building on these approaches, this thesis investigates how different input representations influence the information captured in the latent space, using a dataset of hip joint anatomy, with OA-related factors. We define a set of requirements to generate different representations of data. We explore how each representation affects the structure and interpretability of the latent space. We study this space using two techniques: principal component analysis (PCA) [6], which provides a simple, linear form of dimensionality reduction, and a variational autoencoder (VAE) [7], which learns a nonlinear, compressed representation of the data. With these tools, we explore the information captured in the latent space through latent space traversal: by varying individual latent dimensions and observing the corresponding changes in the reconstructed anatomy we can interpret what aspects of the shapes each dimension encodes. In addition to visualizing the information captured in the latent space we also use the latent space vectors as input to a simple classifier. This allows us to see if the information captured in each representation's latent space is useful for classifying the presence of features such as joint-space narrowing or KL grade. Through these comparisons, we aim to develop a deeper understanding of how the choice of input representation shapes the organization of latent spaces and, in turn, influences both interpretability and downstream classification performance.

2 Related work

Deep learning approaches for KL grade classification Various studies have described the use of deep learning approaches to assign a KL grade to an X-ray image, but performance remains inconsistent across grades. Pi, SW. et al. [8] used an ensemble network to achieve an accuracy of 74.21% on multiclass KL-grade classification. Confusion matrices show a lower accuracy for KL-grade 1, as the difference between grade 0 and 1 can be hard to discern, also found by others [9], [10].

Because of this, other studies focused on a binary classification problem: OA vs non-OA. Üreten, Kemal, et al. [11] achieved an accuracy of 90.2% of predicting OA vs non-OA cases. Reported results vary across studies and datasets, with achieved accuracies of 92.8 [12] and 82.2% [13]. While image-based CNNs can provide good performance, their inconsistency across grades and dependence on subtle brightness or contrast variations suggests that additional or alternative representations may be valuable. Our study builds on this motivation by exploring non-image representations.

Autoencoders in OA classification A specific kind of approach for osteoarthritis severity classification is using an Autoencoder in combination with a classification head. Farooq, Muhammad Umar, et al. [14] use a so-called Dual-Channel Adversarial Autoencoder. They used a dual-channel design to encode and decode both knees. The latent space is also used to predict KL-grade. The accuracy for this prediction was 75.53%. Similar to this study we also use the latent vector as input to a classification method, but in our work this classifier is not part of the model architecture.

Another approach to using autoencoders for OA classification is to introduce additional loss terms that encourage the latent space to separate cases of interest. Such discriminative losses can guide the network to form representations that are not only compact but also clinically meaningful for classification. For example, Nasser, Yassine, et al. [15] used a Discriminative regularized auto-encoder (DRAE). In their approach they introduced a discriminative loss term that helps the network separate non-OA cases from OA cases. The study shows that the latent space can be formed in such a way to help classify OA cases. These works highlight that autoencoder spaces can carry clinically relevant information, when combined with specialized objectives such as classification heads or discriminative losses. In our study we also rely on the latent space. However we keep the autoencoder architecture standard and investigate how different input representations themselves influence the latent space, rather than optimizing classification accuracy through architectural constraints.

Shape models and OA Although image-based approaches can achieve decent classification performance, they also face limitations as mentioned in Section 1: sensitivity to brightness and noise can introduce variations we do not want to learn. This sensitivity has motivated interest in alternative representations. For osteoarthritis imaging Van Buuren, M. M. A., et al. [16] found that different hip shape features can be linked to the development and progression of hip OA and the possibility of a total hip replacement. Shape-based models therefore demonstrate that geometry alone carries clinically relevant information. Our study extends this idea by comparing how different input representations of shape behave when embedded into a latent space

Point-based models Models exist that are designed to work with point data. A notable line of work that directly processes point sets is PointNet, introduced by Qi, Charles R et al.[17]. It is a network architecture that can take entire point clouds (unordered sets of points) as input. It is designed for tasks like object classification and segmentation. Due to its design the network offers permutation invariance: the ordering of the points does not matter. Our dataset, however, consists of ordered anatomical landmarks, where each index corresponds to a specific anatomical location. In this context, permutation invariance would discard meaningful information. Furthermore, we do not use this model, as our emphasis is not on leveraging specific network architectures, but rather on studying and comparing different forms of input representations. In contrast to PointNet, the ordering of points in our data is informative, making permutation invariance unsuitable. The VAE provides a natural framework for exploring the latent structure of those representations, without the added complexity of architecture specific feature learning.

Latent space exploration Latent space exploration provides a means of interpreting the internal representations of generative models. By traversing latent dimensions, we can investigate whether these abstract variables correspond to meaningful and domain-relevant properties. For example, K. Swannet et al. [18] use a VAE to encode and decode a set of points that describe an airfoil shape. In this work they traverse the latent space to show that the dimensions of that latent space correlate with known properties such as airfoil thickness. They show that they can assign meaning to the black-box nature of the latent space for generative models. This demonstrates that VAEs enable interpretable exploration of latent spaces in point-based domains. Inspired by this, we apply similar principles in the OA setting, aiming to uncover clinically relevant structure in the latent spaces of different input representations.

3 Representations

When designing input representations for OA analysis, not all information contained in the raw data is equally valuable for distinguishing clinically meaningful variation. Some aspects are irrelevant variations, such as orientation, position or scaling of the bones. Those do not reflect meaningful anatomical differences, but can vary across images and patients. Other aspects are essential features, such as relative positioning of the bones, or the bone shapes themselves. Those features carry clinically relevant information for downstream tasks like OA classification. This distinction motivates the need for a set of requirements for our representations. Without these requirements the model would waste learning capacity on learning the irrelevant transformations. We categorize the requirements into two categories: invariances, which describe transformations the representation should be insensitive to, and preservation properties, a term we use to denote information that must be retained.

3.1 Invariances

Invariance is desirable because, in medical imaging, these extrinsic factors are often irrelevant to the clinical question but can vary significantly between scans. For example, two X-ray images of different patients may differ in orientation due to changes in positioning during imaging, or in scale due to differences in image magnification. Without invariance, a model must spend part of its learning capacity "discovering" that such transformations do not alter the underlying anatomy, which can reduce efficiency. It also risks overfitting: the model might associate dataset-specific imaging conditions with clinical outcomes.

When considering input representations for osteoarthritis classification it is important to identify to which variations in the data the model should be insensitive. Based on our knowledge of OA we define the following invariances as desirable properties for a representation.

1. **Translation Invariance:** The model should be insensitive to the absolute position of the bones within the image. A bone located at the top-left of an image should be interpreted identically to the same bone placed in the bottom-right corner.

2. **Rotation Invariance:** The orientation of the bone should not affect its interpretation. For example, a femur bone presented vertically or horizontally should still be recognized as the same anatomical structure.

3. **Scale Invariance:** The size of the bone in the image should not influence its interpretation. Whether a bone appears small or large, it should still be understood as the same anatomical entity. Scale invariance in this context refers to invariance with respect to the imaging scale (e.g., magnification or zoom), not the actual anatomical scale.

3.2 Preservation Properties

In addition to invariances it is equally important to specify what information should be preserved in the representation. While irrelevant information should be suppressed, clinically meaningful variation such as bone shape and positioning should remain accessible to the model. We refer to these requirements as preservation properties and define them below.

1. **Bone-to-bone Relationship Preservation:** The relative positioning between bones should be preserved so that their anatomical relationships are not distorted. The way they are ordered and distanced relative to each other offers information that is useful for downstream tasks (for example joint space narrowing).

2. **Bone Shape Preservation:** Similarly, for downstream tasks it is important that the representation encodes the relevant shape characteristics of the bones. Shape variations can be indicators of properties such as joint space narrowing or osteophyte presence.

4 Methods

In this section we describe the methods used to investigate how different input representations of anatomical landmarks on bone shapes affect the information captured in latent spaces. We begin by outlining the three input types: raw coordinates, pairwise distances and polar coordinates. Each satisfy different invariance requirements introduced in Section 3. We then describe how these representations are projected into latent spaces. Finally we explain how latent space traversal is used to interpret the dimensions of these spaces and assess whether they correspond to meaningful anatomical variation relevant for OA.

4.1 Input Types

- **Raw 2D Landmark Coordinates** The first representation is simply taking the x,y coordinates as input. In their raw form, these coordinates are not invariant to the variances described in Section 3. Alignment methods exist to mitigate these variances, such as Procrustes alignment [19]. Procrustes alignment is a method that removes non-shape variations by optimally translating, rotating and scaling the shapes to a reference shape. We in-

Table 1: Satisfaction of representation requirements for each input type.

Representation Type	**Req. 1 (Translation)**	**Req. 2 (Rotation)**	**Req. 3 (Scale)**	**Req. 4 (Bone Relations)**	**Req. 5 (Shape Preservation)**
Raw 2D Landmark Coordinates	No	No	No	Yes	Yes
Pairwise Distances Between Landmarks	Yes	Yes	No*	Yes	Yes
Polar Coordinates	Yes	No*	No*	Partial	Yes

*Rotation and scale invariance can be enforced via alignment and normalization, respectively.

clude this aligned version in our comparisons, since it provides a useful baseline against the other representations, which encode these invariances directly into the feature representation rather than relying on a preprocessing step. The number of features for a sample with n landmark points is given by $N_{\text{points}} = 2n$.

- **Pairwise Distances Between Landmarks** A second representation is generated by taking the pairwise distances of the 2D landmark points: $d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$. In this representation, the distances remain the same whether the points are at an angle or in a different place in the image. This makes this representation rotation and translation invariant.

  The pairwise distance matrix $D \in R^{n \times n}$ is symmetric, since $d_{ij} = d_{ji}$, and its diagonal entries are zero ($d_{ii} = 0$). Consequently, all unique distances are contained in either the upper or lower triangular part of the matrix. For a set of n landmarks, the number of unique distances is given by $N_{\text{distances}} = \frac{n(n-1)}{2}$. These form the features for a single sample. We normalize pairwise distances by the median training-set distance to place inputs in a unit scale, improving numerical stability during training.

- **Polar Coordinates** The third representation is derived by expressing each landmark in polar coordinates relative to a chosen origin. Each landmark is described by a radius (distance from the origin) and an angle (orientation with respect to a reference axis), which together capture the shape's global geometry and local variations. In our case, the origin is defined as the center of the femoral head, and the reference axis is taken as the line through this center and the corner of the acetabular roof. This choice ensures anatomical consistency across samples. To avoid discontinuities of the angle θ_i at 2π, we represent it by its sine and cosine. This representation is advantageous because it preserves the circular nature of angles: values close to 0 and 2π radians remain close in Euclidean space, whereas using raw angle values would incorrectly treat them as being numerically far apart. Each landmark point is represented as $(x_i, y_i) \mapsto (r_i, \sin\theta_i, \cos\theta_i)$.

  This representation is invariant to global translations of the joint within the image, because all landmarks are defined relative to the femoral head center. For a sample of n landmarks the amount of features is equal to $N_{\text{polar}} = 3n$.

The extent to which each of these input types satisfies the requirements introduced in Section 3 is summarized in Table 1.

4.2 Latent Space Generation

We evaluate how the latent space structure changes depending on the type of input that is provided. We learn a mapping to a latent space using the different input representations, using two methods. By comparing classical statistical techniques and modern deep-learning based methods we aim to assess how well different approaches preserve interesting clinical variation. We first establish a baseline using principal component analysis (PCA) and compare it to nonlinear latent representations generated by a variational autoencoder (VAE).

4.2.1 Principal Component Analysis

Principal Component Analysis (PCA) is a statistical technique for dimensionality reduction that projects high-dimensional data onto a lower-dimensional subspace while preserving as much variance as possible. It works by computing orthogonal basis vectors, called principal components, which are ordered by the amount of variance they explain in the data. The first component captures the largest possible variance, the second component captures the largest remaining variance orthogonal to the first, and so on.

In the context of our experiments, PCA serves as a baseline for understanding how much of the clinically relevant structure in the data can be captured without nonlinear transformations.

4.2.2 Variational Autoencoder

A VAE is a type of generative model that learns to encode input data into a low-dimensional latent space and then reconstruct it from that space. Figure 1 illustrates the flow of information in a variational autoencoder, from the input data through the encoder to the latent representation, and subsequently through the decoder to the reconstructed output. Unlike standard autoencoders, VAEs introduce a probabilistic framework: instead of encoding each input into a fixed point, the encoder outputs the parameters of a Gaussian distribution, namely a mean μ and standard deviation σ. A latent vector z is then sampled from this distribution using the reparameterization trick, which ensures differentiability during training. The decoder maps z back to the input space to produce a reconstruction $\hat{x}$. The model is trained by minimizing a loss that combines two terms: a reconstruction loss, which measures how well $\hat{x}$ matches the original input x, and a Kullback–Leibler (KL) divergence term, which regularizes the learned latent distribution to be close to a standard normal prior. By mapping high-dimensional shape data into a lower-dimensional latent space, the VAE offers a compact and potentially interpretable representation of structural joint variations.

4.3 Latent Space Traversal

Once the latent space has been learned we can interpret it using latent space traversal. This traversal allows us to see how changes in the latent space translate to changes in reconstructions. To do this, we gradually change the values of one latent variable while keeping the others fixed to a mean latent vector, and then reconstruct the corresponding bone shapes.

Visualizing the reconstructions is straightforward when using raw coordinates, as the output can directly be plotted as landmark positions, but for pairwise distances we require extra steps. We first have to convert the encoded triangle of the matrix to a full distance matrix. From this we recover approximate landmark coordinates using multidimensional scaling [20]. The resulting shape is then aligned to a reference shape using Procrustes alignment. In our case we align to the mean of the raw coordinates. This ensures a correct orientation that allows for meaningful comparison between samples and representations.

The polar coordinates are visualized by converting them back to x,y coordinates. Similarly the resulting shape is centered and rotated to a common orientation such that it enables meaningful comparisons.

By observing how these shapes change along a smooth path in latent space, we can see which anatomical features are being captured by specific dimensions, and whether these changes are meaningful in the context of osteoarthritis.

4.4 Classifier

To quantify the information captured in the latent space of a generative model we utilize a Logistic Regression model to perform a downstream classification task. We use this to evaluate whether the information captured in the latent spaces is clinically meaningful. The rationale is that if a low-capacity linear model can successfully separate OA from non-OA samples in the latent space, then the learned representation encodes features relevant to the disease.

For the PCA baseline, the principal component scores for each sample serve directly as the embedding. For the VAE, each input is encoded into a latent distribution, and we use the mean vector of this distribution as the latent representation. This is essentially the output of the encoder. These embeddings are then provided as input to the logistic regression model to predict the severity scores.

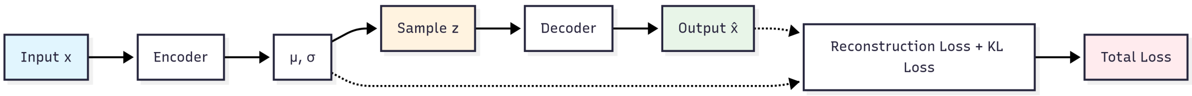


Figure 1: Schematic of a variational autoencoder (VAE). The encoder maps the input x to latent parameters μ and σ. A latent vector z is sampled and passed through the decoder to reconstruct $\hat{x}$. The training objective combines reconstruction loss with a Kullback–Leibler (KL) divergence term, forming the total VAE loss.

5 Experiments

To explore the effect of different input representations on the latent space we have designed a set of experiments that answer the following questions:

1. How does the choice of input representation affect the structure and semantics of the learned latent space in point-based shape models?
2. To what extent does the latent space encode semantically relevant or clinically meaningful information and does this depend on the input representation?

To answer the first research question, we use both principal component analysis (PCA) and the variational autoencoder (VAE) described in Section 4.2.2. For each input representation, we feed the data to each dimensionality reduction method and generate latent vectors. We can traverse the latent dimensions to see what characteristics are captured in each dimension. We use a traversal step of five standard deviations for most representations, as this magnitude is sufficient to reveal clear differences between latent dimensions. For polar coordinates, however, a smaller step of two standard deviations already produces discernible variation. Using a smaller step size in this case reduces visual distortion and results in reconstructions that are easier to interpret.

To quantify the discriminative information we use the latent vectors generated by PCA and the VAE and feed those to the classifier described in Section 4.4. This enables us to evaluate whether the structure learned by each method preserves information relevant to downstream prediction tasks. We want to measure classification performance using the latent vector as input to classify the following features: joint space narrowing scores, osteophyte scores and OA severity scores. We also compute correlation scores between individual latent dimensions and each target feature, allowing us to quantify the strength and nature of the relationships present in the latent space for both methods.

5.1 Dataset

The dataset used in this research is the Cohort Hip and Cohort Knee (CHECK) [21] dataset. It is a longitudinal set, as it contains data obtained from patients at multiple time points. It contains X-ray images of hips of patients with varying OA scores. For each sample we also have expert-derived ordinal scores (0–3) describing radiographic features of osteoarthritis. These include:

- **Joint-space narrowing (JSN)**
  - Medial joint-space narrowing
  - Superior joint-space narrowing
  - Posterior joint-space narrowing
- **Osteophytes**
  - Inferior acetabular osteophytes
  - Superior acetabular osteophytes
  - Inferior femoral osteophytes
  - Superior femoral osteophytes
- **Kellgren-Lawrence grade**

Joint space narrowing quantifies the reduction of space between the femoral head and the acetabulum. Osteophyte scores indicate the presence of bony protrusions at specific acetabular or femoral locations. To illustrate the anatomical regions where these features are assessed, Figure 2

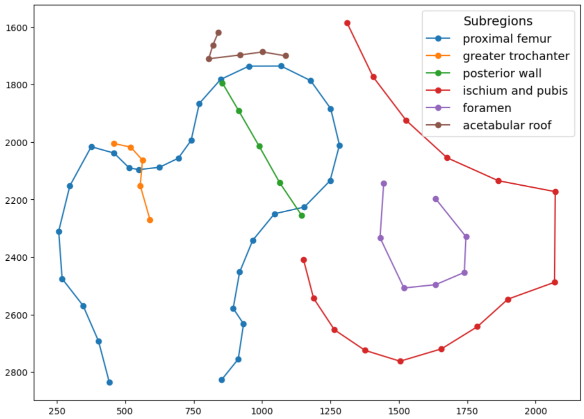


Figure 2: Illustration of the mean landmark configuration across all samples in the dataset. Landmarks are grouped into anatomical subregions, including the proximal femur, greater trochanter, posterior wall, ischium and pubis, foramen, and acetabular roof.

shows the mean landmark configuration with points grouped by subregion.

Because the distribution of these ordinal labels is highly imbalanced (see Table 2), we convert the scores to binary targets by mapping 0 to negative and ≥ 1 to positive. For kellgren we map 0 and 1 to negative and ≥ 2 to positive, since grade 2 marks the presence of definite radiographic osteoarthritis. This binarization strategy is commonly adopted in related work to address imbalance and to reflect clinically meaningful cut-offs in OA diagnosis [22], [23]. The dataset contains 2061 samples for which at least one of the scores is available.

The images in the CHECK dataset have been converted to a set of 2D coordinates using BoneFinder [24], a fully automatic software tool designed to outline and segment skeletal structures from 2D radiographs by placing a set of points along the bone contour or at key landmark positions. This results in a set of 80 landmark points per hip side, outlining the femoral head and acetabular region. Landmarks were distributed across anatomically meaningful subregions, including the proximal femur, greater trochanter, posterior wall, ischium and pubis, foramen, and acetabular roof.

Table 2: Distribution of ordinal (0–3) and binarized (0 vs. $\geq$1) scores for each feature in the CHECK dataset.

Feature	**0**	**1**	**2**	**3**	**0 (bin)**	**$\geq$1 (bin)**
jsn_medial	1208	688	95	1	1208	784
jsn_superior	1410	535	40	7	1410	582
jsn_posterior	1897	35	5	–	1897	40
osteo_acet_inf	1515	276	93	10	1515	379
osteo_acet_sup	536	343	127	15	536	485
osteo_fem_inf	1302	504	102	3	1302	609
osteo_fem_sup	1011	683	257	27	1011	967
kellgren*	639	808	508	38	1447	546

*For Kellgren, binarization is defined as 0–1 $\mapsto$ 0 and $\geq$2 $\mapsto$ 1.

5.2 PCA Dimensionality Analysis

Before fixing the dimensionality of the latent spaces used in subsequent experiments, we used PCA to determine the amount of variance explained by the different dimensions.

Results showed that the first three principal components already captured over 95% of the variance in the data, with subsequent components contributing only marginally. To balance compactness with flexibility, we therefore fixed the latent dimensionality at eight dimensions for all models (both PCA and VAE). This choice ensures consistency across methods while allowing additional latent capacity to capture variations beyond the dominant modes.

5.3 VAE Design

We use a symmetric multilayer perceptron VAE tailored to the sample size and landmark dimensionality. For each sample the input is flattened. The encoder maps this input through two hidden layers (both 64 units) before branching into separate linear layers that output the mean (μ) and log-variance ($\log \sigma^2$) of the latent distribution. From these parameters, an 8-dimensional latent vector z is sampled using the reparameterization trick. The decoder mirrors this structure, expanding from the latent dimension back through 64 and 64 units before reconstructing the output.

The model is optimized using Adam with a learning rate of 1×10^{-3} and trained for 500 epochs. To ensure robustness of the evaluation and to mitigate variance due to train–test splits, we employ 5-fold cross-validation throughout all experiments. We utilize a mean squared error loss.

It is worth noting that we deliberately employ a relatively shallow, fully connected VAE rather than deeper or convolutional architectures often used in image-based studies. Since our inputs are landmark coordinates rather than pixel grids, convolutional layers would not exploit local spatial structure.

5.4 Evaluation Metrics

We use two evaluation metrics to assess the quality of latent representations: root mean squared error (RMSE) and the area under the receiver operating characteristic curve (AUC). RMSE evaluates reconstruction fidelity, while AUC assesses the predictive utility of latent features in a downstream classification task.

5.4.1 Root Mean Squared Error (RMSE)

The Root Mean Squared Error (RMSE) is the square root of the mean squared difference between reconstructed and original features. It expresses reconstruction accuracy in the same units as the input data, making the errors more interpretable than raw MSE. Since the VAE is trained with an MSE-based reconstruction loss, RMSE provides a natural and directly comparable evaluation metric for assessing how closely the model's reconstructions match the input features. For each representation we construct the set of points and compare the original set of points to the set of points reconstructed by the encoding method.

5.4.2 Area Under the ROC Curve (AUC)

To evaluate the quality of the logistic regression probes, we use the AUC. The ROC curve plots the true positive rate against the false positive rate across different classification thresholds. The AUC provides a threshold-independent measure of separability, with values close to 1 indicating strong discrimination for OA features, while values near 0.5 indicate random performance.

We chose AUC over simple accuracy because OA labels can be imbalanced across the dataset, and accuracy can be misleading in such settings. AUC is more robust to class imbalance, as it considers the model's ranking ability across all thresholds rather than its performance at a single decision boundary. Reporting AUC provides a fairer and more interpretable assessment of how well the latent spaces capture clinically meaningful information relevant to OA.

5.5 Correlation Analysis of Latent Features

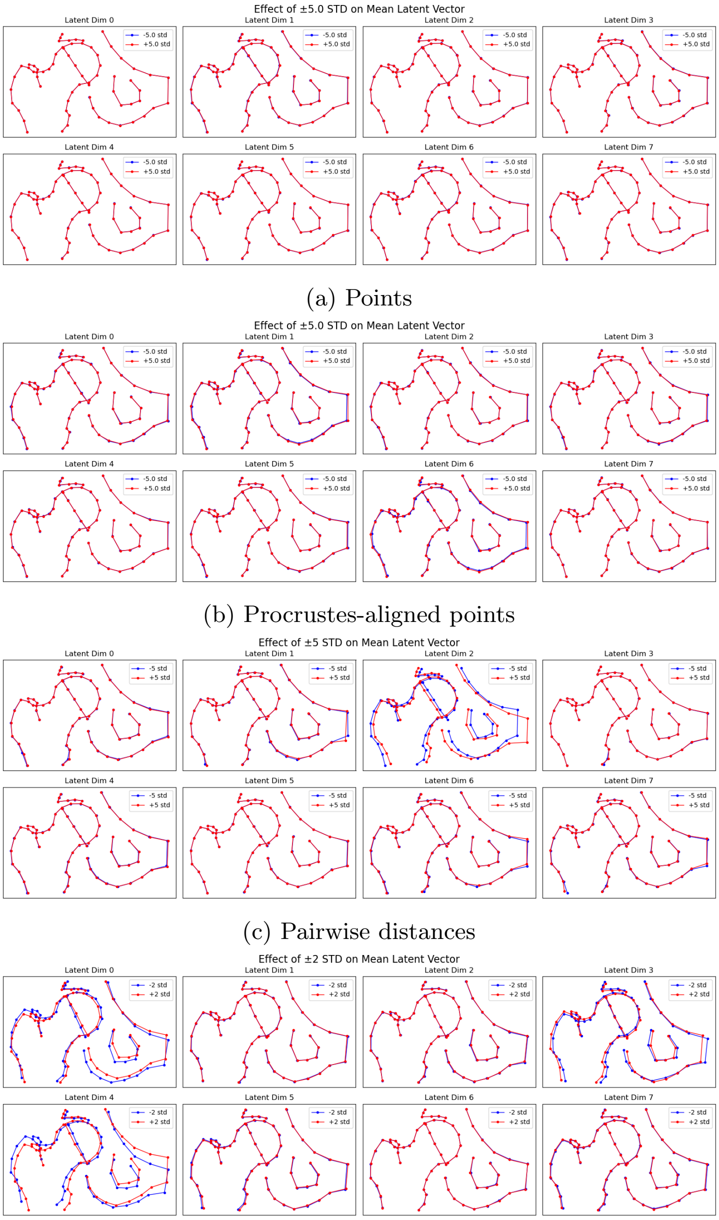
In addition to reconstruction fidelity (RMSE) and predictive performance (AUC), we also investigate how individual latent dimensions relate to clinical variables. To do this, we extract the latent vectors for each sample using either the VAE encoder (taking the mean μ of the latent distribution) or PCA projections. For each representation type (raw points, Procrustes-aligned points, pairwise distances, and polar coordinates), we compute correlations between latent dimensions and clinical outcome measures.

Correlations are calculated using Spearman's rank coefficient, which captures monotonic associations and is robust to non-linear effects. This choice is motivated by the fact that clinical scores are ordinal rather than continuous, making rank-based correlation more appropriate than Pearson correlation, which assumes linear relationships. To aid interpretation, results are visualized in correlation heatmaps, with statistically significant associations (after multiple-testing correction) highlighted. The rationale is that if specific latent dimensions consistently correlate with known radiographic markers of osteoarthritis, this can complement the reconstruction and classification analyses.

6 Results

6.1 Latent Space Visualisation

Figure 3 shows traversal of the latent space of a VAE for the different input representations. The figure shows that across input types, there are always latent dimensions that show no differences when traversed. For raw and Procrustes-aligned points we see almost no difference across all latent dimensions. If we look at pairwise distances we see more pronounced differences in global structure for dimension 2, with more local differences for dimension 1. Note that due to the non-deterministic nature of the VAE the ordering of these dimensions



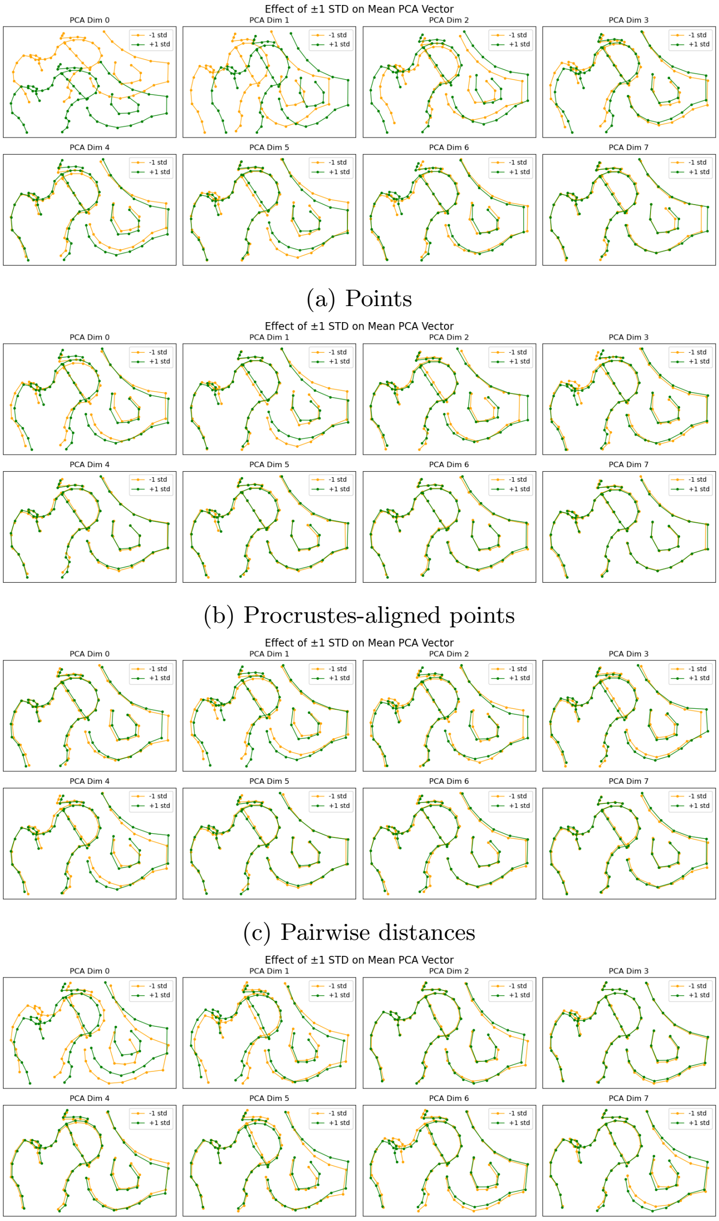
(a) Points

(b) Procrustes-aligned points

(c) Pairwise distances

(d) Polar coordinates

Figure 3: Traversal of the latent space of the VAE for different input representations. Each subplot shows the effect of varying a single latent dimension while holding the others fixed, with rows corresponding to different input types: (a) raw point coordinates, (b) Procrustes-aligned points, (c) pairwise distances, and (d) polar coordinates. For each representation, the reconstructed bone shapes reveal which kinds of variation are captured by the latent dimensions. For example, distance- and polar-based inputs yield more visible global and angular changes than raw or Procrustes-aligned points, where traversals show little or no variation. This illustrates how the choice of input representation influences the interpretability of the learned latent space.



(a) Points

(b) Procrustes-aligned points

(c) Pairwise distances

(d) Polar coordinates

Figure 4: Traversal of the latent space of PCA for different input representations. As in Figure 1, rows correspond to (a) raw point coordinates, (b) Procrustes-aligned points, (c) pairwise distances, and (d) polar coordinates. Here, traversals show that PCA components often capture global transformations, such as x- and y-translations when using raw coordinates, whereas Procrustes alignment reduces such effects and shifts the components toward local shape variation. Polar coordinates emphasize angular changes, with dimension 0 in particular reflecting rotational effects. This comparison highlights how linear embeddings distribute variance differently across representations.

is arbitrary. Using polar coordinates shows more pronounced differences across different dimensions. Dimensions 0, 3 and 4 show relatively large differences in global structure. In contrast to the other input representations we see that dimension 0 also captures angular variation when the polar coordinate representation is used.

Figure 4 shows traversal of the latent space of PCA for the different input representations. We used a step of one standard deviation, which was sufficient to illustrate variation without overwhelming the visualisation. We see that using points as input results in latent dimensions that seem to capture the variations in the data well. Dimension 0 encodes y translation, dimension 1 encodes x translation. There is no dimension that is left out, as they all seem to capture some form of shape variation. When using Procrustes alignment we see that these effects are mitigated, as the alignment already removes the translation variances. The dimensions become less descriptive of global structure and focus more on local variation. For pairwise distances this is quite similar. The dimensions mostly capture variances that translate points, but for polar coordinates we also see that dimension 0 captures angular variation.

Table 3 summarizes the reconstruction performance of both the Variational Autoencoder (VAE) and Principal Component Analysis (PCA) across the four feature representations considered. Across all representations, PCA consistently achieves lower reconstruction error than the VAE. The Procrustes representation produces extremely small errors for both methods. However, these values should not be directly compared to those of the other representations, as Procrustes alignment removes absolute scale information. Consequently, the reported RMSE reflects only residual shape differences after optimal superimposition, which by construction are on a different numerical scale and therefore not directly interpretable against the other feature sets.

6.2 Classification

Figure 5 shows classification performance across clinical features and input representations, using the AUC as a metric. The raw input types overall achieved higher AUC scores than their PCA or VAE counterparts. The PCA representations generally maintained comparable performance. The VAE representations however consistently underperform across all feature sets, indicating poor suitability of VAE latent encodings for this downstream classification task.

Representation	**VAE RMSE**	**PCA RMSE**
Raw x,y coordinates	4.05	2.30
Procrustes*	0.0071	0.0038
Pairwise Distances	14.81	1.95
Polar coordinates	4.92	2.34

Table 3: Reconstruction errors (RMSE) for different feature representations using VAE and PCA. Lower values indicate better reconstruction fidelity. *Procrustes errors are reported on a different scale, since absolute size information is lost during alignment and unrecoverable in reconstruction.

Overall classification performance was not that strong, except for JSN posterior, suggesting that posterior narrowing is the most robustly captured and discriminative feature when considering points as input representation.

Raw and PCA based approaches exhibited relatively low variance, whereas VAE based approaches showed higher variances, for example for using distances to predict JSN Posterior. This variance highlights a lack of stability in nonlinear latent spaces when applied to a clinically relevant classification task.

6.3 Correlation

Figure 6 shows a correlation matrix between the latent dimensions of the VAE based on Procrustes-aligned points and clinical features. Values are generally close to zero, indicating weak or no associations between the learned latent space and clinical outcomes. This result aligns with the underperformance of VAE encodings in classification tasks (Figure 5), underscoring their limited interpretability and clinical utility in the present context. While only one matrix is shown here for illustration, the correlation patterns observed for other input representations and methods were qualitatively similar,

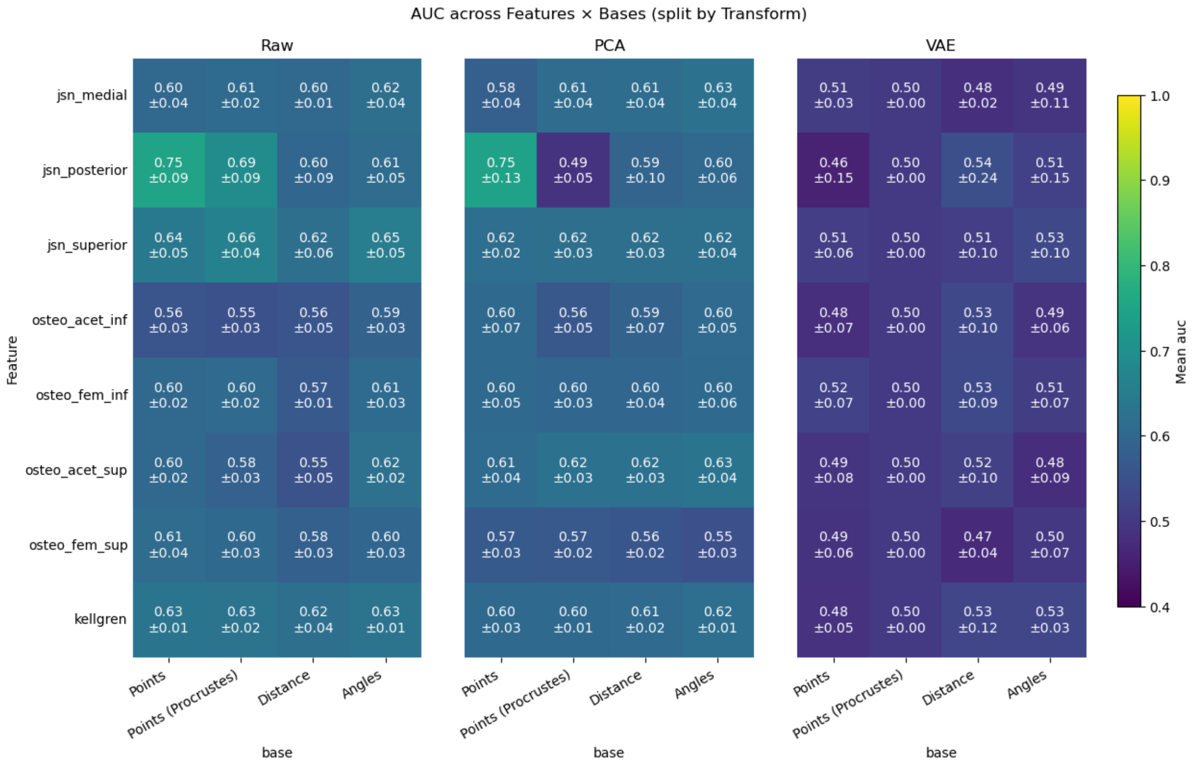


Figure 5: Classification performance (AUC, mean ± standard deviation) across clinical features and input representations. Columns correspond to representations derived from raw points, Procrustes-aligned points, pairwise distances, and angles, each in raw, PCA, or VAE-transformed space. Rows represent clinical features including joint space narrowing (JSN, medial/posterior/superior), osteophytes (acetabular/femoral, inferior/superior), and Kellgren–Lawrence grade.

with no consistent strong associations emerging.

7 Discussion

The results of this thesis provide insight into how different geometric representations of bone structures affect the performance of latent variable models in osteoarthritis (OA) feature classification. By systematically comparing raw point coordinates, Procrustes-aligned points, polar coordinates, and pairwise distances, and evaluating these under both linear (principal component analysis, PCA) and nonlinear (variational autoencoder, VAE) frameworks, we assessed not only the predictive utility of each representation but also the interpretability and consistency of the learned models.

The traversal analyses demonstrate that the kinds of variation present in each representation of the input geometry are directly reflected in the latent space, regardless of whether the embedding method is linear (PCA) or nonlinear (VAE). Each representation emphasizes different geometric properties of the joint, and these biases are carried over into the resulting latent dimensions, shaping their interpretability and informativeness.

Raw Cartesian coordinates offer the most direct encoding of point locations but also embed large amounts of nuisance variation. In both PCA and VAE traversals, the dominant dimensions captured simple global transformations such as translations, rather than meaningful shape differences. This is especially clear in PCA, where the first two

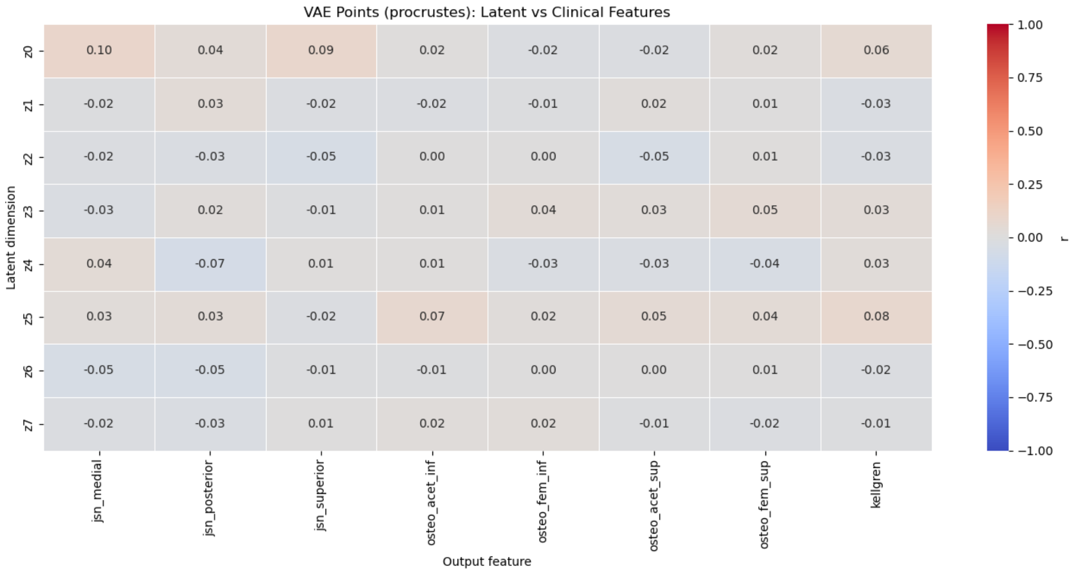


Figure 6: Correlation matrix between VAE latent dimensions (Procrustes-based points) and clinical features.

components aligned almost perfectly with vertical and horizontal shifts. For the VAE, traversals across raw point dimensions produced almost no visible change, suggesting that the nonlinear model either failed to exploit these raw signals or collapsed them into a degenerate representation. This interpretation is consistent with the reconstruction errors in Table 3: although VAEs achieved a moderate RMSE on raw coordinates, this fidelity did not translate into meaningful latent variation, highlighting the disconnect between reconstruction accuracy and interpretability.

Procrustes-aligned points were designed to remove translation and scaling variance prior to modeling. As expected, this reduced global transformation effects, with PCA components shifting toward more local shape variations. In the VAE, however, traversals of Procrustes aligned points remained largely uninformative, producing little to no variation across dimensions. Procrustes alignment successfully removed these irrelevant variations, but in doing so it also eliminated residual global differences that might have carried clinical signal, leaving even less variation for the VAE to organize into interpretable latent dimensions.

Pairwise distances provide a representation that discards absolute coordinate information. With this representation, the VAE latent space exhibited more variation across dimensions. In the VAE, distance inputs yielded at least some latent dimensions that showed more granular shifts. However, the overall weak performance suggests this is less a strength of the distance encoding and more a mismatch between the VAE architecture and the geometric structure of the input, which limited the model's ability to organize variability into clinically meaningful dimensions. Table 3 shows that VAEs produced very high RMSE on distance matrices (14.81), suggesting difficulty in reconstructing this representation faithfully, whereas PCA achieved relatively low error (1.95). This mismatch likely explains the weak and unstable latent organization in the VAE.

Polar coordinates provided the most distinctive and interpretable latent traversals across both PCA and VAE. By explicitly encoding angular information relative to a central point, this representation made rotational differences directly accessible to the embedding methods. In the VAE, several latent dimensions (notably dimensions 0, 3, and 4) produced more distinctive traversals, showing angular variations. In PCA we can see the same phenomenon, with dimension 0 strongly tied to rotational effects. At the same time, these clear rotational patterns should be interpreted with caution. It is not obvious that such rotations correspond to genuine anatomical variation in the patient population. For instance, patients are not systematically rotated around the femoral head in reality. It may be that polar coordinates simply reparameterize the same positional variability that appears as x/y shifts in the Cartesian representation, making it easier for the models to express this as rotation. Thus, while polar coordinates yield more interpretable traversals, it remains uncertain whether they provide a more faithful reflection of clinically meaningful variation.

The visualisation results illustrate the kinds of variation each representation makes accessible; the next question is whether these differences translate into meaningful improvements in classification. One of the most important findings that follows from the classification is that raw point coordinates, which represent the direct spatial locations of anatomical landmarks, often perform as well as or better than all latent representations. While this may partly be explained by the fact that the raw representation retains the full geometric information, it also suggests that the discriminative signal for OA-related features is not less accessible in the original landmark configurations than in the transformed representations, meaning that a linear classifier can already exploit much of this information without additional invariance constraints.

The strong baseline performance of raw coordinates implies multiple things. First, it indicates that the complexity introduced by non-linear transformations, such as VAEs, may not always lead to benefits for downstream tasks. Instead, the introduction of such methods in an unsupervised way can even decrease discriminative power if they fail to preserve clinically relevant variability during compression. For example, regularization can smooth out or suppress small but clinically meaningful variations if those variations do not contribute much to reconstruction quality.

Second, it highlights that datasets with consistent landmarks may already be powerful enough for diagnostic tasks without requiring alternative ways to represent them. This is particularly relevant in clinical settings where interpretability and reproducibility are essential: if raw points already make for competitive classification results, research can focus on making the methods more interpretable.

While raw points performed well as a baseline overall, the Procrustes-aligned representations provided subtle improvements in certain contexts. By removing translation, rotation, and scale differences, Procrustes alignment ensures that classification relies only on shape differences rather than variability related to positioning or orientation. For features such as joint space narrowing (JSN superior) Procrustes-aligned points offered slightly higher AUC scores compared to unaligned raw points. This suggests that removing irrelevant global transformations can bring minor improvement at least for shape-specific tasks.

Having examined how different input representations influence classification, we now turn to the embedding methods themselves. PCA-based representations provided another perspective on variability by expressing the data along orthogonal axes of maximum variance. PCA vectors showed more variation than their VAE counterparts and were inherently ordered by their amount of variation. In terms of classification performance, PCA-based models performed comparable to raw coordinates. This demonstrates PCA's strength as both a dimensionality-reduction and interpretability tool. By emphasizing variance structure in the data, PCA can highlight clinically meaningful modes of variability that might otherwise be obscured. However, PCA's reliance on variance rather than discriminative power means that not all principal components necessarily align with clinical outcomes. While PCA captures structure in the dataset, it does not explicitly optimize for prediction. Thus, while its outputs are interpretable in the sense that each dimension seems to capture a certain kind of variance, they may not always translate to the highest possible classification performance. This is however not a PCA-specific limitation but rather one for unsupervised dimensionality reduction techniques in general, including VAEs.

VAEs consistently performed worse than PCA across all input representations. Classification results based on VAE encodings were close to random (AUC values around 0.48–0.54), and correlations between latent dimensions and clinical features were negligible. This suggests that the VAE failed to encode clinically meaningful variability into the latent dimensions. Instead, the learned encodings appeared diffused, capturing global structural variability without mapping it effectively to clinically relevant outcomes. There were no clear dimensions that could be correlated to single clinical features. VAEs are optimized to reconstruct input data while maintaining a smooth latent space, but this does not guarantee that their latent dimensions align with disease features. Indeed, the lack of interpretability observed in the correlation analysis confirms that the VAE encodings were not clinically meaningful. Although this may seem surprising, it is less unexpected when considering that PCA can be interpreted as a special case of a linear VAE. Unlike PCA, however, VAEs balance reconstruction with latent regularization rather than maximizing explained variance, which in small datasets with subtle anatomical differences can lead to diffused and clinically uninformative encodings. It should also be noted that this underperformance may depend on the VAE configuration: factors such as model depth, regularization strength, and the choice of reconstruction loss can strongly affect the ability of VAEs to capture subtle anatomical variability, and alternative architectures may perform differently.

The results also reveal clear differences in the difficulty of predicting specific clinical features. Among the evaluated outcomes, joint space narrowing (particularly the posterior and superior measures) consistently achieved the highest AUCs, with values up to 0.75 in some representations. This suggests that changes in joint space geometry are reliably captured by landmark-based methods. By contrast, osteophyte features yielded lower AUCs (generally between 0.55 and 0.62), reflecting the subtler and more localized nature of osteophyte growth. Landmark-based representations may not capture these features as effectively. The Kellgren–Lawrence (KL) grade also presented a challenge, with all methods yielding relatively medium performance. This aligns with prior literature, which consistently reports the difficulty of automated KL grading due to its composite nature. The KL score integrates multiple radiographic features into a single grade.

Taken together, these findings also highlight more general lessons about the role of invariances and assumptions in representation learning. Whether a representation emphasizes translation, rotation, or angular structure is not a neutral choice: it determines which forms of variability dominate the latent space and, consequently, what the model can learn. In our case, PCA primarily surfaced global translations, echoing earlier work showing that unsupervised methods often capture high-variance but clinically uninteresting transformations. This is not necessarily "wrong," but it underscores that the usefulness of a representation depends on the task: for supervised classification, raw points already carry the necessary discriminative signal, whereas for unsupervised exploration, invariance choices can decide whether meaningful or trivial modes of variation are recovered. Whether it actually helps to enforce invariances really depends on how the data were preprocessed, how much nuisance variation is still left, and what specific clinical question has to be answered.

Finally, several methodical limitations of this work should be acknowledged. First, the analyses were performed on a relatively modest dataset of 2,061 samples, all drawn from a single cohort (CHECK). Although we used five-fold cross-validation to mitigate overfitting, the absence of external validation means that the generalizability of our findings to other populations, imaging protocols, or landmarking procedures remains uncertain. Studies with other datasets would be necessary to determine whether the trends observed here are generalizable.

Second, the study depends entirely on landmarks provided by a single automatic annotation tool (BoneFinder). While this ensured anatomical consistency across samples, it also introduces the risk that systematic errors or biases in landmark detection propagate throughout all downstream analyses. Small misplacements of landmarks can accumulate when transformed into distance or angular representations, potentially distorting the geometry in ways that affect both reconstruction quality and latent interpretability. Future work could benefit from comparing multiple annotation pipelines or incorporating manual validation to quantify the reliability of the input landmarks.

Third, we limited ourselves to relatively simple latent variable models: PCA as a linear baseline and a shallow VAE as a nonlinear counterpart. This was a deliberate design choice to prioritize interpretability and to isolate the effect of different input representations. However, it also means that we cannot rule out that more expressive models such as deeper VAEs or graph neural networks (GNN) could better preserve subtle shape variations relevant to osteoarthritis. A GNN could work better in this setting by representing landmarks as nodes connected through anatomical or geometric adjacency, allowing the model to capture localized dependencies through message passing, rather than flattening all coordinates into a global vector where such local structure is easily lost.

Fourth, our evaluation metrics come with important notes. Reconstruction error (RMSE) provides a measure of reconstruction ability but does not necessarily align with clinical relevance, since small geometric discrepancies may be visually or clinically inconsequential. Correlation analyses between latent dimensions and clinical scores were restricted to linear or monotonic associations. This means that more complex nonlinear relationships, which could exist between anatomy and OA progression, may have been overlooked. The logistic regression probes we used are intentionally simple, offering a clean test of whether information is already linearly accessible in the latent space, but they cannot see more complex relationships that other classifiers can use.

Finally, there are challenges in interpreting latent traversals. While they offer intuitive visualisations of how latent dimensions correspond to variation, there is no guarantee that the observed changes are explained by anatomical changes. For example, the rotations observed in the polar representation may reflect how the model is parameterized rather than actual rotational variation in patient anatomy. Similarly, the absence of variation in VAE traversals for point-based inputs may reflect model collapse, entanglement of features, or insufficient clinical signal in the representation, and these possibilities cannot be disentangled with certainty in this work.

8 Conclusion

This thesis investigated how different geometric representations, ranging from raw 2D coordinates to engineered latent features, interact with latent variable models (PCA, VAE) to shape the resulting latent space and influence downstream OA classification performance. Our experiments demonstrate that, for the task of classifying radiographic OA features from hip landmarks, the raw point coordinates remain a strong baseline, often outperforming or matching more complex latent representations. This indicates that when landmark annotations are consistently defined, much of the clinically relevant signal is already embedded in the original configurations, and linear classifiers can access it without additional transformations.

Latent representations based on pairwise distances or polar coordinates did not yield consistent improvements in supervised classification tasks. While these representations offered more interpretable latent traversals (distances capturing relative geometry and polar coordinates emphasizing angular variation), they did not translate into stronger discriminative power. This highlights an important distinction: interpretability does not guarantee discriminative power. Some variations are clinically meaningful but not discriminative (anatomical size), while others are neither (orientation). Polar rotations belong to the latter, reflecting artifacts rather than anatomy.

The comparison of PCA and VAEs in this setting showed that nonlinear embeddings, at least in the form of a shallow VAE applied to 2D landmark data, did not provide advantages over PCA. In this unsupervised setting, the VAE's compression and KL regularization sometimes suppressed subtle but clinically meaningful variation, leading to latent spaces that were less discriminative than their linear counterparts. This suggests that the complexity introduced by nonlinear generative models does not necessarily improve classification, and in some cases may even reduce performance if the models fail to preserve clinically relevant features.

Taken together, these findings underscore both the promise and the limits of geometric encodings in osteoarthritis research. Point-based inputs remain competitive for supervised diagnostic tasks, but relational encodings may be more suitable for unsupervised tasks that aim to discover new shape patterns, and generative models may add value in data augmentation or scenarios where invariance to positioning is essential. At the same time, the study also has limitations: reliance on a single dataset and landmarking tool, the use of relatively shallow latent models, and the difficulty of ensuring that latent traversals correspond to true biological processes. These limitations primarily constrain the generalizability of the findings rather than invalidating the main conclusions. Future work should address these limitations by validating results across larger and more diverse datasets and testing on different architectures (e.g., graph neural networks or point-cloud models). Furthermore, advancing evaluation strategies beyond reconstruction error and simple correlations will be beneficial to directly assess clinical utility.

References

[1] V. L. Johnson and D. J. Hunter, "The epidemiology of osteoarthritis," *Best practice & research Clinical rheumatology*, vol. 28, no. 1, pp. 5–15, 2014.

[2] J. H. Kellgren, J. Lawrence, *et al.*, "Radiological assessment of osteo-arthrosis," *Ann Rheum Dis*, vol. 16, no. 4, pp. 494–502, 1957.

[3] P. M. Van Der Kraan and W. B. Van Den Berg, "Osteophytes: Relevance and biology," *Osteoarthritis and cartilage*, vol. 15, no. 3, pp. 237–244, 2007.

[4] B. C. Dharmani and K. Khatri, "Deep learning for knee osteoarthritis severity stage detection using x-ray images," in *2023 15th International Conference on COMmunication Systems & NETworkS (COMSNETS)*, IEEE, 2023, pp. 78–83.

[5] A. Tiulpin, J. Thevenot, E. Rahtu, P. Lehenkari, and S. Saarakkala, "Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach," *Scientific reports*, vol. 8, no. 1, p. 1727, 2018.

[6] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, vol. 2, no. 11, pp. 559–572, 1901.

[7] D. P. Kingma, M. Welling, *et al.*, *Auto-encoding variational bayes*, 2013.

[8] S.-W. Pi, B.-D. Lee, M. S. Lee, and H. J. Lee, "Ensemble deep-learning networks for automated osteoarthritis grading in knee x-ray images," *Scientific Reports*, vol. 13, no. 1, p. 22 887, 2023.

[9] S. M. Ahmed and R. J. Mstafa, "Identifying severity grading of knee osteoarthritis from x-ray images using an efficient mixture of deep learning and machine learning models," *Diagnostics*, vol. 12, no. 12, p. 2939, 2022.

[10] S. B. Kwon, H.-S. Han, M. C. Lee, H. C. Kim, Y. Ku, and D. H. Ro, "Machine learning-based automatic classification of knee osteoarthritis severity using gait data and radiographic images," *IEEE Access*, vol. 8, pp. 120 597–120 603, 2020.

[11] K. Üreten, T. Arslan, K. E. Gültekin, A. N. D. Demir, H. F. Özer, and Y. Bilgili, "Detection of hip osteoarthritis by using plain pelvic radiographs with deep learning methods," *Skeletal Radiology*, vol. 49, pp. 1369–1374, 2020.

[12] Y. Xue, R. Zhang, Y. Deng, K. Chen, and T. Jiang, "A preliminary examination of the diagnostic value of deep learning in hip osteoarthritis," *PloS one*, vol. 12, no. 6, e0178992, 2017.

[13] R. Gebre, J. Hirvasniemi, R. van der Heijden, *et al.*, "Detecting hip osteoarthritis on clinical ct: A deep learning application based on 2-d summation images derived from ct," *Osteoporosis International*, vol. 33, no. 2, pp. 355–365, 2022.

[14] M. U. Farooq, Z. Ullah, A. Khan, and J. Gwak, "Dc-aae: Dual channel adversarial autoencoder with multitask learning for kl-grade classification in knee radiographs," *Computers in Biology and Medicine*, vol. 167, p. 107 570, 2023.

[15] Y. Nasser, R. Jennane, A. Chetouani, E. Lespessailles, and M. El Hassouni, "Discriminative regularized auto-encoder for early detection of knee osteoarthritis: Data from the osteoarthritis initiative," *IEEE transactions on medical imaging*, vol. 39, no. 9, pp. 2976–2984, 2020.

[16] M. Van Buuren, N. K. Arden, S. Bierma-Zeinstra, *et al.*, "Statistical shape modeling of the hip and the association with hip osteoarthritis: A systematic review," *Osteoarthritis and cartilage*, vol. 29, no. 5, pp. 607–618, 2021.

[17] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[18] K. Swannet, C. Varriale, and N. A. K. Anh Khoa Doan, "Latent space correlation for interpretable airfoil parameterization using variational autoencoders," 2024.

[19] J. C. Gower, "Generalized procrustes analysis," *Psychometrika*, vol. 40, no. 1, pp. 33–51, 1975.

[20] J. B. Kruskal and M. Wish, *Multidimensional scaling*. Sage, 1978.

[21] J. Bijlsma and J. Wesseling, *CHECK (Cohort Hip Cohort Knee) data of baseline and 6 to 8 years follow-up*, version V3, 2015. DOI: `10.17026/dans-zc8-g4cw`. [Online]. Available: `https://doi.org/10.17026/dans-zc8-g4cw`.

[22] A. S. Mohammed, A. A. Hasanaath, G. Latif, and A. Bashar, "Knee osteoarthritis detection and severity classification using residual neural networks on preprocessed x-ray images," *Diagnostics*, vol. 13, no. 8, p. 1380, 2023.

[23] E. Vaattovaara, E. Panfilov, A. Tiulpin, *et al.*, "Kellgren-lawrence grading of knee osteoarthritis using deep learning: Diagnostic performance with external dataset and comparison with four readers," *Osteoarthritis and Cartilage Open*, vol. 7, no. 2, p. 100 580, 2025.

[24] D. C. Lindner, *Bonefinder®*. [Online]. Available: `https://bone-finder.com/`.