Exploring the Efficacy of Synthetic Data in MRI-Based Brain Tumor Classification

Meher Aisha, Anar Alimzade, Hadi Sulaiman, Emily Vorderwülbeke

University of Passau, Germany

Abstract

Deep learning has advanced medical image classification but remains heavily reliant on large, diverse datasets, posing ethical and practical challenges. This study investigates the role of synthetic data, generated using Denoising Diffusion Probabilistic Models (DDPM), to address data scarcity in brain tumor MRI classification. Synthetic images were evaluated for fidelity and diversity using Frechet Inception Distance (FID) and Inception Scores (IS), demonstrating high quality for specific classes. A modified VGG-19 CNN classified MRIs into glioma, meningioma, pituitary, and no tumor classes. Experiments with varying real-to-synthetic data ratios revealed that synthetic data can enhance precision and recall for certain classes, though often at the cost of accuracy and generalization. Performance peaked at specific ratios, indicating an optimal balance between real and synthetic data. Fine-tuning with combined datasets improved metrics for underrepresented classes but yielded results comparable to models trained solely on real data. These findings underscore the potential of synthetic data to augment medical imaging datasets and address data scarcity while emphasizing the importance of balanced integration. Future research should focus on validating synthetic data through expert review, refining its quality, and testing its applicability across diverse datasets.

Keywords: Data Synthesis, Data Augmentation, Diffusion Model, DDPM, Brain Tumor Classification, CNN, Deep Learning

1. Introduction

In oncology, particularly in the context of brain cancer, early tumor detection, diagnosis, and management are critical to preventing complications and relapses [53]. While medical practitioners previously relied on symptoms and various tests to diagnose brain tumors, the advent and widespread adoption of Magnetic Resonance Imaging (MRI) have made this process significantly more efficient. MRIs are non-invasive and provide detailed anatomical views of the brain, making them highly effective for tumor detection.

In recent years, the application of machine learning (ML) models in medical imaging has risen substantially. These models excel in pattern recognition and detection and are increasingly being employed for tumor classification. However, ML models require extensive datasets for training. The sensitive nature of medical data, combined with stringent privacy regulations such as the General Data Protection Regulation (GDPR) and Health Insurance Portability and Accountability Act (HIPAA), further complicates data acquisition. This scarcity impedes the development of advanced diagnostic algorithms, which depend on diverse and extensive datasets [19].

To mitigate these challenges, data augmentation techniques have been introduced [3]. Common approaches include affine transformations such as scaling, flipping, cropping, and adding noise, which expand the dataset by inflating existing data. However, these transformations do not generate new images [10]. Instead, they merely alter existing data, limiting the ability to simulate extreme, hypothetical, or abnormal cases.

To address these limitations, augmentation via synthesized [21] data have been proposed as a means to bridge the gap. Using generative models, these approaches produce realistic images that can supplement real datasets. Beyond increasing dataset size, synthetic data also offers opportunities to understand and predict tumor behavior under various conditions, thereby supporting medical research.

The subsequent sections examine the effects, quality, and distribution of synthesized data in training new ML models. This analysis addresses two key research questions:

- **RQ1. Performance Improvement:** Does the integration of synthetic data enhance the performance of brain MRI tumor classification models compared to using real data alone?
- **RQ2. Balance Optimization:** What is the optimal balance between real and synthetic data that maximizes model performance while maintaining cost-effectiveness for brain MRI tumor detection?

This report examines the impact of synthetic data on the effectiveness of brain MRI tumor detection compared to using real data alone. It evaluates whether incorporating synthetic data enhances diagnostic accuracy and reliability, contributing to advancements in medical imaging research and applications.

2. Related Work

Recent advancements in data synthesis have been significantly driven by the development of deep generative models. These models have become a transformative force, providing diverse approaches for generating realistic data across a wide range of applications. They are broadly categorized into Likelihood-based models and Implicit Generative Models. This section offers a comprehensive review of these models, emphasizing their application in synthesizing brain MRI data and their current state of the art.

Likelihood-based models excel in modeling the probability distribution of training data, treby enhancing generalization capabilities and facilitating comparative analysis with unseen data. Their design focuses on maximizing the probability for each data example, avoiding issues like lack of diversity and mode collapse, which are often encountered in Implicit Generative Models like Generative Adversarial Networks (GANs). he

2.1. Likelihood-based models

Two predominent types of Likelihood-based models are flow-based models and autoregressive models.

Flow-based models operate on the principle that data distribution is well-represented when it is easy to model. These models employ a non-linear deterministic transformation involving independent latent variables and utilize the Jacobian and its inverse to simplify and understand data transformations. The resulting images and transformations are evaluated using log-likelihood metrics to ensure high similarity to real images [8].

Autoregressive models, such as PixelCNN, focus on capturing sequential data dependencies. They calculate the conditional distribution for each data bit, hence the name autoregressive. PixelCNN, using a Convolutional Neural Network (CNN) for each conditional distribution, only accesses information about adjacent pixels. An important aspect is, that the three (R, G, B) color channels also depend on each other. G is conditioned on R, where B is conditioned on (R, G). Notably, PixelCNN adeptly handles sharp, coherent image generation and manages both local and long-range spatial correlations [58]. Despite their effectiveness, these models are time-intensive, with synthesis speed dependent on input size [26]. Nonetheless, their utility as decoders in image autoencoders is noteworthy [58].

2.2. Implicit Generative Models

Implicit generative models on the other hand, with GANs as its predominant example try to optimize a zero-sum game between two neural networks: a generator and a discriminator. The generator creates images from random noise, while the discriminator discerns between generated and real samples, defining the generator's loss function in the process. However, GANs often struggle with mode collapse, diversity and generalization measurement [12].

In the realm of medical image synthesis, Skandarani et al. [54] analyzed various GAN architectures, with an extensive hyperparameter optimization over 500 GPU-days. The study, employing datasets like ACDC, SLiver07, and IDRiD, demonstrated that while most GANs were sensitive to hyperparameters, SPADE and StyleGAN exhibited superior stability and FID scores, underscoring their efficacy in medical imaging [42, 32, 22, 39].

2.3. Autoencoder and Diffusion models

The realm of generative modeling has also been enriched by Autoencoder-based models and diffusion models. Variational Autoencoders (VAEs), for instance, have been instrumental in data representation and generative tasks. The Vector Quantised VAE, in particular, demonstrates a remarkable ability to generate high-fidelity samples, surpassing previous VAEs and rivaling state-of-the-art GANs, without succumbing to model collapse or diversity loss [44].

On the other hand, diffusion models try to strike a balance between model flexibility and tractability. They transform input data into a simple distribution, then reverse this process to define a generative model distribution. The Semantic Diffusion Guidance (SDG) framework, integrating language and image guidance into diffusion models, has revolutionized image synthesis, allowing for nuanced control and efficient, self-supervised fine-tuning [55, 29].

3. Data Analysis

In this section, we describe our dataset as well as conduct a comprehensive data analysis to detect and showcase the patterns and features within our dataset of brain tumor MR-images, fostering a profound understanding that will lay the foundation for the subsequent development and evaluation of our generative model.

3.1. Dataset

Since class-imbalanced datasets are common in the medical domain and they can affect a learning algorithm in a way such that it is biased towards the most common features having an overview of the dataset present is of importance [2]. In this study, we use a comprehensive dataset in total consisting of 7022 MR images of the human brain, assembled from three distinct sources by prior research [37]. The dataset includes one dataset from the Figshare repository, along with the SARTAJ dataset and the Br35H collection, each contributing unique and valuable imaging data for analysis. The dataset is separated into a training (**TR**; 5712 images) and testing (**TE**; 1311 images) set. Each

of the sets consists of four classes, which contain images of different tumor types. These tumor-types are *glioma* (TR: 1321 images, TE: 300 images), *meningioma* (TR: 1339 images, TE: 306 images), *pituitary* (TR:1457 images, TE:300 images) as well as *normal* brain MRIs without tumorous cells (*notumor*; TR: 1595 images, TE: 405 images). *Glioma* and normal MR-images are T1-weighted, while the remaining images are a combination of T1, T2 and FLAIR types [37].

While examining the images, different MRI planes were noticed, which resulted in the development of a clustering model, aimed at segregating images into distinct groups based on shared patterns and features [34]. This model successfully identified three prevalent planes: sagittal, axial, and coronal (Fig. 1).

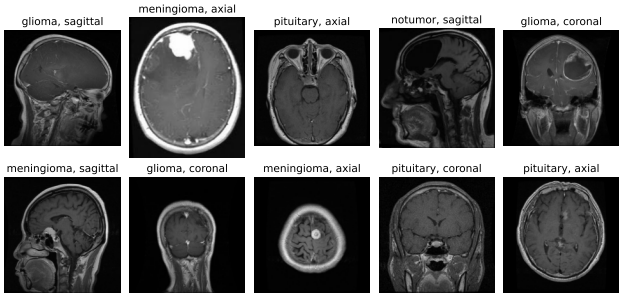
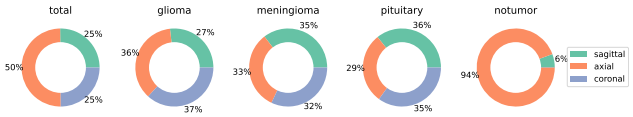


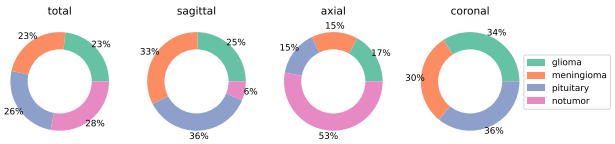
Figure 1: Sample images of the dataset annotated with tumor-types and plane.

A comprehensive analysis of all classes based on their respective planes was conducted to outline the data distribution (Fig. 2). The distribution of the training set across all planes (Fig. 2a) indicate that half (50%) of the images belong to the axial plane, which is the predominant plane for MR-images in the given set. Sagittal and coronal planes each constitute 25% of the data. For the three classes with tumor images the distribution of planes remains relatively consistent, hovering around one-third for each. However, in the no-tumor class most images (94%) are of the axial plane, whereas the rest (6%) are of the sagittal plane.

The class-wise distribution (Fig. 2b) manifests a balanced representation of each class within the entire training set. Nevertheless, this differs for the three planes. The sagittal plane has an equal distribution of images in the classes with tumors, but lacks representation from the *notumor* class (6%). In the axial plane, more than half of the images (53%) are from the no-tumor class, while the tumor-classes demonstrate a near-equal distribution. In the coronal plane, all tumor-classes are equally distributed, whereas the *notumor* class is not represented. This diversification in terms of imaging planes reflects the spectrum of imaging conditions typically encountered within clinical settings.



(a) Distribution of MRI planes in total and for every tumor-type.



(b) Distribution of tumor-types in total and for every MRI plane.

Figure 2: Distribution of MRI planes in total and per class.

3.2. Image Assessment

Ensuring homogeneity in the MRI brain data is a relevant step in order to achieve reliable and consistent data synthesis. The images were assessed based on uniformity in terms of color (Sec. 3.2.1) as well as dimensions (Sec. 3.2.2), and their clarity (Sec. 3.2.3).

3.2.1. Color Homogeneity

As a first step, we evaluated the color features within the images. Given that MR images are typically in grayscale, the homogeneity analysis focused on ensuring consistent intensity distributions across the dataset. Our analysis is essential to confirm that all images were captured and processed under similar conditions, preventing any color-based discrepancies which could lead to biases in the synthesis process [11]. As a result of this initial step, 92 images were found to be in the RGB format and converted into the grayscale format subsequently.

3.2.2. Dimensional Uniformity

The homogeneity of image dimensions was assessed by examining the height and width of each image, since having images with the same size is important for the quality of the generative model [31, 19]. A total of 382 distinct image dimensions are present within the training set, with 512 x 512 being the most common dimension with 3955 images (69,2 %). The next most common dimension, 225x225 is already down by only 268 occurrences (4,7 %) (Fig. 3). This shows an already quite uniformly distributed dataset in terms of dimensions, with a small amount of outliers.

3.2.3. Blur Detection

In real-world scenarios, MR images can be affected by non-ideal conditions, leading to potential imperfections, such as blurry images. Consequently, the pres-

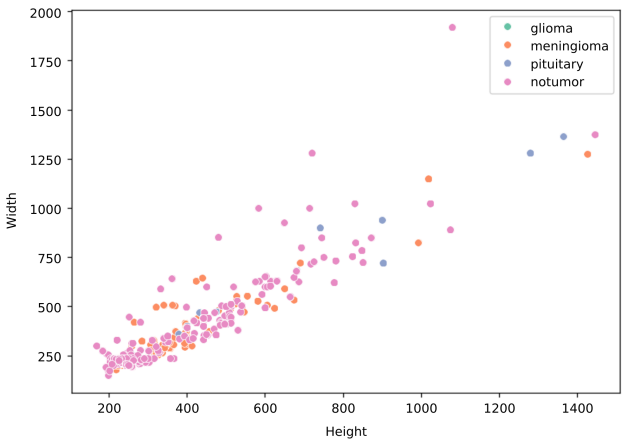


Figure 3: Distribution of image dimensions for every class.

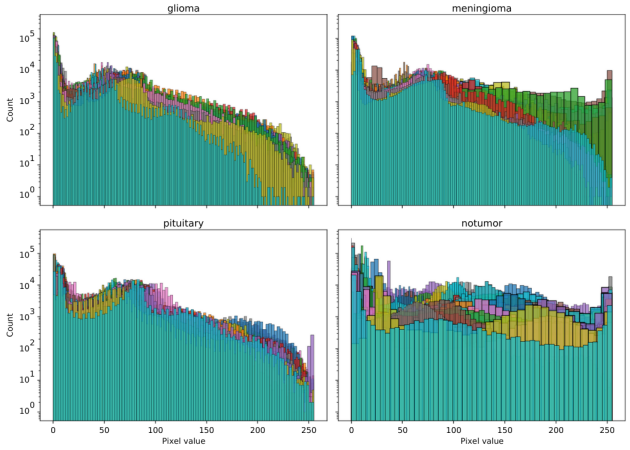


Figure 4: Sample of 100 histograms for every class.

ence of a certain degree of blur in images within the dataset proves beneficial. Employing the Laplacian variance, with a threshold of 50, revealed the identification of 1389 instances (24,3 %) of blurry images within the training set, which seems to be a reasonable amount of blurry images in the dataset [9].

3.3. Pixel value intensities

The pixel intensities of MR images can exhibit significant variations, even when acquired by the same machine. This variability arises from the dependence on proton density and tissue relaxation properties, in contrast to CT scans, where intensities are influenced by electron density. To illustrate the intensity distribution across multiple images, histograms are employed to showcase such variations within the MR images [56].

To gain a comprehensive insight into these distributions across all four classes, histograms derived from 100 sample images, presented in grayscale, are depicted in Fig. 4. In these histograms, pixel values range from 0 to 255, 0 represents an entirely black pixel and 255 denotes an entirely white pixel. Notably, all sample images demonstrate a tendency towards higher concentrations of darker pixel values in comparison to brighter pixel values. While the classes with tumors exhibit relatively aligned distributions of pixel intensities, the pixel intensities within the no-tumor class are not as uniformly aligned, suggesting greater variability in pixel values for images in this class.

4. Data Preprocessing

The quality of the output generated by a generative model is intricately tied to the quality of its input data. Consequently, the domain of data preprocessing plays a pivotal role in optimizing conditions for these models. The refinement of data preprocessing are inherently constrained by the specific characteristics of the dataset and the contextual nuances of its application.

In the following section, each step taken in the data preprocessing pipeline is described and explained, providing a comprehensive understanding of the methods employed to refine and enhance the input data for subsequent model training.

Since MRI images are typically grayscale images, all steps taken from here on consider grayscale versions of all images, regardless of their original color scheme.

4.1. Duplicates

Any duplicate images in the dataset could potentially lead to overfitting and difficulty in learning novel features. To ensure a high quality of the dataset, duplicates were detected by using the *DifPy*[1] python package, which revealed *155* duplicate images in the training set (TE: *21* images). Since some images have multiple duplicates, in total 207 images from the training set (TE: *30* images) were removed, to ensure data integrity and prevent redundancy.

4.2. Image Sizes

As shown in Fig. 3, the dataset consists of images with varying sizes and dimensions. For giving the generative model ideal input data with a consistent dataset, all images should have the same dimensions and sizes [31, 19]. Although 512 x 512 is the most commonly used image size in the dataset, this big of a size is typically not used with generative models, since it requires much more computing power and time [46]. 1047 out of all remaining 1750 images have dimensions in the range of $width, height = [200, 300]$, which is why we decided on resizing all images to the commonly used 256 x 256 dimension.

However, reshaping images where width and height are not equal could have a significant effect on the shape of the brain [60]. Consequently, images with disparate dimensions should potentially be removed from the dataset, but the *notumor*-class then only contains

[1] `https://pypi.org/project/difPy/`

444 images, whereas the other three classes consist of 1321 (*glioma*), 1218 (*meningioma*) and 1408 (*pituitary*) images (Table 1).

	glioma	*meningioma*	*pituitary*	*notumor*	$\sum$
Training	1321	1216	1438	444	4419
Testing	299	207	300	218	1024

Table 1: Data distribution for $Width = Height$ across each class, illustrated for both the training and testing set.

This data imbalance might lead to a bias in the subsequent model [2]. To address this issue, images in the class of *notumor* with a size of $width = [0.9 * height, 1.1 * height]$ and $height = [0.9 * width, 1.1 * width]$ were selected, since these factors seem to result in still reasonable good results when reshaped to $Width = Height$, yielding 817 images. Despite this, a potential bias persisted due to a still slightly imbalanced set. To mitigate this, a final decision was made to take a random sample in the range of $[700, 900]$ for each class with tumor-images, resulting in the data distribution outlined in Table 2.

	glioma	*meningioma*	*pituitary*	*notumor*	$\sum$
Sagittal	232	315	256	60	863
Axial	322	233	223	757	1535
Coronal	313	254	242	0	809
$\sum$	867	802	721	817	3207

Table 2: Final data distribution in training set.

This step aims to achieve a more balanced representation across classes. The testing set on the other hand, already has a reasonably balanced distribution without the need for additional removing of images (Table. 1).

4.3. Data Augmentation

Data Augmentation is one common step in data preprocessing for generative models, because it enhances the generative model to generalize better to unseen variations, improving its ability to generate realistic and diverse outputs during training. Within the array of techniques employed for augmentation, rotational transformations stand out as a notable methodological approach [57]. In the domain of medical imagery, there is a potential of images being flipped and turned by $\pm$ 10°. Consequently, this results in the augmentation of the dataset, where half of the images undergo modification of rotations ($\pm$ 10°), to enhance the overall diversity and richness of the training data and subsequently foster improved model generalization.

4.4. Histogram normalization

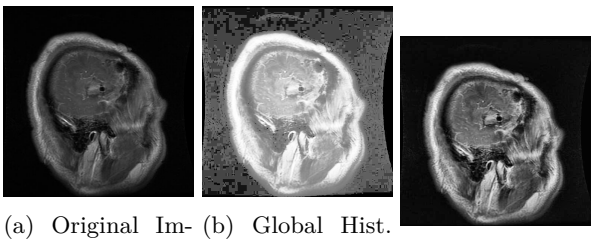
As mentioned in Section 3, even images taken by the same machine, are not necessarily aligned in terms of intensity distribution. For ensuring a uniform training set, with little variations in their intensity distributions different histogram normalization techniques can be applied [56]. These include normalization of each image within itself, as well as of all images in total.

4.4.1. Histogram Equalization

One approach to normalize histograms is through histogram equalization, enhancing image quality and subsequently benefiting model performance.

Histogram equalization is a technique that adjusts the contrasts of an image, leading to a more balanced distribution of intensities [47]. While global histogram equalization is effective for images with histograms confined to a region, it can be detrimental in cases, where both bright and dark pixels are present, such as in MR images.

To address this issue, Contrast Limited Adaptive Histogram Equalization (CLAHE) is often preferred. In CLAHE[2], an image is divided into multiple small blocks, each with a histogram confined to a small region, and equalized accordingly [45]. This adaptive approach proves beneficial in scenarios where global equalization might yield undesirable results. Consequently, it enhances the contrast dynamics between dark and bright pixels, leading to improved visibility of tumors (Fig. 5).



(a) Original Image. (b) Global Hist. Eq. (c) CLAHE.

Figure 5: Example of histogram equalization techniques.

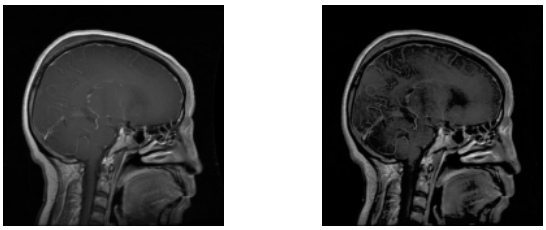
4.4.2. Histogram Matching

Histogram Matching is an alternative approach for normalizing images, aligning an image to a reference image or histogram using its cumulative distribution function.

In our case, intra-class histogram matching was implemented by aligning images to the average histogram of 100 randomly selected images from the specific class (Fig. 6). The histograms of 100 random sampled post-matching are illustrated in Fig. 7

The pixel values were standardized to the range [-1, 1] during the normalization process. This choice is common in deep learning applications as it aligns well with activation functions and weight initialization, contributing to better convergence and stability during model training [11].

[2] https://docs.opencv.org/4.x/d5/daf/tutorial_py_histogram_equalization.html



(a) Original image. (b) Matched.

Figure 6: Example of Histogram Matching.

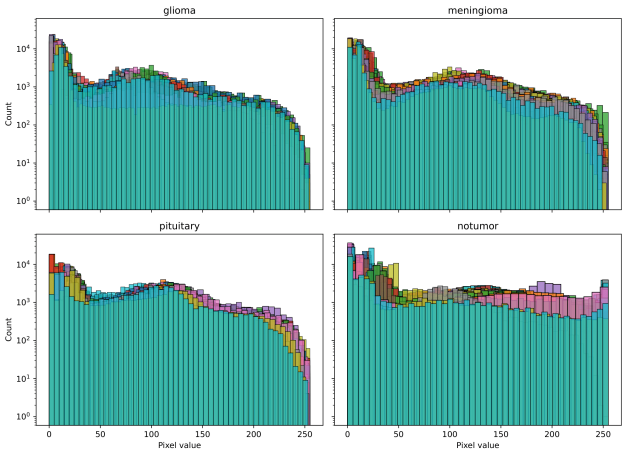
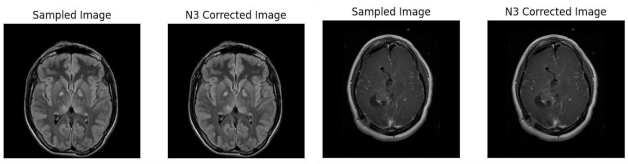


Figure 7: Sample of 100 Histograms for every class after Histogram Matching.

4.5. Limitations in preprocessing

Additional standard preprocessing techniques in the realm of brain MRI scans are skull stripping and bias field correction [1, 48]. Our evaluation of bias field correction methods, particularly N3 [27], on randomly selected images from our dataset suggested that, in our specific scenario, this correction method might not be necessary, since there is no significant difference visible compared to the original images, as can be seen in Fig. 8.



(a) Example 1. (b) Example 2.

Figure 8: Example of Bias Field Correction.

On the other hand the application of skull stripping faced limitations due to the use of JPEG images in our dataset. JPEGs, being standard 2D images without the depth and metadata typical of medical imaging formats, rendered the application of skull stripping impractical for our non-medical image dataset [16].

5. Methodology

The proposed methodology consists of two distinct components: a generative model and a classifier. Initially, the classifier categorizes images without including synthetic data. Subsequently, we train generative models to generate realistic and high-fidelity MR-images for all classes. These synthetic images are then used to evaluate their potential impact on the classification accuracy. This approach enables us to not only create realistic images but also assess their influence on the accuracy of image classification when included in the training process. A detailed visualization of this methodology, illustrating the complete process from generating synthetic images to the comparison of classifier performances, is provided in Fig. 9.

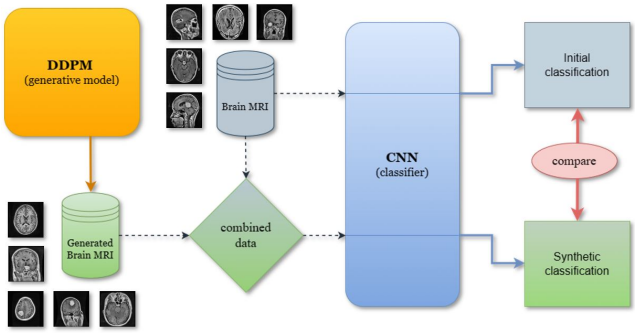


Figure 9: Process flow for classifier performance evaluation with synthetic MRI data integration.

Both the generative model and the classifier are integral to the study, each playing a distinct yet interconnected role. The following sections will detail both components, starting with the generative model, which lays the groundwork for our experimental setup.

To ensure reproducibility, all code related to data preprocessing, generative modeling, and classifier training is publicly available.[3]

5.1. Generative Model

Due to the need of high-fidelity images in the context of medical imaging, where precise details and characteristics are essential [36], we employed a Denoising Diffusion Probabilistic Model (DDPM) in this study. A DDPM is a diffusion model that generates images through gradually denoising a random noise distribution [15]. DDPMs operate by first adding noise to the training data over a series of steps, effectively transforming the data into pure noise. During model training, the reverse process is learned, where the model gradually learns to denoise, or reverse this process, thereby generating new images from a noise distribution [30].

We selected the MONAI [4] as the foundation for our model due to its exceptional capabilities in handling

[3] https://github.com/Alimzade/synthetic_data_efficacy
[4] https://github.com/Project-MONAI/MONAIframework

medical imaging data [41]. MONAI provides a robust toolkit that simplifies the implementation of complex models, making it an ideal choice for our research.

5.1.1. Data Preparation

To achieve a balanced dataset for the tumor classes of *glioma*, *meningioma* and *pituitary*, we ensured equal representation from all three MRI planes - *axial*, *sagittal* and *coronal* by aligning the number of images from each plane to match the count of the least represented plane among them. However, this balancing method, was not applied to the *notumor* class since the majority of these images were from the *axial* plane.

Due to encountering time and computational limitations, all images were downsized from their original dimension of 256 x 256 to 128 x 128 pixels. This resizing enabled us to use a batch size of 16 during model training, significantly enhancing the efficiency of our computations [33]. Stratified split of 80/20 was employed to maintain a balanced representation of each tumor class in both the training and validation sets.

5.1.2. Model Training

For each of the four classes (*glioma*, *meningioma*, *pituitary*, and *notumor*) a separate model is trained with the same parameter settings. These parameters include the diffusion inferer, the DDPM scheduler with an according number of time steps, the optimizer with an according learning rate and the number of epochs required for training.

The Diffusion Inferer is utilized for a model's noise management and image generation process. This component is vital in the DDPM's architecture, as it manages the gradual reduction of noise from the images, guiding the model in generating images that are both clearer and more similar to the original data [36]. The DDPM scheduler, set to 1000 timesteps, plays a crucial role in controlling the incremental introduction and reduction of noise [40], since it ensures that noise levels are appropriately adjusted throughout the training, facilitating a smoother learning curve for the model. The Adam optimizer coupled with a learning rate of $2.5e^{-5}$ serves as the models optimizer, to enhance it's ability to converge and effectively handle the complexities of the dataset [13]. The training is conducted over 750 epochs, enabling the model to capture the intricate details and variations present within the brain MR-images [40].

5.1.3. Generation and Evaluation

For evaluation and further classification purposes 600 images of each class are generated and resized to the original dimension of 256 x 256 pixels (Fig. 10).

To evaluate the quality of all generated images, the Fréchet Inception Distance (*FID*) and Inception Score (*IS*) are utilized, since these metrics are common approaches in literature for evaluating the diversity and fidelity of generated images [28, 38, 50, 25].

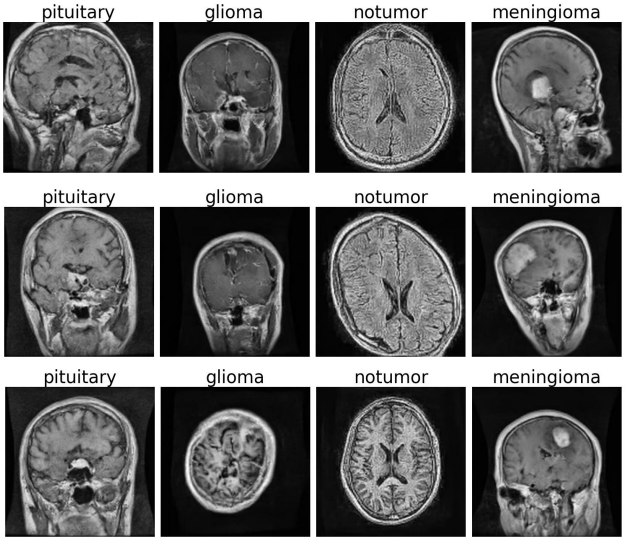


Figure 10: Sample of generated images from each class.

FID measures the similarity between the distribution of generated images and real images in a feature space, where a lower score suggests, that the generated images closely resemble the distribution of real images [59, 5], whereas *IS* evaluates the quality and diversity of the generated images, with higher scores indicating better results [49, 14].

5.2. Classification

Deep learning, particularly convolutional neural networks (CNNs), has become a cornerstone of image recognition, outperforming other classification models like Deep Neural Networks (DNN) and deep Boltzmann machines in accuracy [43]. CNNs are especially effective in brain tumor detection and classification tasks [6, 7, 51, 52]. For this study, we adapted a CNN-based classification model using a pre-existing VGG-19 architecture [35].

5.2.1. CNN Architecture

CNNs employ a weight-sharing architecture that reduces complexity while maintaining robust feature extraction capabilities. To address the high computational demands of training CNNs, transfer learning is widely used, allowing pre-trained models to serve as a foundation for fine-tuning with domain-specific data [17, 24]. VGG-19, a 19-layer CNN architecture trained on millions of labeled images, is a commonly used model in brain tumor classification due to its high accuracy in similar tasks [23, 51].

For our model, we removed the last three layers of VGG-19 and fine-tuned it on our dataset to classify MRI images into four tumor classes: *glioma, meningioma, pituitary,* and *no tumor*. The Adam optimizer, known for its efficiency in MRI-based classification tasks [24], was employed with a learning rate of 0.0001 [18, 20]. A batch size of 32 and 5 epochs

were selected based on empirical results, yielding optimal performance on the dataset. Detailed performance metrics are discussed in Section 6.2.1.

5.2.2. Testing Effects of Synthetic Data

The main focus of this work is to analyze the effects of synthetic data on a classification model. For this purpose six different Scenarios (Sc.) are tested: (1) adding images to underrepresented classes, (2) adding different proportions of synthetic *meningioma* images, (3) adding different proportions of synthetic *meningioma* and *glioma* images, (4) adding a specific amount *(200, 400, 600)* of images to each class, (5) fine-tuning initial classifier on a combined dataset that includes original *meningioma* and *glioma* images, along with all synthetic images and (6) training with synthetic data only.

In the first Scenario, we augment underrepresented classes (*meningioma*, *pituitary*, *notumor*) with synthetic images to eastablish a balanced dataset, with 867 images in each class. For Scenarios 2 and 3, different data distributions of real (R) and synthetic images (S) were taken into consideration: (1) R: 90%, S: 10%, (2) R: 80%, S: 20% and (3) R: 66%, S: 33%. In the fifth Scenario, the classifier is initially trained on all original images and evaluated. Following this, it is further fine-tuned for three more epochs on a combined dataset that includes both original *meningioma* and *glioma* images, as well as all synthetic images, and subsequently re-evaluated. This allows for a later comparison between the different training phases. The idea behind using both real and synthetic data for fine-tuning is to help the model learn additional features or variations present in the synthetic data while reinforcing the patterns learned from the real data.

6. Results

Experimental results for all conducted experiments are outlined in the following sections

6.1. Synthetic Image Evaluation

Before integrating synthetic images into the classification model, a comprehensive evaluation was performed to assess their quality and diversity.

6.1.1. Fréchet Inception Distance

The average *FID* score across all four classes was **74.43**, with the *meningioma* class achieving the best score of **66.42**, as shown in Table 3.

	glioma	*meningioma*	*pituitary*	*notumor*
FID	76.15	66.42	74.88	80.25

Table 3: Fréchet Inception Distance (FID) scores for all classes.

6.1.2. Inception Score

To further evaluate synthetic image quality, *IS* was calculated for both generated and real images, enabling a direct comparison. This approach provides a more accurate class-wise assessment of generated images [4].

The *meningioma* class achieved the highest *IS* for generated images (**3.52**), while the *notumor* class demonstrated the closest similarity to real images with an *IS* of **3.26**, as shown in Table 4.

	glioma	*meningioma*	*pituitary*	*notumor*
(IS) real	4.38	4.24	3.07	3.43
(IS) generated	3.34	3.52	2.36	3.26

Table 4: Inception Scores (IS) for real and generated images.

6.2. Classification

The following section evaluates classification models trained on datasets with varying distributions.

6.2.1. Initial Classification

The initial classification model, trained without synthetic data, achieved an average accuracy of 0.93, as detailed in Table 5. These results are based on the average of three classification reports generated using different random seeds for NumPy and TensorFlow operations.

	Precision	Recall	F1-Score	Training Data
glioma	0.93	0.91	0.93	867
meningioma	0.83	0.88	0.86	802
pituitary	0.98	0.94	0.96	721
notumor	0.98	0.99	0.98	817
Macro Average	0.93	0.93	0.93	-

Table 5: Initial average classification report.

6.2.2. Synthetic Data Classification

The following results provide a detailed analysis of the six previously described scenarios. Each classification task was repeated three times using different random seeds, same as for initial classification, and the results presented here represent the averaged outcomes across all runs.

Scenarios 1 & 6. In Scenario 1, the use of a balanced dataset, as opposed to the unbalanced dataset used in the initial classification, resulted in only marginal changes to model performance. Notable differences included a reduction in precision scores for the *glioma* and *meningioma* classes, both decreasing by 0.04, and a slight reduction in overall accuracy to 0.91 (Fig. 11). These results suggest that balancing the dataset had a limited impact on performance.

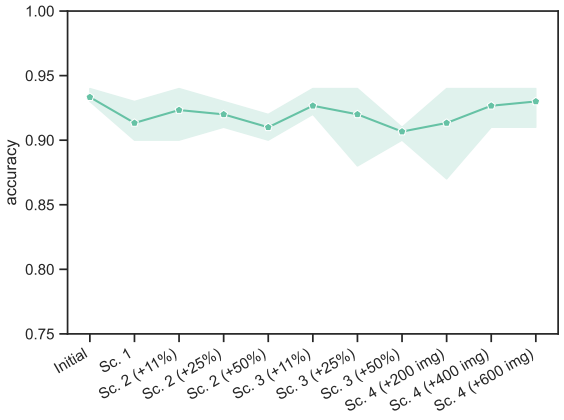
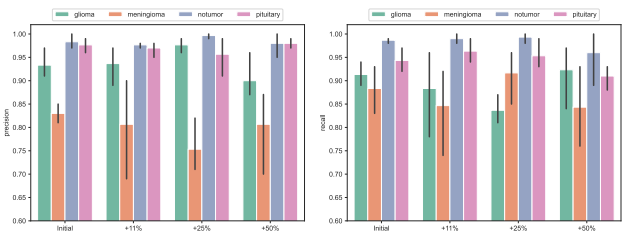


Figure 11: Accuracy over Scenarios 1, 2, 3 & 4.

In Scenario 6, where classification was performed using only synthetic images, all performance metrics were significantly affected. The overall accuracy decreased to 0.78, and the precision scores for all classes showed notable declines. Among the classes, the *notumor* class was the least impacted, while the *meningioma* class experienced a substantial drop in precision, reaching 0.53. Similarly, the *glioma* class exhibited the most pronounced reduction in precision, decreasing to 0.67. These findings highlight the limitations of relying solely on synthetic data for classification tasks.

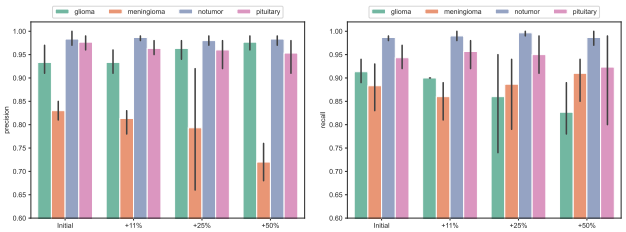
Scenarios 2, 3 & 4. In Scenario 2, different proportions of synthetic *meningioma* images are added to the dataset. For both the *notumor* and *pituitary* classes, precision and recall remain largely consistent. However, in the final case, where the highest proportion of synthetic images is added, recall drops slightly, as shown in Fig. 12. In contrast, the *glioma* and *meningioma* classes are significantly affected. For the *glioma* class, precision improves with the addition of +11%(0.94) and +25%(0.98) synthetic images but declines at +50%(0.90). This trend is reversed for recall, which decreases initially but rises above its baseline value at higher proportions. Similarly, the precision of *meningioma* decreases at +11%(0.81) and +25%(0.75) but increases at +50%(0.81). For recall, the score improves at +25%(0.92) but declines afterward.



(a) Precision (*avg* with *std*) (b) Recall (*avg* with *std*)

Figure 12: Class-wise precision and recall score for Initial Classification compared to Scenario 2.

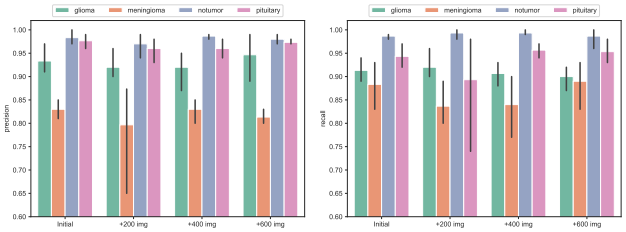
For Scenario 3, synthetic *meningioma* and *glioma* images were added in equal proportions. This addition does not significantly impact the *notumor* and *pituitary* classes. For *glioma*, the precision score increases slightly, but the recall drops to 0.83. In contrast, *meningioma* maintains a stable recall, but its precision score decreases to 0.72 (Fig. 13).



(a) Precision (*avg* with *std*) (b) Recall (*avg* with *std*)

Figure 13: Class-wise precision and recall score for Initial Classification compared to Scenario 3.

In Scenario 4, the addition of a specific number of images to each class had minimal impact on the precision and recall for the *notumor*, *pituitary*, and *glioma* classes. For the *meningioma* class, both precision and recall showed a slight decline initially but increased to 0.81 with the inclusion of 600 additional images (Fig. 14).



(a) Precision (*avg* with *std*) (b) Recall (*avg* with *std*)

Figure 14: Class-wise precision and recall score for Initial Classification compared to Scenario 4.

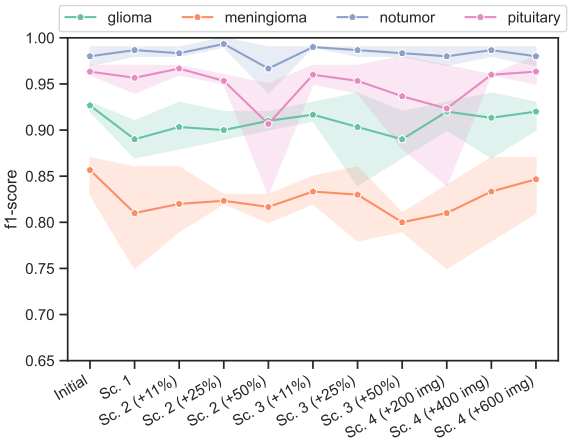


Figure 15: Class-wise F1-scores over Scenarios 1 to 4.

The F1-score decreases in Scenarios 2, 3, and 4, while

it increases in Scenario 6 as more synthetic images are added across all classes (Fig. 15). Overall, the score remains relatively stable for the *notumor* and *pituitary* classes, whereas it shows fluctuations for the *glioma* and *meningioma* classes.

The accuracy remains largely unaffected across different scenarios, with slight decreases observed in Scenarios 2 and 3 as more images are added. It reaches its highest value of 0.93 in Scenario 4 with the addition of 600 images.

Scenario 5. In Scenario 5, a different approach was implemented by fine-tuning the initial classifier on a combined dataset comprising both original *meningioma* and *glioma* images, along with all synthetic images. Initially, the classifier was trained solely on the original images, achieving average values of 0.93 for accuracy, precision, recall, and F1-score. Following fine-tuning, the overall accuracy remained consistent at 93

After fine-tuning, the *glioma* class demonstrated an improvement in precision, increasing to 0.95 (+0.04), while its recall decreased by (−0.05). Conversely, the *pituitary* and *meningioma* classes experienced an increase in recall and a decrease in precision (±0.03). The *notumor* class was not significantly affected (Table 6).

	Initial		Fine-tuned	
	Precision	Recall	Precision	Recall
glioma	0.91	0.95	0.95	0.90
meningioma	0.86	0.82	0.84	0.84
notumor	0.98	0.99	0.98	0.99
pituitary	0.97	0.95	0.94	0.97

Table 6: Class-wise precision & recall in Scenario 5.

Analysis of the confusion matrices revealed notable changes in the model's ability to classify and differentiate between the *glioma* and *meningioma* classes. The classification performance for both classes improved slightly, with an increase of 0.01 in their respective metrics. Additionally, the misclassification rate for the *glioma* class decreased from 0.10 to 0.07, as shown in Fig. 16. Furthermore, the fine-tuned model demonstrated an enhanced ability to identify the *notumor* class, with a 0.02 increase in accuracy, reaching 0.99.

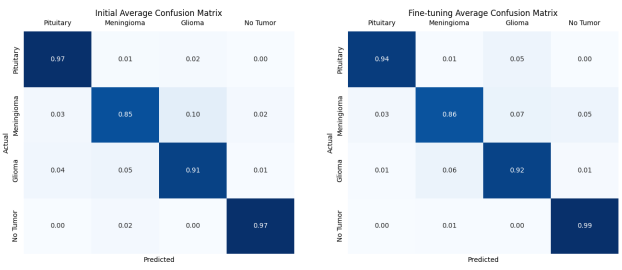


Figure 16: Confusion Matrices for Scenario 5.

6.2.3. Mini-Classification

In image classification tasks, particularly in the context of medical imaging, using a limited dataset for initial model training is a common and effective approach. This method facilitates the assessment of model performance and feasibility when working with constrained datasets. The mini-classification process described below provides a preliminary evaluation of the model's capabilities and highlights areas requiring improvement.

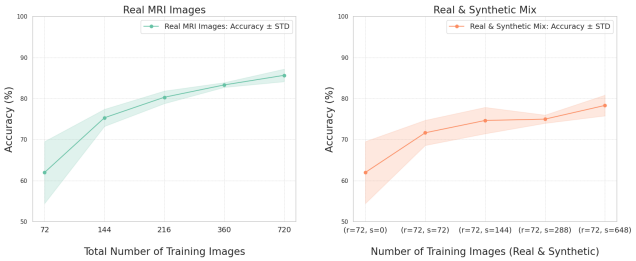


Figure 17: Effect of real and synthetic images on classification accuracy in a limited dataset.

In Fig. 17, the impact of incrementally adding training images on model classification accuracy is examined through two distinct approaches. The first approach involves gradually increasing the number of real MRI images, while the second supplements a fixed number of real images with an increasing quantity of synthetic ones. For all experiments, class-wise balance was maintained, ensuring consistent representation across categories in both real and synthetic image sets.

In the first scenario, only real images were used, with 72, 144, 216, 360, and 720 images. The results demonstrate a clear progression in model accuracy as the number of real images increases. In the second scenario, each case included 72 real (r) images and varying quantities of synthetic (s) images: 72, 144, 288, or 648.

A comparison of these two scenarios reveals distinct patterns in accuracy improvement. The first scenario exhibits a consistent increase in accuracy with additional real images, emphasizing the importance of expanding real image datasets. In contrast, the second scenario shows a different trend in accuracy growth with the inclusion of synthetic images. This comparison provides valuable insights into the effectiveness of synthetic data for augmenting training, particularly when real images are limited.

7. Discussion

This study explores the potential of synthetic images to enhance brain tumor classification models, focusing on their generation and integration into training datasets. The performance of the generative model was evaluated through Frechet Inception Distance (FID) and Inception Scores (IS), assessing both the fidelity of synthetic images compared to real ones and the diversity among synthetic images. The *meningioma* class achieved the

highest scores across both metrics, indicating that the model can generate high-quality synthetic data, though variability in FID across classes highlights room for improvement.

The baseline classification model, trained exclusively on real data, achieved consistently high performance, with accuracy, precision, recall, and F1-scores all reaching 93%. This reliability established a strong foundation for evaluating the impact of incorporating synthetic data. In scenarios where synthetic images were mixed with real ones, a slight decline in accuracy and precision was observed, particularly for specific classes. These results underscore the importance of maintaining an optimal ratio of real to synthetic images in the training set. Conversely, scenarios using only synthetic data performed significantly worse, affirming the limitations of relying solely on generated data for classification tasks.

The experiments revealed nuanced insights into the interplay between real and synthetic data. Varying proportions of synthetic data showed inverse trends between precision and recall for some classes, with performance peaking at specific ratios before declining. This indicates the presence of an optimal threshold for synthetic data integration, beyond which the model's ability to generalize diminishes. These findings emphasize the need for careful calibration of dataset composition to maximize the benefits of synthetic data while mitigating its drawbacks.

Scenario 5 demonstrated that fine-tuning the classifier on a combined dataset of real and synthetic images could enhance performance for certain classes, improving precision and reducing misclassification rates. However, the overall metrics remained comparable to those achieved by fine-tuning on real data alone. This suggests that while synthetic data can supplement real data, its utility may be constrained by factors such as the inherent complexity of the task and the quality of the synthetic images.

To address *RQ1*, our results indicate that incorporating synthetic data does not universally improve classification performance. While it introduces variability that can enhance model robustness in specific cases, the overall accuracy and key metrics often remain unchanged or slightly decline, particularly in simpler tasks where the baseline model already performs well.

Regarding *RQ2*, the study identifies an optimal balance between real and synthetic data, demonstrated in scenarios where precision and recall trends suggest a trade-off. Achieving this balance requires careful experimentation with different ratios of synthetic data, tailored to the specific class distributions and task complexity.

Future research should validate these findings on larger, more diverse datasets, incorporate clinical validation of synthetic images, and explore alternative generative models to enhance their applicability in medical imaging.

7.1. Limitations & Threats to Validity

When working with MRI data, caution is needed due to the lack of a calibrated intensity scale. Unlike CT scans, MRI signals depend on proton density and tissue relaxation properties, which vary across datasets. This variability can influence the generalizability of our findings.

Another limitation is the absence of clinical validation for the generated synthetic images. Without expert review, the true diagnostic relevance of these images remains uncertain. Addressing this limitation in future work will be crucial to ensure the clinical utility and reliability of synthetic data in medical applications.

8. Conclusion

This study investigated the use of synthetic data to enhance brain tumor MRI classification models, addressing challenges posed by limited real-world datasets. The synthetic images, generated using Denoising Diffusion Probabilistic Models (DDPM), demonstrated high quality, particularly for specific tumor classes such as *meningioma*. However, when these synthetic images were integrated into the training pipeline, the overall classification performance did not consistently surpass that of models trained solely on real data.

Our findings reveal that while synthetic data can enhance model robustness and provide value in data-constrained scenarios, its impact depends heavily on the balance between real and synthetic data. Scenarios with mixed datasets showed fluctuations in metrics such as precision and recall, indicating the presence of an optimal ratio of synthetic to real data. Conversely, models trained exclusively on synthetic data exhibited significant declines in performance, reaffirming the importance of real data as the foundation for effective classification.

Despite these limitations, the use of synthetic data shows potential, particularly in augmenting datasets for underrepresented classes and mitigating data scarcity issues. Fine-tuning with combined real and synthetic datasets demonstrated slight improvements in certain metrics, though these gains were often comparable to models fine-tuned with real data alone.

In conclusion, this research underscores the promise and challenges of synthetic data in medical imaging. Achieving the right balance between real and synthetic data is crucial for maximizing its utility. Future work should explore the generalizability of these findings across different datasets and medical imaging tasks, alongside efforts to improve the quality and diversity of synthetic images. Such advancements could pave the way for more robust and scalable solutions in AI-driven medical diagnostics.

Acknowledgements

The authors would like to thank the University of Passau and Prof. Dr. Michael Granitzer for facilitating this research. Special thanks go to Sahib Julka for his invaluable guidance and constructive feedback throughout the project.

References

[1] Zeynettin Akkus, Alfiia Galimzianova, Assiyah Hoogi, Daniel L Rubin, and Bradley J Erickson. Deep learning for brain mri segmentation: State of the art and future directions. *Journal of Digital Imaging*, 30(4):449–459, 2017.

[2] Adamu Ali-Gombe and Eyad Elyan. Mfc-gan: Class-imbalanced dataset classification using multiple fake class generative adversarial network. *Neurocomputing*, 361:212–221, 2019.

[3] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks, 2018.

[4] Shane Barratt and Rishi Sharma. A note on the inception score, 2018.

[5] Eyal Betzalel, Coby Penso, Aviv Navon, and Ethan Fetaya. A study on the evaluation of generative models. *arXiv*, 2206.10935, 2022.

[6] Sunanda Das, OFM Riaz Rahman Aranya, and Nishat Nayla Labiba. Brain tumor classification using convolutional neural network. In *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, pages 1–5. IEEE, 2019.

[7] S Deepak and PM Ameer. Brain tumor classification using deep cnn features via transfer learning. *Computers in biology and medicine*, 111:103345, 2019.

[8] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

[9] Muhammad Fajar Azka Fadillah, Dewinda Julianensi Rumala, Mauridhi Hery Purnomo, and I Ketut Eddy Purnama. The effect of noisy and blurry data on deep learning: Application in brain image classification. In *TENCON 2022-2022 IEEE Region 10 Conference (TENCON)*, pages 1–7. IEEE, 2022.

[10] Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Synthetic data augmentation using GAN for improved liver lesion classification. *CoRR*, abs/1801.02385, 2018.

[11] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[13] Mark S. Graham, Walter H.L. Pinaya, Petru-Daniel Tudosiu, Parashkev Nachev, Sebastien Ourselin, and Jorge Cardoso. Denoising diffusion models for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2947–2956, June 2023.

[14] Shuyang Gu, Jianmin Bao, Dong Chen, and Fang Wen. Giqa: Generated image quality assessment. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 369–385, Cham, 2020. Springer International Publishing.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[16] Andrew Hoopes, Jocelyn S Mora, Adrian V Dalca, Bruce Fischl, and Malte Hoffmann. Synthstrip: Skull-stripping for any brain image. *NeuroImage*, 260:119474, 2022.

[17] Benjamin Q Huynh, Hui Li, and Maryellen L Giger. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *Journal of Medical Imaging*, 3(3):034501–034501, 2016.

[18] A Victor Ikechukwu, S Murali, R Deepu, and RC Shivamurthy. Resnet-50 vs vgg-19 vs training from scratch: A comparative analysis of the segmentation and classification of pneumonia from chest x-ray images. *Global Transitions Proceedings*, 2(2):375–381, 2021.

[19] Talha Iqbal and Hazrat Ali. Generative adversarial network for medical images (mi-gan). *Journal of medical systems*, 42:1–11, 2018.

[20] Jennifer Jepkoech, David Muchangi Mugo, Benson K Kenduiywo, and Edna Chebet Too. The effect of adaptive learning rate on the accuracy of neural networks. 2021.

[21] James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N. Cohen, and Adrian Weller. Synthetic data – what, why and how?, 2022.

[22] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019.

[23] Amjad Rehman Khan, Siraj Khan, Majid Harouni, Rashid Abbasi, Sajid Iqbal, and Zahid Mehmood. Brain tumor segmentation using k-means clustering and deep learning with synthetic data augmentation for classification. *Microscopy Research and Technique*, 84(7):1389–1399, 2021.

[24] Md Sajjad Mahmud Khan, Mahiuddin Ahmed, Raseduz Zaman Rasel, and Mohammad Monirujjaman Khan. Cataract detection using convolutional neural network with vgg-19 model. In *2021 IEEE World AI IoT Congress (AIIoT)*, pages 0209–0212. IEEE, 2021.

[25] Bardia Khosravi, Pouria Rouzrokh, John P. Mickley, Shahriar Faghani, Kellen Mulford, Linjun Yang, A. Noelle Larson, Benjamin M. Howe, Bradley J. Erickson, Michael J. Taunton, and Cody C. Wyles. Few-shot biomedical image segmentation using diffusion models: Beyond image generation. *Computer Methods and Programs in Biomedicine*, 242:107832, 2023.

[26] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

[27] Christian Thode Larsen, J Eugenio Iglesias, and Koen Van Leemput. N3 bias field correction explained as a bayesian modeling method. In *Bayesian and grAphical Models for Biomedical Imaging: First International Workshop, BAMBI 2014, Cambridge, MA, USA, September 18, 2014, Revised Selected Papers*, pages 1–12. Springer, 2014.

[28] Minhyeok Lee and Junhee Seok. Score-guided generative adversarial networks. *Axioms*, 11(12), 2022.

[29] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*. WACV, 2023.

[30] Zheping Liu, Andy Song, Nasser Sabar, and Wenkai Li. Evolving a better scheduler for diffusion models. In Fenrong Liu, Arun Anand Sadanandan, Duc Nghia Pham, Petrus Mursanto, and Dickson Lukose, editors, *PRICAI 2023: Trends in Artificial Intelligence*, volume 14326 of *Lecture Notes in Computer Science*, pages 398–409. Springer, 2024.

[31] Kiran Maharana, Surajit Mondal, and Bhushankumar Nemade. A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, 3(1):91–99, 2022.

[32] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks, 2017.

[33] Dominic Masters and Carlo Luschi. Revisiting small batch training for deep neural networks, 2018.

[34] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205, 2017.

[35] MD Mushfirat Mohaimin. Brain tumor classification.

[36] Quoc Nguyen, Trung Le, Thang Nguyen, and Minh Nguyen Nhat. Class label conditioning diffusion model for robust brain tumor mri synthesis. *Journal of LaTeX Class Files*, 14(8), 2021.

[37] Msoud Nickparvar. Brain tumor mri dataset, 2021.

[38] Artem Obukhov and Mikhail Krasnyanskiy. Quality assessment method for gan based on modified metrics inception score and fréchet inception distance. In Radek Silhavy, Petr Silhavy, and Zdenka Prokopova, editors, *Software Engineering Perspectives in Intelligent Systems*, pages 102–114, Cham, 2020. Springer International Publishing.

[39] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization, 2019.

[40] Jincheng Peng, Guoyue Chen, Kazuki Saruta, and Yuki Terata. 2d brain mri image synthesis based on lightweight denoising diffusion probabilistic model. *Medical Imaging Process & Technology*, 6(1), 2023.

[41] Walter HL Pinaya, Mark S Graham, Eric Kerfoot, Petru-Daniel Tudosiu, Jessica Dafflon, Virginia Fernandez, Pedro Sanchez, Julia Wolleb, Pedro F da Costa, Ashay Patel, et al. Generative ai for medical imaging: extending the monai framework, 2023.

[42] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2016.

[43] Champakamala Sundar Rao and K Karunakara. A comprehensive review on brain tumor segmentation and classification of mri images. *Multimedia Tools and Applications*, 80(12):17611–17643, 2021.

[44] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

[45] Ali M Reza. Realization of the contrast limited adaptive histogram equalization (clahe) for real-time image enhancement. *Journal of VLSI signal processing systems for signal, image and video technology*, 38:35–44, 2004.

[46] Olivier Rukundo. Effects of image size on deep learning. *Electronics*, 42(4):985, 2023.

[47] Ayumi Sada, Yuma Kinoshita, Sayaka Shiota, and Hitoshi Kiya. Histogram-based image pre-processing for machine learning. In *2018 IEEE 7th Global Conference on Consumer Electronics (GCCE)*, pages 272–275. IEEE, 2018.

[48] Kazi Sailunaz, Samer Alhajj, Tansel Özyer, Jon Rokne, and Reda Alhajj. A survey on brain tumor image analysis. *Medical and Biological Engineering and Computing*, 2023.

[49] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[50] Hiroshi Sasaki, Chris G. Willcocks, and Toby P. Breckon. Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models, 2021.

[51] Pooyan Sedigh, Rasoul Sadeghian, and Mehdi Tale Masouleh. Generating synthetic medical images by using gan to improve cnn performance in skin cancer classification. In *2019 7th International Conference on Robotics and Mechatronics (ICRoM)*, pages 497–502. IEEE, 2019.

[52] J Seetha and S Selvakumar Raja. Brain tumor classification using convolutional neural networks. *Biomedical & Pharmacology Journal*, 11(3):1457, 2018.

[53] E. M. Senan, M. E. Jadhav, T. H. Rassem, A. S. Aljaloud, B. A. Mohammed, and Z. G. Al-Mekhlafi. Early diagnosis of brain tumour mri images using hybrid techniques between deep and machine learning. *Computational and Mathematical Methods in Medicine*, 2022:8330833, 2022.

[54] Youssef Skandarani, Pierre-Marc Jodoin, and Alain Lalande. Gans for medical image synthesis: An empirical study. *Journal of Imaging*, 9(3), 2023.

[55] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[56] Xiaofei Sun, Lin Shi, Yishan Luo, Wei Yang, Hongpeng Li, Peipeng Liang, Kuncheng Li, Vincent CT Mok, Winnie CW Chu, and Defeng Wang. Histogram-based normalization technique on human brain magnetic resonance images from different acquisitions. *Biomedical engineering online*, 14(1):1–17, 2015.

[57] Sampada Tavse, Vijayakumar Varadarajan, Mrinal Bachute, Shilpa Gite, and Ketan Kotecha. A systematic literature review on applications of gan-synthesized images for brain mri. *Future Internet*, 14(12):351, 2022.

[58] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016.

[59] Yu Yu, Weibin Zhang, and Yun Deng. Frechet inception distance (fid) for evaluating gans. 09 2021.

[60] Shizhe Zhou, Hongbo Fu, Ligang Liu, Daniel Cohen-Or, and Xiaoguang Han. Parametric reshaping of human bodies in images. *ACM transactions on graphics (TOG)*, 29(4):1–10, 2010.