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

Poli Dimieva – CSE3000 Research Project – 2025

Supervisors: Jesse Krijthe, Gijs van Tulder

TU Delft

INTRODUCTION

Osteoarthritis (OA) is a widespread degenerative joint disease diagnosed through radiographic images. Manual grading requires expert input, making large-scale labeled datasets scarce.

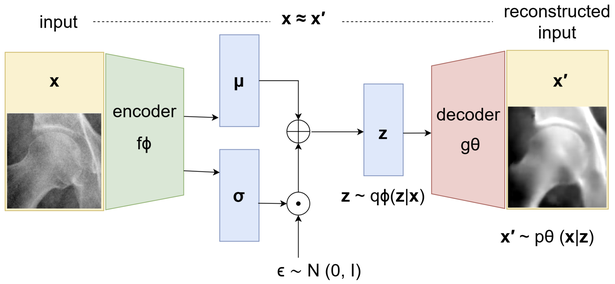
To address this, self-supervised learning has emerged as a promising approach to extract meaningful features from medical images without relying on large labelled datasets.

This study explores one such method - a Variational Autoencoder (VAE).

MOTIVATION

- OA affects 500M+ globally
- Diagnosis requires manual expert-annotated X-ray radiographs, therefore:
  - Little labelled data, which means limited utility of existing deep learning methods
  - Subjective assessment leads to noisy training data
- **Goal: Can a Variational Autoencoder (VAE) enable stable and accurate OA diagnosis with few or noisy labels?**

METHOD OVERVIEW

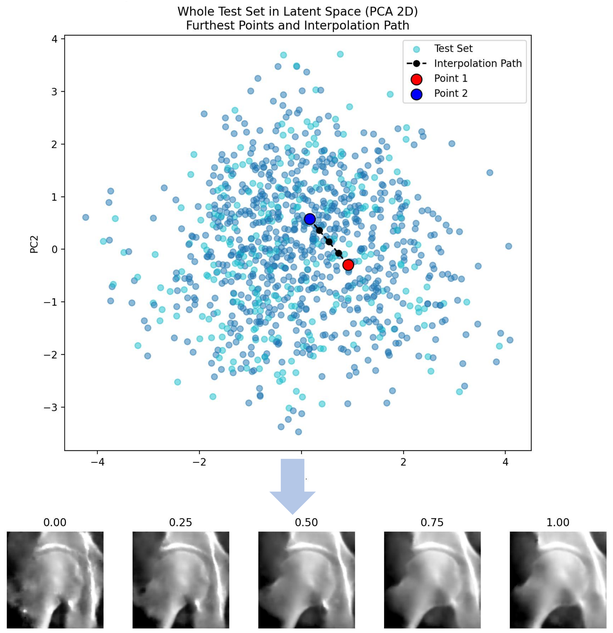


VAE learns compressed representations of hip X-rays without labels. The encoder maps input images to a smooth latent space, which can later be used for classification.

EXPERIMENTS and RESULTS

Reconstruction Quality

Trained a VAE to reconstruct preprocessed hip X-rays and used linear interpolation to verify whether the VAE captures anatomically meaningful structures in the latent space.



The VAE encodes coherent and realistic representations of hip anatomy.

Latent Space Structure

Measured average pairwise Euclidean distances between latent vectors for OA vs. non-OA groups and visualized latent space using t-SNE, UMAP, and PCA dimensionality reduction techniques to assess whether the latent space reflects diagnostic group separation without supervision.

Inter-group distance (OA vs. non-OA)**: 6.56**

Intra-group distances: **6.50** (OA), **6.52** (non-OA)

2D projections show no clear separation with t-SNE, UMAP and PCA

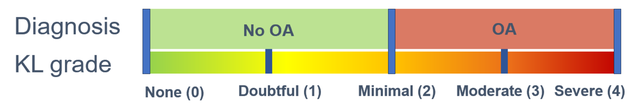
Comparison with Random and Raw Pixel Features

Used the learned VAE latent vectors, random noise and raw pixel values as inputs to identical logistic classifiers to evaluate how informative the VAE's features are for OA classification.

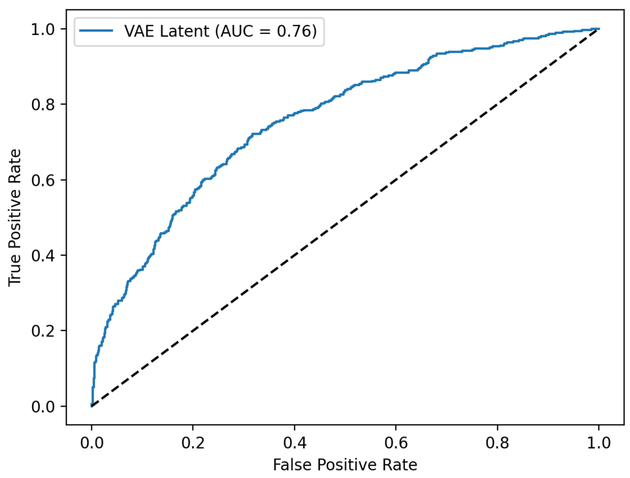
AUC results:

Random = **0.50**
Raw = **0.64**
VAE = **0.78**

Classification task



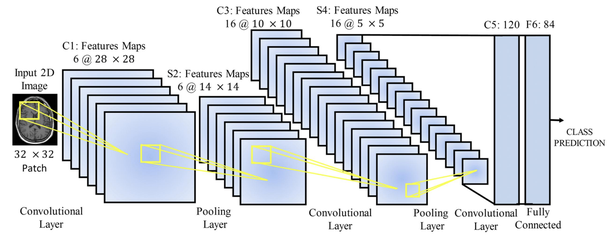
ROC curve of the classifier trained on VAE latent features for a single run

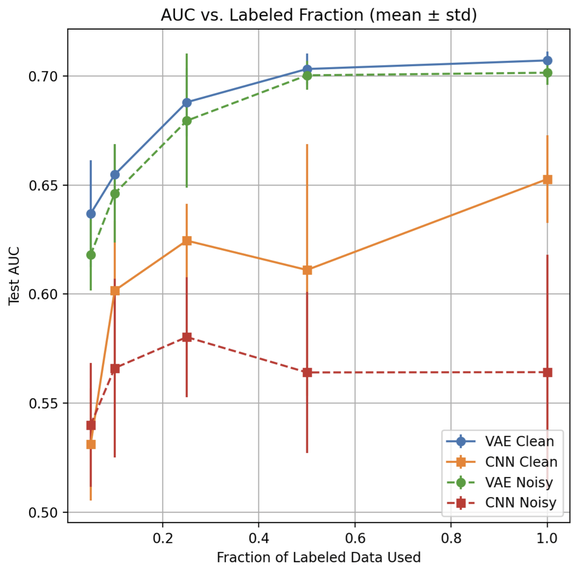


Classification Under Limited Labels and under Label Noise

Trained both VAE-based classifiers and supervised CNNs with 5%, 10%, 25%, 50%, 100% labelled data (and same splits) to test which model generalizes better with fewer labels.
Added 10% random label noise to the training set, and re-evaluated VAE and CNN models to assess sensitivity to mislabelling, a common issue in medical data.

CNN architecture





Key insights:

VAE generalizes better in low-supervision settings

VAE representations are more stable and noise-tolerant than fully supervised CNNs

DISCUSSION and CONLUSIONS

The VAE learned anatomy-aware features from unlabelled hip X-rays and enabled stable OA classification under limited or noisy labels. It outperformed a supervised CNN in these settings, showing strong robustness and generalization. Despite slightly blurry reconstructions, the model preserved key structures and captured clinically relevant variation, supporting the use of self-supervised models for scalable, label-efficient medical imaging.