## The 7-Layer Perceptron

*Deepali Bidwai, Homero Esmeraldo, Mahfuz Ahmed Anik, Mennat Allah Khalifa, Mohamed Gamal, Pawan Thapaliya, Tulika Khargonkar*

[pawanthapaliya09@gmail.com](mailto:pawanthapaliya09@gmail.com),

[gemy1231999@gmail.com](mailto:gemy1231999@gmail.com),

[mennat.abdelfattah04@eng-st.cu.edu.eg](mailto:mennat.abdelfattah04@eng-st.cu.edu.eg),

[bidwaideepali@gmail.com](mailto:bidwaideepali@gmail.com),

[tkhargonkar@gmail.com](mailto:tkhargonkar@gmail.com),

[homero.esmeraldo@gmail.com](mailto:homero.esmeraldo@gmail.com)

## Abstract

Airtable link: <https://airtable.com/appoh6RKyBvxgiJ89/shrUeDqzGe8Cplk8u>

High dimensional data can be well approximated by low-dimensional manifolds (*manifold hypothesis*) and neural networks trained on such data will have features that depend on the intrinsic dimensionality; quantifying this latent dimensionality can help understanding.

Currently, there exists no universal approach to quantify the dimensionality of the input space. We investigated if we can find a useful definition of dimensionality while working with dimensionality reduction techniques (DRTs) applied in classification tasks. To explore this, we assessed the behaviour of two DRTs across changing latent dimensionality.

We employed PCA and VAE in parallel on MNIST and CIFAR-10 datasets, generating dimensionality-reduced image datasets, using latent spaces ranging from very low-dimensional (10 PCs for PCA; 10 bottleneck units for VAE) to approaching the original input dimensionality (up to 500; for reference, original space for MNIST is 784). We measured the classification accuracy of these dimensionality-reduced images using a classifier trained with the original images.

For PCA, as the latent space dimensionality increases, the image reconstruction error gets smaller, and the classification accuracy increases, plateauing around 80 dimensions.

In contrast, with the VAE classification accuracy and reconstruction error reach a plateau earlier, at around 50 dimensions. At dimensions higher than 100, these metrics deteriorate. We verified that these results are not due to overfitting, and persist even on modifying the Gaussian regularization of the VAE (beta VAE).

Our initial results indicate that VAE’s reconstructions reach higher classification accuracy at a lower latent dimensionality compared to PCA. This suggests that VAEs are better suited to be used for measuring the dimensionality of the input space, provided that regularization is appropriate.

Our follow-up experiments are oriented towards understanding this behaviour and defining the manifold dimensionality.

—

Higher dimensional data can be accurately described by a lower-dimensional manifold, according to the manifold hypothesis (Causin & Marta, 2025). In the context of computer vision, knowing this intrinsic, lower-dimensionality of image representation can reveal information about model behaviour, in terms of task accuracy.

In this project, we approach this definition problem by observing the behaviour of two dimensionality reduction techniques: the Variational Auto-encoder (VAE), contrasted against the Principal Component Analysis(PCA) across increasing dimensions. We employ a classification task in this study, presuming that classification accuracy can serve as a proxy for identifying this minimum dimensionality that still retains class-discriminative features.

We trained a VAE and a classifier on images from the MNIST dataset and evaluated classification performance on reconstructed output images (and latent representations?), across varying latent dimensions. Parallely, we compared its behaviour to a traditional PCA-pipeline, aiming to understand which dimensionality-reduction technique could better preserve discriminative structure.

We found that VAE’s performance on classification tasks peaks, and begins to degrade at higher dimensions. In contrast, PCA’s performance plateaus after an initial high. (how much should we expand on the results in the abstract?)

Our results highlight whether VAEs yield more meaningful low-dimensional representations for classification than PCA, and how dimensionality reduction techniques, both linear and non-linear in their construction, affect classification performance.

—

**Intro and background**

High dimensional data such as natural images can be accurately described by a lower-dimensional manifold (*manifold hypothesis*; Causin & Marta, 2025). Hence, quantifying the dimensionality of this space can help us better understand its characteristics.

**Problem**

There is no universal approach to quantify the dimensionality of an input space, although dimensionality reduction techniques (DRTs) are frequently recruited. We investigated if we can find a useful way to quantify dimensionality while working with DRTs applied in classification tasks.

**Results/solution**

We employed in parallel, for comparison, VAE and PCA to reduce dimensionality of images in MNIST and CIFAR-10 datasets. We created dimensionality-reduced image datasets generated using latent spaces ranging from very low-dimensional (around 10 PCs, for PCA; 10 bottleneck units, for VAE) to approaching the original dimensionality of the input space (around 500; original space for MNIST is 784, for example).

Then, we measured the classification accuracy of these dimensionality-reduced images using a classifier trained with the original images.

We observed that for PCA, as the latent space dimensionality increases, the image reconstruction error gets smaller and the classification accuracy increases progressively. The classification accuracy plateaus already at around 80 dimensions in the latent space. On the other hand, classification accuracy and reconstruction error with VAE reach a plateau earlier, at around 30 dimensions, and latent spaces higher-dimensional than 100, these metrics continually deteriorate. These results are not due to overfitting, and when we introduce beta parameter to control the Gaussian regularization of the VAE nor dependent on beta (Higgins et al, 2017)

**Impact**

Our results indicate that VAEs are better suited to be used for measuring dimensionality of spaces.

Defining the behaviour of DRTs on changing latent space dimensionality can increase our understanding of the manifold dimensionality.

—

This project investigates how the size of the latent space in a Variational Autoencoder (VAE) affects classification accuracy on reconstructed images. A VAE is trained on the MNIST dataset to learn compressed representations of digit images at varying latent dimensions (ranging from 0 to 600 in steps of 20). To evaluate the quality and utility of these representations, a classifier trained on original images is tested on the VAE-reconstructed outputs. The goal is to identify the **optimal latent dimensionality that maximizes classification performance on reconstructed data**.

For each latent size, we record VAE training and validation losses (total, BCE, and KLD), classifier accuracy and loss, and reconstruction errors. The learned latent spaces are visualized using PCA and t-SNE, and results are compared with a traditional PCA-based dimensionality reduction approach. Additional evaluations include confusion matrices, classification reports, and original-vs-reconstructed image comparisons.

Our findings demonstrate

## Questions Draft

(*what is the question/idea that we are trying to solve? 3-4 sentences about the project that gives a general idea of what we’re trying to accomplish)*

***Question: How should the dimensionality of the manifold structure of images be quantified?***

***Motivation: The dimensionality of the input manifold is hypothesized to be a cause for representational similarity across different networks trained in classification tasks. Appropriately measuring dimensionality of the input is essential to be able to test this hypothesis. Furthermore, the dimensionality of the input also potentially influences the number of parameters needed for a network to function well.***

***Solution***

**We use VAE and PCA to project and reconstruct images, then we train a network classifier for those images, determining its classification accuracy. We observe how this accuracy changes across different number of dimensions. The dimensionality that yields best classification accuracy will be defined as the dimensionality of the input manifold.**

*For classification, using a DRT might evade some noise, allowing it to focus on the significant features. We want to study if there is an optimal number of latent dimensions to be kept so as to maximize classification accuracy, and if that number varies depending on the dimensionality reduction technique used.*

**CONTEXT**: The visual world does not occupy the whole dimensionality of the pixel space of images. We are interested in understanding the dimensionality of the manifold structure of images and how it enables class discrimination.

To manipulate dimensionality, we use dimensionality reduction techniques to project and reconstruct images. We approach this through the lens of classification.

*Option 1: try to quantify dim of manifold - use VAE or use PCA*

*Op2: we assume the dim given by dim red techniques, then check how they affect classification*

In this project, we are interested in looking at two key possible ways to define dimensionality of the manifold:

* First, whether VAE fares better than methods like PCA in preserving key image features important for classification with more accuracy,across MNIST and CIFAR-10. Results will highlight whether VAE latent space can better retain discriminative information.
  + If so, what aspect of VAE does the heavy lifting for retaining such information? Bottleneck/Reconstructed image?
* Second, *linear* dimensionality reduction techniques (like PCA) improve reconstruction with increased dimensions. Does VAE replicate this behaviour with increasing dimensions, or does its non-linear nature lead to a different behaviour?
  + If so, how many dimensions are “enough” to capture relevant class-discriminative architecture?

Parallel motivating hypothesis: increasing dimensionality of images, will decrease the similarity of representations of networks used to classify those images. Problem: we need to define dimensionality.

*Suggestion: should we double down on the efficiency of using dimensionality reduction techniques for computer vision tasks (specifically classification) - detail what makes VAEs different from other dimensionality reduction techniques - then suggest exploration of VAE behaviour on modifying dimensions?* (is it convincing enough?)

How we can learn a “meaningful” latent space

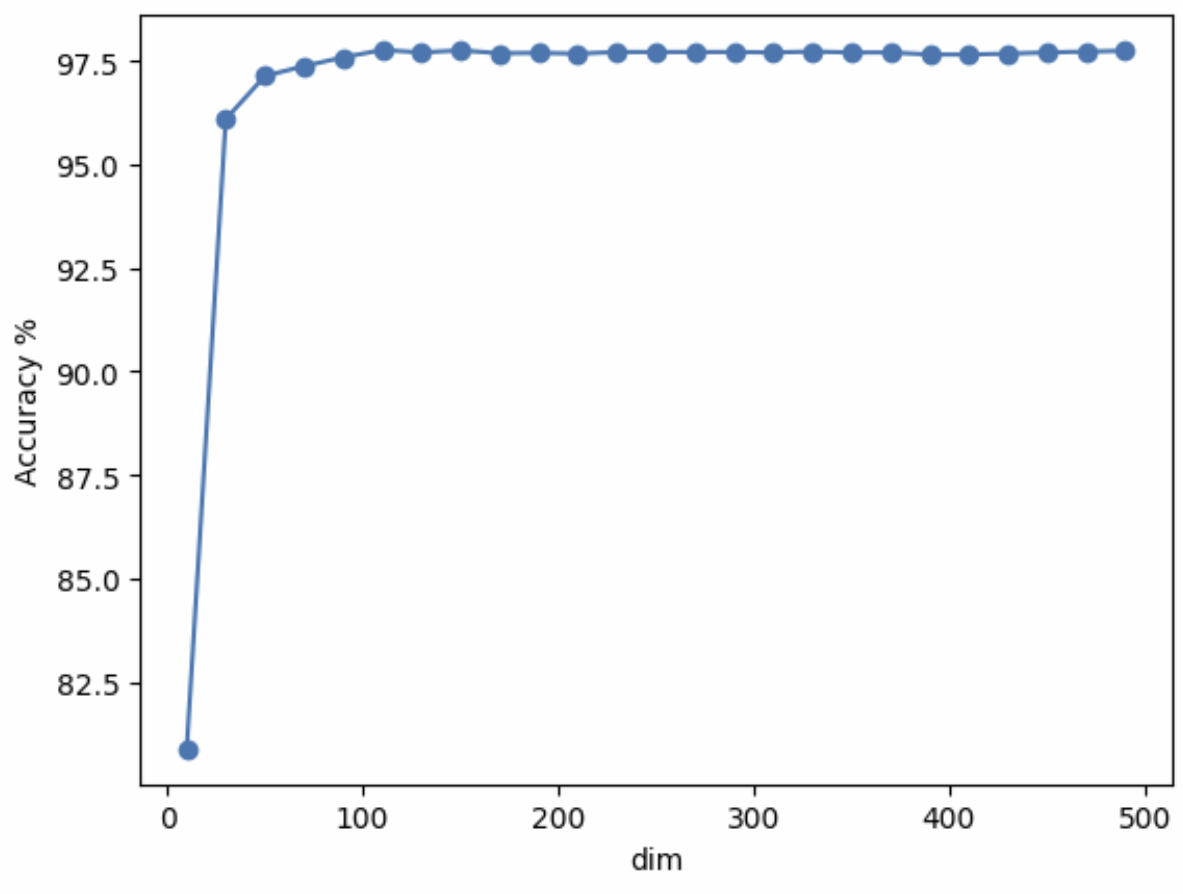
How it can be used for classification

***Old question:***

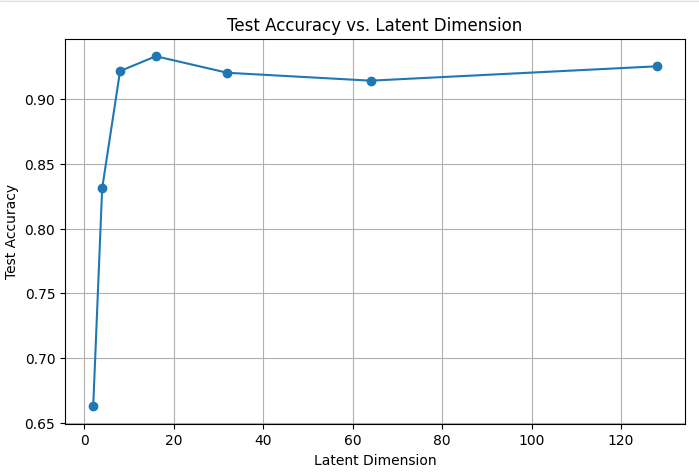
***Can dimensionality reduction techniques retain image features significantly, and if so, what are the optimal parameters(dimensions) at which they can do so?***

## Current results

PCA-reduced images:



VAE result:

 beta=0.1, alpha=0

X values: Training with latent\_dim = 2 Test Accuracy @ z=2: 0.6627 Training with latent\_dim = 4 Test Accuracy @ z=4: 0.8311 Training with latent\_dim = 8 Test Accuracy @ z=8: 0.9218 Training with latent\_dim = 16 Test Accuracy @ z=16: 0.9333 Training with latent\_dim = 32 Test Accuracy @ z=32: 0.9205 Training with latent\_dim = 64 Test Accuracy @ z=64: 0.9144 Training with latent\_dim = 128 Test Accuracy @ z=128: 0.9255

## 

latent\_dims = [1, 10, 25, 50, 100, 250, 300, 400, 500, 600]

First version:

## 

Initial hypothesis:



VAE appears to indeed have much higher accuracy for lower dimensional latent space

## 2027-07-21

\* Homero: train CIFAR classifier

\* Tulika: Experiment with beta-VAE to see if it improves reconstruction quality and classification accuracy > can help explain VAE behaviour over higher dimensions (try alpha=0, beta [0.5, 1]

\* Pawan:

\* Mohamed: check if VAE on CIFAR gets better reconstruction quality with bigger latent space (e.g. DST\_DIMS = [10, 35, 500])

\* Deepali:

## 2025-07-19

\* Deepali: the best classifier model and store it for later evaluation

\* Homero: plot acc vs dim for PCA

\* Pawan: train CIFAR classifier

\* Tulika: Experiment with beta-VAE to see if it improves reconstruction quality and classification accuracy > can help explain VAE behaviour over higher dimensions (try alpha=0, beta [0, 0.5, 1]

\* Mohamed: check if VAE on CIFAR gets better reconstruction quality with bigger latent space (e.g. DST\_DIMS = [10, 35, 500])

VAE Loss: Reconstruction + beta\*KLdiv with Gaussian + alpha\* CELoss(1-layer MLP classifier)

## 2025-07-16 notes

The classifier is currently trained on original images and evaluated on reconstructed images. Do we want to keep it that way?

We want to measure dimensionality of the input manifold. If we train the classifier on the reconstructed images the VAE could build a fake reconstruction with pixels that have a one-hot representation of the labels (for example, if alpha is very big and the reconstruction loss is small). If VAE has a focus only on reconstruction, then the classifier will try its best based on the reconstructed image. So, yes, let’s train it on the original images. Get the model that performs best on original test set (without projection and subsequent reconstruction)

The importance of being VAE instead of just a simple AE is that the latent space will be “meaningful”. But how does that make it more meaningful to the definition of dimensionality? Maybe it is geometrically more accurate than an arbitrary base that a regular AE would learn?

How should we argue for measuring classification accuracy in the reconstructed images instead of the bottleneck? One reason is because it makes the classification space the same, and the classifier the same too. The bottleneck varies in size, so the classifier would change too.

Doing it on the bottleneck makes us more independent of the bottleneck

**todo:**

1. Integrate Deepali’s code on the baseline into Pawan’s plot for original images
2. Integrate training curves into Pawan’s code as well - make sure nothing is overfitting
3. Use diff classifier: [resnet18\_cifar10 · Hugging Face](https://huggingface.co/edadaltocg/resnet18_cifar10) for CIFAR
4. Run classifier on PCA reconstructed images

## 2025-07-15 Notes

Discussion with Mobin:

* Try training VAE with latent dimensions in the step of 50 up till 500
* Check learning curves for overfitting AGAIN

**Thinking:**

Why is the accuracy going down for more dimensions?

* MNIST data is not that complex, might be overfitting

**Can we plot PCA and VAE together, considering PCA is linear, and VAE is not - what values will be on the x axis for a plot of Accuracy vs Dimensions?**

**todo:**

Use train, val, test split

Use early stopping (*this isn’t needed now, right?)*

*- no, provided we check the training curve and make sure that both the VAE and the classifier are neither overfitting, nor underfitting*

Plot training curves for both classifier and VAE to make sure it is not overfitting

Apply classifier on the original data and include it as a baseline (plot as a dashed line in the bar plot (accuracy vs dims)

Run PCA on the same dataset too

Make CIFAR dataset work

## 2025-07-09 Todo

Try different datasets

Create dimensionality-customized datasets. I.e. dim = arange(10, 500, 25)

* Plot reconstruction error over the course of learning for training and test splits: reconstruction error vs epochs
* Max epochs=20, but do early stopping if test error increases for 3 epochs

Train classifier on bottleneck (z) neurons

Train classifier on reconstructed outputs

Train classifier on reconstructed/bottleneck after silencing bottleneck (z) neurons

–

Group 1:

* Pawan ([pawanthapaliya09@gmail.com](mailto:pawanthapaliya09@gmail.com))
* Mohamed ([gemy1231999@gmail.com](mailto:gemy1231999@gmail.com))
* Mennat ([mennat.abdelfattah04@eng-st.cu.edu.eg](mailto:mennat.abdelfattah04@eng-st.cu.edu.eg))

Group 2:

* Deepali ([bidwaideepali@gmail.com](mailto:bidwaideepali@gmail.com))
* Tulika ([tkhargonkar@gmail.com](mailto:tkhargonkar@gmail.com))
* Homero ([homero.esmeraldo@gmail.com](mailto:homero.esmeraldo@gmail.com))

## Project Questions

1. **How dimensionality-reduction in images affect accuracy of classification** 
   * Can dimensionality reduction with VAE better preserve features important for image classification than other common dimensionality reduction algorithms?
   * How does reducing image dimensionality using VAEs (or PCA, ICA, MDS, …) impact classification accuracy?
   * Task: object classification in datasets like MNIST, CIFAR-100, COCO, Places 365, Imagenet
   * classification accuracy metric: percentage of correctly labeled inputs
   * If we have time:
   * (object detection / segmentation)
   * Have a working Brain classification task
   * Adapt VAE to be 3D to use in brain data
2. (OLD) How does the dimensionality of the input statistical structure affect the representational similarity of the networks trained on them?

* Have some data for PCA; how does VAE behave in this content? Why are we interested in VAE? (the nonlinearity could deal with the data better)
* How does it help explain the representational similarity of networks?
* How does changing dimensionality affect this? What are we expecting?

## 

## Project ideas

### Can a VAE compress and Reconstruct Brain MRI slices in a way that preserves class-discriminative anatomical features for downstream classification tasks (e.g., disease vs. healthy)?

**Q: Since Mobin mentioned that people analysing medical data steer away from dimensionality reduction – do we have an indication/citation to suggest that VAE might somehow alleviate those concerns?**

[**https://doi.org/10.1016/j.media.2020.101952**](https://doi.org/10.1016/j.media.2020.101952) **—-- this paper has compared several auto encoder**

Do we have a simple dataset and task for this one?

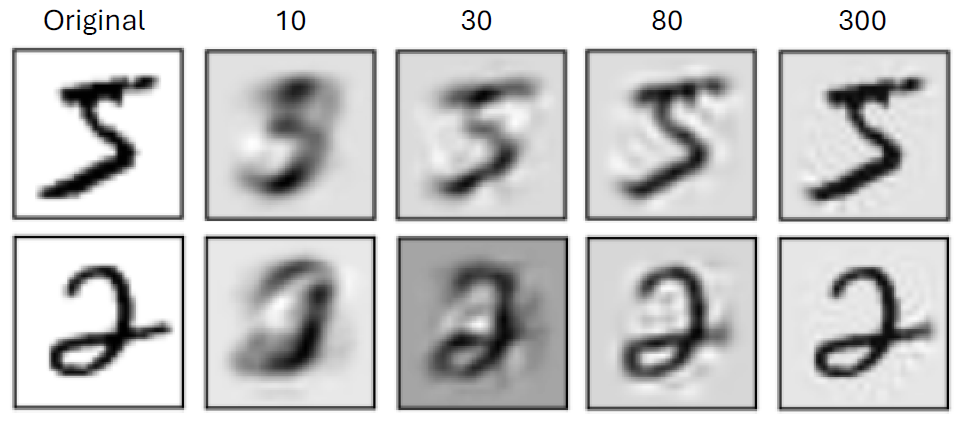
This can be done in parallel with classification tasks in images —> Yes, kaggle has dataset.

<https://www.kaggle.com/datasets/navoneel/brain-mri-images-for-brain-tumor-detection>

### How dimensionality-reduction images affect (object detection / segmentation /or classification)

Use VAEs to perform dimensionality reduction and understand how that affects the accuracy of (object detection / segmentation /or classification)

* Comparative study for common dimensionality reduction algorithms



MNIST images projected into PCA space of 10, 30, 80 or 300 dimensions and reconstructed

* Apply VAE through the dimensionality reduction lens on EEG data; fine tune it on limited amount of data



### VAE

train VAEs on CIFAR-100 or ImageNet and silence some of the bottleneck neurons to see how that would affect the images.

Check how the dimensionality of the input images affect representational similarity of trained networks. I.e. Can the dimensionality of the images in the training dataset be related to [The Platonic Representation Hypothesis](https://r.jordan.im/download/technology/huh2024.pdf) (Huh et al, 2024)?

**Question**: how does the dimensionality of the input statistical structure affect the representational similarity of the networks trained on them?

Other ideas:

* Denoising tools for EEG data where noise is prevalent
* Generalising VAEs - is it possible to measure/improve upon generalisation capabilities?
* VQ (Vector Quantised) VAE: will use less data/weight.

### Object detection, Image segmentation

YOLOV5, YOLO8, Unet, Segment anything

YOLOv12 + SAM 2

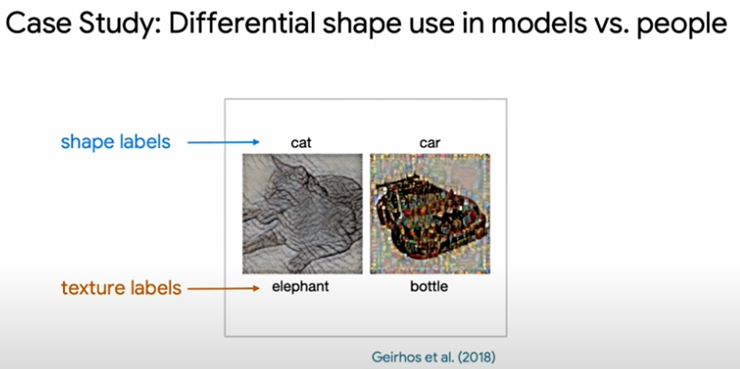
Qn: do what with this?

* *Use case: Can we automatically identify and isolate regions of interest (ROIs) in brain scans?*

### Shape vs Texture bias in vision networks

[GAC 2 talk from Katherine Hermann](https://www.youtube.com/watch?v=LSczBZBHVkM&t=260s&ab_channel=CognitiveComputationalNeuroscience) (minute 40)

[Hermann et al (2020) The Origins and Prevalence of Texture Bias in Convolutional Neural Networks](https://arxiv.org/abs/1911.09071)



**Question:** Are there any representational signatures of texture bias or shape bias in networks? I.e. just looking at representations in middle layers, can we predict if the network has shape vs texture bias?

## Datasets

* <https://www.kaggle.com/code/pawanthapaliya/brain-mri-tumor-classification-cnn-90-percent>
* MRi/Spectrogram EEG data
* <https://www.kaggle.com/datasets/cdeotte/brain-eeg-spectrograms/code>
* <https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset>
* <https://openneuro.org/>
* [Ideas — Neuromatch Academy: Deep Learning](https://chatify.deeplearning.neuromatch.io/projects/Neuroscience/ideas_and_datasets.html) - Neuroscience
* <https://arxiv.org/abs/2103.01071>
* [Natural Scenes Dataset](https://naturalscenesdataset.org/)

Allen et al (2021) [A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence | Nature Neuroscience](https://www.nature.com/articles/s41593-021-00962-x)

large-scale fMRI dataset - 8 healthy adult subjects while they each viewed 9,000–10,000 distinct, color natural scenes (22,500–30,000 trials) over the course of 30–40 scan sessions.

Aggregated across participants, NSD includes responses to 70,566 distinct natural scene images—this is more than an order of magnitude larger than similar datasets involving fMRI sampling of many images

4.2G ./dorsal\_visual\_data

8.5G ./lateral\_visual\_data

1.8M ./masks

597M ./V1v\_data

15G ./ventral\_visual\_data

581M ./v4\_data

693M ./V2v\_data

570M ./V3v\_data

2.1G ./pycortex\_files

706M ./V1d\_data

568M ./V2d\_data

531M ./V3d\_data

34G .

We may be able to use just V1 data. If we use ventral and dorsal it is 1 Gb, which may be doable. Ask Reza

* Kay al et Gallant dataset

fMRI responses to images

Paper: [Identifying natural images from human brain activity | Nature](https://www.nature.com/articles/nature06713)

\*[load\_kay\_images - Colab](https://colab.research.google.com/github/NeuromatchAcademy/course-content/blob/master/projects/fMRI/load_kay_images.ipynb)

\*video about dataset [Projects Dataset: Human fMRI responses while viewing natural images (Kay/Gallant)](https://www.youtube.com/watch?v=LdJkLyw4yzg&ab_channel=NeuromatchAcademy)

\* [Moving beyond Labels: Finetuning CNNs on BOLD response — Neuromatch Academy: Deep Learning](https://chatify.deeplearning.neuromatch.io/projects/Neuroscience/finetuning_fmri.html)

* Algonauts dataset

fMRI responses to videos

[algonauts\_videos - Colab](https://colab.research.google.com/github/NeuromatchAcademy/course-content-dl/blob/main/projects/Neuroscience/algonauts_videos.ipynb)

[Load algonauts videos — Neuromatch Academy: Deep Learning](https://chatify.deeplearning.neuromatch.io/projects/Neuroscience/algonauts_videos.html)

* Brain MRI Images for Brain Tumor Detection (2020)

Modality: 2D T1-weighted grayscale MRIs Data: 253 normal and 1,550 tumor images Format: Pre-extracted 2D PNG images (no need for slicing)

Use Case: Directly train VAE on normal images; detect tumors via recon error

Source: Kaggle - Brain Tumor MRI

* [Tri-VAE: Triplet Variational Autoencoder for Unsupervised Anomaly Detection](https://openaccess.thecvf.com/content/CVPR2024W/VAND/papers/Wijanarko_Tri-VAE_Triplet_Variational_Autoencoder_for_Unsupervised_Anomaly_Detection_in_Brain_CVPRW_2024_paper.pdf)

[in Brain Tumor MRI](https://openaccess.thecvf.com/content/CVPR2024W/VAND/papers/Wijanarko_Tri-VAE_Triplet_Variational_Autoencoder_for_Unsupervised_Anomaly_Detection_in_Brain_CVPRW_2024_paper.pdf)

Hansen Wijanarko, Evelyne Calista, Li-Fen Chen, Yong-Sheng Chen

National Yang Ming Chiao Tung University

Taiwan

Summary of this Paper:

**Summary:** Extends standard VAE by incorporating a triplet loss to better separate normal and anomalous brain MRI latent features. Enhances tumor localization accuracy.

**Builds on:** Classic VAE latent space anomaly detection but adds discriminative latent loss

## 

## Reading about VAEs

[Tutorial - What is a variational autoencoder? – Jaan Lı 李](https://jaan.io/what-is-variational-autoencoder-vae-tutorial/)

[Understanding Vector Quantized Variational Autoencoders (VQ-VAE) | by Shashank Yadav | Medium](https://shashank7-iitd.medium.com/understanding-vector-quantized-variational-autoencoders-vq-vae-323d710a888a)

CNN-VAE example: [pytorch-vae/vae-cnn.ipynb at master · sksq96/pytorch-vae · GitHub](https://github.com/sksq96/pytorch-vae/blob/master/vae-cnn.ipynb)

VAEs are ideal for:

* **Compressing and reconstructing images** (e.g., chest X-rays, MR slices)
* **Learning meaningful latent spaces** (e.g., clustering or progression modeling)
* **Detecting anomalies** (e.g., out-of-distribution tumors, fractures)
* **Generating synthetic medical images** (for data augmentation or privacy)

## Other resources

* [*Videos for Modelling principles and practice*](https://deeplearning.neuromatch.io/projects/modelingsteps/ModelingSteps_1through2_DL.html)
* [Ideas webpage](https://deeplearning.neuromatch.io/projects/ComputerVision/ideas_and_datasets.html)
* <https://universe.roboflow.com/>

<https://arxiv.org/abs/2103.01071>