VAE FOR MEDICAL IMAGES
CS 675 / ECE 685 FINAL PROJECT

Maximilian Holsman, Teo Feliu & Emmanuel Mokel

ABSTRACT

Variational Autoencoders (VAEs) (Kingma & Welling, 2022) are a deep generative model used to produce realistic synthetic data in the same theme as training data. Collecting data in a medical context can be a time consuming and expensive task, making it difficult to make predictive and interpretable machine learning models in medical contexts. VAEs can be used to improve the quality of medical image datasets, which may be of low resolution or contain lots of noise. We utilize several VAE variants in order to generate images from several medical datasets. We then train a generator to distinguish synthetic training data created by our VAEs from real medical data, similar to how Generative Adversarial Networks (GANs) work, in order to improve our images. We hope this method can be applied to generate synthetic medical images for real world use, and improve existing data.

1 VAE INTRODUCTION

Generative machine models have seen incredible success in a variety of domains in recent years due to the success of deep learning. One particular model are Variational Autoencoders (VAEs) (Kingma & Welling, 2022), which have found applications in many diverse domains, from image generation, to anomaly detection, to drug discovery. VAEs are shaped into an encoder and decoder neural network, where the encoder seeks to find a low dimensional latent representation. The decoder then maps the latent space back to an original representation of the data. VAEs are probabilistic generative models, where data and latent variables are parametrized by probability distributions (typically Gaussian) in both the encoder and decoder networks. A central problem in VAEs is finding $p_\theta(z|x)$, or the conditional posterior distribution of the latent variables given the original input data. However, this distribution is computationally intractable, and a variational approach is instead taken. By modeling the above posterior with an approximate probability distribution (variational inference), we are able to circumvent the intractability of our desired posterior distribution. It has been shown that maximizing the Evidence Lower Bound (ELBO) is an appropriate loss function for backpropagation in VAEs. Variational Autoencoders have fluorished since their first introduction in 2013, and many variants building upon the original literature have since been introduced (van den Oord et al., 2018; Larsen et al., 2016; Dupont, 2018; Higgins et al., 2017)

1.1 MEDICAL APPLICATIONS

One application area where VAEs have found particular success is in the medical field. This is because they generate realistic images, which can have a variety of implications. Most importantly, they can be used for the generation of synthetic data. Many image datasets for medical applications are quite small, particularly for areas where data is difficult and expensive to collect. This dearth of data makes it impossible to train quality predictive models for medical professionals to use in conjunction with their own expertise. For cancer and other medical datasets, VAEs and other high quality generative models can make a serious difference in a practitioner's ability to train a model.

Another critical application is the the possibility to denoise and impute data. Due to the nature of VAEs, as well as encoder-decoder architectures in general, a rich latent representation of input data needs to be found. For certain types of medical images, measurement may be a difficult process that is succeptible to noise. Finding and reconstructing this latent representation makes it possible to remove noise from image and provide high quality views of data for practitioners to use.

2 MEDICAL DATASETS

In order to test the quality of Variational Autoencoders for medical image generation, we used three datasets that each focus on some medical aspect.

2.1 PROSTATE GLEASON

The Prostate Gleason dataset (https://github.com/MicheleDamian/prostate-gleason-dataset) provides about 70,000 patches of 256x256 pixel prostate cancer biopsy images, split into classes based on their Gleason Score. It is commonly used for transfer learning for pathology, and can used for supervised learning. We apply our VAE models to this dataset to reconstruct and sample realistic looking prostate biopsy data. A snapshot from this dataset is provided below:

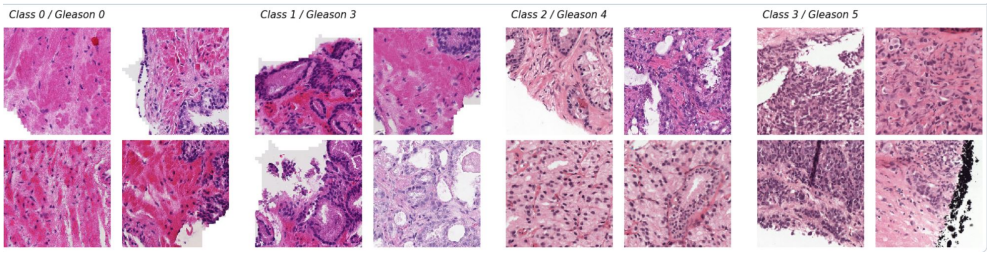

Figure 1: A snapshot of the Prostate Gleason Dataset

2.2 CHESTMNIST

The ChestMNIST dataset is one of 18 MNIST-like datasets in the MedMNIST (Yang et al., 2023) collection, a set of 2d and 3d datasets that focus on medical applications. It contains 112,120 total samples of Chest-Xray samples in greyscale format, that are each 28x28 pixels. An example is provided below:

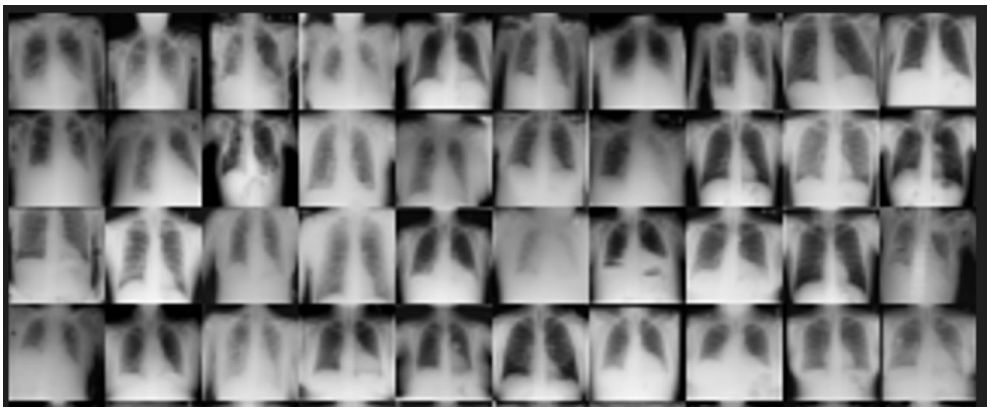

Figure 2: A snapshot of the ChestMNIST Dataset

2.3 RETINAMNIST

Similar to the ChestMNIST dataset, the RetinaMNIST dataset is a set of 1600 images of retinae taken on a Fundus Camera (Yang et al., 2023). These images are not greyscale, and are each 28x28 pixels with 3 color channels. An example is provided below:

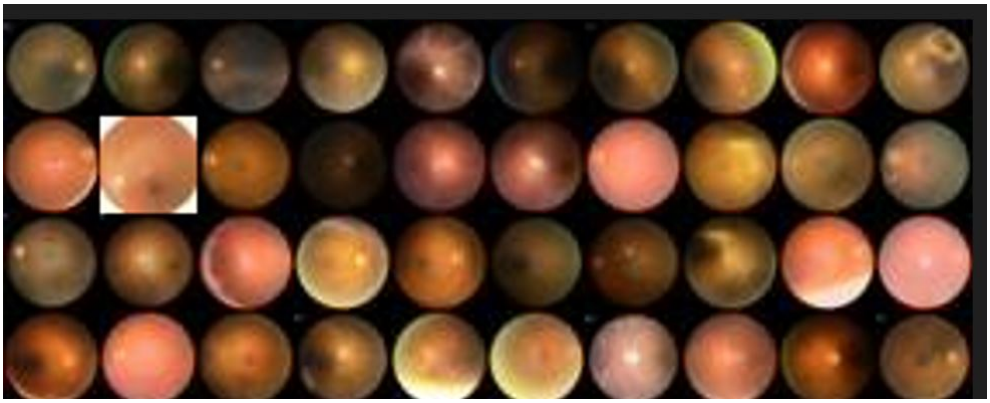

Figure 3: A snapshot of the RetinaMNIST Dataset

3 VAE VARIANTS FOR MEDICAL IMAGE GENERATION

All below networks were trained in a Google Colabatory environment with Colab Pro on a T4 GPU.

3.1 VANILLA VAE

We began our investigation by implementing a vanilla VAE (Kingma & Welling, 2022) in order to get a sense of what the performance for the simplest VAE variant would look like. We transformed the prostate dataset to a resolution of 64x64 and normalized with a mean of 0.5 and a standard deviation of 0.5. Our encoder implementation used 5 convolutional layers of hidden dimension [32, 64, 128, 256, 512] followed by a linear layer for each the mean and variance vectors. Our decoder consisted of a linear layer, followed by the reverse of the encoder's convolutional pipeline and a Tanh activation function. We ran the variant with a learning rate of 0.005 and a KL divergence weight of 0.00025 and observed that the reconstruction loss quickly stoped decreasing (at around epoch 10) and both the reconstructed and sampled images were of very low quality. An example of the sample images are provided below:

After testing several variations of this vanilla VAE in which we slightly modified the hyperparameters and encoder-decoder architecture, we continued to observe the same results. Thus, we attempted to run our implementation on a different dataset in order to determine if the implementation itself is incapable of learning, or if the poor performance was a result of the dataset. After running the original variant on the CelebA dataset, we observed that already after one epoch the VAE was producing much higher quality reconstruction and sample images, and thus concluded that the poor performance must be specific to the prostate dataset we used.

After observing the differences between the prostate dataset and the CelebA dataset, we hypothesized that this decrease in performance likely originated from two primary differences in the two datasets. Firstly, all images in the CelebA dataset had approximately the same visual structure, with each sample being an image of a face of a consistent size centered in the image. This likely made it much easier for the VAE to learn latent representations of the images, as they all adhered to a similar overall structure, and thus made the resulting reconstructions and samples much higher quality. The other major difference was that most pixels in the image in the prostate dataset had around the same color, as they all were various shades of pink and purple. Therefore, even if two prostate images might be depicting very different semantic content, these two images would still be relatively close in the pixel space, as on average their pixels all had around the same value. This was less so true for the CelebA dataset, as semantic differences in facial features were often mean much larger

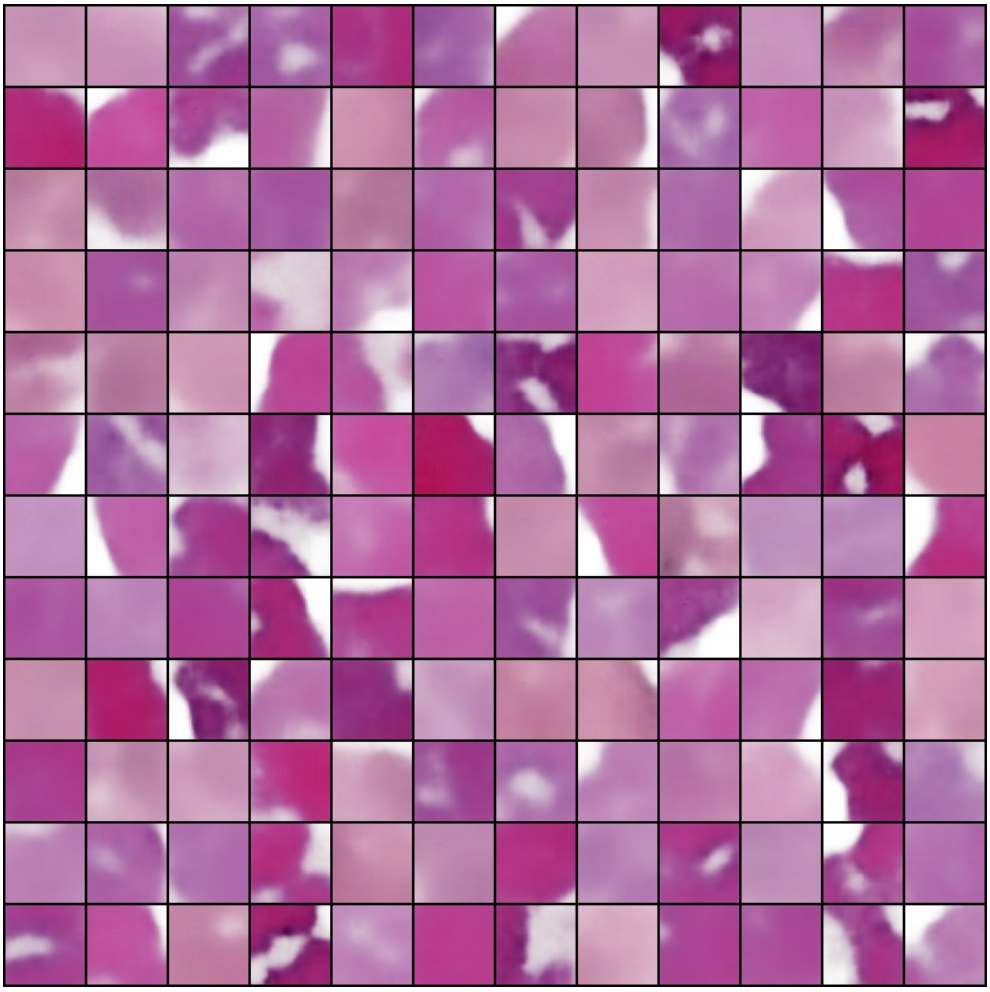

Figure 4: Poor generated samples (reconstructions not shown) for Vanilla VAE on Prostate Gleason Dataset

differences in pixel space. Since the vanilla VAE's loss function used the pixel space different to guide the model, it seemed likely that this pixel-space difference simply didn't properly capture the differences between images in the prostate dataset, and therefore the model was unable to learn the different features needed to generate high quality samples. Thus, to test this hypothesis and address the low quality images we were generating, we tested VAE variants with different loss function to determine whether this would lead to improvements.

On the other hand, this model performed quite well immediately when applied to the RetinaMNIST and ChestMNIST datasets, particularly for the RetinaMNIST dataset. Generated samples for the RetinaMNIST dataset are provided below after training with 100 epochs with the initial specifications listed above:

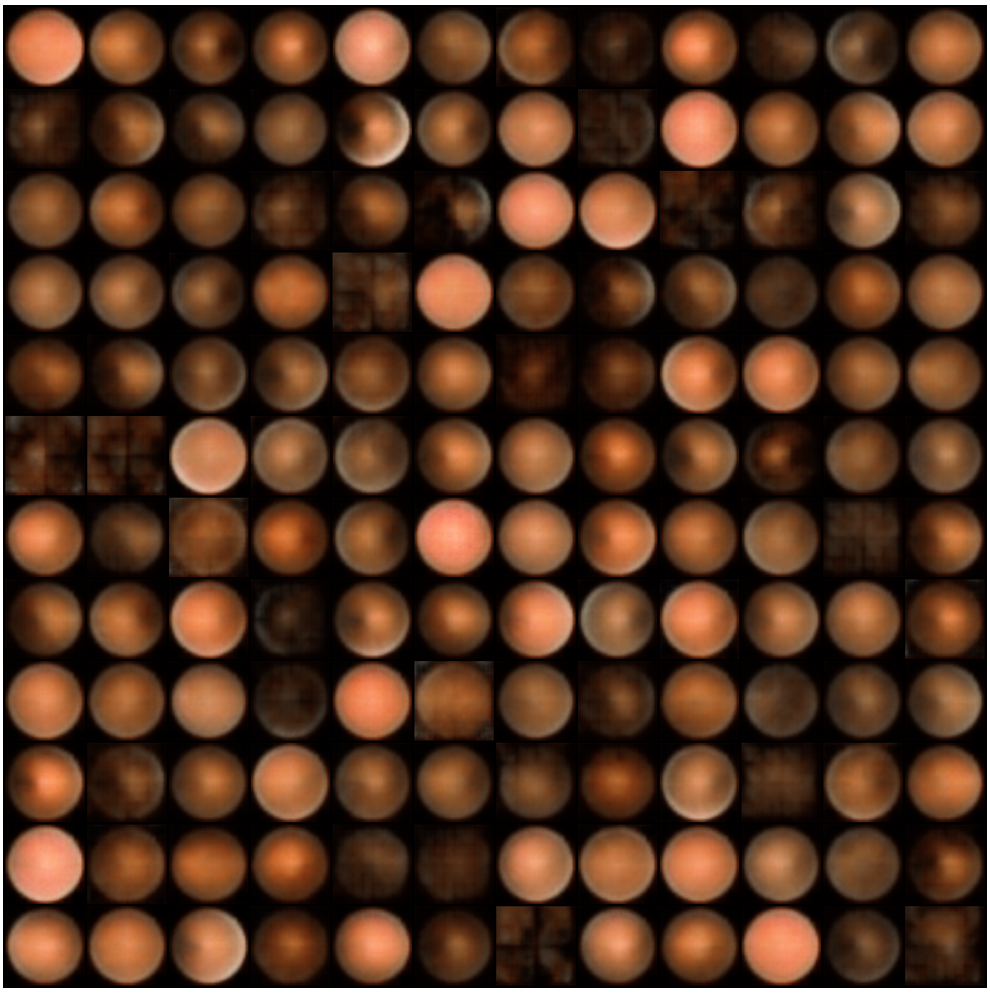

Figure 5: High quality generated retina samples with VanillaVAE

3.2 MSSIM VAE

After implementing and adjusting many hyperparameters of our vanilla VAE, we were still not seeing proper performance. We tested several VAE variants (Conditional-VAE, β-VAE, and more) to no avail. We began thinking about what might be causing the poor performance of our model, and decided to experiment with adjusting our loss function. This fundamentally changes how a VAE works, and leads to different VAE variants. One such variant focused on using perceputal similarity metrics, and we decided to stick with this implementation (Snell et al., 2017)

The Multiscale Structural Similarity Score (MS-SSIM) (Nilsson & Akenine-Möller, 2020) is a differentiable loss metric that is specifically meant to be calibrated to human perceptual judgements. This is important for optimizing for models that would be used for medical applications, and whose end users in some way would be humans making decisions. We also felt that the reconstruction loss alone would not be a particularly helpful loss metric due to how similar the training images were across samples. After normalizing our data, we trained a VAE with the same model specifications as above for 50 epochs, and tuned the latent space over a variety of sizes, eventually settling on 128. We saw the below results for random samples from our decoder:

Similarly, this network performed well when applied to the ChestMNIST dataset, which was normalized with a mean and standard deviation of 0.5, and rescaled to be 64x64. It was not noticeably better than the VanillaVAE, and some samples from our trained network can be found below:

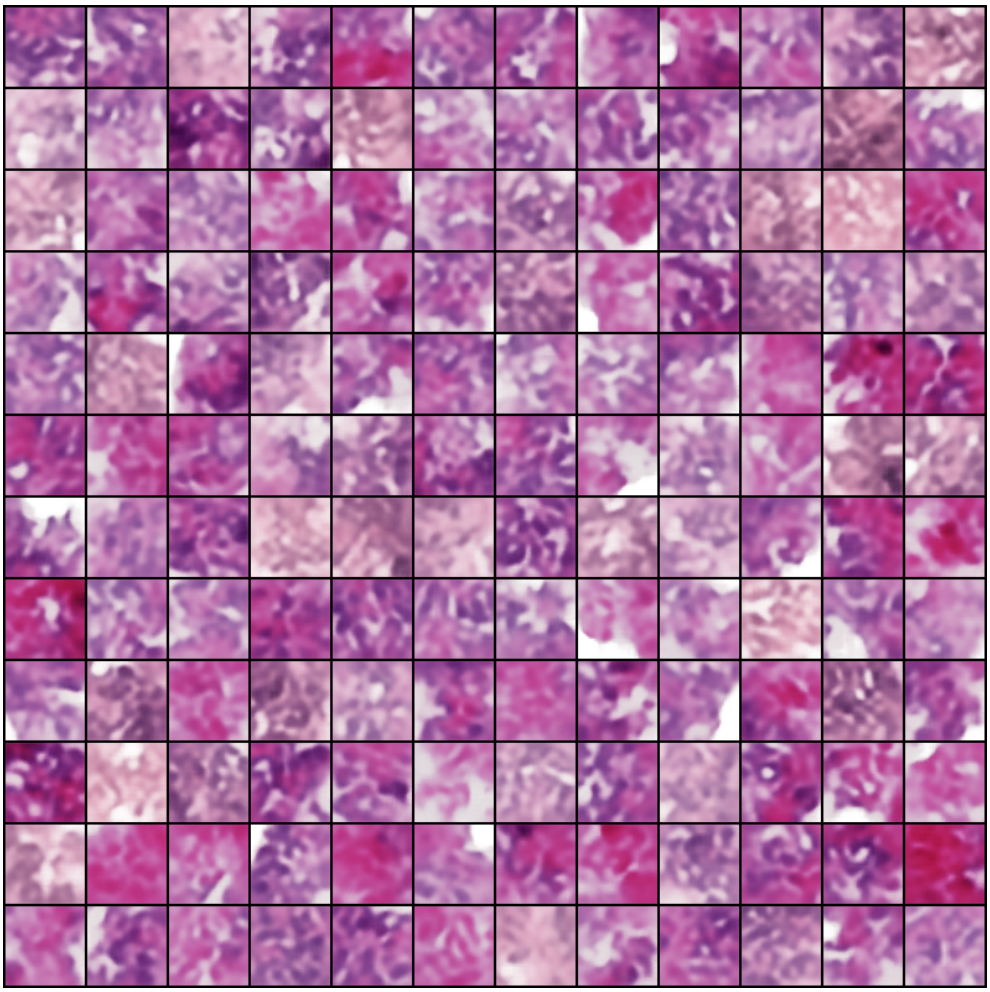

Figure 6: Generated Prostate Gleason Images with MS-SSIM Based VAE

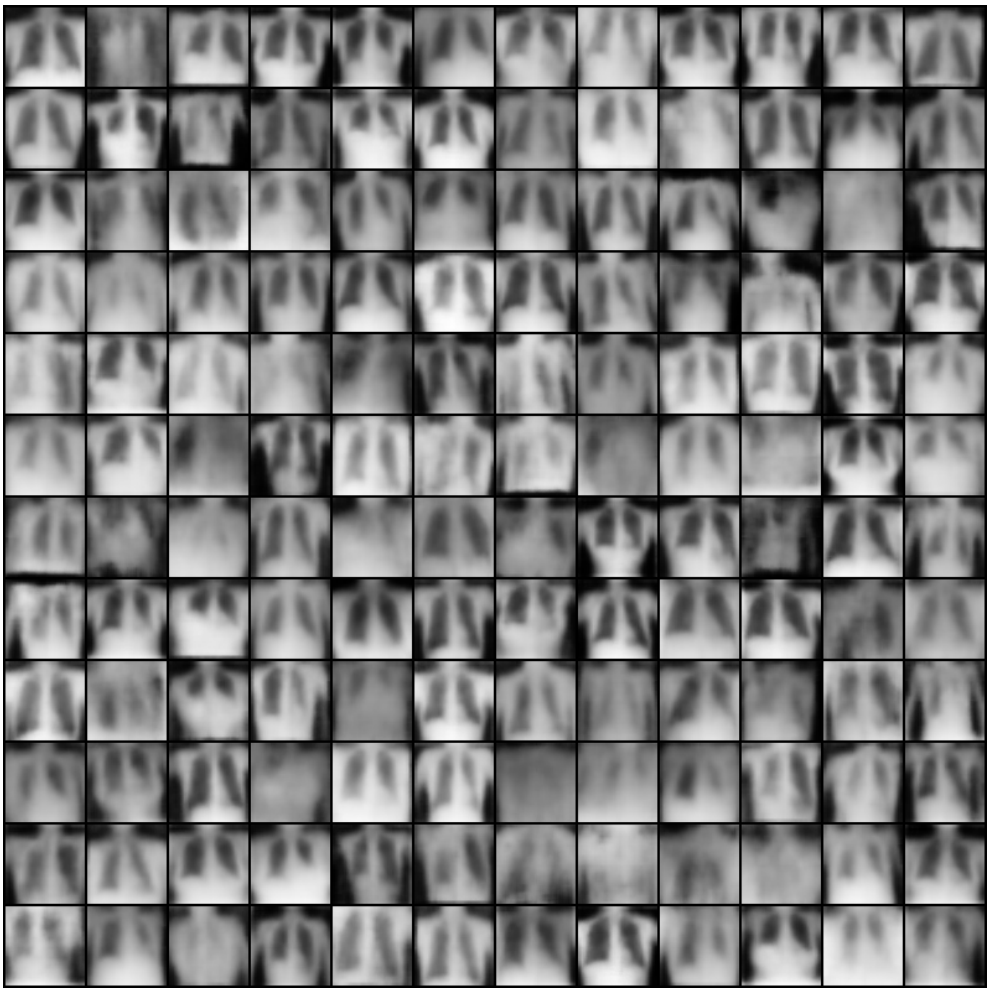

Figure 7: Generated ChestMNIST Images with MS-SSIM Based VAE

3.3 VAE-GAN

In an attempt to further improve the quality of our reconstructed and sampled images by improving the model's ability to compare reconstructed images with the originals during training, we implemented a VAE-GAN. This variant combines the VAE encoder-decoder architecture with the generator-discriminator architecture of a Generative Adversarial Network (GAN) (Larsen et al., 2016). Essentially, an input image is fed through the VAE encoder-decoder network to produce the reconstructed image. This reconstruction is then fed to a discriminator, which tries to differentiate between real and generated (reconstructed) samples. The idea here is that by nature of the adversarial network, the discriminator is forced to learn a similarity metric that effectively differentiates the reconstructed samples from the original images. This makes VAE-GANs suitable for our application, as they should be able to learn a similarity metric that is insensitive to the fact that images from the dataset are all structurally very different and are all of similar color.

We implemented a VAE-GAN whose encoder-decoder architecture was very similar to that of our vanilla VAE. Our discriminator consisted of a series of convolutional layers identical to those of the decoder, followed by a series of linear layers and a sigmoid that feed the input to a single prediction for each image.

We trained our VAE-GAN for 10 epochs with a learning rate of 0.0003. This produced our highest quality results, which can be seen in the figure below. The high quality of the reconstructed and sampled images leads us to believe that the discriminator is able to learn a better similarity metric for the data than both the pixel-space similarity metric used by the vanilla VAE and the Multiscale Structural Similarity Metric used by the MS-SSIM VAE.

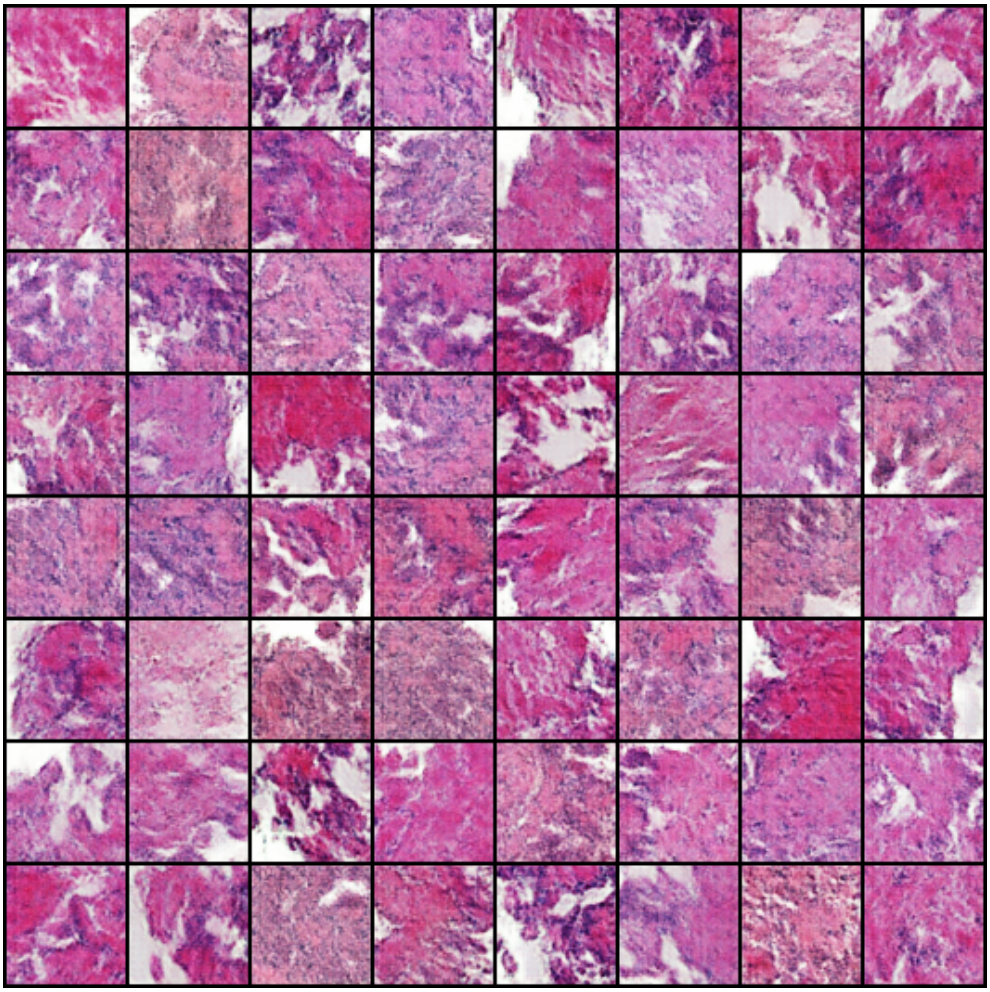

Figure 8: High-quality results after training VAE-GAN for prostate data

4 Conclusions and Next Steps

Variational Autoencoders, while no longer are necessarily state of the art in many applications, serve as a important building blocks for generative models. They have seen moderate success in medical applications, and hopefully can continue to be improved upon. They are theoretically elegant, and easier to train than other deep neural networks such as GANs. For the prostate dataset, we were able to generate high quality synthetic images of prostate data, which could perhaps be used as training data for Gleason score classification in a CNN. These different VAE variants also performed well on both MedMNIST datasets tested, and hopefully across a variety of medical datasets. The potential for VAEs and other generative models in medical applications are huge, and it is a very active area of research.

On our end, given more time and computing resources we would have continued to train these models. The VAE-GAN model worked quite well for the prostate dataset, but was very computationally expensive and time consuming to run. We wanted to run the prostate dataset on 256x256 size pixels, but ran into memory issues and had to downscale to 64x64 in order for our environment not to crash. Given more computing resources, we would be able to experiment further with different configurations of hyperparameters, as well as higher resolution data. We were also limited to approximately 10 epochs for this model, and it would be interesting to see performance when training for a larger number of epochs. Experimenting with other GAN / VAE hybrid variants as well would be interesting, considering how much the VAE-GAN model worked.

Code is provided with submission in a zipped file of .py and .yaml configuration files with submission, as well as a Jupyter Notebook where experiments were run using Google Colab Pro. This code is based heavily on the repository https://github.com/AntixK/PyTorch-VAE. The zipped file also contains loss plots for some of the neural networks described above.

REFERENCES

Emilien Dupont. Learning disentangled joint continuous and discrete representations, 2018.

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*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1558–1566, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/larsen16.html.

Jim Nilsson and Tomas Akenine-Möller. Understanding ssim, 2020.

Jake Snell, Karl Ridgeway, Renjie Liao, Brett D. Roads, Michael C. Mozer, and Richard S. Zemel. Learning to generate images with perceptual similarity metrics, 2017.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning, 2018.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

A APPENDIX

Below are some additional results from our running and tuning of various Variational Autoencoder Basis Neural Network architectures.

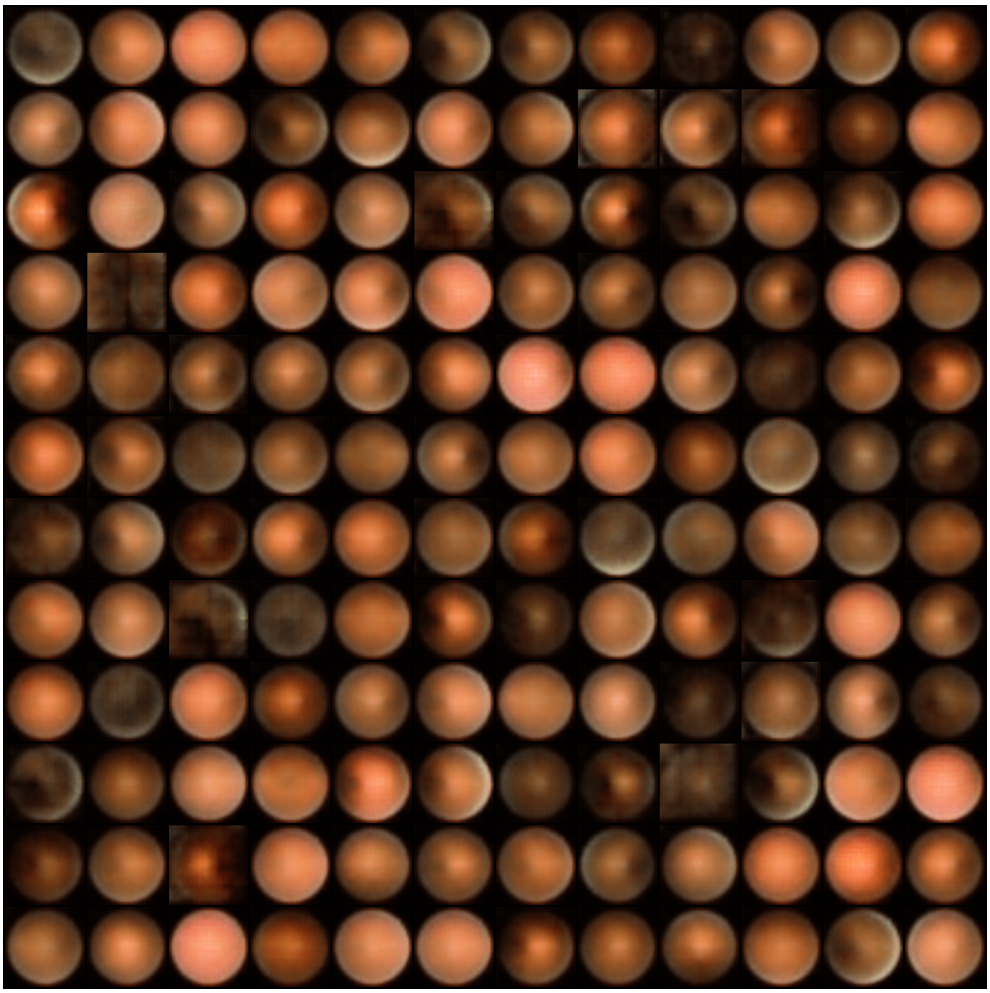

Figure 9: Samples generated using the JointVAE architecture for the RetinaMNIST dataset. We were satisfied with our retina results, and decided not to train a VAE-GAN for this dataset due to its costly nature

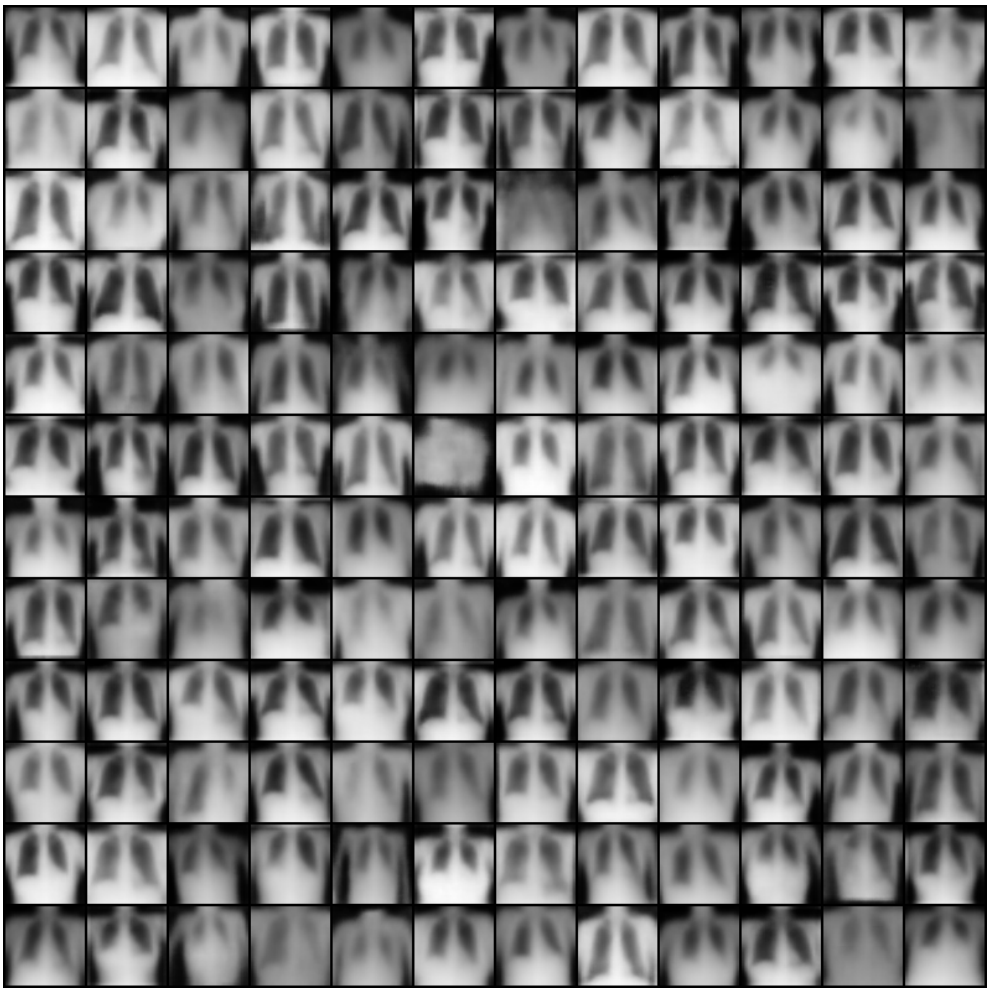

Figure 10: VanillaVAE results for ChestMNIST dataset after training for 50 epochs