Self-supervised feature learning for diagnosing hip osteoarthritis in X-ray images

How effectively can a VAE's latent space reflect osteoarthritis severity and enable diagnostic accuracy under label scarcity and label noise?

Poli Dimieva[1]

Supervisors: Jesse Krijthe[1]**, Gijs van Tulder**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Poli Dimieva
Final project course: CSE3000 Research Project
Thesis committee: Jesse Krijthe, Gijs van Tulder, Michael Weinmann

Abstract

Osteoarthritis (OA) is a prevalent and progressive joint disease whose diagnosis from radiographs often requires expert-labeled data, which is expensive and time-consuming to obtain. Variational Autoencoders (VAEs) offer a way to learn compact, unsupervised representations that may be reused for downstream classification in low-label scenarios. In this work, we assess whether a VAE can learn latent features from hip radiographs that support OA classification with minimal supervision. We evaluate the model's reconstruction quality, latent space structure, and diagnostic utility under label scarcity and label noise. Results show that VAE-derived features outperform raw pixel and random baselines, suggesting the latent space captures disease-relevant structure. These findings underscore the potential of VAEs as scalable, label-efficient tools for clinical imaging tasks like OA diagnosis.

1 Introduction

Osteoarthritis (OA) is a chronic joint disease that affects over 500 million people globally and is a major cause of disability in older adults [1]. Hip OA, in particular, significantly impairs mobility and often necessitates joint replacement in advanced stages [1]. Diagnosis is typically performed by radiologists using standardized grading systems such as the Kellgren–Lawrence (KL) scale [2], which assesses disease severity based on structural features in X-ray images. However, developing automated OA diagnostic systems remains challenging due to the high cost of obtaining large volumes of expert-annotated radiographs. Deep learning methods generally require extensive labeled data to achieve strong performance, limiting their utility in low-resource clinical settings [3].

In addition to data scarcity, another practical challenge in medical imaging is label noise - inaccuracies in expert annotations. OA grading, for example, relies on subjective assessment of radiographic features, which often leads to inter-rater variability and occasional mislabeling [4, 5]. Such noise can significantly impair the performance of supervised models, which rely on clean labels to learn accurate decision boundaries. Models that remain stable in the presence of mislabeled data are therefore crucial in real-world diagnostic pipelines.

Self-supervised learning (SSL) addresses these limitations by enabling feature learning from unlabeled medical images [3, 6]. Within SSL, contrastive methods have shown strong results by learning representations that distinguish between augmented views of images, while generative methods learn to reconstruct images, thereby capturing the underlying data distribution [6]. Variational Autoencoders (VAEs), as a generative approach, are especially promising in medical imaging because they produce latent spaces that may reflect anatomical and pathological variation relevant to diagnosis [7].

A Variational Autoencoder (VAE) learns a probabilistic latent representation by reconstructing input data while constraining the latent space to follow a smooth prior distribution [8]. This makes VAEs well-suited for learning compact representations from large amounts of unlabeled data. Unlike supervised models, which often degrade with limited or noisy labels, VAEs can exploit the full dataset and may thus offer improved robustness [9, 10].

In this study, we evaluate whether a VAE trained on hip radiographs can support robust OA classification in regimes where supervised models struggle - specifically, when annotations are scarce or affected by label noise. First, we assess the structure of the learned representations by measuring intra- and inter-class latent distances [11, 12]. Second, we visualize the latent space with t-SNE, UMAP and PCA to examine whether disease severity is implicitly reflected [13, 14]. Third, we test the utility of VAE-derived features for classification by comparing them to random and raw-pixel baselines [11, 14]. Finally, we benchmark the VAE-based model against a fully supervised convolutional neural network (CNN) across a range of label availability and noise levels, simulating real-world constraints [3, 6, 8, 10].

Our central question is: How effectively can a VAE's latent space reflect osteoarthritis severity and enable diagnostic accuracy under label scarcity and label noise? By answering this, we aim to demonstrate the potential of generative self-supervised learning as a more resilient alternative for automated medical diagnosis in low-resource settings.

The paper is organized as follows: Section 2 reviews related work; Section 3 describes the methodology; Section 4 details the experiments; Section 5 presents the results; Section 6 discusses findings and future directions; Section 7 reflects on responsible research aspects; and Section 8 concludes.

2 Related Work

By modeling the underlying data distribution, VAEs can learn compact latent spaces that encode meaningful variation without requiring manual labels. Prior studies, such as Chartsias et al. [7], have demonstrated their utility in tasks like segmentation, synthesis, and anomaly detection. However, these works emphasize reconstruction performance or visual interpretability, without evaluating how well the learned features support diagnostic classification tasks - especially in clinically realistic scenarios with limited or noisy annotations.

Anomaly detection is one of the most common applications of VAEs in this domain. For example, Uzunova et al. [15] used a Conditional VAE (C-VAE) to detect abnormalities in brain MRIs by comparing reconstructed and original images. Although their method effectively distinguished normal from pathological scans, it relied on anatomical conditioning variables during training, introducing a degree of supervision. This limits generalizability to settings where such metadata are unavailable.

In parallel, contrastive self-supervised learning (SSL) has gained traction in medical imaging. These methods learn representations by distinguishing between augmented views of the same image and others in the batch. While contrastive SSL has shown strong performance in classification tasks with limited labels [6], it often requires large batch sizes,

carefully crafted augmentations, and lacks the generative interpretability of VAEs. In contrast, VAEs explicitly model the full image distribution, making them especially promising for capturing subtle radiographic patterns such as those linked to osteoarthritis.

In summary, while VAEs have demonstrated their capacity to learn informative representations from medical images, their diagnostic utility under real-world constraints remains underexplored. Existing studies often rely on auxiliary supervision or focus on anomaly detection without assessing classification robustness. This research addresses these gaps by evaluating a VAE-based classifier for binary hip osteoarthritis diagnosis and benchmarking its performance against a supervised CNN under varying levels of label availability and noise.

3 Methodology

This section outlines the VAE framework used to learn unsupervised representations of hip radiographs. We describe the model architecture, latent space formulation, and ELBO training objective.

3.1 Variational Autoencoder Framework

Let $X = \{x_i\}_{i=1}^N$ denote a dataset of hip X-ray images. A VAE models the underlying distribution of the data using a latent variable $z \in R^d$, where d is the latent dimensionality. The model consists of two probabilistic components:

- An **encoder** $q_\phi(z|x)$, which approximates the intractable posterior distribution over latent variables given input x. This distribution is modeled as a multivariate Gaussian with mean $\mu(x) \in R^d$ and diagonal covariance defined by $\sigma^2(x) \in R^d$.
- A **decoder** $p_\theta(x|z)$, which reconstructs the input image from the latent code z, defining the likelihood of the data given the latent representation.

The prior over latent variables is assumed to be a standard normal distribution:

$$p(z) = \mathcal{N}(0, I).$$

To allow for gradient-based optimization through the stochastic sampling of z, the model employs the **reparameterization trick**. Instead of sampling $z \sim \mathcal{N}(\mu, \sigma^2)$ directly, we reparameterize as:

$$z = \mu + \sigma \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I),$$

where $\odot$ denotes element-wise multiplication. This formulation separates the deterministic transformation from the stochasticity, making the sampling operation differentiable.

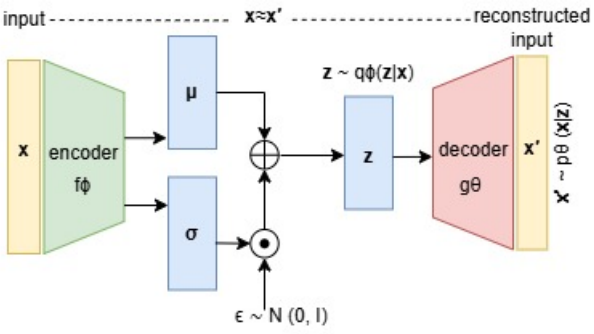

Figure 1: Schematic of a Variational Autoencoder (VAE). The encoder network f_ϕ maps the input image x to the parameters of a latent Gaussian distribution. The latent code z is sampled via the reparameterization trick and decoded back to a reconstructed image x' via g_θ.

3.2 Training Objective: Evidence Lower Bound (ELBO)

The VAE is trained by maximizing the Evidence Lower Bound (ELBO) on the marginal log-likelihood $\log p_\theta(x)$, which decomposes as:

$$\log p_\theta(x) \geq E_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{\text{KL}}(q_\phi(z|x)\|p(z)).$$

The first term, $E_{q_\phi(z|x)}[\log p_\theta(x|z)]$, encourages accurate reconstruction of the input from the latent code. It is approximated in practice using the mean squared error:

$$\text{REC}(x, \hat{x}) = \|x - \hat{x}\|_2^2.$$

The second term, $D_{\text{KL}}(q_\phi(z|x)\|p(z))$, is the Kullback–Leibler divergence, which regularizes the approximate posterior to remain close to the prior $p(z)$. This constraint ensures that the latent space is smooth, continuous, and aligned with a known distribution.

To control the trade-off between reconstruction fidelity and latent regularization, we introduce a scaling factor $\beta > 0$, leading to the final training objective:

$$\mathcal{L}_{\text{total}} = \text{REC}(x, \hat{x}) + \beta \cdot D_{\text{KL}}(q_\phi(z|x)\|\mathcal{N}(0, I)).$$

The parameter β allows flexibility in balancing the information content of the latent codes and the level of disentanglement in the learned representation [16]. A higher value places more emphasis on regularization, which may improve robustness but reduce reconstruction quality.

This formulation enables the VAE to learn compact and generalizable image representations from unlabeled data, which can later be evaluated for their clinical relevance in downstream tasks.

4 Experiments

In this section, we detail the dataset used, preprocessing steps, evaluation metrics and experimental setup for each experiment used to assess the diagnostic utility of VAE-learned representations.

4.1 Dataset and Preprocessing

The dataset used was the CHECK (Cohort Hip and Cohort Knee) - a Dutch longitudinal study focused on early osteoarthritis in patients. This is a dataset of anterior-posterior (AP) hip radiographs annotated with Kellgren-Lawrence (KL) grades, which range from 0 (no OA) to 4 (severe OA). For binary classification, we binarize the labels into *no/mild OA* (grades 0–1) and *moderate/severe OA* (grades 2–4), consistent with prior work.

Preprocessing was carried out designed to standardize image characteristics and reduce inter-patient variability. Radiographs were first resampled to a uniform pixel spacing of 0.4 mm to ensure spatial consistency. Each image was then cropped around the femoral head based on anatomical landmarks, yielding a 224×224 pixel region of interest. Percentile-based intensity normalization was applied to mitigate differences in brightness and contrast, and left hip images were horizontally flipped to align with the orientation of right hips.

The resulting preprocessed images were stored alongside their KL grade annotations and relevant metadata, ensuring consistency and reproducibility in downstream analyses involving both unsupervised and supervised tasks.

The dataset is split into training and test subsets using predefined subject IDs from a text-based split file. The training set includes both labeled and unlabeled images, while the test set is used only for evaluation.

4.2 Evaluation Metrics

For unsupervised latent analysis, we compute latent distances between disease groups to quantify structural separation and employ dimensionality reduction for visualization.

For the classification experiments, we use two primary metrics:

- **Area Under the ROC Curve (AUC)**: Measures the ability of the classifier to distinguish between classes, especially important in imbalanced settings.
- **Accuracy (ACC)**: The proportion of correct predictions.

4.3 Reconstruction Quality

Before evaluating the latent space and classification performance, we inspect the VAE's reconstructions of input radiographs. Visual comparisons of original and reconstructed images help confirm that anatomical structures - such as femoral head shape and joint space - are preserved. This supports the claim that the VAE's latent space captures clinically relevant information.

4.4 Latent Space Structure

To assess whether the VAE's latent space captures diagnostic structure, we analyze how similar the latent representations are within and across clinically defined groups. Specifically, we compute average pairwise Euclidean distances among:

- **Intra-OA:** samples with osteoarthritis (KL grade ≥ 2),
- **Intra-no-OA:** samples without osteoarthritis (KL grade < 2),
- **Inter-group:** pairs spanning the OA and no-OA groups.

Let $\mathcal{Z} = \{z_i\}_{i=1}^{N}$ be the set of latent vectors obtained from the VAE encoder. We define $\mathcal{D}$ as the index set of disease cases and $\mathcal{N}$ for non-disease cases. We compute:

$$\text{Intra}_{\mathcal{D}} = \frac{2}{|\mathcal{D}|(|\mathcal{D}|-1)} \sum_{i<j,\, i,j\in\mathcal{D}} \|z_i - z_j\|_2, \qquad (1)$$

$$\text{Intra}_{\mathcal{N}} = \frac{2}{|\mathcal{N}|(|\mathcal{N}|-1)} \sum_{i<j,\, i,j\in\mathcal{N}} \|z_i - z_j\|_2, \qquad (2)$$

$$\text{Inter}_{\mathcal{D},\mathcal{N}} = \frac{1}{|\mathcal{D}||\mathcal{N}|} \sum_{i\in\mathcal{D}} \sum_{j\in\mathcal{N}} \|z_i - z_j\|_2. \qquad (3)$$

A well-structured latent space should exhibit higher inter-group distances compared to intra-group distances, suggesting that the VAE captures clinically relevant separation in an unsupervised manner. This experiment evaluates the extent to which latent representations reflect diagnostic groupings without direct supervision.

4.5 Latent Space Visualization

To complement quantitative analysis, we visualize the latent space using dimensionality reduction techniques such as t-SNE, UMAP, and PCA. Latent vectors are projected to two dimensions and color-coded by KL grade indicating OA and KL grade indicating healthy images to reveal clustering patterns. These plots offer visual evidence of the VAE's capacity to organize disease-relevant variation.

4.6 Comparison with Random and Raw Pixel Feature Extraction

To contextualize the value of the learned VAE representations, we compare them against two alternative feature extraction methods that do not involve meaningful learning:

- **Random features**: Latent vectors are sampled from a standard Gaussian distribution $\mathcal{N}(0, I)$ and treated as feature inputs for classification. This baseline simulates uninformative representations and serves as a lower bound on performance. It tests whether the classifier can succeed using purely random, structureless features.
- **Raw pixels**: Each image is flattened into a one-dimensional vector of raw pixel intensities and passed directly into the classifier. This setup reflects a naive encoding of the image and helps assess how much signal is present in the data without learned feature transformation.

For all three feature types (VAE, random, raw pixels), we use the same downstream logistic regression classifier. This classifier is trained on the same labeled training data and evaluated on the same test set using AUC as the primary metric, ensuring a fair comparison.

In the case of the VAE, we isolate the encoder after unsupervised training and use the mean latent vectors (μ) as fixed representations. The decoder is discarded, and no further fine-tuning occurs. This design tests whether unsupervised pretraining leads to useful feature representations for classification, independent of reconstruction quality.

Superior performance by the VAE-based classifier would indicate that the learned representations capture more discriminative structure than those derived from unstructured baselines.

4.7 Comparison Setup

Baseline Models

We compare two modeling approaches:

- **Supervised CNN.** A convolutional neural network trained end-to-end on labeled X-ray images. The architecture includes 4 convolutional layers with ReLU activations, followed by a fully connected output layer (256×14×14) and Sigmoid output for binary classification. It is trained using binary cross-entropy loss to directly predict disease presence. This training process is fully tailored to the diagnostic task.
- **VAE + Classifier.** The VAE comprises a convolutional encoder and a mirrored decoder. The encoder includes five convolutional layers (32–512 filters) followed by two fully connected layers that output the mean and log-variance of a 64-dimensional latent space. Latent sampling is performed using the reparameterization trick to allow backpropagation through stochastic nodes. The decoder uses transposed convolutions to reconstruct the original input.

  The model is trained using the Evidence Lower Bound (ELBO) objective, which combines a pixel-wise L2 reconstruction loss with a Kullback-Leibler (KL) divergence term, weighted by $\beta = 4.0$ to encourage disentangled and semantically meaningful representations. Importantly, classification is not performed during VAE training. Only the encoder is used post-training to extract mean latent vectors from the labeled subset. The decoder is excluded from this stage, ensuring that evaluation focuses solely on the learned feature space. These latent vectors are then used to train a downstream logistic regression classifier, enabling assessment of how well the unsupervised latent space captures disease-relevant structure.

Hyperparameters Setup

Table 1 summarizes the key hyperparameters used for training both the VAE and CNN models. These values were chosen based on standard practice in medical image analysis and validated via preliminary experiments on the validation set.

Parameter	VAE / CNN	Value
Image size	both	224×224
Latent dimension	VAE	64
Batch size	both	32
Optimizer	both	Adam
Learning rate	both	1×10^{-3}
Max epochs	both	100
Train/Val/Test split	both	80% / 20%

Table 1: Summary of key hyperparameters.

These hyperparameter choices were grounded in standard practice and initial validation:

These hyperparameter choices were guided by conventions in deep learning for medical imaging and validated through preliminary experiments:

- Latent dimension (64): Chosen to ensure the latent space is compact yet expressive. Prior work on VAEs in medical imaging has used similar sizes to encode clinically relevant features without overfitting [17, 18].
- Batch size (32) and learning rate (1×10^{-3}): Default values commonly used in training both CNNs and VAEs, shown to provide stable convergence in unsupervised and supervised deep learning models [8].
- Optimizer (Adam): Selected for its adaptive learning rate and robustness across deep learning applications, especially effective in scenarios with noisy gradients.
- Max epochs (100): Sufficient to ensure convergence based on early stopping checks during validation, without risk of overfitting. This also aligns with typical training durations in similar applications [17].
- Train/Val/Test split (80% / 20%): Follows common practice to ensure a robust estimate of generalization while preserving enough data for training.

4.8 Classification Under Limited Labels

To simulate real-world scenarios of label scarcity, we vary the proportion of labeled data used for training (5%, 10%, 25%, 50%, 100%). These fractions were selected to represent a range from extreme label scarcity to full supervision. This type of fractional supervision experiment is standard practice in label-efficient and self-supervised learning literature, particularly for benchmarking model performance under realistic annotation constraints. Prior studies in both medical imaging and general vision have used similar setups, training or fine-tuning with subsets ranging from 1% to 50% of labeled data to reflect limited resource scenarios [19, 20].

In our setting, for each label fraction, both the VAE-based and CNN-based models are trained and evaluated on the same held-out test set. To account for variability and assess stability, we repeat all experiments five times using different random seeds. These seeds determine the random subsampling of training data and initialization of model parameters.

To ensure a fair comparison, the same data splits and seeds are used for both models in each run. This way, differences in performance can be attributed to the modeling approach rather than differences in data exposure or randomness. We report the mean and standard deviation of Accuracy and AUC across the five runs.

This experiment tests whether unsupervised VAE representations enable effective and stable classification under conditions of limited supervision, compared to fully supervised learning.

4.9 Robustness to Label Noise

We further evaluate model robustness by injecting 10% symmetric label noise into the training labels — simulating imperfect annotations commonly encountered in clinical datasets. Both the VAE-based classifier and the supervised

CNN are retrained on this corrupted data while the test set remains clean.

As in the previous experiment, we perform five runs for each label fraction using consistent seeds and splits between models. By introducing controlled randomness and measuring variability across repetitions, we assess both the average performance and the sensitivity of each model to label corruption. Metrics are again reported as mean ± standard deviation of Accuracy and AUC.

This experiment reveals the resilience of each model type to label noise and helps identify which approach maintains stability and diagnostic reliability under realistic annotation imperfections.

5 Results

This section reports the outcomes of our experiments designed to evaluate the representational quality and diagnostic utility of VAE-learned latent features. We organize the results by type of assessment, beginning with qualitative reconstruction analysis, followed by structure in the latent space, visual inspection, and finally, classification performance in multiple experimental settings.

5.1 Reconstruction Quality

The VAE reconstructions are slightly blurry but successfully preserve key anatomical structures of the hip, such as the femoral head and joint space. This suggests that the model has captured essential features necessary for image synthesis. A linear interpolation between two latent codes (Figure 2) shows a smooth and continuous transition, indicating that the latent space encodes gradual anatomical variation in a coherent manner.

Furthermore, samples generated from random latent vectors (Figure 6) appear structurally plausible and diverse. This implies that the decoder has learned a meaningful mapping from the latent space back to the image domain, consistent with a well-trained generative model.

Figure 6 presents randomly sampled images generated from the prior distribution over the latent space. The samples appear realistic and diverse, indicating that the decoder has learned a meaningful mapping from latent space to image space.

5.2 Latent Space Structure

To quantify how well the VAE encodes disease-related structure, we computed average pairwise Euclidean distances between latent vectors:

Distance Type	Mean Distance
Intra-OA group	6.50
Intra-no-OA group	6.52
Inter-group (OA vs. no-OA)	6.56

Table 2: Mean Euclidean distances between latent vectors of samples in the same or different diagnostic groups.

Inter-group distances slightly exceeded intra-group distances, indicating that the latent space captures some disease-relevant separation without supervision.

5.3 Latent Space Visualizations

We apply three dimensionality reduction techniques—t-SNE, UMAP, and PCA - to visualize the latent representations of test samples. Figures 3, 7, 8 show no clearly separated clusters. However, this may be due to the complexity of the underlying image data and the loss of information when reducing from a 64-dimensional latent space to 2D.

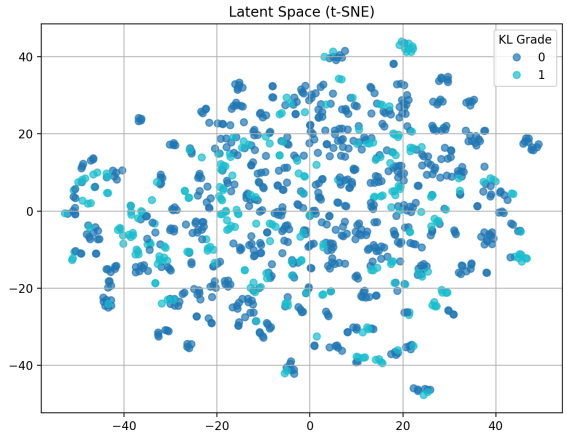

Figure 3: t-SNE projection of latent vectors.

5.4 Comparison with Random and Raw Pixel Feature Extraction

The random baseline simulates uninformative latent features drawn from a standard normal distribution, while the raw pixel baseline reflects a non-learned, high-dimensional representation. The VAE-derived representations yield the highest classification performance (mean AUC = 0.78 ± 0.02), compared to 0.64 ± 0.032 for raw pixels and 0.50 ± 0.014 for random features. These results are averaged over 10 independent runs to ensure robustness and account for variability. Figure 4 shows the ROC curve comparison between the random baseline and the VAE latents.

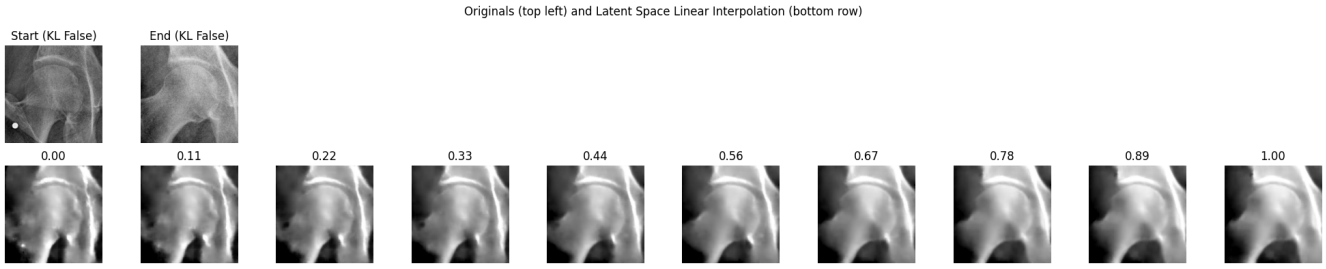

Figure 2: Original images (top row) and VAE reconstructions generated by linear interpolation between two latent vectors (bottom row).

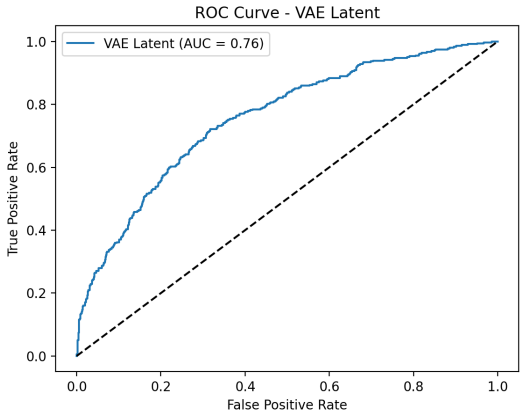

Figure 4: ROC curve comparison of the classifier trained on VAE latent features and random noise for a single run.

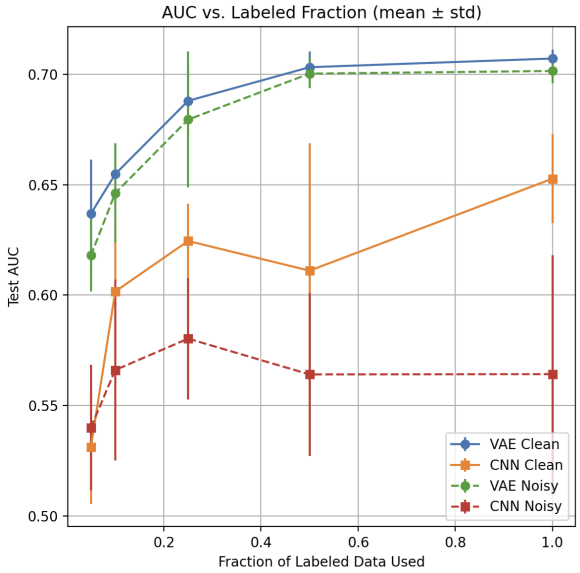

Figure 5: Test AUC as a function of labeled data fraction, comparing VAE-based classifiers and end-to-end CNNs under both clean and noisy label conditions. Each point represents the mean of 5 independent runs.

5.5 Classification Under Limited Labels

The VAE consistently outperformed the CNN across all data fractions. At 5% labels, the VAE achieved an AUC of 0.63 ± 0.03 compared to the CNN's 0.53 ± 0.02. With full labels, the VAE reached 0.71 ± 0.01, while the CNN reached 0.65 ± 0.04.

In contrast, the CNN achieved higher accuracy at every level. At 5%, it reached 0.72 ± 0.03 versus the VAE's 0.63 ± 0.02. At 100%, the CNN reached 0.72 ± 0.01 while the VAE plateaued around 0.66 ± 0.01.

5.6 Robustness to Label Noise

To assess robustness, we repeated the same experiment with 10% label noise added to the training labels.

The VAE maintained stable performance despite noise, showing minimal degradation. For instance, at 100% labels, AUC dropped marginally from 0.71 ± 0.01 to 0.70 ± 0.01. The CNN, in contrast, was more sensitive: its AUC dropped from 0.65 ± 0.04 to 0.56 ± 0.07 under the same conditions.

Both models saw reduced accuracy under noisy labels, but the VAE remained more stable. At 5% labels, the VAE dropped from 0.63 ± 0.02 to 0.60 ± 0.03, whereas the CNN remained around 0.72 ± 0.03, likely due to overfitting to noisy labels.

6 Discussion

This section provides an interpretation of the experimental results, discusses the limitations of the current framework, and outlines directions for future research.

6.1 Interpretation of Results

The results indicate that the VAE is capable of learning meaningful representations of hip radiographs in an unsupervised setting. Reconstructions were slightly blurry but consistently preserved key anatomical structures such as the femoral head and joint space, suggesting that the model captures clinically relevant structure even without supervision. Latent interpolations between samples revealed smooth, gradual transitions in appearance, indicating that the latent space encodes coherent anatomical variation.

This behavior reflects a well-known trade-off in VAE models: enforcing a regularized and structured latent space often compromises reconstruction fidelity. Our model uses a relatively high value of $\beta = 4.0$, which increases the weight

of the KL divergence term in the ELBO objective. This encourages the encoder to produce compressed and disentangled latent codes, which may reduce pixel-level accuracy but enhance representation learning. In this context, blurrier reconstructions are not a flaw but rather an artifact of stronger latent structure - beneficial for downstream tasks like classification.

Regarding latent space structure, the inter-group distances were slightly higher than intra-group distances, offering modest support that the latent space aligns with OA severity. However, the absolute differences were small, indicating weak class separation. This is consistent with the absence of clear clustering in dimensionality-reduced visualizations (t-SNE, UMAP, PCA), likely due to projection loss and the subtlety of disease signals.

An important factor may be the high anatomical similarity across patients in the dataset. Many individuals - with OA or not - share similar skeletal morphology, which the VAE may model more strongly than subtle disease differences. Moreover, the dataset contains repeated scans from the same patient across time points. This encourages the model to prioritize subject identity over diagnostic differences, reducing class separability in both distance metrics and 2D projections. Future work could mitigate this by incorporating metadata, introducing constraints on class separation, or exploring contrastive training objectives.

The comparison between feature extractors provides further insight. A classifier trained on VAE-derived features significantly outperformed those trained on random or raw pixel inputs. This confirms that unsupervised pretraining, even without labels, produces more expressive and diagnostic feature representations. Since all three classifiers use the same architecture and are trained on the same data, this result isolates the encoder as the key variable.

The results from the classification experiments offer several insights into the behavior of unsupervised and supervised models under varying label availability and noise conditions. In low-label regimes (e.g., 5%–10% labeled data), the VAE-based classifier consistently outperformed the CNN in terms of AUC (fig.5). This suggests that the VAE's unsupervised pretraining allows it to learn meaningful features even when supervision is scarce, leading to more robust generalization from limited labeled examples.

As the fraction of labeled data increased, the accuracy of the CNN gradually improved, eventually surpassing the VAE at 100% label availability (fig. 9). This aligns with expectations: the CNN is directly optimized for the classification objective and can fully leverage large labeled datasets to learn discriminative patterns. In contrast, the VAE is trained to reconstruct input images and does not directly optimize for class separation, limiting its ceiling performance when labels are abundant.

Interestingly, the VAE also demonstrated greater resilience to label noise. While both models saw a decline in AUC under noisy conditions, the performance drop was consistently smaller for the VAE (fig. 5, 9). This suggests that the encoder's latent features, shaped by unsupervised learning, are more stable and less sensitive to noise in the supervisory signal - an important property in real-world medical applications where label noise is common.

Beyond this specific setting, several broader lessons emerge. First, unsupervised learning methods like VAEs can extract medically meaningful features from imaging data, even without labeled supervision - highlighting their utility in data-scarce domains. Second, evaluating latent representations via a frozen encoder + simple classifier setup offers a reproducible and modular way to assess feature quality across architectures and tasks. Third, this study illustrates the importance of dataset structure: repeated scans of the same subjects and inter-patient similarity can bias model learning and impact downstream generalization.

These insights extend beyond hip OA classification. For instance, similar techniques could be applied to other imaging modalities (e.g., chest X-rays, MRI) or diseases where annotation is costly or ambiguous. The methodology also provides a blueprint for isolating and evaluating the utility of pretraining pipelines in larger clinical models.

Based on these findings, we hypothesize that augmenting the VAE with contrastive learning objectives or supervised regularization could further improve the disease-separability of latent representations - an idea worth exploring in future work.

6.2 Limitations and Future Work

While the results are promising, several limitations must be acknowledged. First, the evaluation of latent space separability relied primarily on pairwise distances and 2D visualizations. While informative, these analyses are exploratory; more rigorous metrics such as silhouette scores, clustering accuracy, or supervised probing tasks could offer deeper insights into latent geometry.

Second, the VAE was optimized for reconstruction rather than classification. As a result, it cannot outperform supervised CNNs when large amounts of labeled data are available. Future work could explore hybrid training objectives that combine generative and discriminative signals, such as semi-supervised learning or contrastive losses tailored to clinical class boundaries.

Third, the CHECK dataset is relatively small and homogeneous. Its single-cohort nature and inclusion of repeated scans from the same patients limit generalizability. Extending the evaluation to multi-center datasets would improve robustness and clinical relevance.

Despite the promising results, this study has several limitations. First, the CNN and VAE models were not extensively tuned. Hyperparameters such as learning rate, number of epochs, and architecture size were selected based on standard practice and held constant across experiments to ensure comparability. While this setup avoids biasing the results toward one model, it may not reflect each model's best possible performance.

Second, classification was conducted using a frozen VAE encoder followed by a simple logistic regression model. This isolates the learned representations from the generative component, but may underutilize the potential of fine-tuning or integrating the classifier into the VAE training loop.

Third, the comparison is limited to a single dataset and binary classification task (OA vs. non-OA). Although the

findings provide useful insights, further evaluation on other datasets and conditions (e.g., multi-class grading, external cohorts) is needed to assess generalizability.

Finally, the number of training epochs (100) was fixed for all models, which may favor some architectures over others depending on their convergence dynamics. In future work, more rigorous tuning and early stopping criteria could provide a better performance ceiling for each model.

More broadly, this study reinforces the promise and limitations of unsupervised learning in healthcare. While not a replacement for fully supervised approaches, VAEs and similar models offer a scalable, label-efficient method to capture structure in medical data. By combining them with targeted supervision or clinically-informed objectives, future systems could offer more robust and interpretable diagnostic support.

7 Responsible Research

This study addresses an important medical challenge - early and accurate diagnosis of osteoarthritis - using methods that reduce dependency on costly annotated datasets. However, the use of generative models in healthcare also carries responsibilities. While VAEs reduce reliance on expert labeling, they risk encoding and amplifying biases present in training data, such as demographic imbalances or imaging artifacts. As such, careful auditing of latent representations is necessary before clinical deployment.

Furthermore, although this study did not involve real-time predictions or patient-facing applications, downstream misuse of the model (e.g., applying it outside validated populations) could lead to diagnostic errors. To mitigate this, transparent reporting of training data composition and model limitations is essential. Environmental impacts were minimal given the modest scale of experiments, but future large-scale training should consider energy efficiency and carbon footprint.

Overall, this work supports the responsible development of AI tools in radiology by emphasizing data efficiency, model transparency, and reproducibility.

8 Conclusion

This study demonstrates that VAEs can learn structured representations of hip X-rays that support reliable osteoarthritis classification, even when labeled data is scarce or noisy. While latent space projections do not show clear visual KL grade separation, the quantitative performance of VAE-based classifiers - particularly their robustness to label noise - validates the diagnostic relevance of the learned features. While this work focused on hip osteoarthritis, the same framework-combining a VAE encoder with a lightweight classifier - could be applied to other radiographic diagnosis tasks, such as identifying fractures, tumors, or joint degeneration. By learning anatomy-aware features without supervision, the method offers a scalable approach to pretraining models across diverse clinical domains.

References

[1] GBD 2021 Musculoskeletal Disorders Collaborators. Global burden of osteoarthritis 1990–2020: a systematic analysis. *The Lancet Rheumatology*, 3(6):e385–e397, 2021.

[2] JH Kellgren and JS Lawrence. Radiological assessment of osteo-arthrosis. *Annals of the rheumatic diseases*, 16(4):494–502, 1957.

[3] S. C. Huang et al. Self-supervised pretraining for medical image analysis: A survey. *Medical Image Analysis*, 84:102680, 2023.

[4] Xiaoyan et al. Wang. Inconsistency in grading severity of knee osteoarthritis between clinical and imaging-based assessments: a systematic review. *Osteoarthritis and Cartilage*, 28(3):388–397, 2020.

[5] Danielle Aparecida de Oliveira et al. Castro. Inter-rater reliability of the kellgren–lawrence classification for osteoarthritis of the knee: A systematic review. *Seminars in Arthritis and Rheumatism*, 50(3):478–485, 2020.

[6] Q. Liu et al. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[7] Aris Chartsias, Thomas Joyce, Rohan Dharmakumar, and Sotirios A. Tsaftaris. Chapter 8 - autoencoders and variational autoencoders in medical image analysis. In *Biomedical Image Synthesis and Simulation: Methods and Applications*, pages 129–162. Elsevier, 2022.

[8] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2014.

[9] Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. *Advances in neural information processing systems*, 27:3581–3589, 2014.

[10] Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11:625–660, 2010.

[11] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. *International conference on machine learning*, pages 478–487, 2016.

[12] Anna Rydhmer, Jimmy Larsson, Niklas Wängberg, and Anders Eklund. Measuring latent space separability in unsupervised image embeddings. *Scientific Reports*, 11(1):14460, 2021.

[13] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11):2579–2605, 2008.

[14] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. *European Conference on Computer Vision (ECCV)*, pages 132–149, 2018.

[15] H. Uzunova, S. Schultz, H. Handels, and J. Ehrhardt. Unsupervised pathology detection in medical images using conditional variational autoencoders. *International Journal of Computer Assisted Radiology and Surgery*, 14(3):451–461, 2019.

[16] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*, 2017.

[17] A. Chartsias, T. Joyce, R. Dharmakumar, and S. A. Tsaftaris. Chapter 8 - autoencoders and variational autoencoders in medical image analysis. In *Biomedical Image Synthesis and Simulation: Methods and Applications*, pages 129–162. Elsevier, 2022.

[18] Wenjun Li et al. Explainable variational autoencoder for anomaly detection in medical imaging. *Medical Image Analysis*, 75:102299, 2022.

[19] Soheil Azizi, Basil Mustafa, Frances Ryan, Zachary Beaver, Justin Freyberg, Joshua Deaton, Allen Loh, Alan Karthikesalingam, Simon Kornblith, and Ting Chen. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3478–3488, 2021.

[20] Mahmoud Assran, Mathilde Caron, Piotr Bojanowski, Ishan Misra, Armand Joulin, Nicolas Ballas, and Michael Rabbat. Masked siamese networks for label-efficient learning. *European Conference on Computer Vision (ECCV)*, 2022.

A Appendix

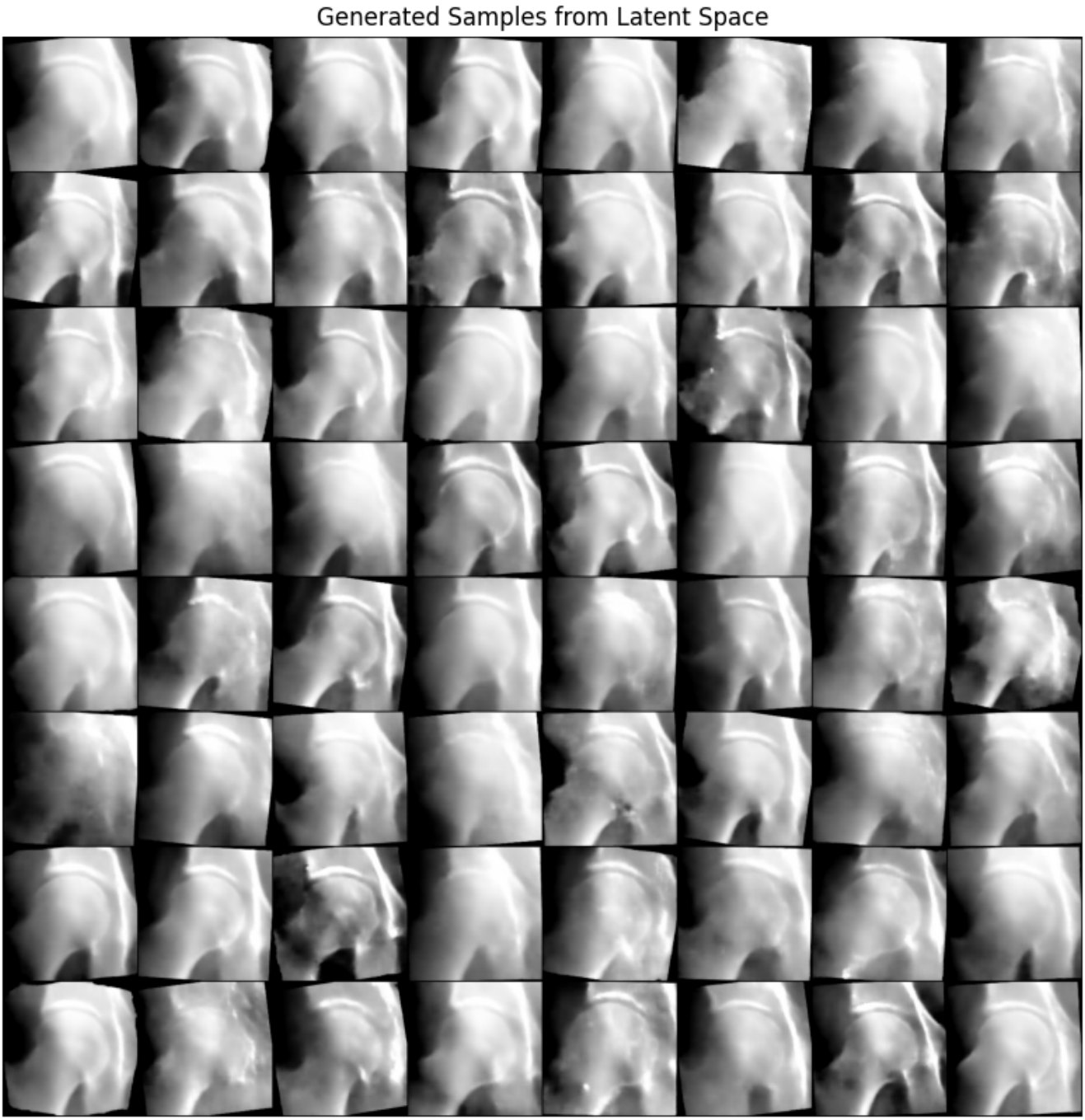

Figure 6: Random samples generated from the VAE latent space.

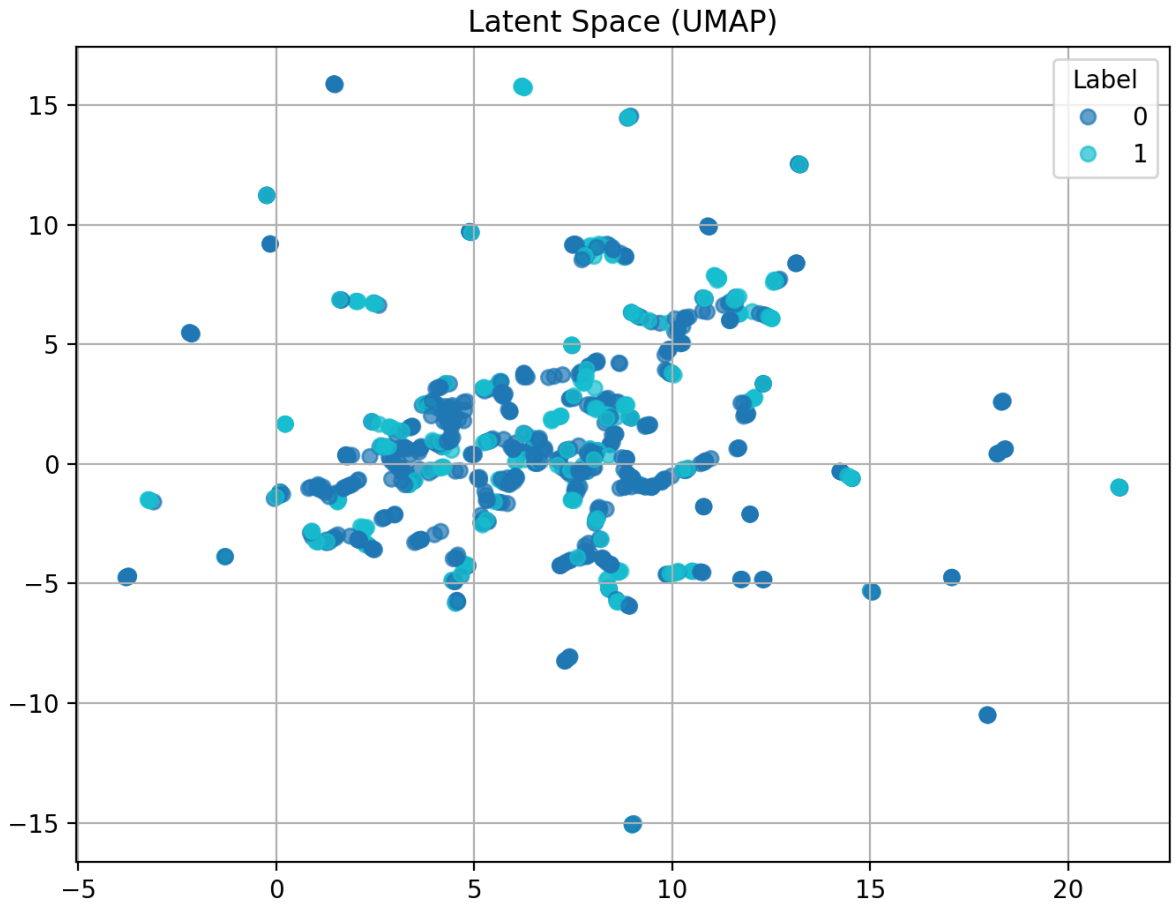

Figure 7: UMAP projection of latent vectors.

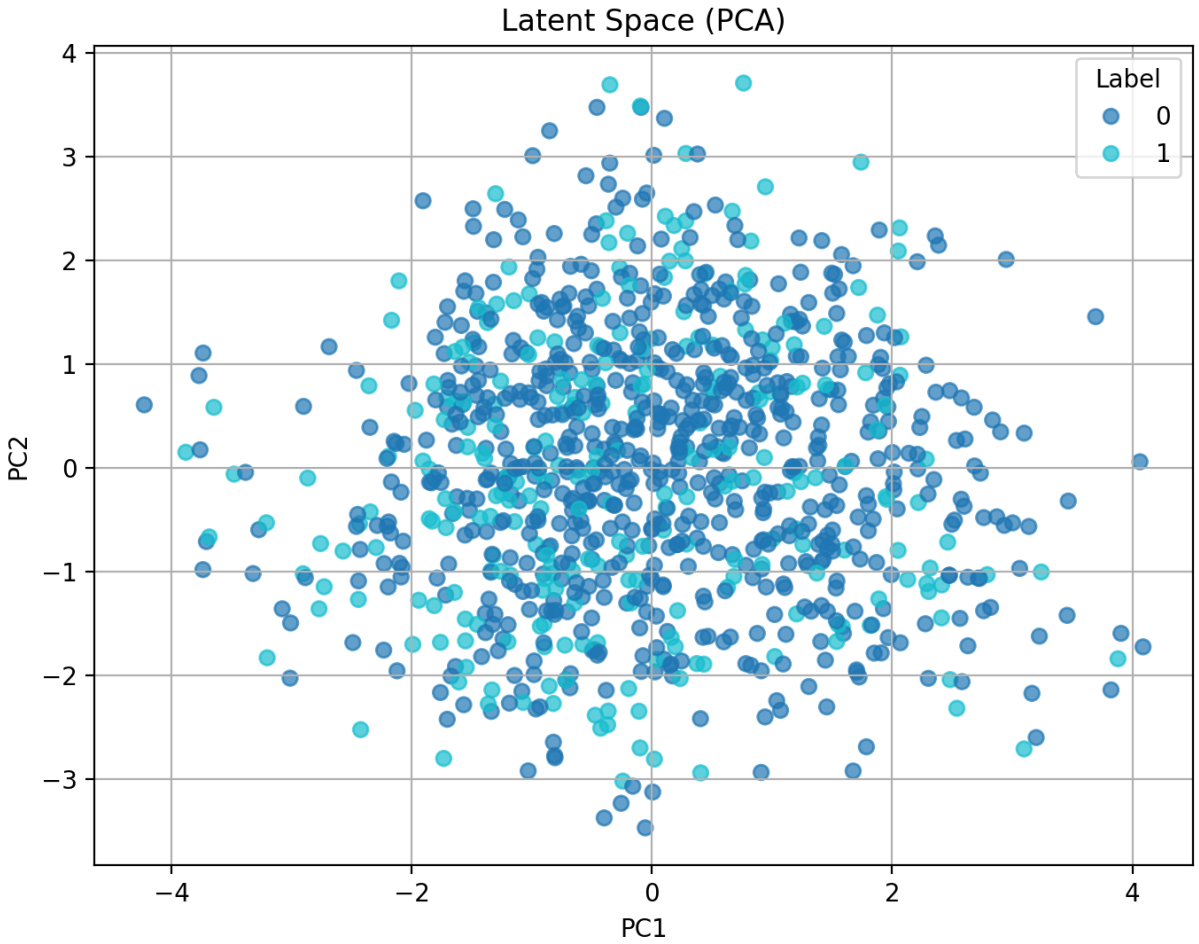

Figure 8: PCA projection of latent vectors.

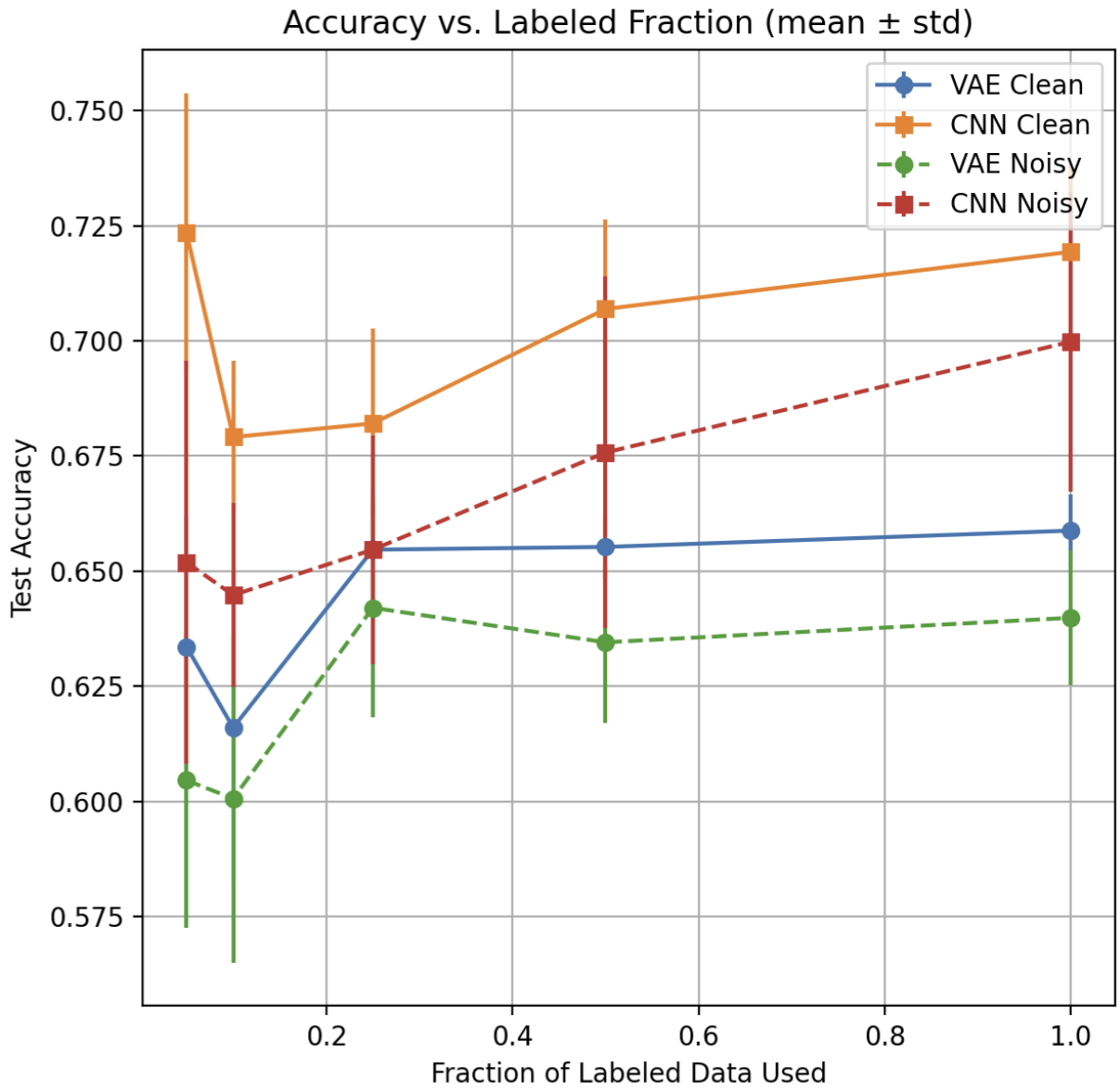

Figure 9: Test accuracy as a function of labeled data fraction, comparing VAE-based classifiers and end-to-end CNNs under both clean and noisy label conditions. Each point represents the mean of 5 independent runs.