Generating pediatric brain cancer images using cross disease conditions

Introduction:

Medical image classification and segmentation holds immense promise for revolutionizing healthcare. Its potential applications span disease diagnosis, treatment planning, and patient prognosis, offering tools to improve patient outcomes and advance medical understanding. However, this potential is hindered by a significant challenge: data scarcity. Limited availability of labeled data plagues medical image classification due to several factors. Ethical considerations around patient privacy and data security often restrict data collection and sharing. The cost of acquiring medical imaging equipment and the expertise needed for data interpretation further limit data generation. Additionally, the rarity of certain diseases creates an imbalance in class distribution, where some classes have significantly fewer data points compared to others.

In recent years Computer aided diagnosis has been advanced significantly this is an interdisciplinary technology combining elements of Artificial intelligence and computer vision with radiological and pathological image processing. A typical application is the detection of a tumor. MRIs contain valuable information regarding the type, size, shape, and position of brain tumors without subjecting the patient to harmful ionizing radiation. MRIs provide higher contrast of soft tissues compared with computerized tomography (CT) scans. Thus, coupled with a CAD system, MRIs can quickly help identify the location and size of tumors. Studies show physicians who relied on CAD were able to detect much more information about the MRI rather than just visually examining the MRI. With the rapid advancements in deep learning. CAD systems were able to integrate various types of advancements including tumor detection, segmentation, and also image generation.

Brain cancer mainly diffuse glioma is one of the most common diseases in the modern age among children aged 5-10 years due to number of reasons. Once a child is diagnosed the survival generally only ranges from 8 to 11 months. Detecting DG in early stage is quite hard and expensive. One of the biggest bottlenecks for detecting early stages of diffuse glioma is data scarcity. Generative adversial Network, a recent generative modeling technique, offer a potential solution by generating synthetic and realistic data.

In this research thesis I am going to tackle this problem to use generative AI techniques (GAN's, LDM's, etc;) to make use of a broader general cancer type (glioblastoma) and train a CGAN (conditional generative adversial network) to generate conditioned glioblastoma images in such a way that the model makes use of the general dataset and generate diffuse glioma images.

Literature Review:

There has been good amount of research done on generative AI techniques in the medical domain. But pediatric cancer types remain underexplored. With the rapid advancements in deep learning. Much research is focused on tumor detection. Before 2012 most detections models included traditional machine learning based models like SVM(Support vector machines), KNN (K-means nearest neighbor). But after 2012 - CNN (Convolutional Neural network) has been regarded as one of the best model with utmost accuracy to successfully detect any tumors. Early detection was seen challenging implicating people get their health checked only if they seem something is feeling off.

One other problem people were facing was detecting the overlapped region of the tumors. for example it is hard to know if the tumor has affected any other regions within the organ with the likelihood of cancer cells spreading to other tissues. Since the MRI's are bulky in volume, it is physically impossible to segment the images layer by layer. To overcome this problem professionals have introduced U-NET based segmentation architecture.

But the main problems with detection and segmentation of these models included data scarcity. Most of the research has been done on widely available datasets which included Brain and Brest cancers. Other cancer types like rectal and skin with limited datasets have been tackled by generating data using GANs, LDMs which was seen as a success.

Many of the other cancer types, especially pediatric cancer types, have been underexplored or not explored at all. One of the leading factors for this data scarcity. This is because of high costs of image acquisition and expert annotation, strict patient privacy regulations that hinder data sharing, the diverse and complex nature of medical images requiring specialized collection, and the low prevalence of rare diseases. These factors limit the size and generalizability of datasets, making it difficult to train robust deep learning models. Deep learning frameworks like Alex-NET which uses CNN, U-NET for segmentation require huge amounts of data to detect specific tumors and areas of interest. With dataset scarcity also comes the problem of dataset imbalance. Most of the cancer images are labelled as progression from early to late-stage types. But most of the underexplored cancer types only contain late-stage cancer images which makes it harder to diagnose in the early stages.

Attempts have been made to overcome this problem. One such way is to augment the existing data. This included steps like cloning the existing data images, rotating the data images in clockwise direction and in the anticlockwise direction, manually displacing the tumors into other regions and changing the sizes, shapes of tumors. But these methods did not actually address the underlying problem.

Generative adversial network GANs were introduced in 2014 and have had a profound impact on designing deep learning models. GANs integrate two neural networks which are trained by competing with each other. The trained generator can then create realistic samples from complex distributions 1) GANs maximize the probability density over the data generating distribution by exploiting density ratio estimation [3] in an indirect fashion of supervision; (2) GANs can discover the high dimensional latent distribution of data, which has led to significant performance gains in the extraction of visual features. Professionals only started adapting GANs at the end of 2015.

As GANs can produce high-quality medical images even in the face of limited datasets, they have revolutionized diagnostic precision and image enhancement in medical imaging. Over time, GANs have evolved significantly, beginning with the basic adversarial process in 2014 but facing challenges in fully covering data distribution. The introduction of DCGAN in 2015 improved image quality, followed by WGAN in 2017 addressing mode stability. CycleGAN enabled image-to-image translation without paired training data, while PGAN introduced a progressive training approach. Later, SAGAN in 2019 focused on relevant image regions and long-term relationships. Recent advancements include RANDGAN, emphasizing segmentation for anomaly detection and outperforming traditional GANs in medical imaging.

However, in the medical domain, GANs and other generative AI techniques have not been widely applied across diverse cancer types. Most existing work focuses on adult cancers, leaving pediatric diseases such as *diffuse glioma* largely unexplored. Moreover, even when generative models are used, they typically rely on well-balanced datasets within a single disease domain. Few studies have investigated **cross-disease generation**, where knowledge from one cancer type could help synthesize another related type with limited data.

This insight motivates the present research: to explore whether conditioning on a more abundant dataset—such as glioblastoma—can enable the generation of realistic synthetic images for underrepresented pediatric cancers like diffuse glioma.

In summary, while generative models such as GANs and LDMs have shown strong potential in medical image synthesis, their application to rare pediatric cancers remains underexplored. This research aims to bridge that gap through CGANs or other frameworks that are widely available.

Research design and methods:

Approached methodology:

- The aim is to generate diffuse glioma dataset to address data scarcity for better detection and diagnosis using another broad dataset of glioblastoma. To achieve this I propose the following method.

Dataset:

The dataset for glioblastoma is available at https://doi.org/10.7937/TCIA.709X-DN49 and the dataset for diffuse glioma is available at https://doi.org/10.7937/tcia.bdgf-8v37. the broader dataset consists roughly of 828k images and diffuse glioma consists of 12k images. The data types for each of them are as follows: MR, Molecular Test, Demographic for glioblastoma and MR, Measurement, Demographic, Follow-Up, Diagnosis for diffuse glioma.

Dataset preprocessing:

Both the glioblastoma and diffuse glioma datasets contain MRI scans stored primarily in NIfTI format, along with associated metadata such as patient demographics, molecular test results, and diagnostic information. Because these datasets originate from different studies and may vary in acquisition parameters, a consistent preprocessing pipeline is required to ensure uniformity and compatibility for training generative models.

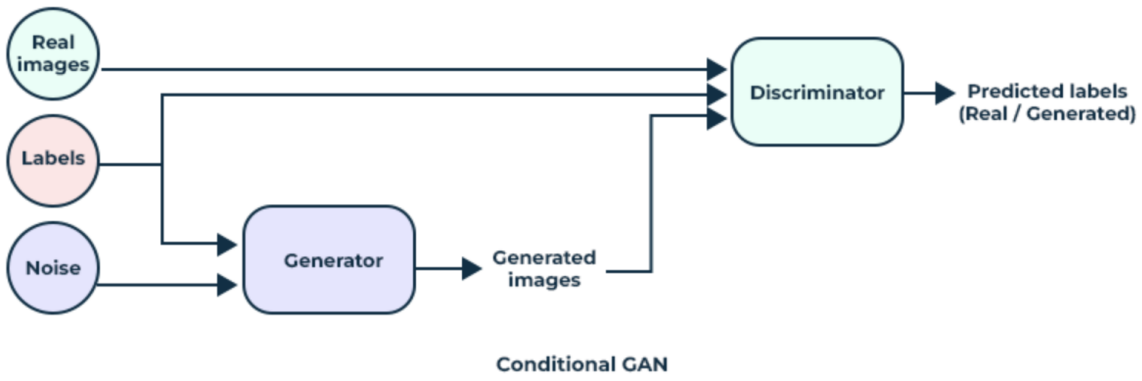
- The MRI scans from the cancer achieve are arranged as T1, T2 and flair. Non-image file contains demogrpahic information which will be stored separately and be made use when training.
- MRI intensities vary across scanners and acquisition protocols. To reduce this variability, voxel intensities will be normalized to a consistent range (e.g., [0, 1] or z-score normalization). This standardization improves model convergence and reduces modality bias.
- All volumes will be resampled to a fixed voxel dimension (e.g., $1 \times 1 \times 1$ mm^3) and cropped or padded to a uniform spatial size (e.g., $64 \times 64 \times 64$ voxels). This ensures uniform input size for the generative network
- For the conditioning part of the CGAN, the images should contain segmentation masks which would be extracted from glioblastoma and be fed into the CGAN model. If the segmentations masks are not available a U-NET may be trained to get the details.

- Data augmentation steps like cloning, flipping, rotating may be performed to keep the data consistent among all classes.

Through these preprocessing steps, the datasets will be standardized and normalized, enabling the conditional generative model to effectively learn mappings between glioblastoma and diffuse glioma domains.

Model Architecture & Training

The core of this research involves training a **Conditional Generative Adversarial Network (CGAN)** to generate diffuse glioma images conditioned on glioblastoma data. Unlike a traditional GAN, which produces random outputs from noise, a cGAN incorporates conditioning variables—such as segmentation masks or tumor class labels—to guide image generation toward specific features.



Conditional Generative Adversarial Network

Loss functions:

- Both the generator and discriminator will be optimized using **Binary Cross-Entropy (BCE)** loss with the **Adam optimizer** (learning rate = 0.0002, $\beta_1 = 0.5$). This configuration provides stable adversarial training and helps the generator produce anatomically plausible images.

The Generator:

- Inputted would be a noise vector and a label
- Reshape and concatenate label embedding with the input image.
- Process noise through dense layers with LeakyRELU activation.
- Reshape and concatenate label embedding with noise features.
- Use Conv2DTranspose layers to up-sample into 32×32×3 images.

- Output layer uses tanh activation to scale pixels between -1 and 1

The Discriminator:

- Use Conv2D layers with LeakyReLU activations from torch framework to extract features.
- A minimum of 4 layers with 32 to 64 neurons
- Flatten features apply dropout to prevent overfitting
- Flatten features, apply dropout to prevent overfitting.
- Final dense layer with sigmoid activation outputs probability of real or fake.

Evalutation

To evaluate the effectiveness of the generated images, a **CNN-based classifier (AlexNet)** will be trained on real diffuse glioma data to detect and localize tumor regions.

- The model will consist of 3–4 convolutional layers with 32–64 filters each, trained with **ReLU** activations and **Adam optimization**.
- Once trained, this classifier will be tested on the **synthetic diffuse glioma images** generated by the cGAN.
- Comparative metrics such as **classification accuracy**, **Structural Similarity Index (SSIM)**, and **Fréchet Inception Distance (FID)** will be used to assess the quality and realism of generated data.

This approach not only validates the realism of the generated MRI scans but also measures their **utility in improving diagnostic model performance**—a crucial step toward developing reliable AI systems for rare pediatric cancers.

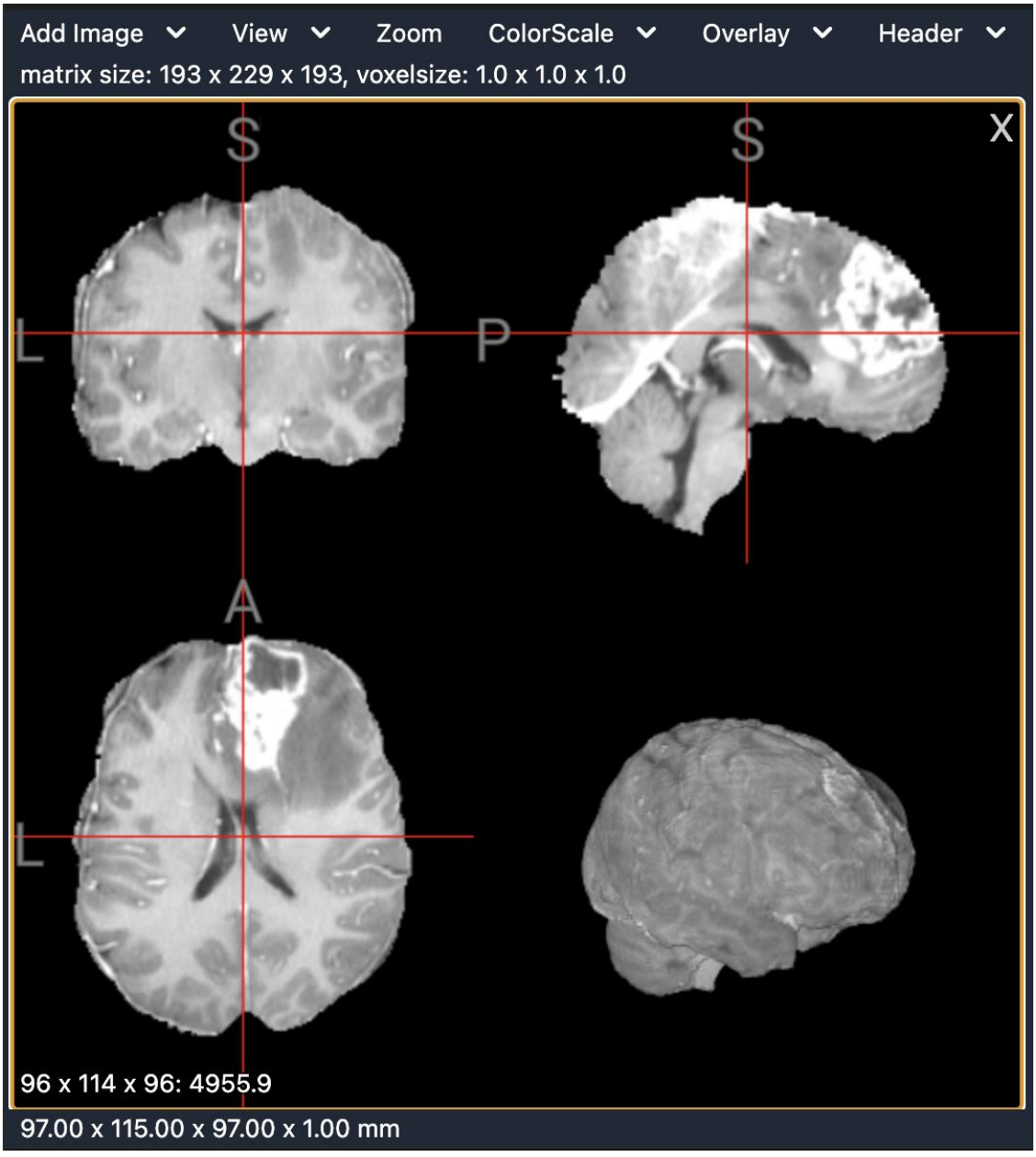
Optimization:
If the generated images or classifier results show signicant signs of instability - such as blurring, mode collapse, or poor anatomical fidelity - hyperparameter tuning will be performed to improve both visual quality and convergence. GAN training is known to be highly sensitive to parameter selection, so iterative experimentation will be essential. And these parameters will be adjusted:

- Learning rate
- Batch size
- Dropout and regularization
- Noise vector dimension

Each configuration will be evaluated using quantitative metrics such as **FID**, **SSIM**, and classifier accuracy, as well as qualitative visual inspection by overlaying generated images against real MRI scans. The best-performing parameter set will then be selected for final training runs.

Test Run:

This is how the real image data is stored for the diffuse glioma dataset:



For testing purposes, I trained a classical GAN model on this dataset which consisted of only around 33 subjects to see if it is possible to generate images of this format: