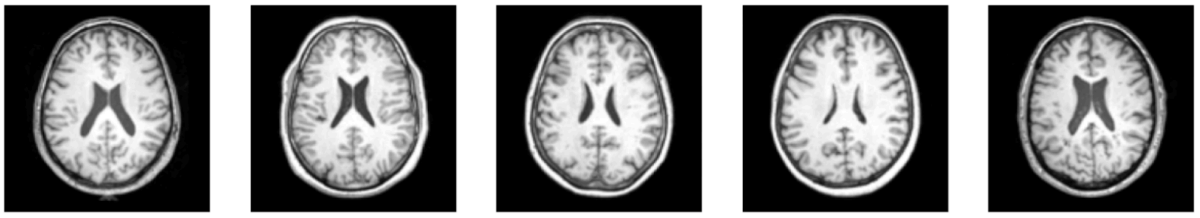
Practical AIM2 - generative models

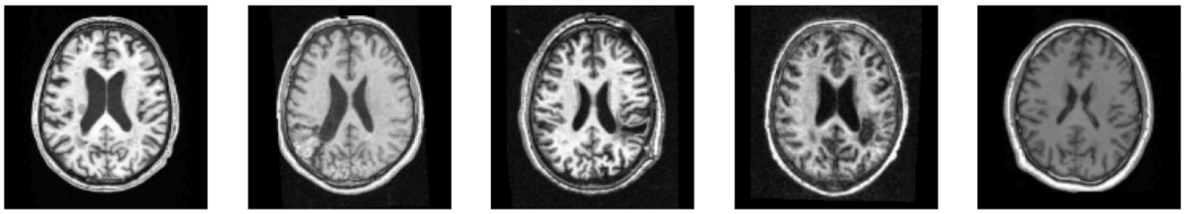
| Mohammed Elbushnaq – 03786474 – go56cuh
| Marina Aoki – 03773384 – ge94naz
| Sama Elbaroudy – 03768259 – ge83muj
| Sarah Berbuir – 03767700 – ge83cak

1. Understanding Data

healthy:



stroke:

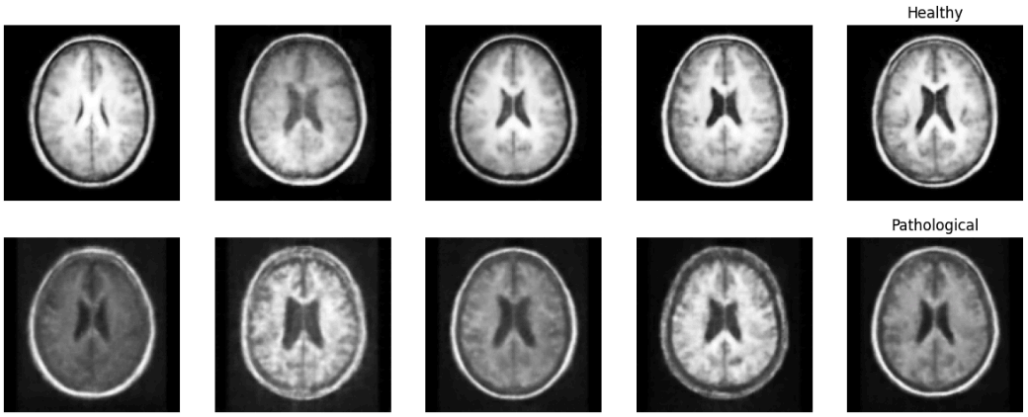


2. Understanding VAEs

T1 (VAE)

Here is the completion of the VAE code:

```
################ TASK 1: Fill out the missing lines (#TODO)
def vae_loss(self, x, x_hat, mu, logvar):
    reconstruction_loss = F.mse_loss(x_hat, x, reduction='sum') # TODO is mse reconstruction loss
    kl_divergence = -0.5 * torch.sum(1 + logvar - mu.pow(2) - logvar.exp())
    return reconstruction_loss + kl_divergence
```



Q1 (VAE)

Describe the quality of the generated images from the VAE model. Are there any noticeable differences between the healthy and pathological (stroke lesion) cases? Provide possible reasons for the observed similarities or differences.

The real healthy scans exhibit no lesions and have consistent, smooth brain tissue. They are generally symmetrical with mirrored structures on both sides and consistent contrast levels. The generated images from the VAE maintain the reflection of typical brain anatomy, appearing quite clear though not as sharp as the real ones. The consistency and symmetry are preserved, and overall, the generated images look realistic and uniform.

Stroke-affected images show irregularities and variations in structure due to the lesions. They are asymmetrical, with inconsistent contrast across the brain image. Lesion areas appear distinctly lighter or darker compared to the surrounding tissue. The reconstructed images from the VAE for stroke cases show some inconsistency. While they tend to be more symmetrical and do not accurately capture the lesions, the variations in texture and contrast are not as pronounced as in the real images. Lesion areas might be somewhat reflected, but not consistently or accurately.

Notable Differences Between Healthy and Pathological Cases are for example the symmetry, while in the healthy case both real and generated image are generally symmetrical, in the pathological cases real images are often asymmetrical due to lesions, but the generated images tend to maintain symmetry and do not capture the asymmetries accurately. Regarding texture and contrast, healthy images (real and generated) show this consistent, while pathological real images show variations in texture and contrast (especially around lesions), but the generated pathological images do not really capture these variations.
Possible reasons for those differences in generating images from healthy images and generating images from stroke images could be that there is more uniformity and less variability in healthy brain structures, which the VAE can capture more

accurately, in contrast the variability in lesion appearance (size, shape, location) poses a challenge for the VAE.

In conclusion the quality of VAE-generated images reflects these differences, with healthy brain images being more consistent and precise compared to the more variable and challenging reconstructions of stroke-affected brains.

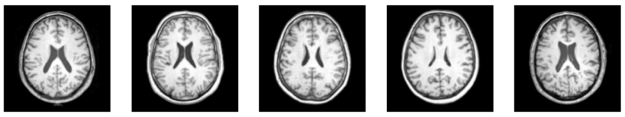
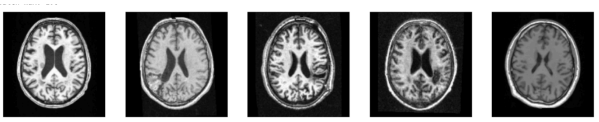
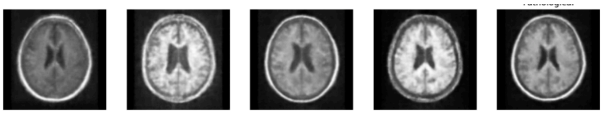
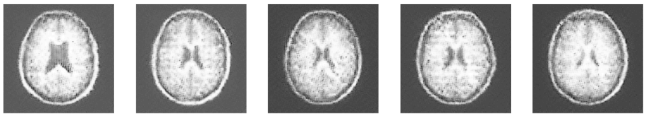
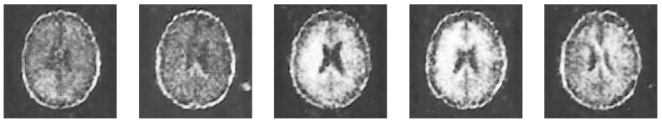
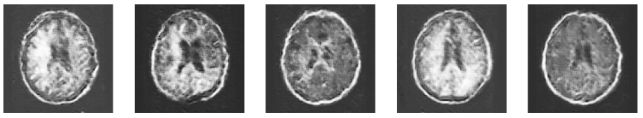
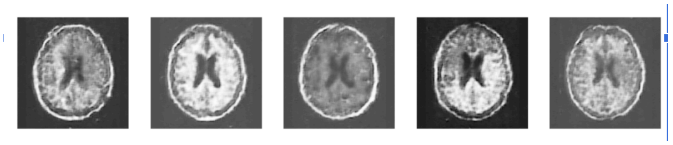
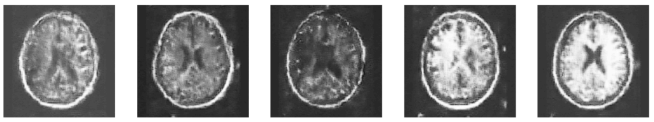
Q2 (VAE)

Suggest potential ways to improve the quality and diversity of the generated images, particularly for the pathological cases.

Improving the quality and diversity of VAE-generated images, especially for pathological cases, could be achieved through various approaches. These include augmenting the training data, improving the model architecture, using advanced training techniques, refining the latent space representations, and using transfer learning and domain adaptation. More specifically, we could utilise data augmentation and apply transformations such as rotation and scaling to artificially increase the diversity of the training dataset if it is possible to introduce variability in lesion size, shape and location to allow the VAE model to better explore the pathology behind the lesions in the stroke data. Additional datasets of pathological brain MRIs may also be possible. Another possibility would be to increase the depth and complexity of the encoder and decoder networks to better capture intricate details of the lesions. Another possibility, which involves more effort but may be worth trying, is to incorporate some form of supervised learning to control the organisation of the latent space so that it is more interpretable and better suited to generating different images.

Another area to explore is gradient steering, which improves the quality and variety of images generated with VAE by using gradient-based information to better control the generation process, especially for capturing complex features such as lesions in pathological cases. Techniques include gradient-based regularisation to enforce latent space smoothing, latent space optimisation to improve realistic features, and guided latent sampling to ensure that the generated images reflect accurate pathological features. Together, these methods could enhance VAE's ability to produce high-quality, diverse and realistic images.

3. Understanding GANs

	healthy	stroke
Real		
VAE		
GAN	epoch 200:  epoch 400:  epoch 600:  epoch 800: 	epoch 200:  epoch 400:  epoch 600:  epoch 800: 

Describe the quality of the generated brain MRI images from the GAN model. How do they compare to the real images in terms of realism and diversity? How do they compare to VAEs?

The following description of the quality comes from GAN results using the provided config and the provided epoch size (200)

The healthy MRI images generated by the GAN are quite realistic, capturing the overall symmetry and anatomical features accurately. However, they suffer from significant blurriness, which detracts from the fine details. The diversity of these GAN-generated healthy images is generally on par with the real images, with a good representation of varied anatomical structures, especially in the central regions.

For the pathological cases, while the generated images also exhibit considerable blurriness, they manage to capture some aspects of the asymmetry typical of stroke lesions. Compared to VAEs, the GAN-generated lesions are somewhat more pronounced, making the pathological features slightly clearer. Additionally, the diversity of pathological images is well-maintained. However, like the healthy images, the pathological images are also very blurry.

Overall, for both healthy and pathological cases, the GAN-generated images are significantly blurrier and more pixelated than those produced by VAEs.

We observed an interesting trend with increasing epoch size. As expected, given the nature of the model, the blurring decreases with more epochs. Notably, lesions become more prominent and the asymmetry in pathological images is highly pronounced. However, despite these improvements, the blurriness in the GAN outputs does not surpass that of the VAE.

Q4 (GAN) **How can me measure the quality of the reconstruction for GANs? Implement a metric and compare the similarity to the pathological and healthy test sets.**

To measure the quality of reconstructions produced by Generative Adversarial Networks (GANs), it's important to evaluate how well the generated data aligns with the real data distribution and maintains fidelity. One common metric is the inception score, which assesses the diversity and quality of the generated images. Another widely used metric is the Frechet Inception Distance (FID), which compares the distribution of real and generated images; this is the metric we implemented. Additionally, traditional metrics such as recall, precision, and Mean Squared Error (MSE) can be employed to measure pixel-wise differences between real and generated images, ensuring high fidelity in the reconstructions. Visual evaluation also plays a role in assessing the overall quality and fidelity of the reconstructions.

See further interpretation and details about FID in T2.

T2 (GAN)

Implement a metric and compare the similarity to the pathological and healthy test sets. (Hint: The metric should evaluate the closeness of two distributions)

The Fréchet Inception Distance (FID) compares the distribution of generated images to that of real images. We use a pre-trained inception v3 network to extract features from the images and then we compare the mean and standard deviation of real and generated images. It looks at the fidelity and how realistic the generated images are as well as how varied the images are. The lower the resulting number is, the better, this indicates that the generated images are similar to real images in terms of distributions of their features.
The function calculate_fid calculates this function:

$$d_F(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\mu', \Sigma'))^2 = \|\mu - \mu'\|_2^2 + \mathrm{tr}\left(\Sigma + \Sigma' - 2(\Sigma\Sigma')^{\frac{1}{2}}\right)$$

First we ran healthy & real against healthy & real to test the implementation, which resulted in -0.0 as expected. In the following table are the FID scores for our different runs using different epochs

(It was not really obvious where to find a test dataset which is why we computed those scores using the val_loader())

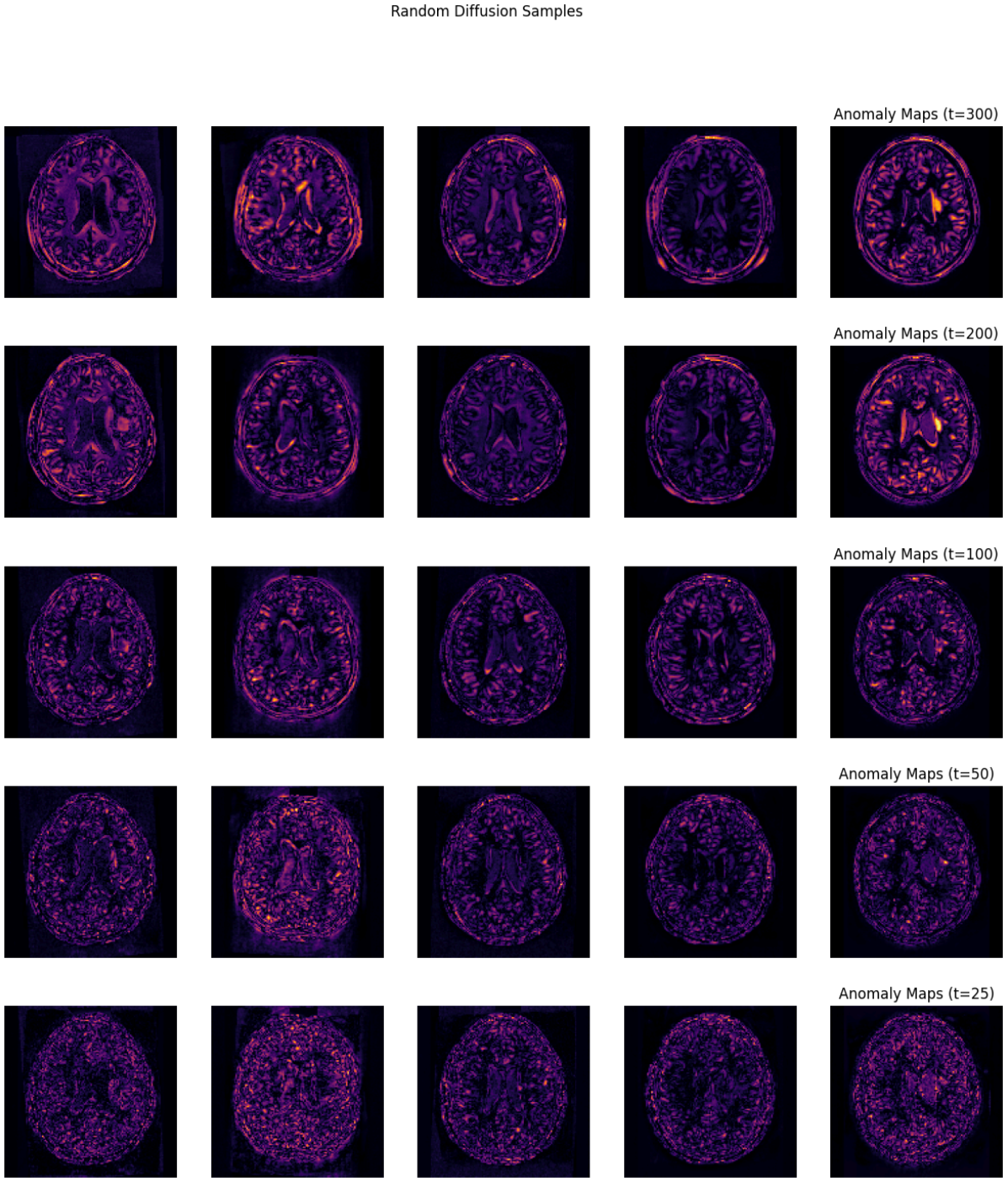
epoch size	healthy: real - generated	stroke: real - generated
200	187	217.966
400	89.365	83.139
600	86.217	82.826
800	175.848	96.499

At epoch 200, the scores are relatively high, with 187 for healthy and 217.966 for stroke, indicating a significant difference between real and generated images. By epoch 400, the scores improve markedly, with 89.365 for healthy and 83.139 for stroke, suggesting better quality and closer resemblance to real images. This trend continues at epoch 600, where the scores further decrease to 86.217 for healthy and 82.826 for stroke. However, at epoch 800, there is a mixed result: while the stroke score improves to 96.499, the healthy score increases substantially to 175.848, implying potential overfitting or other issues affecting the generation quality for healthy images.

4. Understanding Diffusion Models

Q5 (Diffusion)
How does the anomaly detection performance vary with varying noise levels t?

We investigated the effect of different values of t by visualising the anomaly detection for one batch while varying $t \in \{300, 200, 100, 50, 25\}$. The results can be seen below.



For larger values of t, the anomalies that are detected seem to capture more of the structures inherent to the input images. However, for t=300, it appears as if most of the brain region is highlighted as an anomaly. As the value of t decreases, the anomalies seem to

“scatter” and appear more disappeared. They do not seem to correspond to realistic anatomical structures of the brain.

Overall, there seems to be some kind of trade-off between the level of detail that can be reconstructed and the number of anomalies that are erased due to the addition of noise.

Therefore, it would be advisable to choose a value of t that is not too large and not too small.

For tasks T3 and T4, please refer to the attached notebook.

5. Bonus: Vision-Language Generative Models (VLMs)

Q6 (VLMS)

How Vision-Language Generative Models can be used to generate samples.

Vision-Language Generative Models (VLMs) for generating brain MRI scans involve sophisticated methods to integrate visual and textual data. These models can create new MRI scans from text descriptions or generate detailed reports from MRI images, aiding in medical diagnostics and research

Training Process

1. **Data Preparation**
    a. Data Collection: Collecting training data for VLMs is more challenging than collecting data for traditional AI models since it involves acquiring large datasets of brain MRI scans paired with detailed reports.
    b. Data Preprocessing: Normalize MRI images and tokenize text. Images might be resized and augmented, while text is cleaned, tokenized, and maybe embedded using pre-trained language models. Segmenting MRI scans into relevant brain regions can further enhance training.
2. **Model Architecture**: Modern frameworks can understand complex relationships between different modalities and deliver cutting-edge outcomes.
    a. **Contrastive learning** is a method that helps understand how data points differ from each other by calculating a similarity score between data instances and reducing contrastive loss. An example is the CLIP model, which uses contrastive learning to find similarities between text and image embeddings through text and image encoders.
        i. During pretraining, it teaches text and image encoders to learn from image-text pairs.
        ii. It changes the training dataset classes into captions.
        iii. It finds the best caption for a given input image for zero-shot prediction.
    b. Masked language-image modeling: is another approach where parts of the input (either text or image) are masked, and the model is trained to predict the masked parts. This helps the model learn context and improves its understanding of the relationship between different modalities.

    c. **Encoder-decoder modules with transformers:** by leveraging the strengths of transformers in handling sequential data and learning long-range dependencies. In the context of Brain MRIs, the encoder module processes and encodes the input MRI images into a dense representation, capturing essential features and patterns. Simultaneously, the textual data, such as medical reports, are encoded into another representation. The decoder module then integrates these representations to generate outputs that can range from detailed captions describing the MRI findings to highlighting specific regions of interest in the images. The use of transformers in both encoding and decoding stages ensures that the model can effectively manage the complexity and variability inherent in medical imaging and textual data.

3. **Loss Functions**
    a. **Contrastive Loss/Cross-Modal Matching Loss:** Aligns MRI scan features with corresponding textual descriptions in a shared embedding space, improving the coherence of generated samples.
    b. **Generation Loss:** Measures the accuracy of generated text or images against ground truth data, essential for high-fidelity output. Examples: **Structural Similarity Index (SSIM), KL divergence, Perceptual Loss** by using a pre-trained network (e.g., VGG) to compare high-level features between the generated and ground truth images

Generation Process

1. **Generating Text from MRI Scans:**
    a. Image Encoding: MRI scans are processed by the image encoder to extract detailed spatial features.
    b. Feature Decoding: These features are decoded into medical reports by the text decoder, sequentially generating sentences that describe the MRI findings.
    c. Attention Mechanisms: Enable the model to focus on relevant parts of the MRI scan while generating specific parts of the report, enhancing accuracy.
2. **Generating MRI Scans from Text:**
    a. Text Encoding: Descriptive or diagnostic text is encoded to produce a fixed-length latent representation.
    b. Image Decoding: Textual features are decoded into synthetic MRI images using generative models like GANs (Generative Adversarial Networks), which iteratively refine the image to enhance detail and accuracy.

Applications

1. **Multi-modal Anomaly Detection:** By integrating both modalities (vision and language), VLMs can potentially improve anomaly detection. For instance, a model might highlight regions in an MRI scan that are indicative of pathology based on visual patterns, and correlate these with specific terms or phrases in the text data to make a more informed diagnosis.
2. **Automated Diagnosis and Reporting:** Automatically generate reports from MRI scans, aiding radiologists by providing initial assessments and create diagnostic summaries from detailed text descriptions, supporting the interpretation of complex cases and enhancing decision-making.

3. **Medical Education and Training:** Generate synthetic MRI scans from descriptive scenarios to help train medical students, providing diverse learning materials and also creating varied case studies with corresponding MRI images and reports, enriching educational resources.
4. **Assistive Tools:** Develop tools that provide text descriptions of MRI findings for clinicians with visual impairments, ensuring they can effectively interpret scans.

Challenges

1. **Data Availability:** High-quality, annotated datasets combining MRI images and associated text are required for effective training.
2. **Complexity:** Integrating and fine-tuning multi-modal models can be computationally intensive and require expertise in both medical imaging and natural language processing.
3. **Ethical and Legal Considerations:** Ensuring that these models meet clinical standards and patient privacy and are ethically deployed in healthcare settings.
4. **Bias and Fairness:** Addressing potential biases in the training data to ensure equitable and unbiased model performance across different patient demographics.