SYNTHETIC DATA FOR DEEP LEARNING MEDICAL APPLICATIONS: GENERATION, EVALUATION, AND UTILIZATION

BY

FAWAD ASADI

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR

THE DEGREE OF DOCTOR OF ENGINEERING IN

BIOMEDICAL ENGINEERING

COLLEGE OF BIOMEDICAL ENGINEERING

GRADUATE SCHOOL, RANGSIT UNIVERSITY

ACADEMIC YEAR 2023

Dissertation entitled

SYNTHETIC DATA FOR DEEP LEARNING MEDICAL APPLICATIONS: GENERATION, EVALUATION, AND UTILIZATION

by

FAWAD ASADI

was submitted in partial fulfillment of the requirements

for the degree of Doctor of Engineering in Biomedical Engineering

Rangsit University

Academic Year 2023

Prof. Chuchart Pintavirooj, Ph.D.	Assoc. Prof. Manas Sangworasil, Ph.D.
Examination Committee Chairperson	Member
Prof. Suejit Pechprasarn, Ph.D.	Assoc. Prof. Nuttapol Tanadchangsaeng, Ph.D.
Member	Member
Asst. Prof. Jamie O'Reilly, Ph.D.	Thanate Angsuwatanakul, Ph.D.
Member and Co-Advisor	Member and Advisor

Approved by Graduate School

(Prof. Suejit Pechprasarn, Ph.D.)

Dean of Graduate School

June 21, 2024

Acknowledgements

First, I would like to express my sincere gratitude to Dr. Jamie A. O'Reilly, my advisor until recently, for providing invaluable support throughout my research. I also thank my current advisor, Dr. Thanate Angsuwatanakul, for offering feedback during the thesis writing process. Furthermore, thanks go to Dr. Nuttapol Tanadchangsaeng, Director of the Graduate Program, for his continuous supervision and support.

I extend my gratitude to the College of Biomedical Engineering for providing the resources that were crucial to the progression of my research. Special appreciation is reserved for the esteemed Dean and the distinguished faculty members whose unwavering dedication illuminated my academic journey. Acknowledgment is also due to the administrative staff, whose kindness and assistance were indispensable.

Finally, I express my heartfelt appreciation to my family: my mother, for enduring the separation from me during this pursuit, and my late father, who is my greatest motivation to expand my knowledge. Lastly, I am forever indebted to my wife, whose constant encouragement fueled my determination throughout this journey, and to my daughter Layla, who had to spend many nights without my presence. Their love and sacrifice have been my source of strength.

Fawad Asadi

Researcher

6204323	:	Fawad Asadi
Dissertation Title	:	Synthetic Data for Deep Learning Medical Applications: Generation, Evaluation, and Utilization
Program	:	Doctoral degree program in Biomedical Engineering
Dissertation Advisor	:	Thanate Angsuwatanakul, Ph.D.
Dissertation Co-Advisor	:	Asst. Prof. Jamie A. O'Reilly, Ph.D.

Abstract

Deep learning algorithms show promise in medicine for tasks like diagnosis, treatment planning, and health monitoring due to their ability to analyze diverse data and detect complex patterns, though their development is hindered by constraints such as the need for large, labeled, high-quality, and unbiased training datasets, posing challenges in their collection and preparation. Generative adversarial networks (GANs) can potentially address challenges. This research involved training progressively growing GAN (PGGAN) to generate synthetic computed tomography (CT) images of the lungs, investigating the impact of weight initialization methods on Inception v3 performance, and employing StyleGAN2 with adaptive discriminator augmentation (ADA) to generate synthetic fluid-attenuated inversion recovery (FLAIR) magnetic resonance (MRI) images with corresponding masks for deep learning training. The initial findings indicated that synthetic images generated by the PGGAN resembled real images, but there was room for improvement in GAN's performance. The subsequent study emphasized the superiority of pre-trained weights over randomly initialized models. Lastly, StyleGAN2-ADA performed well with a Fréchet inception distance (FID) score of 14.39. The addition of synthetic images did not significantly impact U-nets' performance, but geometric augmentation alongside synthetic images enhanced generalization. Overall, despite their modest impact on training, synthetic data holds the potential to address data collection challenges, warranting broader exploration across other deep learning applications beyond segmentation tasks.

(Total 91 pages)

Keywords: Generative Adversarial Networks, Medical Datasets, Evaluation Metric, Deep Learning, Fréchet Inception Distance.

Student's Signature …..... Dissertation Advisor's Signature .....

Dissertation Co-advisor's Signature .....

Table of Contents

Table of Contents (continued)

List of Tables

Page

Tables

List of Figures

Abbreviations and Symbols

Symbol	Meaning
AI	Artificial Intelligence
DL	Deep Learning
ML	Machine Learning
ANN	Artificial Neural Network
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GAN	Generative Adversarial Networks
PGGAN	Progressively Growing Generative Adversarial Networks
ADA	Adaptive Discriminator Augmentation
MRI	Magnetic Resonance Imaging
CT	Computed Tomography
PET	Positron Emission Tomography
VTT	Visual Turing Test
FID	Frechet Inception Distance
IS	Inception Score
P and R	Precision and Recall
k-NN	k-Nearest Neighbors
t-SNE	t-Distributed Stochastic Neighbor Embedding
LCTSC	Lung CT Segmentation Challenge (dataset)
TCIA	The Cancer Imaging Archive (dataset)

Chapter 1

Introduction

1.1 Background

Artificial intelligence (AI) involves the development of advanced computer software capable of emulating human cognitive functions such as learning, adaptation, reasoning, problem-solving, and decision-making (Zhang & Lu, 2021). It enables the execution of complex tasks traditionally performed by humans and leads to increased efficiency and productivity in some domains (Russell & Norvig, 2016). AI adoption, for better or worse, promises changes in industries such as health care, education, entertainment, agriculture, transportation, and more (Anaya-Isaza, Mera-Jiménez, & Zequera-Diaz, 2021; Guha et al., 2021; Guzman & Lewis, 2020; Jaiswal, 2023; Meskó & Görög, 2020; Nti, Adekoya, Weyori, & Nyarko-Boateng, 2022; Sing, Teo, Huang, Chiu, & Xing, 2022). In recent years, Deep Learning (DL), a subset of AI that employs multi-layered artificial neural networks (ANNs) to learn and extract meaningful information from data, has been crucial in advancing the field, facilitating significant breakthroughs in applications such as image recognition, natural language processing, and object classification (Sharifan & Amini, 2023). Some of its most prominent techniques include convolutional neural networks (CNNs), recurrent neural networks (RNNs), long short-term memory (LSTM) networks, and generative adversarial networks (GANs).

Using medical data in the healthcare sector has enhanced medical services and public health (Zhou, Greenspan, & Shen, 2023). Examples of medical data include health records, imaging, genomic, biometric, and clinical trial data (Albahra et al., 2023). Acquiring medical imaging involves using sophisticated image-capturing techniques and technologies to construct visualizations of what is beneath the skin for medical purposes (Albahra et al., 2023). These technologies provide tremendously

valuable insights to physicians about the patient's health. Some of the most common technologies employed in medical imaging include X-ray, Magnetic Resonance Imaging (MRI), Computed Tomography (CT), fundus, and Ultrasound (Goceri, 2023). A commonly used algorithm in deep learning for image data is the convolutional neural networks (CNNs). It is made up of multiple convolutions and pooling layers, with a fully connected layer at last. It is used for tasks such as object detection, recognition, and segmentation (Deng et al., 2009; LeCun, Bottou, Bengio, & Haffner, 1998).

In medicine, deep learning algorithms are being extensively researched and developed for potential use. Their ability to interpret structured and unstructured data, identify complex patterns, and continually learn to enhance their output makes them powerful tools for medical purposes. These include enhancing diagnostic accuracy, detecting illnesses and abnormalities, tailoring treatment plans, monitoring healthiness, and anonymizing data (Diaz et al., 2021; Tang, 2019; Kermany et al., 2018). Moreover, their immediate analysis of new data after training makes them especially helpful for cases where time for diagnosis and treatment is limited (Ostrom et al., 2019; Shaver et al., 2019). A demonstrative example is where deep learning algorithms are deployed in the analysis of brain tumors, such as gliomas, utilizing Magnetic Resonance Imaging (MRI). However, the development of deep learning algorithms is slowed by several constraints, one of which is the availability of adequate training datasets. An ideal dataset is labeled, diverse, bias-free, high in quality, collected ethically, and large in size. This type of dataset, when used, produces a generalized trained model that performs well on new, unseen-before data. Nonetheless, collecting and preparing such a dataset is a challenging task (Anaya-Isaza et al., 2021; Perone & Cohen-Adad, 2019).

Geometric and synthetic data augmentation are two approaches to tackling the challenge of insufficient data availability for deep learning training (Dang, Vo, Ngo, & Ha, 2022; Shorten and Khoshgoftaar, 2019). The first involves altering the images used for training using geometric transformations such as translation, rotation, zooming, and shear. Theoretically, this would increase the model's generalization due to increased unique image pixel arrangements. In tasks such as segmentation, top-performing participants in competitions like the Multimodal Brain Tumor Segmentation Challenge

have utilized techniques like affine, pixel-level, and elastic deformations to enhance segmentation accuracy (Isense, Kickingereder, Wick, Bendszus, & Maier-Hein, 2019; McKinley, Meier, & Wiest, 2019; Myronenko, 2019; Nalepa et al., 2019). On the other hand, synthetic data augmentation focuses on generating entirely new artificial images rather than modifying existing ones. These synthetic images are then added to the training set. This method offers the potential for a more diverse set of augmented images compared to traditional geometric transformations (Basaran, Qiao, Matthews, & Bai, 2022; Shin et al., 2018; Zhang et al., 2023). However, the effectiveness of synthetic data augmentation relies on factors such as the method of synthetic image generation and how the synthetic data is managed and integrated into the training process (Carver, Dai, Liang, Snyder, & Wen, 2021; Cha et al., 2020; Foroozandeh & Eklund, 2020; Larsson, Akbar, & Eklund, 2022). While both methods have their advantages and limitations, it is not uncommon for researchers to integrate them as they can complement one another (McKinley et al., 2019).

Generative adversarial networks (GANs) consist of two CNNs, the generator and the discriminator (Goodfellow et al., 2014; Skandarani, Jodoin, & Lalande, 2023; Yi, Walia, & Babyn, 2019). These networks compete with one another with the overall purpose of generating realistic synthetic images that resembles the distribution of real images. The generator aims to produce synthetic data that can't be recognized as such, while the discriminator tries to differentiate between real and synthetic images (LeCun et al., 1998). The presumption is that synthetic images could be employed to augment the training set if they are indistinguishable from real images(Kora Venu & Ravula, 2020; Sandfort, Yan, Pickhardt, & Summers, 2019).

The process of assessing generated synthetic data to determine their authenticity, which indicates their usefulness, can be done through a visual Turing test (VTT) (Chuquicusma, Hussein, Burt, & Bagci, 2018) or by employing quantitative evaluation metrics. In the first approach, experts individually examine and rate synthetic samples as either real or fake. This method is expensive and prone to bias errors when done by a single evaluator. Therefore, it is recommended that multiple raters be employed (Chuquicusma et al., 2018; Salimans et al., 2016). Nonetheless, this would be time-

consuming and expensive. On the other hand, several evaluation metrics, such as Inception Score (IS) and Precision and Recall (P and R) have been suggested and widely used as an alternative, cost-effective generated synthetic data evaluation approach (Salimans et al., 2016). Among them, the Fréchet Inception Distance (FID) is considered by many to be the most prominent one (Karras et al., 2020). FID utilizes the Inception V3 image classification network trained on the ImageNet dataset to encode images into 2048-element vector embeddings (Heusel, Ramsauer, Unterthiner, Nessler, & Hochreiter, 2017). This process is carried out for both real and generated synthetic data groups. Then, it evaluates the distance between the distributions of the two groups and produces a scalar score. A lower score suggests a shorter distance, indicating that the two groups are similar.

In this research, GAN architectures and their application in generating synthetic images across multiple medical image types were thoroughly examined. The performance of these GANs was evaluated using state-of-the-art metrics, and the influence of dataset size on these metrics was explored. Additionally, the process of vector embedding generation was investigated, and an augmentation and segmentation pipeline was proposed. Lastly, the impact of geometric and synthetic image augmentation on deep learning performance was assessed.

1.2 Research Objective

The aim of this study is to:

1.2.1 Employ state-of-the-art deep learning algorithms to generate realistic synthetic medical images.

1.2.2 Evaluate the generation performance using state-of-the-art evaluation metrics.

1.2.3 Investigate the impact of hyperparameters on the evaluation process.

1.2.4 Gain insights into the efficacy of geometric and synthetic data augmentation techniques for enhancing deep learning model training.

1.3 Research Questions

By the end of this study, we aim to answer the following questions:

1.3.1 Can Generative Adversarial Networks (GANs) effectively generate real-looking synthetic medical images?

1.3.2 Will the generated synthetic images be correctly identified by quantitative evaluation metrics?

1.3.3 What impact, if any, does the size of datasets containing real and synthetic images have on the performance of quantitative evaluation metrics?

1.3.4 Does utilizing pre-trained weights as opposed to randomly initialized weights for the Inception v3 model affect the computation of FID?

1.3.5 To what extent does utilizing exclusively synthetic images for augmentation affect the training of a deep learning segmentation model?

1.3.6 How does incorporating geometric augmentation techniques, in addition to synthetic images, influence the training of a deep learning segmentation model?

1.3.7 Does the performance of the augmented deep learning model justify the GAN's training cost?

1.4 Research Framework

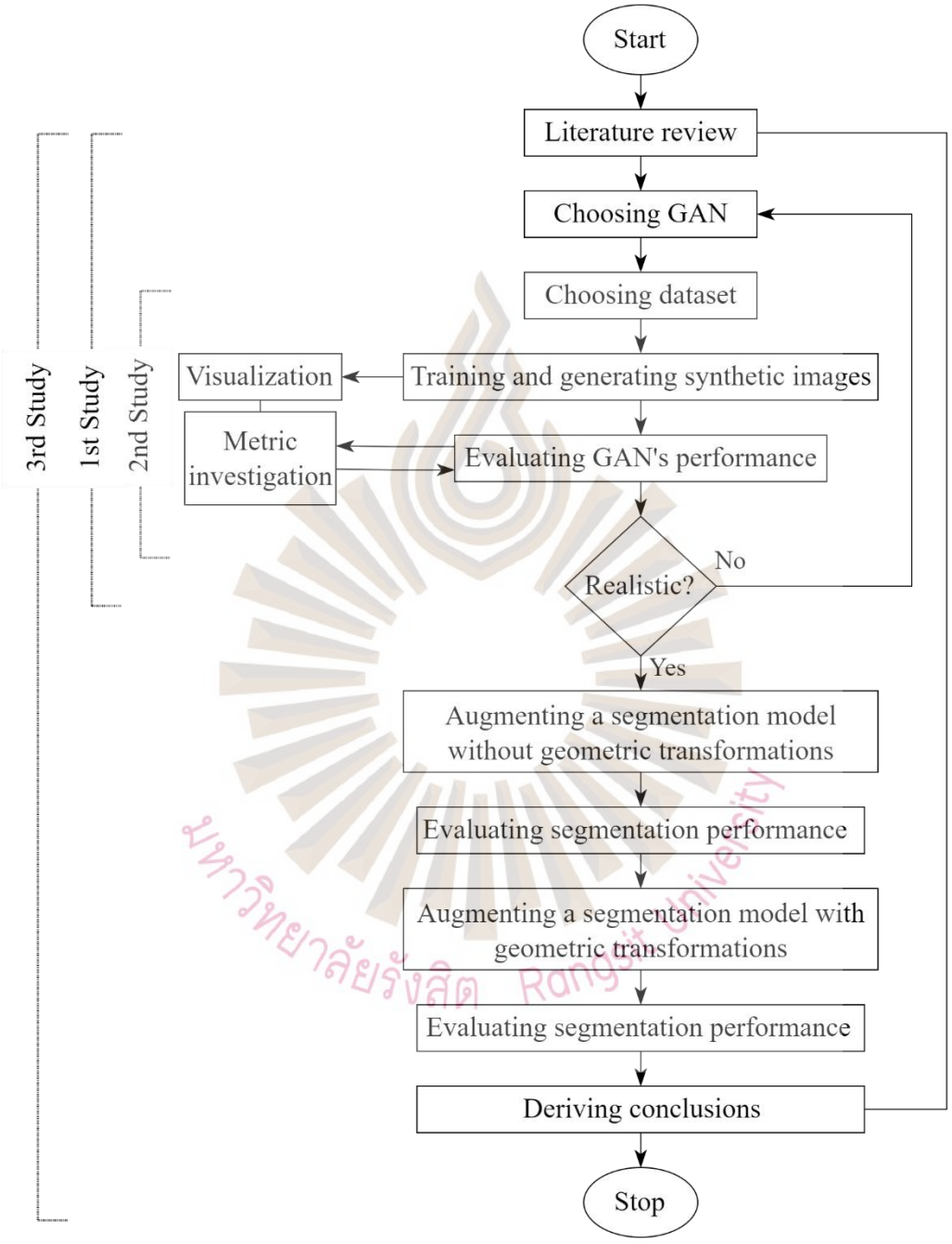


Figure 1.1 Research Conceptual Framework

Source: Researcher, 2024

Chapter 2

Literature Review

2.1 Artificial Intelligence

Artificial intelligence (AI) refers to computer software that aims to replicate human intelligence (Zhang & Lu, 2021). This type of software is considered intelligent due to its ability to learn, reason, and draw conclusions autonomously from given data (Russell & Norvig, 2016). Effectively, it is deemed intelligent when its output passes the Turing test (Gonçalves, 2023). Recently, AI systems have seen a significant increase in usage, driven by improvements in AI algorithm design, data availability, and advancements in computational power manufacturing in conjunction with their affordability (Mondal, 2020). AI is expected to transform several sectors in the near future, including education, health care, finance, manufacturing, retail, communication, entertainment, and more (Guha et al., 2021; Guzman et al., 2020; Jaiswal, 2023; Meskó & Görög; Nti et al., 2022; Sing et al., 2022). On the other hand, machine learning (ML) is a subset of AI that involves the creation of software that leverages data and algorithms in order to learn and adapt to achieve a desired outcome without explicit human intervention (Zhang & Lu, 2021). Multiple ML techniques exist, including supervised, unsupervised, reinforcement, and deep learning (Morales & Escalante, 2022; Sharifani & Amini, 2023). The latter differs in complexity, training data requirements, and application. Generally, deep learning models employ a typical neural network with depth in layers. A typical neural network consists of three layers: input, output and hidden layer. Any more than one hidden layer is considered a deep learning algorithm. The depth of the deep learning algorithm facilitates its ability to perform complex computations to identify sophisticated patterns (Aggarwal, 2018). Deep learning algorithms are power extensive and require large training sets, typically leading to significant training time. On the other hand, deep learning networks require less human intervention.

2.2 Deep Learning in Medicine

Medical images are paramount in the health sector. Physicians and radiologists utilize them to observe soft and hard tissues in the patient's body for screening, diagnostic, treatment planning, and monitoring purposes. Throughout the years, different medical image collection techniques were invented, each for a specific purpose, and some complement each other. Some of the prevalent types of medical images include X-rays, computed tomography (CT) scans, magnetic resonance imaging (MRI), positron emission tomography (PET), eye fundus, and ultrasound (Mallappallil, Sabu, Gruessner, & Salifu, 2020). Deep learning algorithms have successfully aided the medical field in several applications (Diaz et al., 2021; Tang, 2019). Convolutional neural networks (CNNs) are a class of deep learning inspired by the principles of operation of the human's visual cortex. Some CNNs achieved unprecedented results in tasks like segmentation and classification (Brock, Donahue, & Simonyan, 2018; LeCun et al., 1998). However, the application of CNNs and deep learning models remains limited to several constraints. These include the lack of high-quality data, the high cost of labeling, distribution imbalances, potential bias, data collection privacy concerns, and inconsistency in formatting and collection protocols (Alowais et al., 2023; Chen, Lu, Chen, Williamson, & Mahmood, 2021; Piccialli, Di Somma, Giampaolo, Cuomo, & Fortino, 2021; Price & Nicholson, 2019).

2.3 Generative adversarial networks

Generative adversarial neural networks (GANs) are one of the most prominent potential solutions for data scarcity (Chen et al., 2021; Dash, Ye, & Wang, 2024). Ever since their development by (Goodfellow et al., 2014), several architectures have been developed with the aim of generating synthetic images that cannot be distinguished as such (Yi et al., 2019). The possibility of generating synthetic data that can pass the Turing and computational tests would provide a viable solution to the data scarcity problem (Kora Venu & Ravula, 2020; Sandfort et al., 2019). (Frid-Adar, Klang, Amitai, Goldberger, & Greenspan, 2018) employed deep convolutional generative adversarial

networks to generate synthetic liver lesions. They reported enhancement in their classification model performance after training with the augmented training set. (Chuquicusma et al., 2018) The same GAN architecture was used to produce lung cancer nodules and put them to a Turing test. The radiologist concluded that the generated synthetic data were realistic. Another group of researchers experimented with multiple GAN architectures in 2018, including DCGAN, LAPGAN, and PGGAN. The latter generated the most accurate resemblance to the real images (Baur, Albarqouni, & Navab, 2018). The PGGAN was also used by (Korkinof et al., 2018) to generate high-resolution mammograms. And in 2021, a group of researchers also used PGGAN to generate synthetic CT scans. Consequently, the synthetic data in a VTT team comprising 14 trained radiologists. The study showed promising results, with some reservations regarding the anatomical precision of some generated synthetic scans. GANs have also been utilized to overcome privacy-related concerns (Eilertsen, Tsirikoglou, Lundström, & Unger, 2021; Subramaniam et al., 2022). StyleGAN was trained on CT and MR images from patients with pelvic malignancies and achieved promising results, demonstrating effective manipulation of image features and accurate prediction of longitudinal slice positions (Fetty et al., 2020). And in a more recent work, (Hong et al., 2021) employed StyleGAN2 to generate three-dimensional brain MRI images, aiming to address the limitations of current Generative Adversarial Network (GAN) technologies for 3D medical image synthesis.

2.4 Evaluation of Generative Adversarial Networks

Evaluating synthetic images is done through subjective and objective analysis methods. The visual Turing test is considered the best evaluation metric for synthetic images (Chuquicusma et al., 2018). It involves rating each image individually as real or fake by an expert in the field. This could be a radiologist, physician, or an Artificial intelligence and image processing expert. However, humans are influenced by natural inherent bias and vary in experience. This perhaps causes different opinions on image authenticity to be reached. One approach to overcome this subjectivity is employing multiple experts for the evaluation process and declaring the decision on authenticity based on the majority consensus. However, this approach could be unfeasible for its

high cost. Nevertheless, some studies have indeed employed multiple experts to evaluate synthetic data. Others employed single rater for evaluation. In both cases, rating parameters vary in terms of the number of images shown at once, time per image, environment settings, and specific requests given to raters (Dash et al., 2024). (Denton, Chintala, & Fergus, 2015) synthesized and presented CIFAR-10-type natural images to 15 raters to evaluate their authenticity. Synthetic images were mistakenly classified as real images around 40% of the time. (Salimans et al., 2016) followed a different approach by presenting nine images to a group of raters. They were instructed to identify synthetic images out of the group with no time limitation. By the end of each round, the raters were given feedback, which helped improve their ability to distinguish artificial images. 52.4% of MNIST-type images and 78.7% of CIFAR-10-type images were correctly classified from batches with an equal number of real and synthetic images. However, the percentage of artificial images that were mistakenly classified as genuine was not reported. (Chuquicusma et al., 2018) studied their synthetic lung nodule images by presenting two radiologists with 36 batches of all synthetic or half synthetic half real images, in an overall ratio of 3:1 to synthetic data. The radiologists were asked to identify any images that appeared to be artificial. The mean overall accuracy was 46.25%. (Park et al., 2021) conducted similar research; however, the data was presented to ten radiologists with no time limitation and feedback. Moreover, the batch of images consisted of 300 real and synthetic images divided evenly. The radiologists' mean overall accuracy was found to be 59.4% and the experience of the radiologists was found irrelevant.

A cost-effective alternative to VTT is Inception score (IS) (Salimans et al., 2016) and Fréchet Inception distance (Heusel et al., 2017). IS measures the diversity and fidelity of synthetic data, while FID measures the similarity between real and synthetic data (Asadi & O'Reilly, 2021; Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016). Both algorithms rely on a pre-trained Inception v3 model for feature extraction to analyze the distribution of synthetic against real data. However, it has been suggested that relying on these pre-trained weights might not be ideal for evaluating certain types of images, such as medical images, especially when these types of images were not included in the training dataset. This is because the model might have learned biases or features specific

to the types of images it was trained on. However, it is worth mentioning that the authors used MNIST and voice spectrograms with their work, which are fairly simple types of images when compared to everyday color images or medical images. They also employed their own novel density and coverage metrics. Another indirect way to measure GAN's performance is to use the generated synthetic images to train a deep learning model. If the synthetic images are of high in quality and very similar to real images, the performance of the deep learning model would improve. (Yu, Zhou, Wang, Fripp, & Bourgeat, 2018) generated T2 flair images using conditional GAN and then used generated images to train a segmentation model to segment brain tumors. (Hamghalam, Wang, & Lei, 2020) performed comparable work but using different GAN architectures, CycleGAN and cGAN. They reported improved pixel-wise segmentation performance observed in a 0.89 Dice coefficient score. (Carver et al., 2021) extended this idea by training GANs to generate different types of brain MR images along with manually adjusted tumor segmentation masks. They then utilized these synthetic images and masks to augment the training dataset for a U-nets model. They reported an improvement in the Dice coefficient by 4.8%, which indicated the efficacy of this approach in enhancing segmentation performance. However, the introduction of the human element in preparing the masks makes this approach challenging to implement, especially when the training set is augmented with a large number of synthetic images (Chlap et al., 2021; Dash et al., 2024).

2.5 Augmentation for Brain Tumor Segmentation

There are two primary types of data augmentation techniques for brain tumor segmentation. The first type involves transforming original data, which can be categorized into affine, elastic, and pixel-level transformations. Affine transformations refer to geometric changes such as rotation, zooming, cropping, flipping, or translations applied to the training data. However, (Shin et al., 2018) suggested that its effectiveness in enhancing the generalization of deep learning models may be limited due to the generation of highly similar or correlated images. Conversely, elastic transformations involve diffeomorphic mappings, preserving brain shape integrity and resulting in

natural-looking changes. This method has shown promise, particularly when combined with affine transformations. Additionally, in pixel-level augmentation, the transformations occur at the pixel level, including manipulating pixel intensity values, shifting and scaling of pixel-intensity values, gamma correction, sharpening, and blurring. The second type of data augmentation involves the generation and then utilization of synthetic data (Nalepa, Marcinkiewicz, & Kawulok, 2019).

Geometric augmentations played a significant role in the BraTS2018 tumor segmentation challenge. The top-performing algorithms used a mix of affine, pixel-level, and elastic deformation transformations to augment their training data (Isensee et al., 2019; McKinley et al., 2019; Myronenko, 2019). However, it was reported that applying these traditional augmentation techniques causes limited diversity within the training set (Basaran et al., 2022; Shin et al., 2018; Zhang et al., 2023).

Synthetic data augmentation is currently a subject of extensive research. Differences in implementation details may result in varying outcomes (Carver et al., 2021; Foroozandeh & Eklund, 2020; Larsson et al., 2022). (Cha et al., 2020) developed a breast mass detection algorithm for mammography using a deep-learning neural network. He used both real and synthetic images generated by GAN for training. His findings suggest that the algorithm's performance improves based on the number of real and synthetic images used for training. Simply increasing the amount of synthetic data does not guarantee better performance. (Shin et al., 2018) employed a GAN to generate synthetic images to train a brain segmentation model. They explored how different combinations of real and synthetic data affected the model's performance and revealed that incorporating synthetic images alongside real data enhances the model's performance. However, they found that adding geometric augmentations did not further improve the model's performance. Furthermore, (Eilertsen et al., 2021) investigated the impact of ensembled GANs to generate synthetic data for deep-learning models in an effort to overcome the problem of lack of diversity in the synthetic images when generated from a single GAN. They provided evidence supporting using ensembles of independently trained GANs for generating synthetic training data, particularly beneficial for anonymization.

Chapter 3

Research Methodology

Within this research, multiple studies were carried out to investigate synthetic data generation, quality evaluation, and utilization in deep learning training for medical purposes. In the initial study, 512x512 synthetic CT images of the chest were generated using progressively growing generative adversarial networks (PGGAN). The quality of the images was then measured using various quantitative evaluation metrics, including Fréchet Inception Distance (FID), Inception Score (IS), and Precision and Recall (P and R). Additionally, the impact of dataset size on the aforementioned metrics was explored. The FID metric, widely considered the most prominent synthetic data evaluation technique, relies on vector embeddings of real and synthetic images to produce a single score indicating their similarity level. The deep learning architecture responsible for producing those embeddings was investigated in the subsequent study. The final study replaced the PGGAN from the first study with StyleGAN2 with adaptive discriminator augmentation (ADA) to automatically generate fluid-attenuated inversion recovery (FLAIR) magnetic resonance images and corresponding glioma segmentation masks. The effectiveness of the generated synthetic images in training a deep learning algorithm, U-nets, was then evaluated using an augmentation and segmentation pipeline. An overall overview of this research is illustrated in Figure 3.1. Each study is represented in a unique color code and a number indicating its order.

3.1 Preliminary Synthetic Data Generation

In the first study of this research, the PGGAN architecture was utilized to generate high-resolution synthetic CT images of the thoracic region. Several quantitative evaluation metrics, including FID, IS, and P and R, were employed to assess the quality of the generated synthetic images. Furthermore, the influence of dataset size on the aforementioned evaluation metrics was assessed.

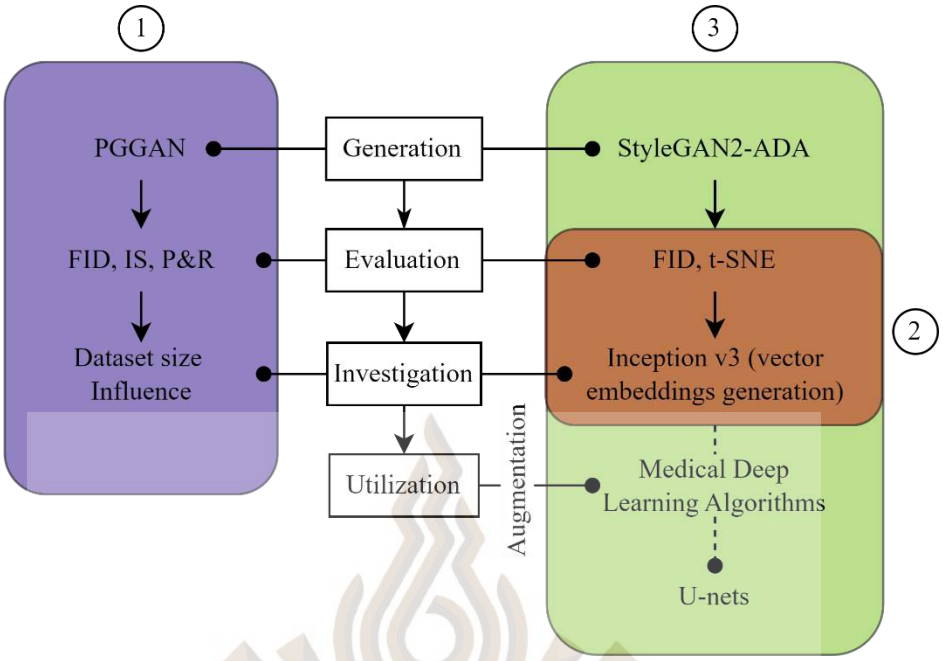


Figure 3.1 Research Overview

Source: Researcher, 2024

3.1.1 Data Collection

The LCTSC public dataset containing a collection of 60 thoracic scans and 9,593 images with a uniform resolution of 512x512 was acquired. The data were collected from 60 patients using CT, RT, and RTSTRUCT modalities. The images were then divided into training and validation sets, with 5,858 and 3,675 images in each set, respectively. To enhance the contrast between soft tissues, pixel values were clipped from −200 to 300 HU. Pixel values were then adjusted to fall within the range of 0 to 255. Subsequently, the images were saved as Joint Photographic Experts Group (.JPEG) files. The final step involved normalizing images to ensure they fall within the range of [-1,1]. This was done before feeding the images into the discriminator network.

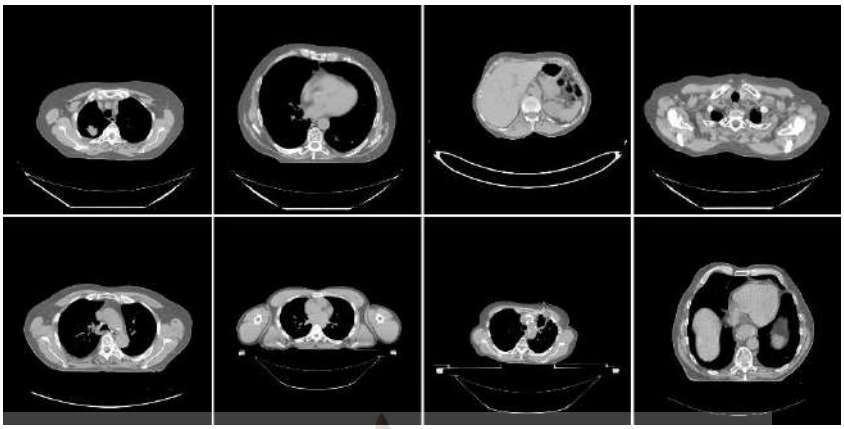


Figure 3.2 A Random CT Images from the LCTSC Dataset

Source: Researcher, 2024

3.1.2 Generative Adversarial Networks

Generative Adversarial Networks (GAN) architecture mainly comprises a generator and a discriminator. The two deep-learning networks work against one another to produce synthetic images that are as similar as possible to real ones (Korkinof et al., 2018). In this setup, the generator takes a random noise vector (z) and constructs synthetic images, while the discriminator binary classifies a given input image into either a real or fake class. This could be a real image from the training set or a synthetic one produced by the generator. The adversarial operation between the two models facilitates the generator's improvement in its ability to generate realistic images over time. Equation 3-1 shows the minimax objective function (V) used to represent the generator (G) and the discriminator's (D) maximum-minimum relationship. The function is split into two sections: The first shows the anticipated value of the logarithm of D(x), which happens when the input image is sampled from the real data distribution. The subsequent part indicates the anticipated value of the logarithm of D(G(z)), which happens when the input image is sampled from the generated data distribution.

$$\min_{\theta_G} \max_{\theta_D} V(D,G) = E_{x\sim P_{data}}[\log D(x)] + E_{z\sim P_z(z)}[\log(1-D(G(z)))] \quad (3\text{-}1)$$

As the discriminator and generator alternate training continues, their weights θ_D and θ_G update to optimize their performances. An optimal training result is achieved when the discriminator classifies its input image as real or fake with 50% confidence. However, GAN's convergence criterion is an active research topic.

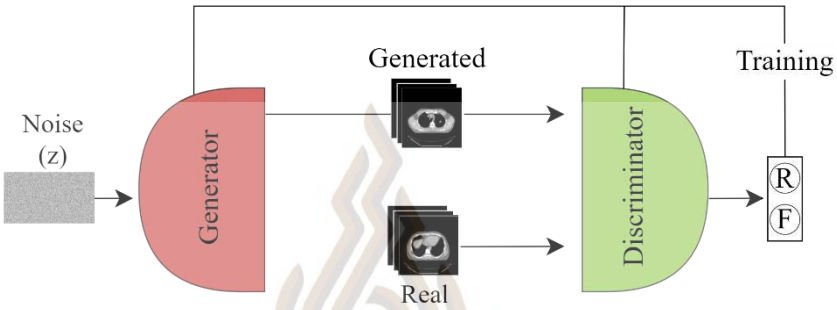


Figure 3.3 Generative Adversarial Networks (GAN)

Source: Researcher, 2024

3.1.3 The Generative Model

The Progressively Growing Generative Adversarial Networks (PGGAN) architecture is known for its ability to generate reliable large synthetic images (Karras, Aila, Laine, & Lehtinen, 2017). It offers training stability as the size of the generated image increases with the model scaling up in several stages. The first image is generated in 4×4 resolution. As the generator and the discriminator double in scale, the output images scale up to 8x8 resolution. The process repeats until a maximum of 1024×1024 images are generated (512×512 in this study). The rescaling process, which involves adding convolutional layers, causes training instability. This was mitigated by a parameter "alpha" and attaching two outputs from the previous block in parallel with the next block. The value of alpha (ranging from 0 to 1) would determine the strength of the connection between the blocks. An increase in Alpha value results in an increase in connection strength and vice-versa (Figure 3.4).

The authors proposed multiple techniques to enhance stability and contribute to generating high-quality images. In "MiniBatch SD" a feature map from the standard

deviation of spatial features in a mini-batch is constructed and included in the penultimate layer of the discriminator. In “Equalized Learning Rate,” the weights of each layer are initialized such that the variance of the outputs of that layer is approximately 1, as opposed to initiating the weights randomly across all layers. “Pixel-wise normalization” is used to decrease the sensitivity to variations in pixel values (Karras et al., 2017; Salimans et al., 2016).

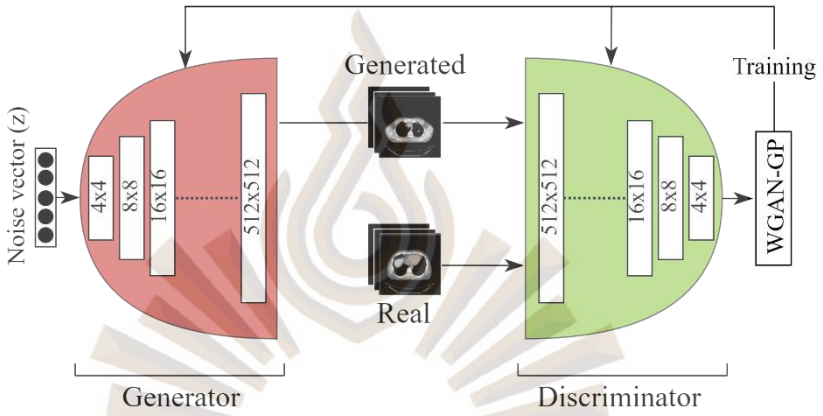


Figure 3.4 Progressively Growing Generative Adversarial Networks (PGGAN)

Source: Researcher, 2024

During training, the model gradually increases image size, and batch sizes adapt in accordance with the size of the image. For example, when images are 4×4, a batch size would be 64, whereas when images scale up to 512×512, the batch size changes to 4. This approach helps prevent overloading our computational resources and optimize the training process. Training employed the Wasserstein GAN gradient penalty loss (WGAN-GP), which is a widely used technique for stabilizing the training of GANs. Adam optimizer was utilized with set initial learning rate of 0.001, and the β1 and β2 values of 0 and 0.99 (Gulrajani, Ahmed, Arjovsky, Dumoulin, & Courville, 2017).

3.1.4 Evaluation

The loss curves of the generator and discriminator during the learning process indicate the stability of the GAN during training. However, they could not be used to

acquire details regarding the similarity level between the generated synthetic and real images. To that end, other metrics are employed for this purpose. They vary in concepts and areas of investigation. Human evaluation through a visual Turing test (VTT) is widely regarded as the gold standard. However, it can be expensive due to the manual evaluation of images and the need for expert evaluators. As a result, researchers are exploring various quantitative techniques to address this challenge (Salehi, Chalechale, & Taghizadeh, 2020). Four metrics were tested to evaluate realism, which included Fréchet Inception distance (FID), Inception score (IS), precision (P), and recall (R).

3.1.4.1 Fréchet Inception Distance

The FID is commonly used to assess the authenticity of generated synthetic images (Heusel et al., 2017). It measures the level of similarity between input images by comparing the disparity between their vector embeddings (g) and real image vector embeddings (x). These embeddings are obtained from the second to last layer (penultimate) of the Inception v3 model or a similar pre-trained deep neural network. A decrease in the FID value indicates a reduction in the disparity between two sets of images, suggesting a high level of similarity between them (LeCun et al., 1998). Equation 3-2 computes the disparity between vector embeddings. In it, μ_x and μ_g represent the mean values of features extracted from real and generated images, respectively. $\sum x$ and $\sum g$ are covariance matrices derived x and g, Tr denotes the trace operator, which sums diagonal elements of matrices, and the ||_|| represents the Euclidean distance.

$$\text{FID}\ (x,g) = \left\| \mu_x - \mu_g \right\|_2^2 + \text{Tr}\left(\sum x + \sum g - 2\left(\sum x \sum g \right)^{\frac{1}{2}} \right) \quad (3\text{-}2)$$

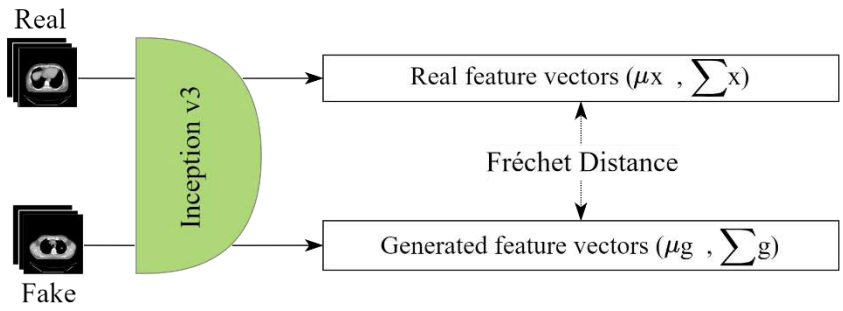


Figure 3.5 Fréchet Inception Distance

Source: Researcher, 2024

3.5 Inception Score (IS)

The IS is another algorithm frequently employed to evaluate synthetic images. It utilizes the pre-trained Inception v3, similar to FID, to compute the marginal class distribution P(y) and the conditional class distribution P(y|x). In Figure 3.6, the P(y) represents the distribution of classes in the dataset. In a best-case scenario, P(y) would be a uniform distribution where each class is equally represented. Meanwhile, P(y|x) represents how likely a generated image belongs to a certain class. In a best-case scenario, a high probability is assigned to the correct class for each image, indicating high confidence in the classification. The Kullback-Leibler divergence (D_{KL}) is then calculated to measure the quality through P(y|x) and the diversity through P(y) for the synthetic images (g). The term 'quality' denotes the general visual fidelity of the images, whereas 'diversity' refers to their variability. When both are high, a high scalar value (score) is outputted, indicating a high similarity between synthetic and real images and vice versa (Salimans et al., 2016). In Equation 3-3, E_x denotes the expected value of D_{KL} between P(y) and P(y|x).

$$IS\ (g) = \exp\left(E_{x\sim p} D_{KL}(p(y|x)||(p(y)))\right) \quad (3\text{-}3)$$

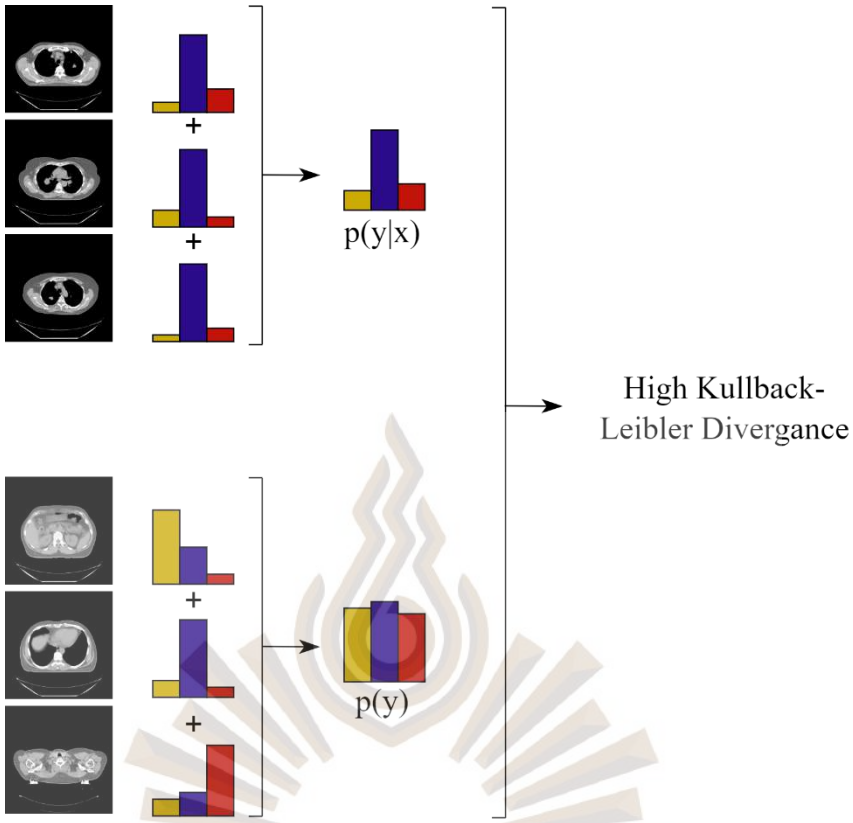


Figure 3.6 Marginal P(y) and Conditional P(y|x) Class Distributions

Source: Researcher, 2024

3.1.4.3 Precision and Recall

Commonly used in classification tasks, P and R can be employed to assess the performance of GANs (Kynkäänniemi, Karras, Laine, Lehtinen, & Aila, 2019). P is a measure of how closely synthetic images resemble real ones. It is calculated by dividing the number of generated synthetic images that fall within the real images manifold by the total number of generated images. R is a measure of the diversity of generated synthetic images compared to the real images. It calculates how many real images are found within the generated images' manifold in relation to the total number of real images (Figure 3.7).

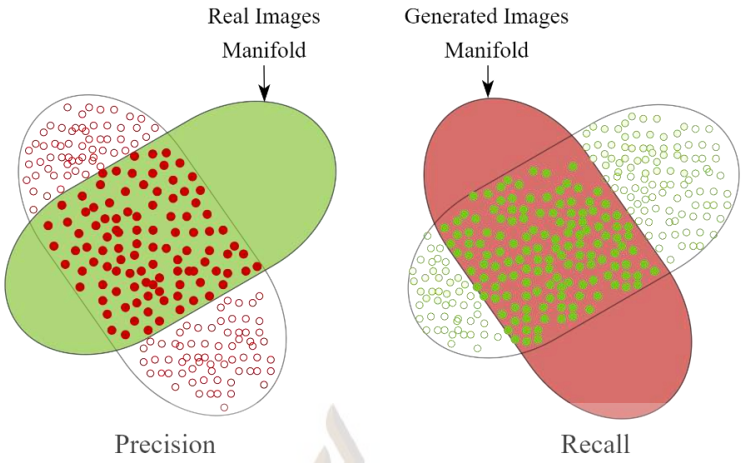


Figure 3.7 Precision and Recall Manifolds: Real (green) – Generated (pink)

Source: Researcher, 2024

Image vector embeddings are first computed from a pre-trained CNN VGG16 model. Both real and generated image vector embeddings are then used to estimate manifolds using a k-nearest neighbors (k-NN) classifier with k set to 3 (Kynkäänniemi et al., 2019). The manifolds are constructed by computing Euclidean distances between feature vectors and creating hyperspheres with a radius equal to the distances to the kth nearest neighbors. The manifolds resulting from this process are volumes formed by the estimated hyperspheres. Equations 3-4 and 3-5 are used to calculate precision and recall, respectively. $\emptyset_r$ and $\emptyset_g$ represent the feature vector of real images and generated images, respectively. Φ_r and Φ_g represents a set of real and generated images feature vectors, respectively. $f(\emptyset, \Phi)$ is calculated from Equation 3-6, where $NN_k(\emptyset', \Phi)$ returns kth nearest feature vector $\emptyset'$ from the set Φ.

$$\text{Precision } (\Phi_r, \Phi_g) = \frac{1}{|\Phi_g|} \sum \emptyset_g \in \Phi_g \, f(\emptyset_g, \Phi_r) \tag{3-4}$$

$$\text{Recall } (\Phi_r, \Phi_g) = \frac{1}{|\Phi_r|} \sum \emptyset_r \in \Phi_r \, f(\emptyset_r, \Phi_g) \tag{3-5}$$

$$f(\emptyset, \Phi) = \begin{cases} 1, \text{ if } \|\emptyset - \emptyset'\|_2 \leq \left\|\emptyset' - NN_k(\emptyset', \Phi)\right\|_2 \\ 0, \text{ otherwise} \end{cases} \tag{3-6}$$

3.2 The Inception v3 Weights Initialization

Two common methods of weight initializing in deep learning architectures include utilizing pre-trained weights and random weights. Pre-trained weights involve using weights learned from training on a different dataset, while random weights involve using random values according to a distribution such as uniform or normal distributions. Some researchers suggest that randomly initializing the weight values could help alleviate any biases that might exist in the training set, unlike pre-trained weights (Naeem, Oh, Uh, Choi, & Yoo, 2020). This has motivated this study to evaluate the reliability of pre-trained (trained on the ImageNet Large Scale Visual Image Recognition Challenge) versus random weights Inception v3 architecture when computing the FID score to assess the quality of generated synthetic images.

The impact of noise on the FID score was investigated. A baseline was established by computing the FID between two matching datasets of X-ray images with no variation. Then, the dataset was divided into two halves and the images in each half were compared with different distortion levels applied. This was done using both pre-trained and random weight initializing models. Comparing the halves of the dataset allowed for a more subtle assessment of the FID metric's performance when there is a similarity but not a complete match between the images, simulating synthetic images compared to real images. Additionally, both models were used to compute the FID score between different types of images (X-ray, CT, Fundus, and dog). Finally, the embeddings extracted by both models were visualized (Heusel et al., 2017).

3.2.1 Data Collection

A total of one hundred images were collected, including chest X-ray, computed tomography (CT) scans of the thoracic region, color fundus photographs, and golden retriever (dog) images (Kermany et al., 2018; Khosla, Jayadevaprakash, Yao, & Li, 2011; Porwal et al., 2018; Yang et al., 2018). Using bicubic interpolation, the images were resized to a standard size of 299x299 RGB pixels and then saved in unsigned 8-bit

integer Joint Photographic Experts Group format. All images were scaled in a range of [−1, 1] to prepare them for the Inception model.

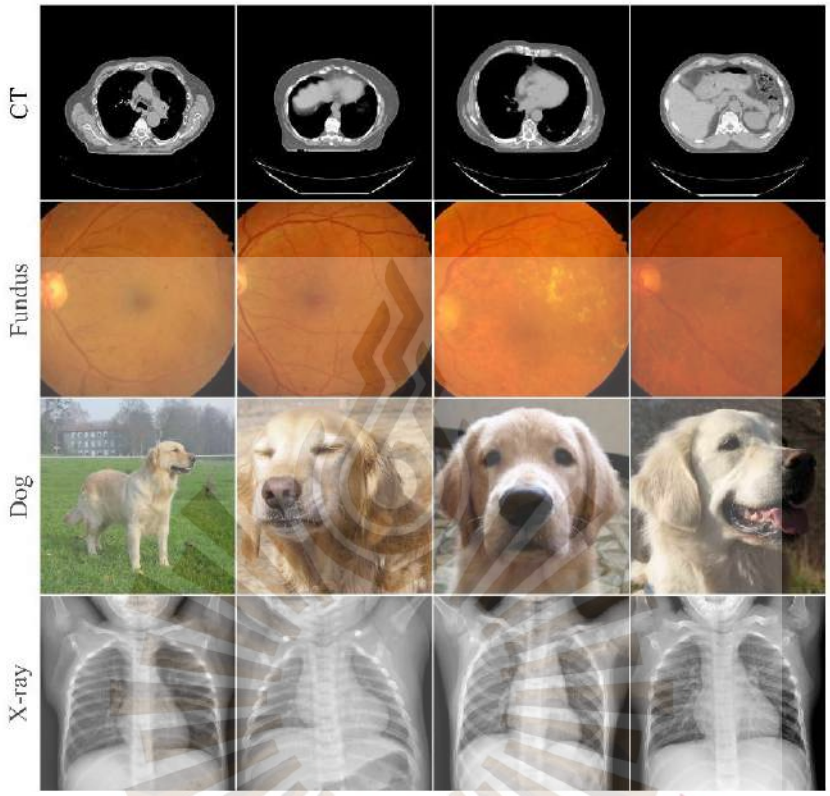


Figure 3.8 Study 2 - Datasets

Source: Researcher, 2024

3.2.2 Image Distortions

Initially, a dataset of 100 unaltered (level 0 distortion) X-ray images was selected. The same dataset was then split into a 1:1 scheme, and different levels of distortion were introduced to one of the two halves incrementally, starting from 1 to 3 as illustrated in Figure 3.9. The FID score was computed to assess how different levels of distortion affect the similarity level between the unaltered and distorted halves reported by FID. Distortions used include:

3.2.2.1 Gaussian Noise

Each image was subjected to a random noise with a mean of zero. The source of the noise was a Gaussian distribution, and the level of distortion was controlled by

adjusting the standard deviation of the noise distribution. Specifically, the standard deviation was set to 10, 20, and 40.

3.2.2.2 Blur

Average filter kernels of varying sizes were applied to blur the images to different degrees. Specifically, kernels of sizes 5, 9, and 15.

3.2.2.3 Swirl

Swirl transformations with a radius of 200 pixels and varying strengths of 2, 4, and 6 were applied to achieve swirl distortion.

3.2.2.4 Impulse Noise

Impulse distortion was applied by randomly converting a certain percentage of pixels to black or white. Specifically, 2%, 10%, or 20% of pixels were converted.

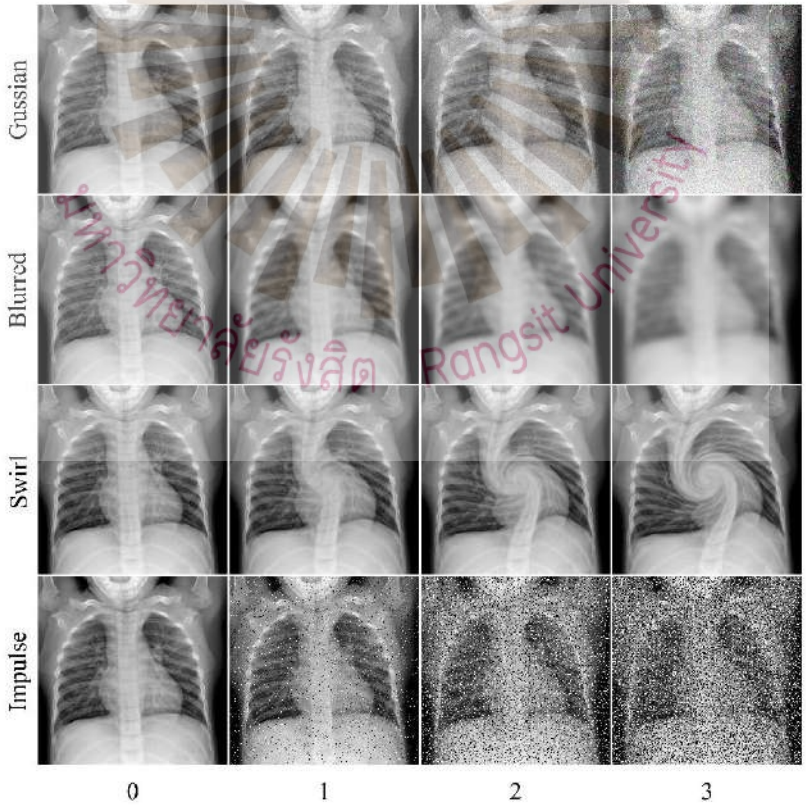


Figure 3.9 Study 2 - Distortions

Researcher: Asadi, 2024

3.2.3 Fréchet Inception Distance

Similar to 3.1.4.1 from the previous study, the FID metric was employed in this study (Heusel et al., 2017). Figure 3.8 illustrates the types of image sets employed in this study. The Inception v3 model is an image recognition algorithm that could also be employed for feature vector extraction (Szegedy et al., 2016). In this study, it was initialized with pre-trained weights derived from the ImageNet dataset and random weights.

3.2.4 Dimensionality Reduction

Visualizing data helps in comprehending the fundamental structure of the data and identifying any existing patterns. To visualize the feature vectors extracted from both real and generated synthetic images, the 2048-dimensional vectors needed to be transformed into a 2-dimensional representation. To achieve this, t-SNE (t-distributed Stochastic Neighbor Embedding) was utilized.

t-SNE is a nonlinear technique that emphasizes the preservation of local similarities between data points. It maps high-dimensional data to a lower-dimensional space by minimizing the difference between their probability distributions according to the Kullback-Leibler divergence, it aims to retain the relative distances between data points (Van der Maaten & Hinton, 2008).

3.3 Synthetic Data Generation and Integration for Deep-learning Training

During the final study of this research, the GAN architecture employed in the first study was replaced with a more advanced, state-of-the-art GAN model: StyleGAN2, with adaptive discriminator augmentation (ADA) to generate synthetic fluid-attenuated inversion recovery (FLAIR) magnetic resonance images and corresponding glioma segmentation masks.

An automated pipeline for a training set augmentation and then object segmentation was aimed to be established. Insights into the usefulness of the generated synthetic images, particularly for medical purposes, would be offered by the performance of the deep learning segmentation model (U-nets). Additionally, another way to assess the quality of the generated images alongside FID would be served as by it. To improve augmentation, established geometric augmentation techniques were combined with synthetic data to determine the most effective augmentation methods for automated applications using deep learning models.

3.3.1 Data

The Cancer Imaging Archive (TCIA) was the source of the dataset used in this study. It contains brain MR images of lower-grade glioma paired with manually annotated segmentation masks and genomic cluster data. However, the latter was abandoned in this study. The masks highlight abnormality regions in FLAIR (Fluid-attenuated inversion recovery) sequences. In total, the dataset consists of 110 scans from various patients. A few patients do not have both MR sequences (T1 and T2), meanwhile all of them have complete FLAIR sequences (Buda, Saha, & Mazurowski, 2019; Clark et al., 2013). Images were preprocessed by loading and standardizing them, then clipping them within the range of [-2, 7], followed by normalization to the range [0, 255]. From 8-bit FLAIR sequence images, 3,929 uniformly sized images, each 256×256 pixels, were extracted. Masks were utilized to distinguish between pixels representing tumors (indicated by a value of 0) and those that did not, indicated by a value of 1. When tumor tissue was not detected in the initial scans, masks of equivalent size were created in a matrix where every element equals zero. These images were then organized sequentially based on their order in the scans to prevent potential data leakage. The first 2,751 (70%) images were allocated for training, followed by 590 (15%) for validation, and the remaining 588 (15%) for testing.

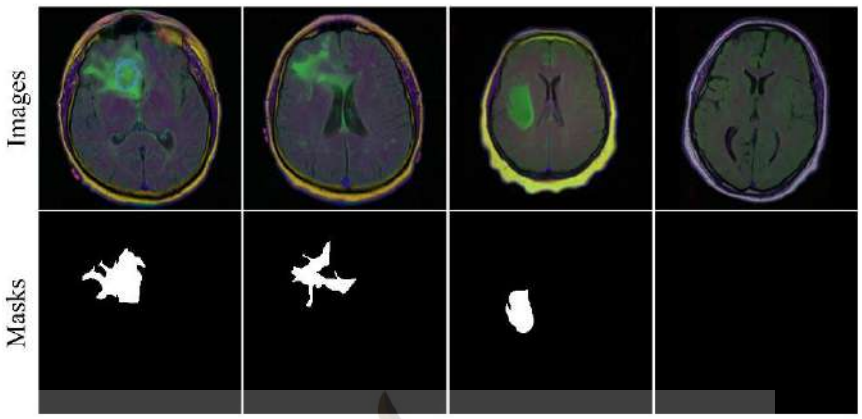


Figure 3.10 Study 3 - Dataset

Source: Researcher, 2024

3.3.1.1 Data Preprocessing for GAN

The training set was exclusively utilized to train the generation model in order to prevent any data leakage that could impact the performance of the segmentation model. The validation and test sets were kept separate from this. The GAN training set comprised three-channel images. Within these images, the FLAIR sequence was denoted in the green and blue channels, while the corresponding masks were present in the red channel. The objective was to generate synthetic labeled images and eliminate the laborious task of manually segmenting the tumors (Carver et al., 2021).

Subsequently, during the data augmentation for training the segmentation model, the three-channel GAN-generated images were divided into grayscale FLAIR images and mask images. The average of the green and blue channels was calculated to generate artificial FLAIR images, whereas the red channels were thresholded at >128 to derive binary segmentation masks.

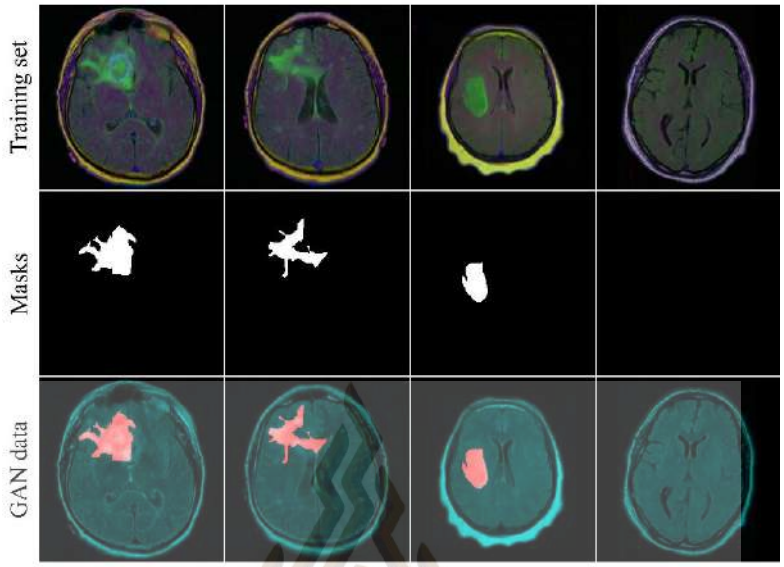


Figure 3.11 GAN Prepared Dataset

Source: Researcher, 2024

3.3.1.2 Data Preprocessing for Segmentation Model

The grayscale FLAIR images and binary segmentation masks were processed to ensure that their pixel values were normalized to fall within the range of [0, 1]. Additionally, these images were kept at a consistent size of 256x256 pixels. This standardization ensured that the images could be effectively analyzed and compared.

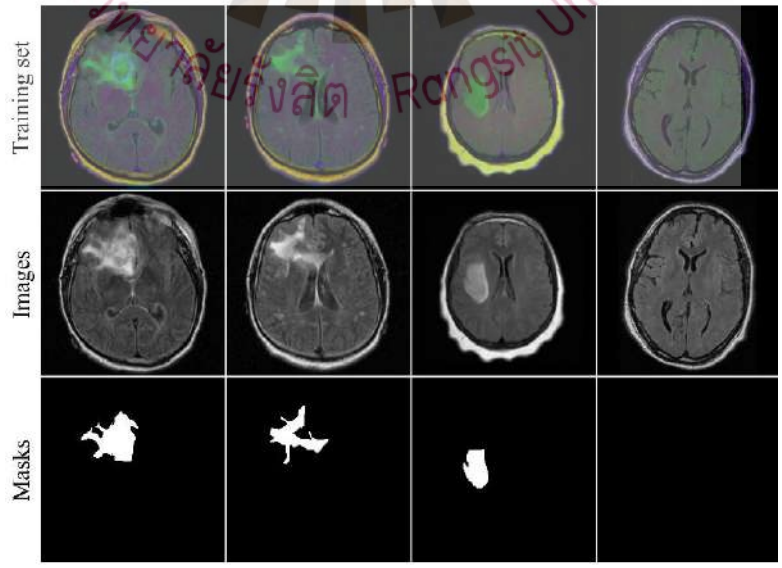


Figure 3.12 U-nets Prepared Dataset

Source: Researcher, 2024

3.3.2 The Generative Model

StyleGAN2 is an innovative variation of Generative Adversarial Network (GAN) that has gained significant recognition for its ability to produce highly realistic images at high resolution. Its remarkable performance has contributed significantly to the development of synthetic image generation (Karras et al., 2020). Its generator network receives a randomly sampled latent vector (Z) from a Gaussian distribution as an input to a learned mapping network. The mapping network comprises eight fully connected layers. The output of this network is a mapped latent vector into a latent space, style vector W, of size 512 that controls the style of generated feature maps by the synthesis network. The synthesis network consists of multiple blocks, the number of which depends on the target resolution of the generated images. Each block consists of multiple components, with the overall objective being to progressively increase the resolution of the feature maps and refine the generated image. These include two convolutional layers for feature extraction, style modulation for fine-grained control over image features, noise Injection to create diverse and realistic images and prevent overfitting, up-sampling to up-sample the feature maps to higher resolutions, skip connections to connect the output of a block to the input of the next block and avoid stage-transitioning artifacts, instance normalization (IN) to normalizes feature maps per instance and an activation function for linearity introduction (Karras et al., 2020).

The discriminator network comprises multiple convolution blocks matching the synthesis network and progressively reduces the resolution of the feature map by multiples of two. Each down-sampling step corresponds to a specific block in the discriminator. Close to the classification stage of the network, a mini-batch standard deviation layer is utilized. This layer computes the standard deviation for each feature map within a mini-batch. The resulting standard deviations are averaged and added as an extra feature to the feature maps. Finally, the last feature map is flattened and passed through a fully connected layer for classification.

StyleGAN2 requires a large training set to avoid overfitting the discriminator. This limits its applications due to data scarcity. However, (Karras et al. 2020) proposed a

solution, StyleGAN2-ada, which introduces adaptive discriminator augmentations. The objective of this method is to avoid the influence of training image augmentations on the generated images. To accomplish this, a series of stochastic discriminator augmentations is exclusively implemented on images before their introduction to the discriminator. This ensures that the discriminator is trained to distinguish between real and fake images using the same augmentations while the generator learns to generate images without such augmentations.

Furthermore, to enhance StyleGAN2-ada's performance, eighteen distinct image manipulations in a predefined order are applied with specific proportions to input images fed into the discriminator. Evidence shows that this does not compromise the quality of generated images. Heuristics are used to detect overfitting in the discriminator and automatically adjust the percentage augmentations applied during training (Karras et al., 2020; Situ, Teng, Liu, Luo, & Zhou, 2021; Asadi, Angsuwatanakul, & O'Reilly, 2024).

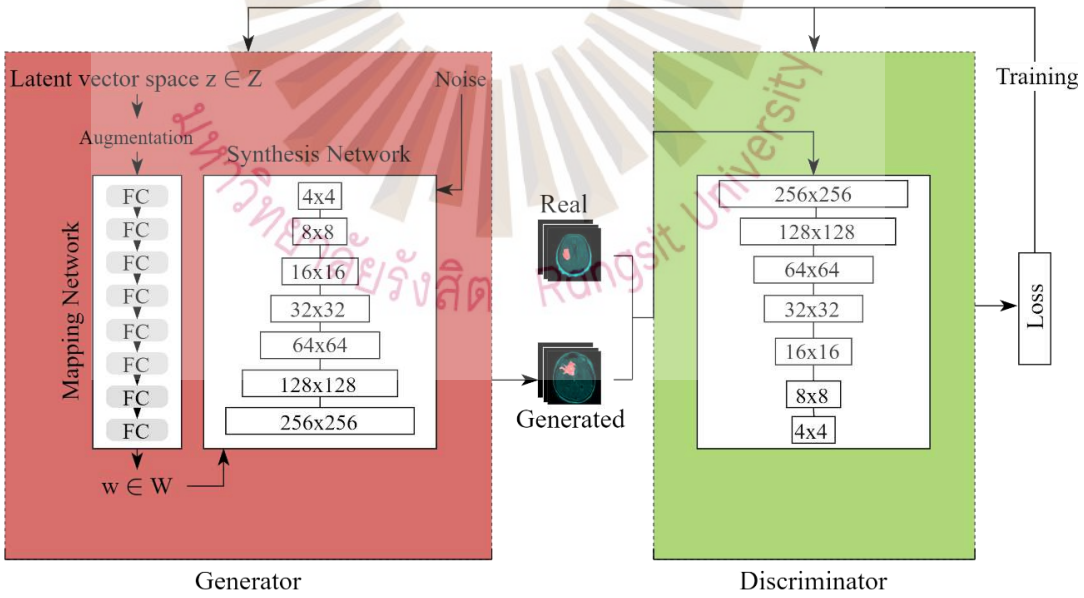


Figure 3.13 StyleGAN2, with Adaptive Discriminator Augmentation (ADA)

Source: Researcher, 2024

We utilized a tailored data augmentation pipeline throughout the training phase, incorporating flipping, rotation, scaling, and color adjustments. Transfer learning was employed using a pre-trained network on FFHQ dataset images at a resolution of 256×256. Batch size 32 was used, and the model was saved every three iterations, with snapshots of the generated images captured for visual evaluation.

3.3.3 Evaluation of the Generative Model

The effectiveness of the GAN was assessed using subjective and objective evaluation techniques. A total of 100 generated images were randomly inspected and their quality was analyzed in the subjective evaluation. Additionally, all generated data was thoroughly analyzed to identify any lower-quality images produced by the GAN. The Fréchet Inception Distance (FID) was used to assess the GAN's performance quantitatively (Heusel et al., 2017). The FID score provides a measure of how similar the generated images are to the real ones, with smaller FID scores indicating better performance by the GAN. The FID metric is based on the activations of an Inception V3 model trained on the ImageNet dataset (Szegedy et al., 2016). The activations are extracted from a pooling layer and are assumed to follow a multivariate normal distribution. The FID score is then computed as the distance between the means of the activations of the real and generated images minus the trace of the product of the covariance matrices of the real and generated images.

3.3.4 Segmentation Model

The U-net architecture is a deep-learning model commonly used for image segmentation. Created by (Ronneberger, Fischer, & Brox, 2015) in 2015, the model features four encoding blocks, each containing multiple convolutional layers, batch normalization, ReLU activation, and max pooling. The encoder feeds into a convolutional block with 1024 filters before transitioning to the decoder. The decoder has four blocks with convolutional transpose layers that double the spatial dimensions. The number of filters in the decoder matches that of the encoder, with 512, 256, 128, and 64 filters respectively. The final output is generated from a convolutional layer with

one filter and a sigmoid activation function. This architecture has been successful in various medical imaging applications and has been shown to be effective in performing automatic image segmentation tasks.

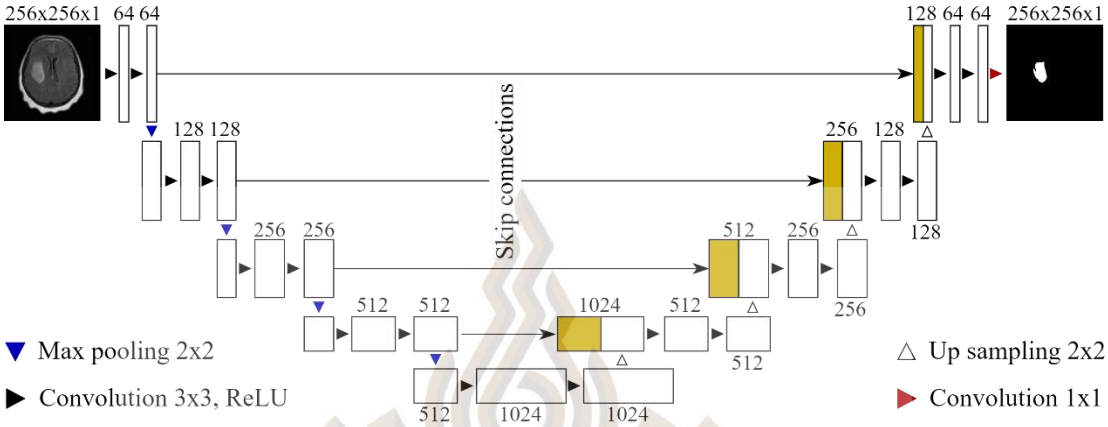


Figure 3.14 Segmentation Model Architecture (U-nets)

Source: Researcher, 2024

The U-nets first underwent fifteen rounds of training. It began with a real dataset for the initial round and then expanded the training set by 1,000 generated synthetic images with their corresponding glioma segmentation masks for each subsequent round. By the fifteenth round, 14,000 generated images were utilized along with their masks. Following this, the process was repeated, this time with the introduction of geometric augmentation techniques in addition to the synthetic data. Randomly applied geometric augmentations included horizontal and vertical flips, translations of up to 30% of the image size in both horizontal and vertical directions, shearing with a range of 0.2, zooming with a range of 0.2, and brightness adjustment with a range of values between 0.5 and 1.05. Any new edge pixels were filled using the 'wrap' mode, which was also employed to populate newly added edge pixels. The validation set was used throughout the training process to track the model's progress by the end of training iterations. The model was set to stop training if no progress was made in the dice coefficient score.

3.3.5 Evaluation of Segmentation Model

The computational costs of U-net training were analyzed by quantifying the total number of completed training iterations, total training time (in hours), and time per iteration (in minutes), which are reported in Tables 4.6 and 4.7. The evaluation metrics used at the end of each training iteration included intersection over union, precision, recall, and the Dice coefficient. However, the Dice coefficient was primarily used to evaluate segmentations, where a higher value indicated segmentations with a high level of accuracy. The learning curves (Figure 4.6) were constructed using values recorded by the end of each iteration for both the training and validation sets. Meanwhile, the test set was employed to evaluate trained models. Additionally, the dice coefficients for the training, validation, and test sets were examined to capture the influence of synthetic and geometric augmentations compared to the baseline, which exclusively relied on real images. Finally, Pearson's correlation coefficient (r) was used to assess the correlation between these metrics and the amount of synthetic data used.

Chapter 4

Research Results and Discussion

The results of each study conducted in the context of this research are presented in this chapter. These findings and their respective contributions to the overarching objectives of this research are thoroughly discussed. The first study involved generating synthetic Computed Tomography (CT) images and investigating the influence of the dataset size on quantitative evaluation metrics. The following study examined the impact of using a randomly initialized Inception V3 network to produce image feature vectors versus pre-trained weights. In the final study, the findings of previous studies were leveraged to evaluate the effectiveness of using synthetic data augmentation to train a deep fully convolutional network for automatically segmenting brain tumors in neuroimaging data.

4.1 PGGAN Data Generation and Metrics Investigation

The PGGAN was trained in 8 stages; at each stage, the model was scaled up by a factor of 2, producing images in higher resolution, starting from 4×4 and going up to 512×512. The training was stable despite the loss fluctuation displayed in Figure 4.1. Visually, most of the generated images resembled real ones to a high degree. However, there were some exceptions. Appendix A illustrates realistic and unrealistic images produced by the PGGAN. Tables 4.1, 4.2, and 4.3 report evaluated metric scores, showing how dataset size affects FID, IS, and P and R metrics, respectively. Best scores were observed with the largest sets. Hence, the rest of this discussion focuses on those results.

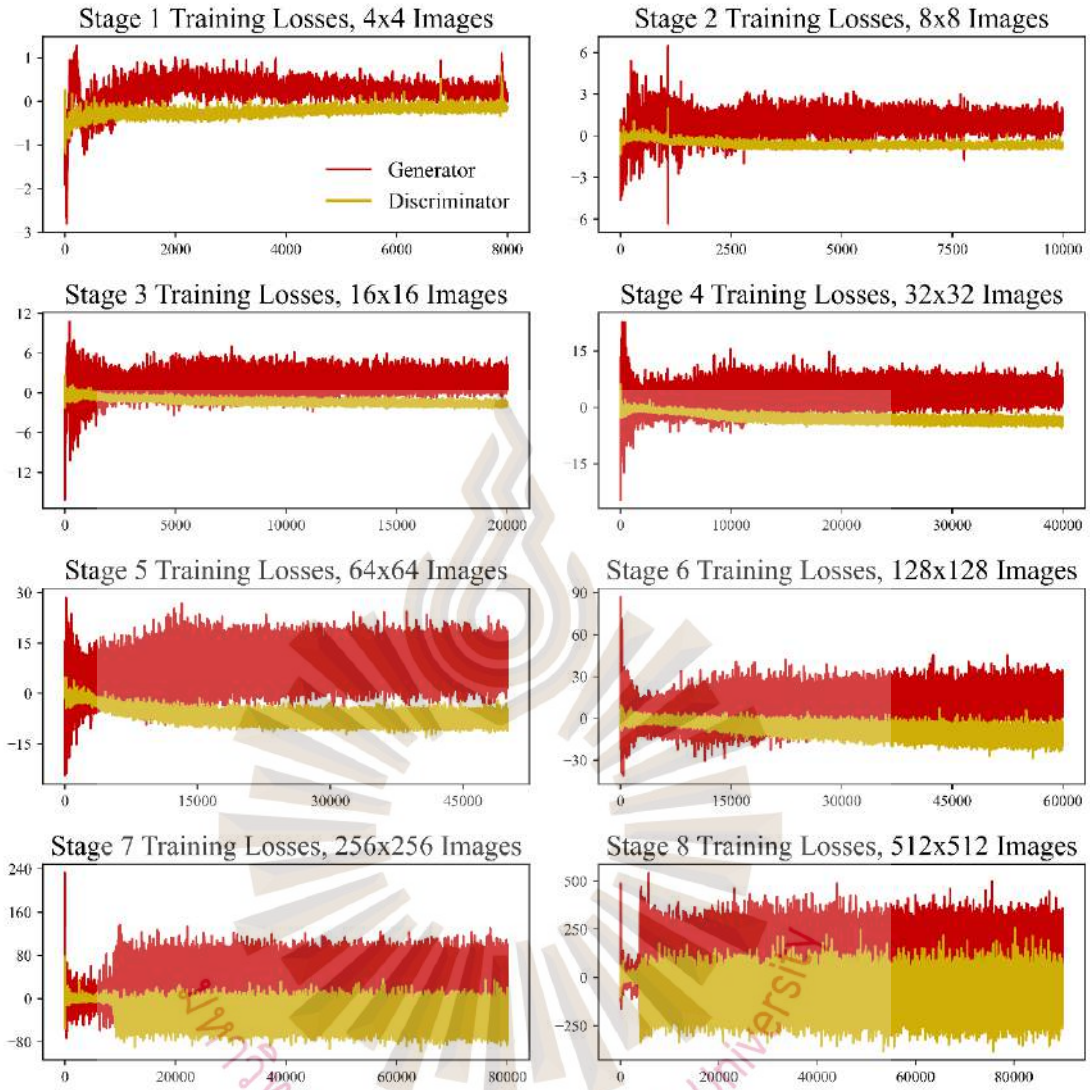


Figure 4.1 PGGAN Training Loss Curves

Source: Researcher, 2024

FID scores in Table 4.1 show an inverse relationship with dataset size when evaluated among real and against generated image sets. The FID score decreased, indicating higher similarity between the two sets, as we incremented the number of images equally for both sets. At 3,675 images for all, the FID score between the generated and validation sets was 42.40, the training and validation sets was 24.06, and the generated and training sets was 31.6. This indicates a potential margin for improvement by 18 between generated and training sets when compared with training and validation score. Unexpectedly, in some studies, the FID computed between the

training and validation sets, both real CT images, exceeded that reported for real and generated images (Skandarani et al., 2021).

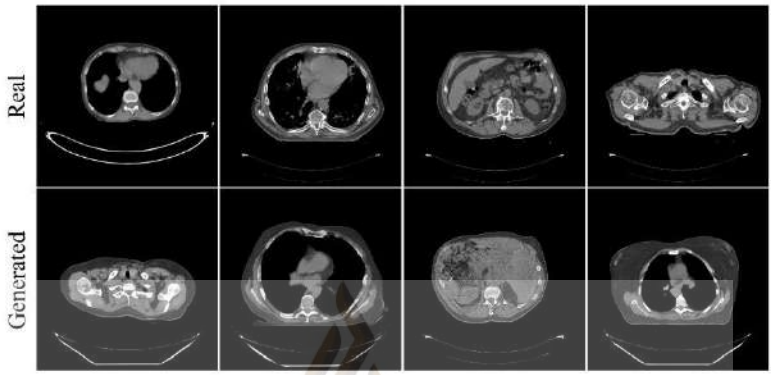


Figure 4.2 Real (Top) and PGGAN Generated Images (Bottom)

Source: Researcher, 2024

Table 4.1 Fréchet Inception Distance (Standard deviation, five replications)

Number of images	Generated VS Validation	Training VS Validation	Training VS Generated
1,225	52.14 (1.30e-05)	32.10 (3.70e-06)	42.03 (7.50e-6)
2,450	44.78 (8.31e-08)	26.41 (7.02e-08)	33.884 (4.50e-08)
3,675	42.40 (2.11e-12)	24.06 (1.62e-12)	31.66 (4.40e-12)

As seen in Table 4.2, the change in inception score did not significantly vary with the number of image variations. Also, the score was almost similar for both real and generated images, indicating high similarity between the two sets. However, the validation set was lower. This perhaps stems from the variations among real images, as the training set consisted of 45 CT scans while the validation set consisted of 15 CT scans.

Table 4.2 Inception Score (Standard deviation, five replications)

Number of images	Generated VS Validation	Training VS Validation	Training VS Generated
1,225	3.02 (0.004)	2.71 (0.004)	2.938 (0.004)
2,450	3.00 (0.002)	2.764 (0.002)	2.992 (0.002)
3,675	3.075 (0.002)	2.766 (0.001)	3.02 (0.001)

In Table 4.3, As the dataset size increased, the precision and recall values generally decreased, indicating a potential trade-off between dataset size and image. The most favorable scores were observed when comparing the generated images to the training set, as anticipated. In contrast, the comparisons with the validation set yielded relatively lower scores. These outcomes are partially attributed to the sensitivity of the k parameter, which significantly impacts the precision and recall metrics.

Table 4.3 Precision and Recall (P, R; k=3)

Number of images	Generated VS Validation	Training VS Validation	Training VS Generated
1,225	0.30, 0.47	0.38, 0.55	0.67, 0.77
2,450	0.13, 0.37	0.16, 0.31	0.46, 0.73
3,675	0.04, 0.32	0.06, 0.18	0.29, 0.68

Overall, this study's findings suggest that although the generated synthetic images appeared realistic, there is still room for improvement in the image generation process. This is evident by comparing the FID score between validation and generated images and validation and real images. Appendix A illustrates both realistic and unrealistic images generated by the PGGAN. Additionally, the study highlights that larger datasets lead to more precise evaluation scores, particularly in FID. This underscores the significant impact of dataset size on quantitative evaluation metrics.

4.2 Pre-trained Inception v3

This study investigated the initialization method of weights for the Inception v3 convolutional neural network model, which produces feature vectors for real and generated synthetic images in the form of 2048-element vectors. These vectors are subsequently employed to compute the FID score. Two common weight initialization methods are explored: pre-trained weights and random weights. Pre-trained weights are derived from training with a subset of everyday color images from the ImageNet Large Scale Visual Image Recognition Challenge (ILSVRC) dataset.

It can be observed from Figure 4.3 that increasing the level of distortion would typically increase the FID. However, the level of incrementation is highly influenced by the type and level of distortion. This is especially true for the randomly initialized inception V3 model. At level 0 of distortion, FID was close to zero as expected. To a certain extent, similar FIDs were observed at level 1 of distortion for some of the noises. The sample size might have caused this. Generally, the pre-trained model produced consistent FID scores that increased in a proportional relationship with the level of distortion. The random model was sensitive to Gaussian and impulse distortions and insensitive to swirl distortion.

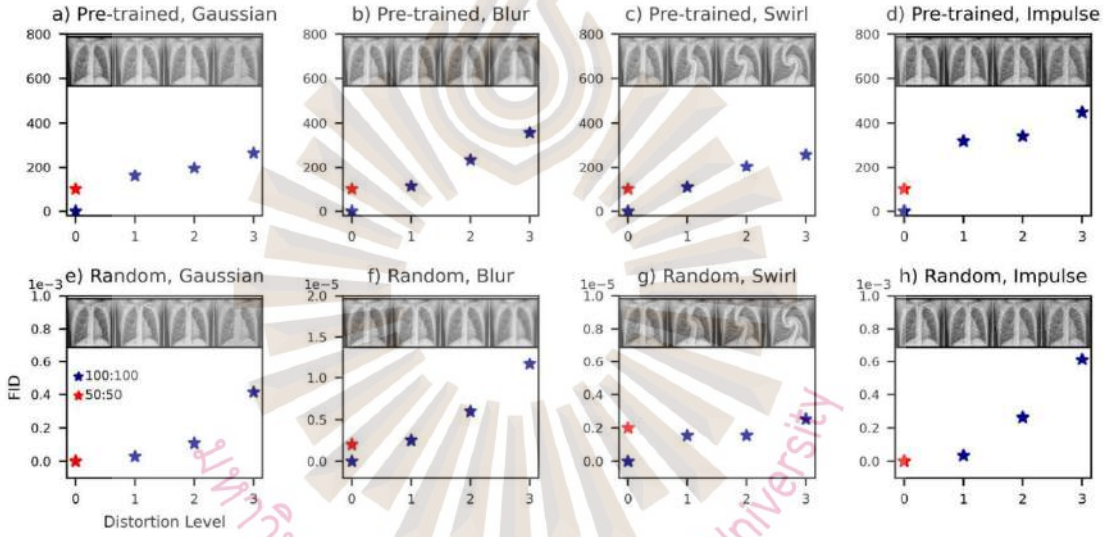


Figure 4.3 Distortions Effect on FID

Source: O'Reilly & Asadi, 2021

Tables 4.4 and 4.5 include normalized FID values computed for feature vectors produced by both the pre-trained and random models. It is observed that the pre-trained model yields high FIDs between different image types, indicating that the feature vectors it produces are more diverse and distinct. When comparing the same image type, the FIDs are slightly larger than those produced by the random model, suggesting that the feature vectors produced by the random model are more consistent with each other. However, the random model also results in lower FID between different types of images, indicating that the feature vectors it produces are less diverse and distinct from each

other, which is unfavorable. Overall, the pre-trained model appears to be better at distinguishing between image types.

Table 4.4 Normalized FID Between Different Types of Images for Pre-trained Model

	X-ray	CT	Fundus	Dog
X-ray	0.093	0.773	1.000	0.78
CT	0.773	0.100	0.803	0.722
Fundus	1.000	0.803	0.000	0.777
Dog	0.789	0.722	0.777	0.030

Table 4.5 Normalized FID Between Different Types of Images for Random Model

	X-ray	CT	Fundus	Dog
X-ray	0.002	1.000	0.467	0.251
CT	1.000	0.000	0.176	0.323
Fundus	0.467	0.176	0.001	0.060
Dog	0.251	0.323	0.060	0.010

t-SNE was employed to transform image feature vectors into a two-dimensional space. Unlike the random model, the feature representations produced by the pre-trained model were linearly separable. This further confirms that the pre-trained model outperforms the randomly initialized model at capturing meaningful image features across different types of images.

4.3 StyleGAN2-ADA Data Generation and Integration for Deep-learning Training

In this study, the significant disparity in FID scores between (real and generated) and (real and validation) sets, observed in the initial study, underscores the necessity for an alternative GAN architecture capable of generating more realistic synthetic images. To that end, StyleGAN2-ADA was trained for three full days, during two of which, the generated synthetic image quality subjectively improved through visual tests and FID

scores. On the third day, the training hit a plateau, and there was no decrease in FID or noticeable visual improvements. The best model produced an FID score of 14.39, falling within the expected range for GAN-generated neuroimaging data based on existing literature (Kossen et al., 2022; Kossen et al., 2021; Subramaniam et al., 2022). However, it is worth noting that factors like image resolution, dataset size, and the specifics of the computational approach can influence the FID score, potentially leading to misleading comparisons (Nunn, Khadivi, & Samavi, 2021; O'Reilly & Asadi, 2021; O'Reilly, 2022). The generated images were observed to be realistic and akin to real ones, yet they don't replicate them exactly, a desirable outcome. There were almost no visually noticeable artifacts. A significant improvement compared the PGGAN generated images (Asadi & O'Reilly, 2021; Carver et al., 2021; Foroozandeh & Eklund, 2020; Karras et al., 2020; Karras et al., 2019; Park et al., 2021).

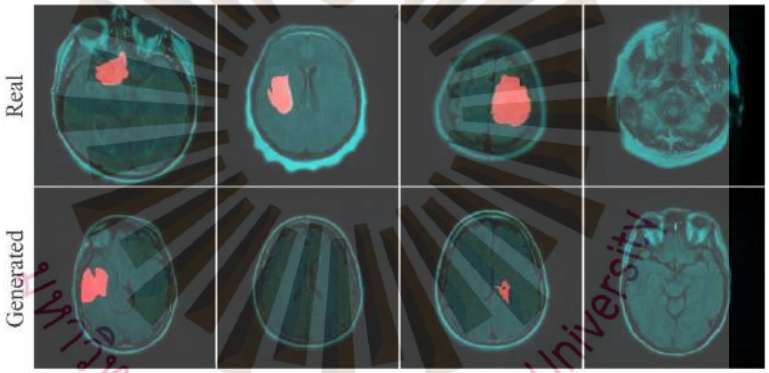


Figure 4.4 Real (Top) and StyleGAN-ADA Generated Images (Bottom)
Source: Researcher, 2024

Further investigation went into assessing generated images by employing t-SNE for feature vector visualization. We sampled 2,751 real and 2,751 random synthetic images and extracted their feature vectors using a pre-trained Inception V3 model. The resulting visualization displayed an overlap between the real and generated synthetic image distributions (Figure 4.5). This further backed our visual assessment and the low FID score, showing that synthetic images are indistinguishable as such, similar to the work of (Woodland et al., 2022). However, a small number of poorly generated images were found through an intensive visual inspection. Those images lacked details in

FLAIR images or had a noise/fade in segmentation masks, which was mitigated by applying thresholding. A sample of poorly generated images can be found in Appendix C.

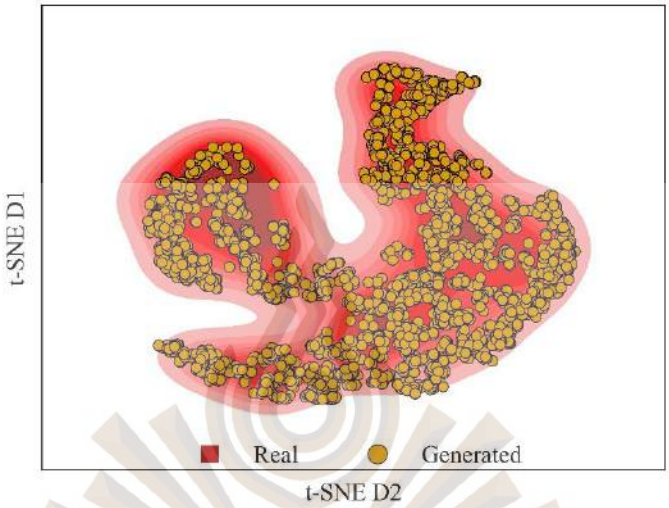


Figure 4.5 t-SNE Visualization for Real (Blue) and Synthetic (Red) Image Feature Vectors

Source: Researcher, 2024

Various elements need to be assessed when considering the tradeoff between the computational costs of training a GAN and generating synthetic images and the benefits of using them. These elements include the time, energy, and memory demands associated with GAN training, as well as the subsequent U-net training with an augmented dataset. As for the benefits, we can analyze the improvements in segmentation performance achieved through synthetic data augmentation.

The GAN model was employed to generate 14,000 synthetic images split into fourteen batches. All images were then split into their corresponding FLAIR images and glioma segmentation masks. Upon separation, generated flair images matched well with extracted glioma masks. Furthermore, the location and morphology of tumor regions within healthy brain tissue varied, a desired outcome. They were then employed across fourteen rounds to augment real images and investigate the impact of synthetic data

augmentation on the efficacy of the U-net model trained for automated glioma segmentation.

The U-nets training rounds, the quantity of added synthetic images, iterations per round, overall training period, and time per epoch are detailed in Tables 4.6 and 4.7 for U-nets trained with and without geometric augmentation techniques. Investigations revealed that the correlation between the number of training iterations and the quantity of added synthetic images was insignificant, with correlation coefficients of -0.201 and a p-value of 0.472 without geometric augmentation, and correlation coefficients of -0.426 and a p-value of 0.113 with geometric augmentation. However, that is not the case when it comes to the overall training period and time per epoch. The analysis of the former produced a correlation coefficient of 0.776 and a p-value of 6.76×10^{-4} without geometric augmentation and a correlation coefficient of 0.937 and a p-value of 2.69×10^{-7} with geometric augmentation. The analysis of the latter resulted in a correlation coefficient of 1.0 and a p-value of 4.65×10^{-24} without geometric augmentation and a correlation coefficient of 1.0 and a p-value of 1.83×10^{-29} with geometric augmentation. Both analyses demonstrated significant correlations with the rounds of synthetic data augmentation. This is reasonable considering that the increase in training data leads to an increase in training batches and, consequently, more time for model parameter optimization.

Overall, the results indicate that while training takes longer with the addition of synthetic data, the U-net achieves its best performance within similar numbers of iterations. However, this does not present the complete story.

The training progression of U-nets without and with geometric augmentation, using varying levels of synthetic image augmentation are illustrated in Figures 4.7 and 4.8. Both training and validation set losses during model training are represented in Figure 4.6. The convergence of training set losses accelerates with the increased usage of synthetic images for augmentation. Upon analyzing the learning curves, the rate of loss change was evaluated, showing a significant correlation between the initial convergence rate of training loss and the number of synthetic images used, with a correlation

coefficient of -0.989 and a p-value of 2.86×10^{-12} without geometric augmentation. Similar behavior was observed with geometric augmentation, with a correlation coefficient of -0.991 and a p-value of 9.18×10^{-13}.

Table 4.6 U-nets Training without Geometric Augmentation

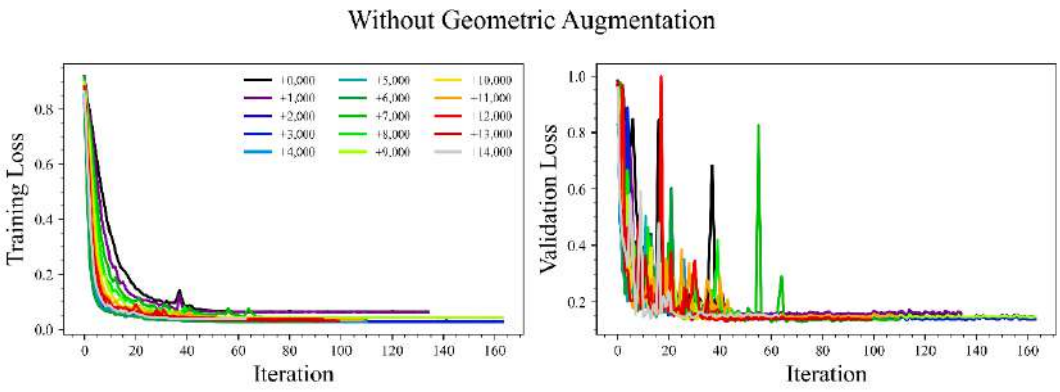
Round	Real + Added Synthetic Images	Iterations	Training Time	Minutes Per Iteration
1	3,751	135	03:05:04	1.37
2	4,751	111	03:11:31	1.73
3	5,751	86	02:57:59	2.07
4	6,751	164	06:33:49	2.4
5	7,751	99	04:32:29	2.75
6	8,751	110	05:46:42	3.15
7	9,751	94	05:27:24	3.48
8	10,751	100	06:26:45	3.87
9	11,751	64	04:29:09	4.21
10	12,751	164	12:24:10	4.54
11	13,751	164	13:23:24	4.9
12	14,751	111	09:43:18	5.25
13	15,751	77	07:12:53	5.62
14	16,751	94	09:21:16	5.97

A comparable correlation was observed between the initial convergence rate of validation loss and the number of synthetic data, with a correlation coefficient of -0.67 and a p-value of 0.00629 without geometric augmentation, and a correlation coefficient of -0.943 and a p-value of 1.4×10^{-7} with geometric augmentation (Figure 4.8). However, by the second training iteration, the correlation between the convergence rates of validation loss and the amount of synthetic data diminished - (correlation coefficient of -0.228 and p-value of 0.413 without geometric augmentation, and correlation coefficient of 0.369 and p-value of 0.176 with geometric augmentation).

Table 4.7 U-nets Training with Geometric Augmentation

Round	Real + Added Synthetic Images	Iterations	Training Time	Minutes Per Iteration
1	3,751	140	02:24:02	1.03
2	4,751	111	02:31:44	1.37
3	5,751	111	03:10:33	1.72
4	6,751	111	03:48:34	2.06
5	7,751	78	03:08:12	2.41
6	8,751	114	05:14:30	2.76
7	9,751	98	05:04:03	3.1
8	10,751	87	05:00:18	3.45
9	11,751	125	07:51:41	3.77
10	12,751	95	06:31:45	4.12
11	13,751	97	07:12:02	4.45
12	14,751	97	07:48:23	4.83
13	15,751	123	10:32:51	5.15
14	16,751	89	08:10:20	5.51

The combined results indicate that initially U-net optimization benefits from synthetic data augmentation. However, optimizing its weights requires similar numbers of iterations through the training data before the early stopping condition is met. These results make it challenging to justify that the advantages of synthetic data augmentation outweigh the added costs.



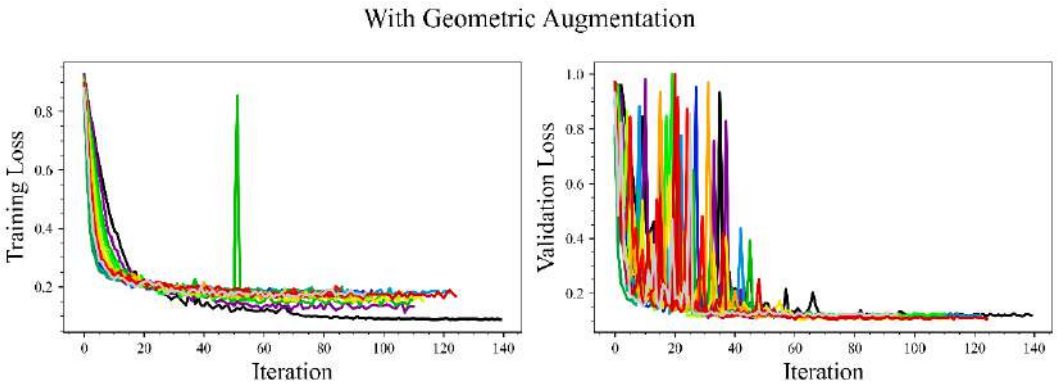


Figure 4.6 Learning Curves: Without (Top) and with Geometric Augmentation (Bottom)

Source: Researcher, 2024

Geometric augmentation involving flipping, rotation, scaling, and color adjustments offers an advantage in enhancing the model's ability to adapt better to varied inputs, eliminating the necessity for the extra computational burden from training and employing a GAN to produce synthetic images.

Figure 4.7 displays dice coefficients for U-nets in 14 rounds of training with increments in synthetic image augmentation without geometric augmentation. It was found that augmenting the training set with synthetic images had no effect on its performance. An absence of correlation was evident across all sets—training, validation, and test observed through low correlation coefficients and p-values that are not statistically significant.

Similar outcomes are seen upon incorporating geometric augmentation. The relationship between segmentation performance and the quantity of synthetic data remains statistically nonsignificant across all datasets. Nevertheless, the geometric augmentation enhanced model generalization and contributed to reducing variations in performance among the training, validation, and test sets present in Figure 4.8, suggesting a more consistent performance of the model across diverse datasets.

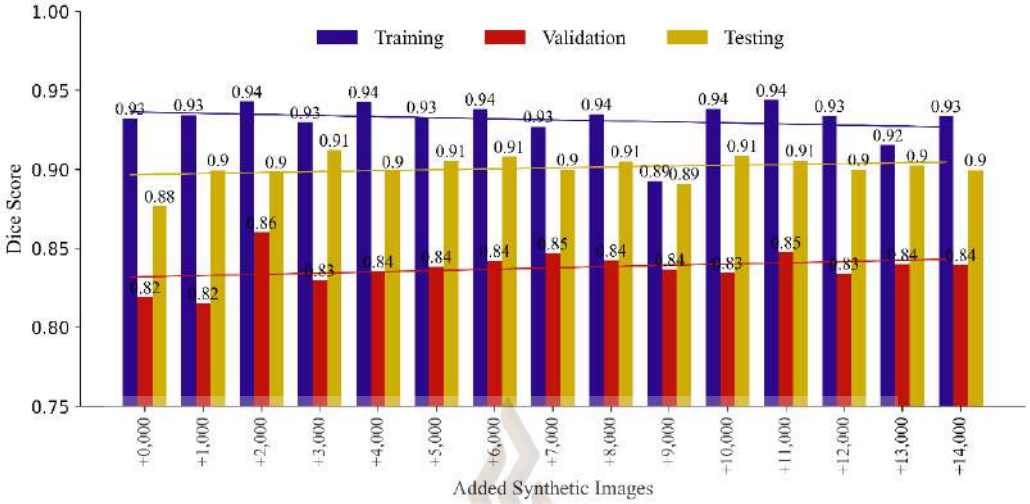


Figure 4.7 Dice coefficients for U-nets training without geometric augmentation

Source: Researcher, 2024

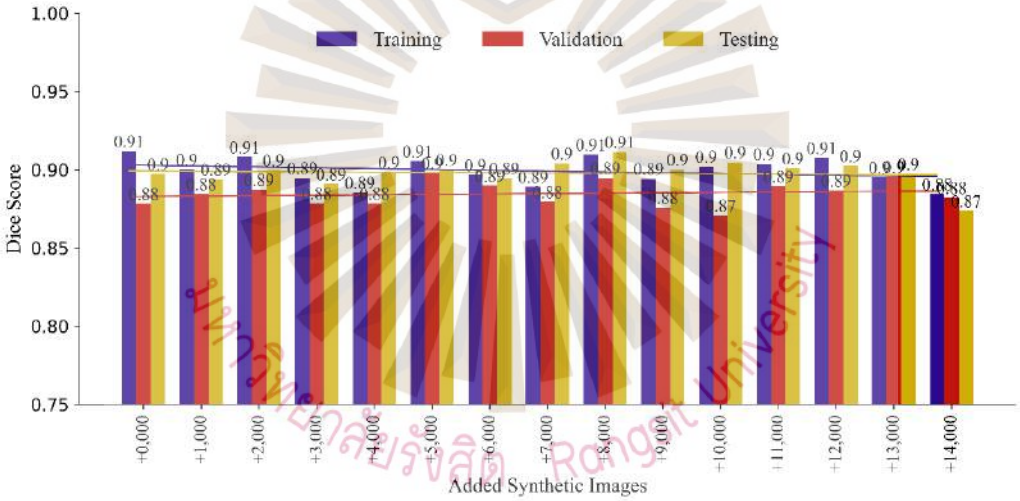


Figure 4.8 Dice coefficients for U-nets training with geometric augmentation

Source: Researcher, 2024

The U-nets model was also trained using only real images and corresponding segmentation masks. The Dice coefficients for the training, validation, and testing sets were 0.932, 0.819, and 0.877, respectively. We then repeated the process, introducing synthetic image augmentation, resulting in Dice coefficients of 0.944, 0.860, and 0.912, indicating progress in the validation and test set scores. When we applied only geometric augmentation, the Dice coefficients were 0.912, 0.878, and 0.897 for the training,

validation, and testing sets. Finally, we applied synthetic image and geometric augmentations, resulting in Dice coefficients of 0.910, 0.898, and 0.911. Notably, the standard deviation of Dice coefficients for U-nets trained without geometric augmentation was 0.0406, while it was 0.0104 for those trained with geometric augmentation, indicating improvement in generalization.

The results align with other studies that used progressively growing GAN and a combination of GANs and found slight enhancements in brain tumor segmentation using GAN-generated synthetic data (Foroozandeh & Eklund, 2020; Larsson et al., 2022). Although an ensemble of GANs resulted in improved U-net segmentations compared to no data augmentation, with an average Dice coefficient of 0.735 versus 0.729, the expense associated with this approach makes it challenging to justify, particularly considering the relatively modest enhancement observed.

Chapter 5

Conclusion and Recommendations

5.1 Conclusion

From the first study conducted in this research, it was found that most of the synthetic images generated by the PGGAN resembled the real images to a high degree. This was supported by quantitative metrics scores with FID of 42.4, IS of 3.0, and P and R of 0.04 and 0.32, respectively. Although these scores approached state-of-the-art back then, there was a large margin for improvement in the GAN performance. We also noticed that the FID metric produced the most accurate representation of the GAN's performance.

The next study indicated that a pre-trained Inception V3 model is favored over a randomly initialized weights model for medical image feature vector extraction. The pre-trained model's ability to produce feature vectors that are separable for the same or different image types makes it valuable for enhancing the evaluation of synthetic medical images and improving the reliability of comparisons across studies.

Building upon the findings from the first two studies, in the last study, a different GAN architecture was employed, specifically StyleGAN2-ADA, and a pre-trained Inception V3 model was used during FID computations. The synthetic images generated by the GAN model were impressive and practically indistinguishable from real images, with an FID score of 14.39. Then, 14,000 synthetic images and their masks were generated and used to augment a segmentation model, U-nets, during training. However, this did not significantly improve the segmentation performance when evaluated using validation and test sets yielding Dice coefficient scores of +0.0409 for the first and +0.0355 for the latter. This could be due to the data distributions in the training, validation and test sets partially not overlapping. Meanwhile, geometric augmentation

improved the generalization with standard deviation among training, validation, and test sets of 0.04 without geometric augmentation and 0.01 with geometric augmentation. These findings suggest that the costs associated with GAN training may be challenging to justify, given the little enhancements observed in augmenting the segmentation model with synthetic data. Nonetheless, it remains plausible that synthetic data augmentation could yield more improvement in segmentation performance under different circumstances. Moreover, realistic synthetic data holds promise in vital domains such as medical data anonymization and safeguarding patient privacy. Ultimately, we hope that the findings of this work contribute in future research aimed at refining and fine-tuning GAN architectures specifically for medical image generation.

5.2 Recommendations

While the incorporation of synthetic images to augment training data for U-net models in brain tumor segmentation did not result in significant performance improvements, it's crucial to recognize that results may vary depending on different scenarios. Further investigation is recommended, including the exploration of alternative deep learning architectures. Additionally, exploring other potential applications of Generative Adversarial Networks (GANs) to address data scarcity issues, such as anonymization, is suggested.

5.3 Research Outputs

List of publications derived from conducted research:

5.3.1 Asadi, F., & O'Reilly, J. A. (2021). Artificial Computed Tomography Images with Progressively Growing Generative Adversarial Network. *2021 13th Biomedical Engineering International Conference (BMEiCON),* 1-5. https://doi.org/10.1109 /BMEiCON53485.2021.9745251

5.3.2 O'Reilly, J. A., & Asadi, F. (2021). Pre-trained vs. random weights for calculating fréchet inception distance in medical imaging. *2021 13th Biomedical Engineering International Conference (BMEiCON)*, 1-4.https://doi.org/10.1109/BMEiCON53485.2021.9745214

5.3.3 O'Reilly, J. A., & Asadi, F. (2022). Identifying Obviously Artificial Medical Images Produced by a Generative Adversarial Network. *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 430-433. https://doi.org/10.1109/EMBC48229.2022.9871217

5.3.4 O'Reilly, J. A., Wehrman, J., Carey, A., Bedwin, J., Hourn, T., Asadi, F., & Sowman, P. F. (2023). Neural correlates of face perception modeled with a convolutional recurrent neural network. *Journal of Neural Engineering*, *20*(2), 026028. https://doi.org/10.1088/1741-2552/acc35b

5.3.5 Asadi, F., Angsuwatanakul, T., & O'Reilly, J. A. (2024). Evaluating synthetic neuroimaging data augmentation for automatic brain tumour segmentation with a deep fully-convolutional network. *IBRO Neuroscience Reports*, 16, 57-66. https://doi.org/10.1016/j.ibneur.2023.12.002

References

Aggarwal, C. C. (2018). *Neural networks and deep learning*. Berlin: Springer.

Albahra, S., Gorbett, T., Robertson, S., D'Aleo, G., Kumar, S. V. S., Ockunzzi, S., ... & Rashidi, H. H. (2023). Artificial intelligence and machine learning overview in pathology & laboratory medicine: A general review of data preprocessing and basic supervised concepts. *Seminars in Diagnostic Pathology*, *40*(2), 71-87. https://doi.org/10.1053/j.semdp.2023.02.002

Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., Badreldin, H. A. J. B. m. e. (2023). Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Medical Education, 23*(1), 689. https://doi.org/10.1186/s12909-023-04698-z

Anaya-Isaza, A., Mera-Jiménez, L., & Zequera-Diaz, M. (2021). An overview of deep learning in medical imaging. *Informatics in Medicine Unlocked*, *26*, 100723. https://doi.org/10.1016/j.imu.2021.100723

Asadi, F., Angsuwatanakul, T., & O'Reilly, J. A. (2024). Evaluating synthetic neuroimaging data augmentation for automatic brain tumour segmentation with a deep fully-convolutional network. *IBRO Neuroscience Reports*, *16*, 57-66. https://doi.org/10.1016/j.ibneur.2023.12.002

Asadi, F., & O'Reilly, J. A. (2021). Artificial Computed Tomography Images with Progressively Growing Generative Adversarial Network. *2021 13th Biomedical Engineering International Conference (BMEiCON)*, 1-5.https://doi.org/10.1109/BMEiCON53485.2021.9745251

Basaran, B. D., Qiao, M., Matthews, P. M., & Bai, W. (2022). Subject-specific lesion generation and pseudo-healthy synthesis for multiple sclerosis brain images. *International Workshop on Simulation and Synthesis in Medical Imaging*, 13570. https://doi.org/10.1007/978-3-031-16980-9_1

Baur, C., Albarqouni, S., & Navab, N. (2018). Generating Highly Realistic Images of Skin Lesions with GANs. *Proceedings of OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis, 5*, 260-267. https://doi.org/10.1007/978-3-030-01201-4_28

References (Cont.)

Brock, A., Donahue, J., & Simonyan, K. (2018). Large Scale GAN Training for High Fidelity Natural Image Synthesis. *arXiv, 1809.11096*. https://doi.org/10.48550/arXiv.1809.11096

Buda, M., Saha, A., & Mazurowski, M. A. (2019). Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. *Computers in biology and medicine*, *109*, 218-225. https://doi.org/10.1016/j.compbiomed.2019.05.002

Carver, E. N., Dai, Z., Liang, E., Snyder, J., & Wen, N. (2021). Improvement of multiparametric MR image segmentation by augmenting the data with generative adversarial networks for glioma patients. *Frontiers in Computational Neuroscience*, *14*, 495075. https://doi.org/10.3389/fncom.2020.495075

Cha, K. H., Petrick, N., Pezeshk, A., Graff, C. G., Sharma, D., Badal, A., & Sahiner, B. (2020). Evaluation of data augmentation via synthetic images for improved breast mass detection on mammograms using deep learning. *Journal of Medical Imaging, 7*(1), 012703-012703. https://doi.org/10.1117/1.JMI.7.1.012703

Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F., & Mahmood, F. (2021). Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering, 5*(6), 493-497. https://doi.org/10.1038/s41551021-00751-8

Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., & Haworth, A. (2021). A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology, 65*(5), 545-563. https://doi.org/https://doi.org/10.1111/1754-9485.13261

Chuquicusma, M. J., Hussein, S., Burt, J., & Bagci, U. (2018). How to fool radiologists with generative adversarial networks? A visual turing test for lung cancer diagnosis. *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, 240-244. https://doi.org/10.1109/ISBI.2018.8363564

References (Cont.)

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., ... & Prior, F. (2013). The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *Journal of digital imaging*, *26*, 1045-1057. https://doi.org/10.1007/s10278-013-9622-7

Dang, K., Vo, T., Ngo, L., & Ha, H. (2022). A deep learning framework integrating MRI image preprocessing methods for brain tumor segmentation and classification. *IBRO Neuroscience Reports, 13*, 523-532. https://doi.org/10.1016/j.ibneur.2022.10.014

Dash, A., Ye, J., & Wang, G. (2024). A Review of Generative Adversarial Networks (GANs) and Its Applications in a Wide Variety of Disciplines, *Medical to Remote Sensing. IEEE Access*, *12*, 18330-18357. https://doi.org/10.1109/ACCESS.2023.3346273

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *IEEE conference on computer vision and pattern recognition,* 248-255. 10.1109/CVPR.2009.5206848

Denton, E. L., Chintala, S., & Fergus, R. (2015). Deep generative image models using a laplacian pyramid of adversarial networks. *Advances in neural information processing systems, 28.* https://doi.org/10.48550/arXiv.1506.05751

Diaz, O., Kushibar, K., Osuala, R., Linardos, A., Garrucho, L., Igual, L., ... & Lekadir, K. (2021). Data preparation for artificial intelligence in medical imaging: A comprehensive guide to open-access platforms and tools. *Physica medica, 83*, 25-37. https://doi.org/10.1016/j.ejmp.2021.02.007

Eilertsen, G., Tsirikoglou, A., Lundström, C., & Unger, J. (2021). Ensembles of GANs for synthetic training data generation. *arXiv preprint,* 2104.11797. https://doi.org/10.48550/arXiv.2104.11797

Fetty, L., Bylund, M., Kuess, P., Heilemann, G., Nyholm, T., Georg, D., & Löfstedt, T. (2020). Latent space manipulation for high-resolution medical image synthesis via the StyleGAN. *Zeitschrift für Medizinische Physik*, *30*(4), 305-314. https://doi.org/10.1016/j.zemedi.2020.05.001

References (Cont.)

Foroozandeh, M., & Eklund, A. (2020). Synthesizing brain tumor images and annotations by combining progressive growing GAN and SPADE. *arXiv, 2009.05946*. https://doi.org/10.48550/arXiv.2009.05946

Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018). Synthetic data augmentation using GAN for improved liver lesion classification. *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, 289-293. https://doi.org/10.1109/ISBI.2018.8363576

Goceri, E. (2023). *Medical image data augmentation: techniques, comparisons and interpretations.* Berlin: Springer.

Gonçalves, B. (2023). *Minds and Machines.* Berlin: Springer.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems,* 27. Retrieved from https://ui.adsabs.harvard.edu/abs/2014arXiv1406.2661G

Guha, A., Grewal, D., Kopalle, P. K., Haenlein, M., Schneider, M. J., Jung, H., ... & Hawkins, G. (2021). How artificial intelligence will affect the future of retailing. *Journal of Retailing*, *97*(1), 28-41. https://doi.org/10.1016/j.jretai .2021.01.005

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of wasserstein gans. *Advances in neural information processing systems*, 30. https://doi.org/10.48550/arXiv.1704.00028

Guzman, A. L., & Lewis, S. C. (2020). Artificial intelligence and communication: A human–machine communication research agenda. *New media & society, 22*(1), 70-86. https://doi.org/10.1177/1461444819858691

Hamghalam, M., Wang, T., & Lei, B. (2020). High tissue contrast image synthesis via multistage attention-GAN: application to segmenting brain MR scans. *Neural Networks*, *132*, 43-52. doi: https://doi.org/10.1016/j.neunet.2020.08.014

References (Cont.)

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30. Retrieved from https://ui.adsabs.harvard.edu/abs/2017arXiv170608500H

Hong, S S., Marinescu, R., Dalca, A. V., Bonkhoff, A. K., Bretzner, M., Rost, N. S., & Golland, P. (2021). 3D-StyleGAN: A style-based generative adversarial network for generative modeling of three-dimensional medical images. *Deep Generative Models, and Data Augmentation, Labelling, and Imperfections Proceedings*, 1, 24-34. https://doi.org/10.1007/978-3-030-88210-5_3

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., & Maier-Hein, K. H. (2019). No new-net. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. *4th International Workshop, BrainLes 2018*, 234-244. https://doi.org/10.48550/arXiv.1809.10483

Jaiswal, S. (2023). Artificial Intelligence And Its Application In Media, Communication And Entertainment. *Multi-Disciplinary Journal*. Retrieved from http://210.212.169.38/xmlui/handle/123456789/12251

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. Retrieved from https://ui.adsabs.harvard.edu/abs/2017arXiv171010196K

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., & Aila, T. (2020). Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33, 12104-12114. Retrieved from https://ui.adsabs.harvard.edu/abs/2020arXiv200606676K

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8110-8119. Retrieved from https://ui.adsabs.harvard.edu/abs/2019arXiv191204958K

References (Cont.)

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., ... & Zhang, K. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, *172*(5), 1122-1131. https://doi.org/10.1016/j.cell.2018.02.010

Khosla, A., Jayadevaprakash, N., Yao, B., & Li, F. F. (2011, June). Novel dataset for fine-grained image categorization: Stanford dogs. *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, 2, 1.

Kora Venu, Venu, S., & Ravula, S. (2020). Evaluation of deep convolutional generative adversarial networks for data augmentation of chest x-ray images. *Future Internet, 13*(1), 8. https://doi.org/10.3390/fi13010008

Korkinof, D., Rijken, T., O'Neill, M., Yearsley, J., Harvey, H., & Glocker, B. (2018). High-resolution mammogram synthesis using progressive generative adversarial networks. *arXiv preprint arXiv,* 1807.03401. Retrieved from https://ui.adsabs.harvard.edu/abs/2018arXiv180703401K

Kossen, T., Hirzel, M. A., Madai, V. I., Boenisch, F., Hennemuth, A., Hildebrand, K., ... & Frey, D. (2022). Toward sharing brain images: Differentially private TOF-MRA images with segmentation labels using generative adversarial networks. *Frontiers in artificial intelligence,* 5, 813842. https://doi.org/10.3389/frai.2022.813842

Kossen, T., Subramaniam, P., Madai, V. I., Hennemuth, A., Hildebrand, K., Hilbert, A., ... & Frey, D. (2021). Synthesizing anonymized and labeled TOF-MRA patches for brain vessel segmentation using generative adversarial networks. *Computers in biology and medicine,* 131, 104254. https://doi.org/10.1016/j.compbiomed.2021.104254

Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., & Aila, T. (2019). Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems,* 32. Retrieved from https://ui.adsabs.harvard.edu/abs/2019arXiv190406991K

References (Cont.)

Larsson, M., Akbar, M. U., & Eklund, A. (2022). Does an ensemble of GANs lead to better performance when training segmentation networks with synthetic images?. *arXiv preprint arXiv*, 2211.04086. https://doi.org/10.48550/arXiv.2211.04086

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86(11)*, 2278-2324. https://doi.org/10.1109/5.726791

Mallappallil, M., Sabu, J., Gruessner, A., & Salifu, M. (2020). A review of big data and medical research. *SAGE open medicine*, 8, 2050312120934839. https://doi.org/10.1177/2050312120934839

McKinley, R., Meier, R., & Wiest, R. (2019). Ensembles of densely-connected CNNs with label-uncertainty for brain tumor segmentation. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018*, 4, 456-465. https://doi.org/10.1007/978-3-030-11726-940

Meskó B, Görög M. A short guide for medical professionals in the era of artificial intelligence. (2020). *NPJ Digit Med*, 3-126. doi: 10.1038/s41746-020-00333-z

Mondal, B. (2020) *Artificial Intelligence: State of the Art*. Berlin: Springer.

Morales, E. F., & Escalante, H. J. (2022). *Biosignal Processing and Classification*. Massachusetts: Academic Press.

Myronenko, A. (2019). 3D MRI brain tumor segmentation using autoencoder regularization. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018*, 311-320. https://doi.org/10.1007/978-3-030-11726-9_28

Naeem, M. F., Oh, S. J., Uh, Y., Choi, Y., & Yoo, J. (2020). Reliable fidelity and diversity metrics for generative models. *International Conference on Machine Learning,* 7176-7185. https://doi.org/10.48550/arXiv.2002.09797

Nalepa, J., Marcinkiewicz, M., & Kawulok, M. (2019). Data augmentation for brain-tumor segmentation: a review. *Frontiers in computational neuroscience*, 13, 83. https://doi.org/10.3389/fncom.2019.00083

References (Cont.)

Nalepa, J., Mrukwa, G., Piechaczek, S., Lorenzo, P. R., Marcinkiewicz, M., Bobek-Billewicz, B., ... & Hayball, M. P. (2019, September). Data augmentation via image registration. *2019 IEEE International Conference on Image Processing (ICIP)*, 4250-4254. https://doi.org/10.1109/ICIP.2019.8803423

Nti, I. K., Adekoya, A. F., Weyori, B. A., & Nyarko-Boateng, O. (2022). Applications of artificial intelligence in engineering and manufacturing: A systematic review. *Journal of Intelligent Manufacturing*, *33*(6), 1581-1601. https://doi .org/10.1007/s10845-021-01771-6

Nunn, E. J., Khadivi, P., & Samavi, S. (2021). Compound frechet inception distance for quality assessment of gan created images. *arXiv preprint*, arXiv:2106 .08575. https://doi.org/10.48550/arXiv.2106.08575

O'Reilly, J. A., & Asadi, F. (2021). Pre-trained vs. random weights for calculating fréchet inception distance in medical imaging. *2021 13th Biomedical Engineering International Conference (BMEiCON)*, 1-4. https://doi.org/10 .1109/BMEiCON53485.2021.9745214

O'Reilly, J. A., & Asadi, F. (2022). Identifying Obviously Artificial Medical Images Produced by a Generative Adversarial Network. *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 430-433. https://doi.org/10.1109/EMBC48229.2022.9871217

O'Reilly, J. A. (2022). Recurrent neural network model of human event-related potentials in response to intensity oddball stimulation. *Neuroscience*, 504, 63-74. https://doi.org/10.1016/j.neuroscience.2022.10.004

Ostrom, Q. T., Cioffi, G., Gittleman, H., Patil, N., Waite, K., Kruchko, C., & Barnholtz-Sloan, J. S. (2019). CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2012–2016. *Neuro-oncology*, 21, 1-100. https://doi.org/10.1093/neuonc/noz150

Park, H. Y., Bae, H. J., Hong, G. S., Kim, M., Yun, J., Park, S., ... & Kim, N. (2021). Realistic high-resolution body computed tomography image synthesis by using progressive growing generative adversarial network: visual turing test. *JMIR medical informatics*, *9*(3), e23328. https://doi.org/10.2196/23328

References (Cont.)

Perone, C. S., & Cohen-Adad, J. (2019). Promises and limitations of deep learning for medical image segmentation. *Journal of Medical Artificial Intelligence*, 2. https://doi.org/10.21037/jmai.2019.01.01

Piccialli, F., Di Somma, V., Giampaolo, F., Cuomo, S., & Fortino, G. (2021). A survey on deep learning in medicine: Why, how and when?. *Information Fusion*, 66, 111-137. https://doi.org/10.1016/j.inffus.2020.09.006

Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., & Meriaudeau, F. (2018). Indian diabetic retinopathy image dataset (IDRiD): a database for diabetic retinopathy screening research. *Data*, *3*(3), 25. https://doi.org/10.3390/data3030025

Price, W., & Nicholson, I. I. (2019). Artificial intelligence in the medical system: four roles for potential transformation. *Yale JL & Tech.*, 21, 122. Retrieved from https://repository.law.umich.edu/articles/2257

Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention*, 9351. https://doi.org/10.1007/978-3-319-24574-4_28

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach*. London: Pearson.

Salehi, P., Chalechale, A., & Taghizadeh, M. (2020). Generative adversarial networks (GANs): An overview of theoretical model, evaluation metrics, and recent developments. *arXiv preprint*, arXiv:2005.13178. Retrieved from https://ui.adsabs.harvard.edu/abs/2020arXiv200513178S

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training gans. *Advances in neural information processing systems*, 29. Retrieved from https://ui.adsabs.harvard.edu/abs/2016arXiv160603498S

Sandfort, V., Yan, K., Pickhardt, P. J., & Summers, R. M. (2019). Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. *Scientific reports*, *9*(1), 16884. https://doi.org/10.1038/s41598-019-52737-x

References (Cont.)

Sharifani, K., & Amini, M. (2023). Machine learning and deep learning: A review of methods and applications. *World Information Technology and Engineering Journal*, *10*(07), 3897-3904. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4458723

Shaver, M. M., Kohanteb, P. A., Chiou, C., Bardis, M. D., Chantaduly, C., Bota, D., ... & Chang, P. D. (2019). Optimizing neuro-oncology imaging: a review of deep learning approaches for glioma imaging. *Cancers*, *11*(6), 829. https://doi.org/10.3390/cancers11060829

Shin, H. C., Tenenholtz, N. A., Rogers, J. K., Schwarz, C. G., Senjem, M. L., Gunter, J. L., ... & Michalski, M. (2018). *Simulation and Synthesis in Medical Imaging*. Berlin: Springer

Shorten, C., & Khoshgoftaar, T. M. (2019). *Journal of big data*, *6*(1), 1-48. https://doi.org/10.1186/s40537-019-0197-0

Sing, C. C., Teo, T., Huang, F., Chiu, T. K., & Xing Wei, W. (2022). Secondary school students' intentions to learn AI: Testing moderation effects of readiness, social good and optimism. *Educational technology research and development*, *70*(3), 765-782. https://doi.org/10.1007/s11423-022-10111-1

Situ, Z., Teng, S., Liu, H., Luo, J., & Zhou, Q. (2021). Automated sewer defects detection using style-based generative adversarial networks and fine-tuned well-known CNN classifier. *IEEE Access*, 9, 59498-59507. https://doi.org/10.1109/ACCESS.2021.3073915

Skandarani, Y Y., Jodoin, P. M., & Lalande, A. (2023). Gans for medical image synthesis: An empirical study. *Journal of Imaging*, *9*(3), 69. Retrieved from https://ui.adsabs.harvard.edu/abs/2021arXiv210505318S

Subramaniam, P., Kossen, T., Ritter, K., Hennemuth, A., Hildebrand, K., Hilbert, A., ... & Madai, V. I. (2022). Generating 3D TOF-MRA volumes and segmentation labels using generative adversarial networks. *Medical Image Analysis*, 78, 102396. https://doi.org/10.1016/j.media.2022.102396

References (Cont.)

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818-2826. http://doi .ieeecomputersociety.org/10.1109/CVPR.2016.308

Tang, X. (2019). The role of artificial intelligence in medical imaging research. *BJR| Open*, *2*(1), 20190031. https://doi.org/10.1259/bjro.20190031

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, *9*(11). Retrieved from http://jmlr.org/papers/v9 /vandermaaten08a.html

Woodland, M., Wood, J., Anderson, B. M., Kundu, S., Lin, E., Koay, E., ... & Brock, K. K. (2022). *Evaluating the performance of StyleGAN2-ADA on medical images*, *International Workshop on Simulation and Synthesis in Medical Imaging*, 142-153. Berlin: Springer International Publishing.

Yang, J., Veeraraghavan, H., Armato III, S. G., Farahani, K., Kirby, J. S., Kalpathy□ Kramer, J., ... & Sharp, G. C. (2018). Autosegmentation for thoracic radiation treatment planning: a grand challenge at AAPM 2017. *Medical physics*, *45*(10), 4568-4581. https://doi.org/10.1002/mp.13141

Yi, X., Walia, E., & Babyn, P. (2019). Generative adversarial network in medical imaging: A review. *Medical image analysis*, 58, 101552. https://doi.org /10.1016/j.media.2019.101552

Yu, B., Zhou, L., Wang, L., Fripp, J., & Bourgeat, P. (2018). 3D cGAN based cross-modality MR image synthesis for brain tumor segmentation. *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, 626-630. https://doi.org/10.1109/ISBI.2018.8363653

Yu, K. H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature biomedical engineering*, *2*(10), 719-731. https://doi.org/10.1038 /s41551-018-0305-z

Zhang, C., & Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. *Journal of Industrial Information Integration*, 23, 100224. https://doi.org/10.1016/j.jii.2021.100224

References (Cont.)

Zhang, X., Liu, C., Ou, N., Zeng, X., Zhuo, Z., Duan, Y., ... & Ye, C. (2023). CarveMix: a simple data augmentation method for brain lesion segmentation. *Neuroimage*, 271, 120041. https://doi.org/10.1016/j.neuroimage.2023.120041

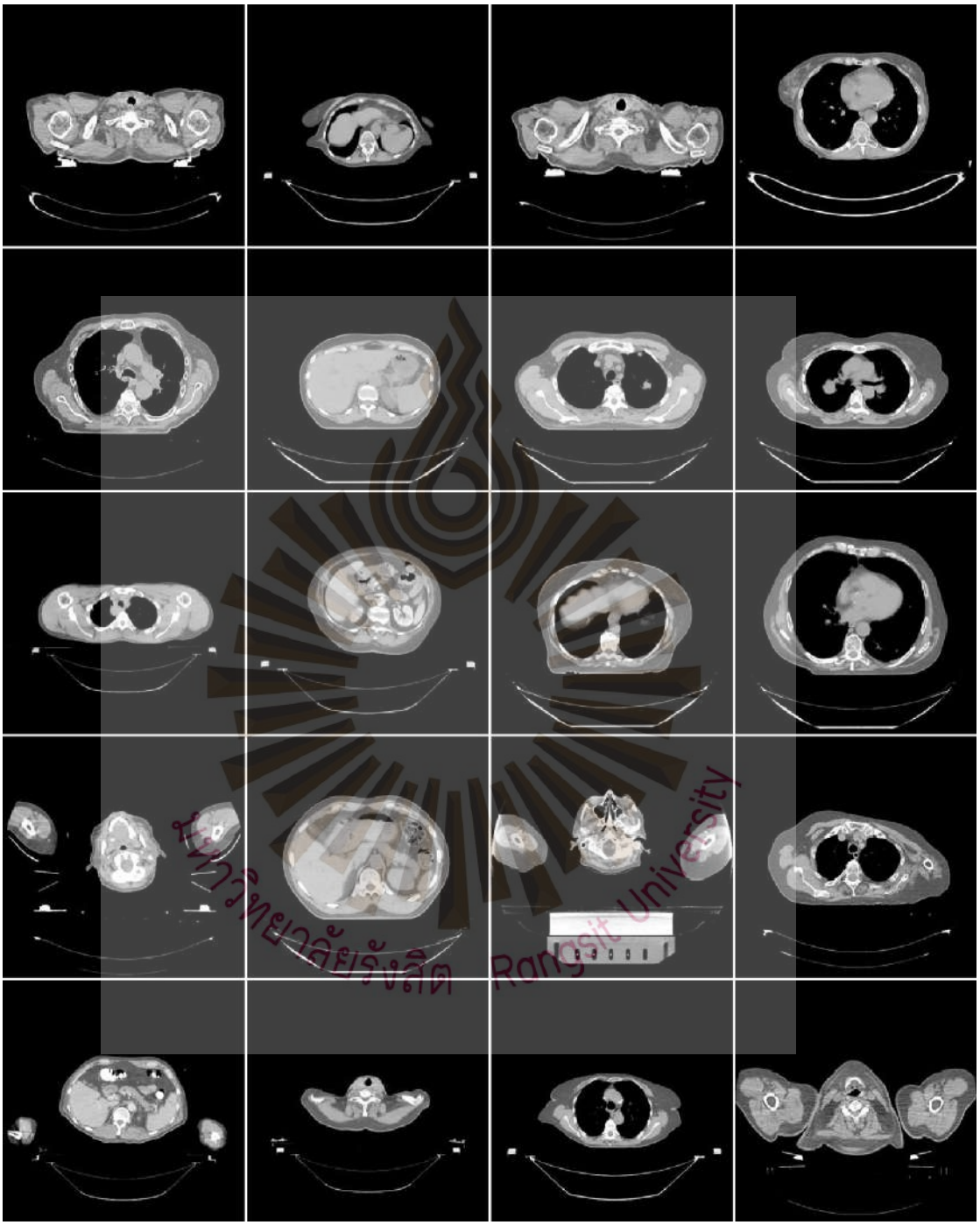
Zhou, S. K., Greenspan, H., & Shen, D. (Eds.). (2023). *Deep learning for medical image analysis*. Massachusetts: Academic Press

APPENDICES

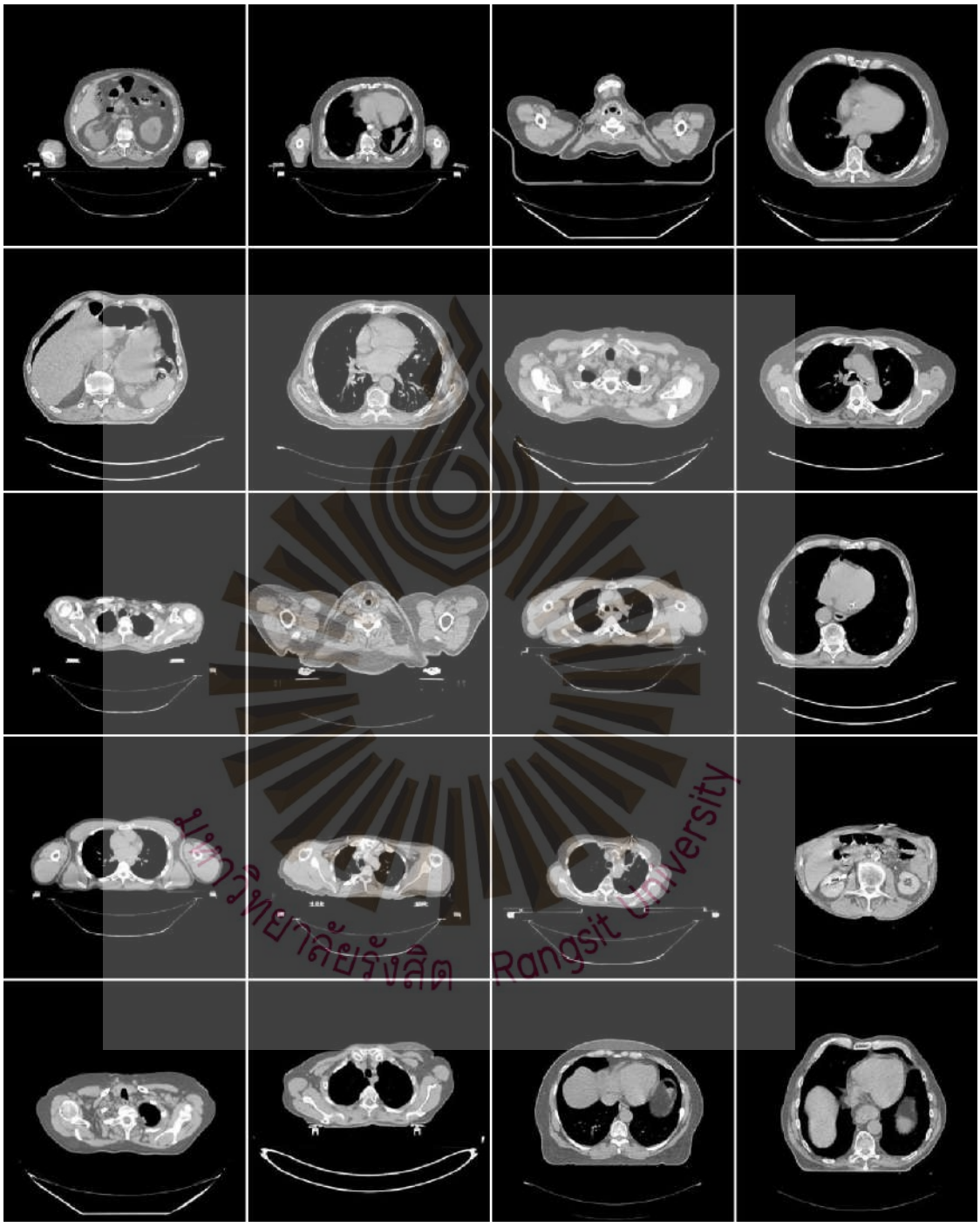
Appendix A

Study 1: Dataset, Realistic, and Unrealistic PGGAN Generated Image

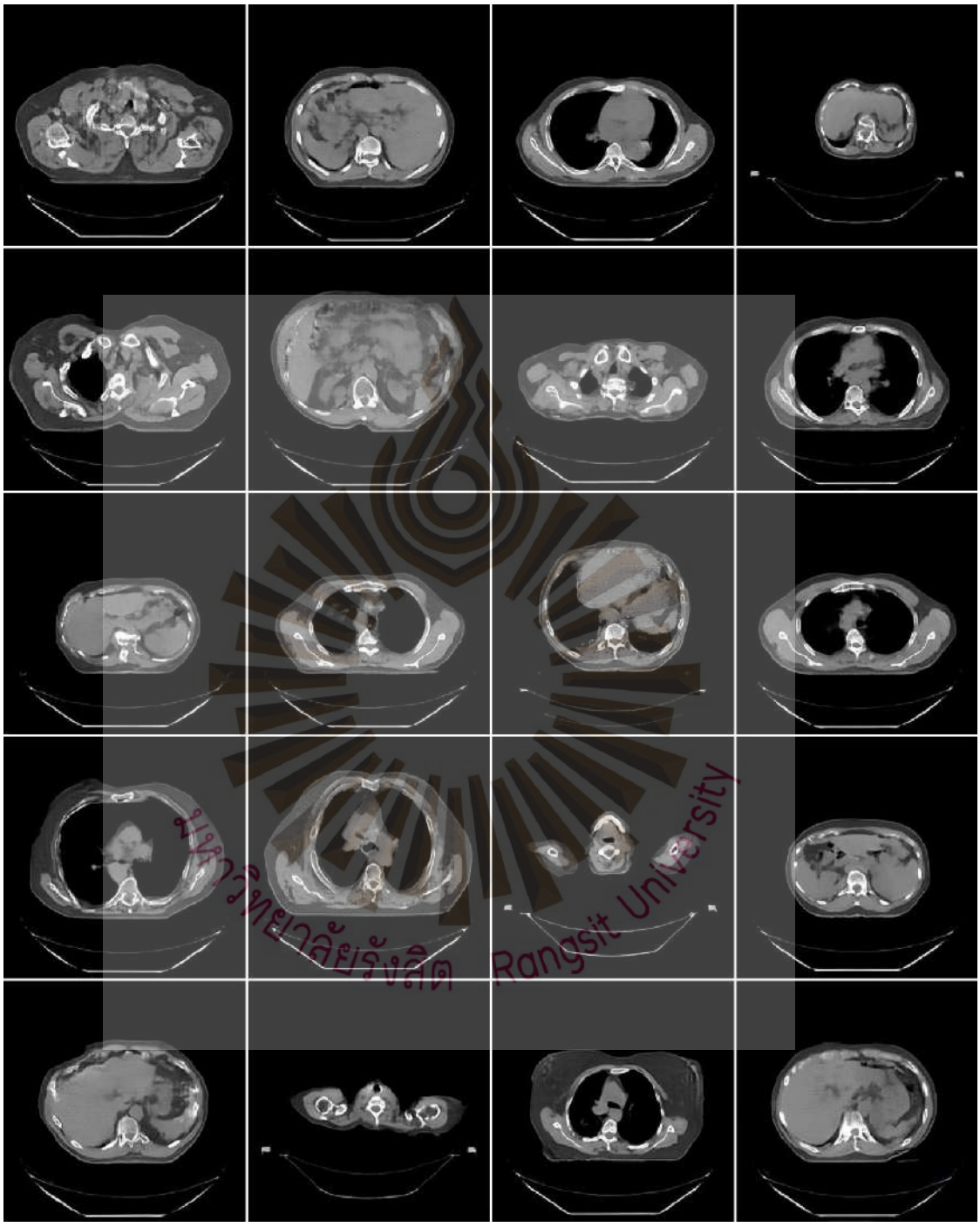
Real Images (batch 1)



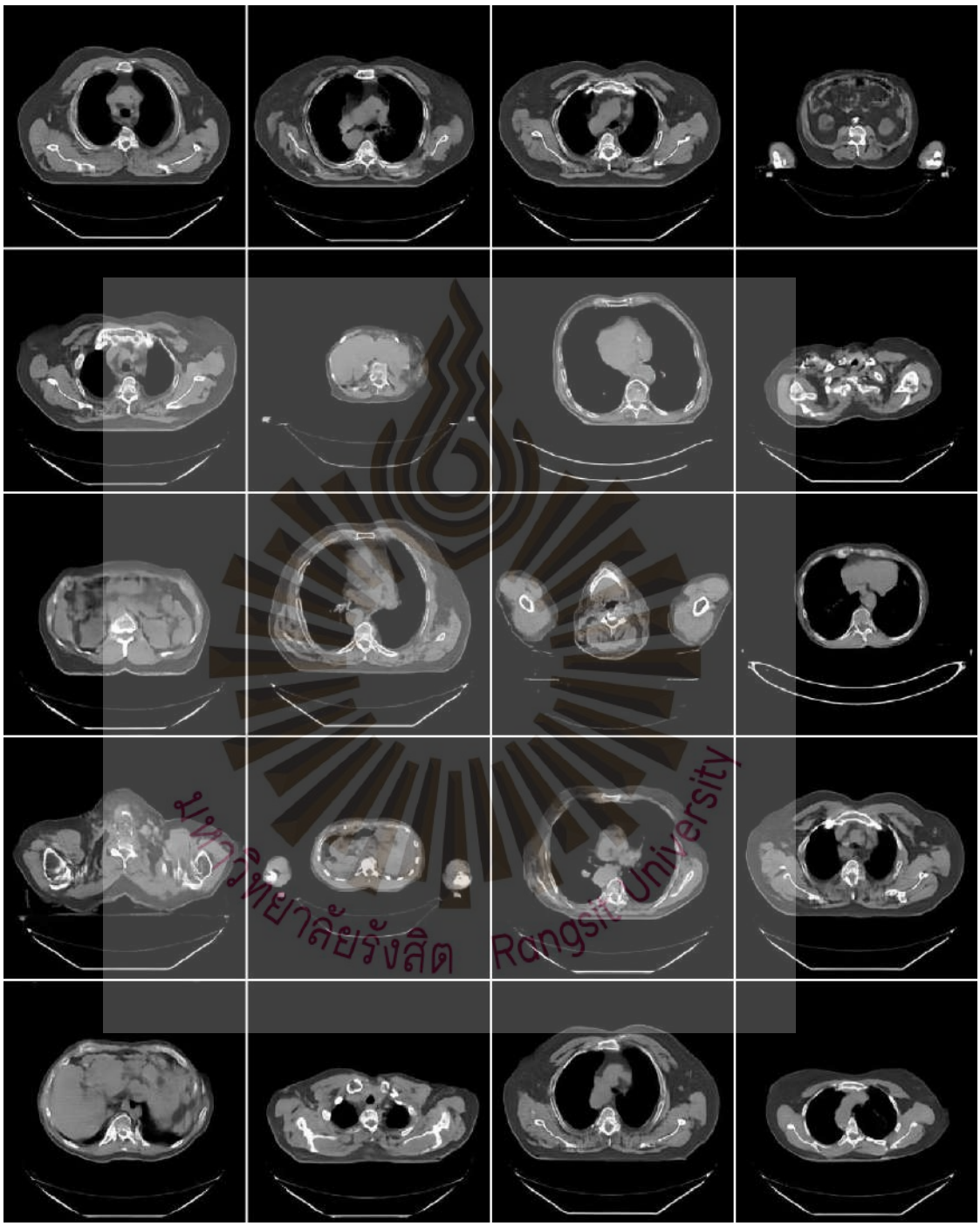
Real Images (batch 2)



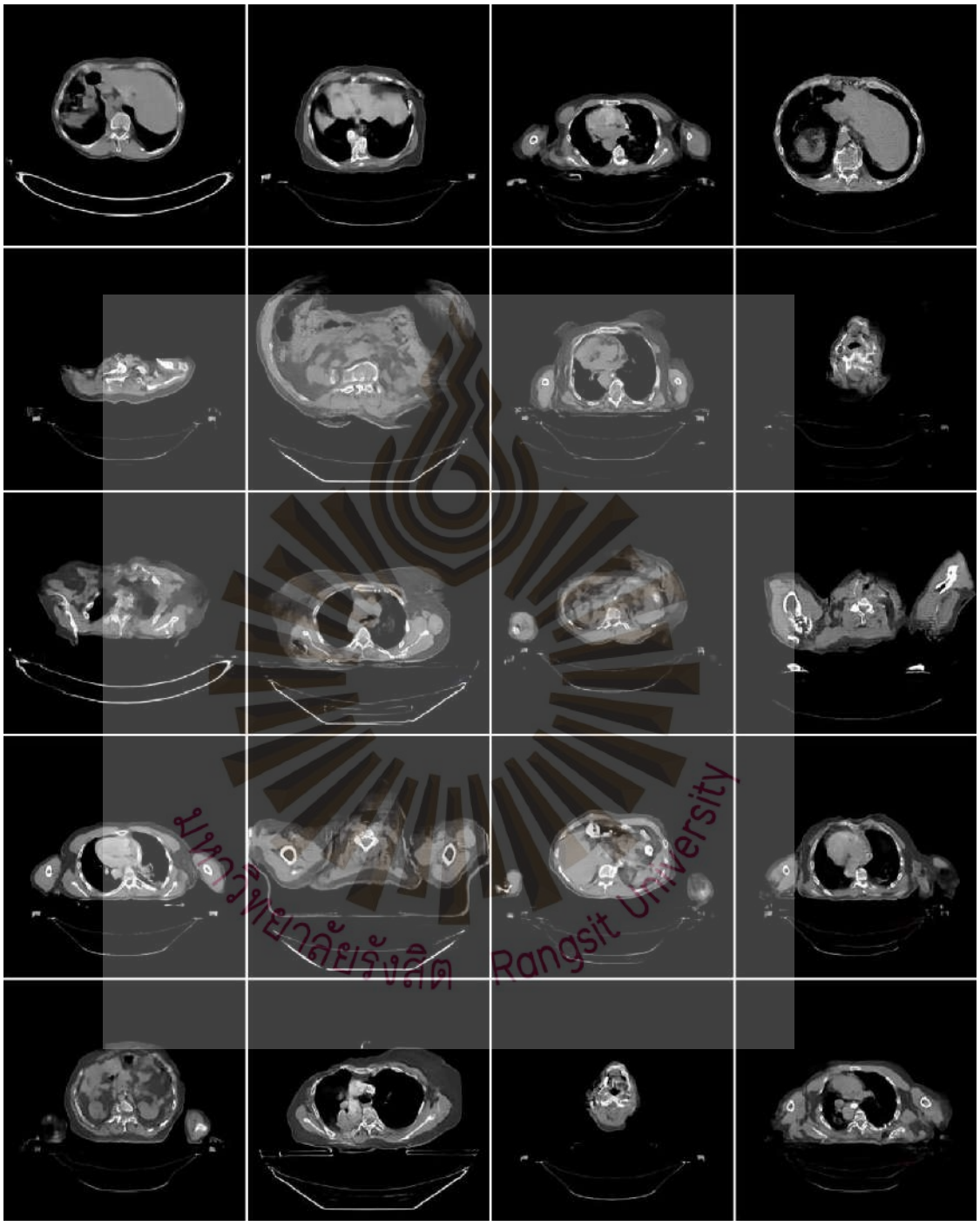
Realistic PGGAN-generated synthetic images (batch 1)



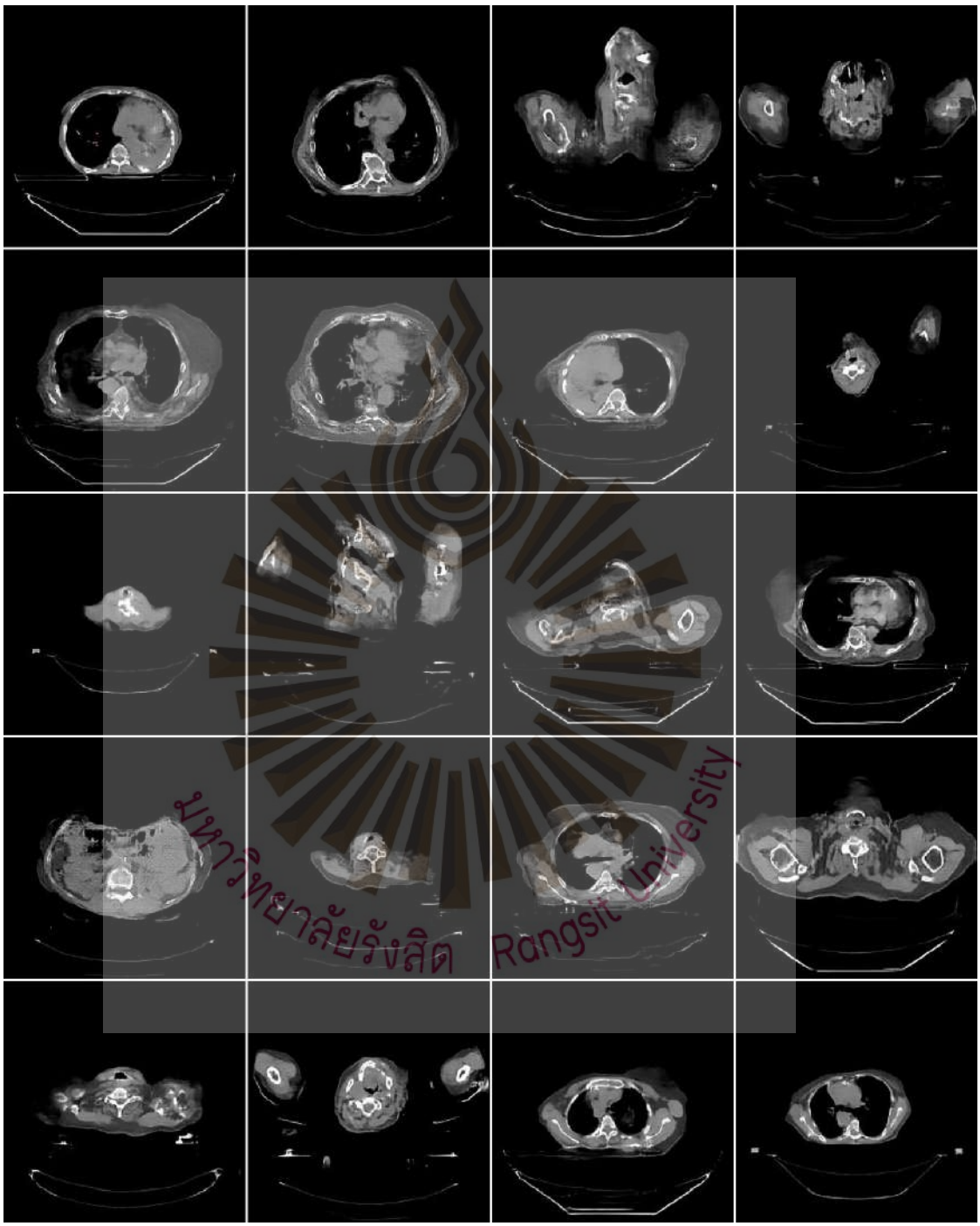
Realistic PGGAN-generated synthetic images (batch 2)



Unrealistic PGGAN-generated synthetic images (batch 1)



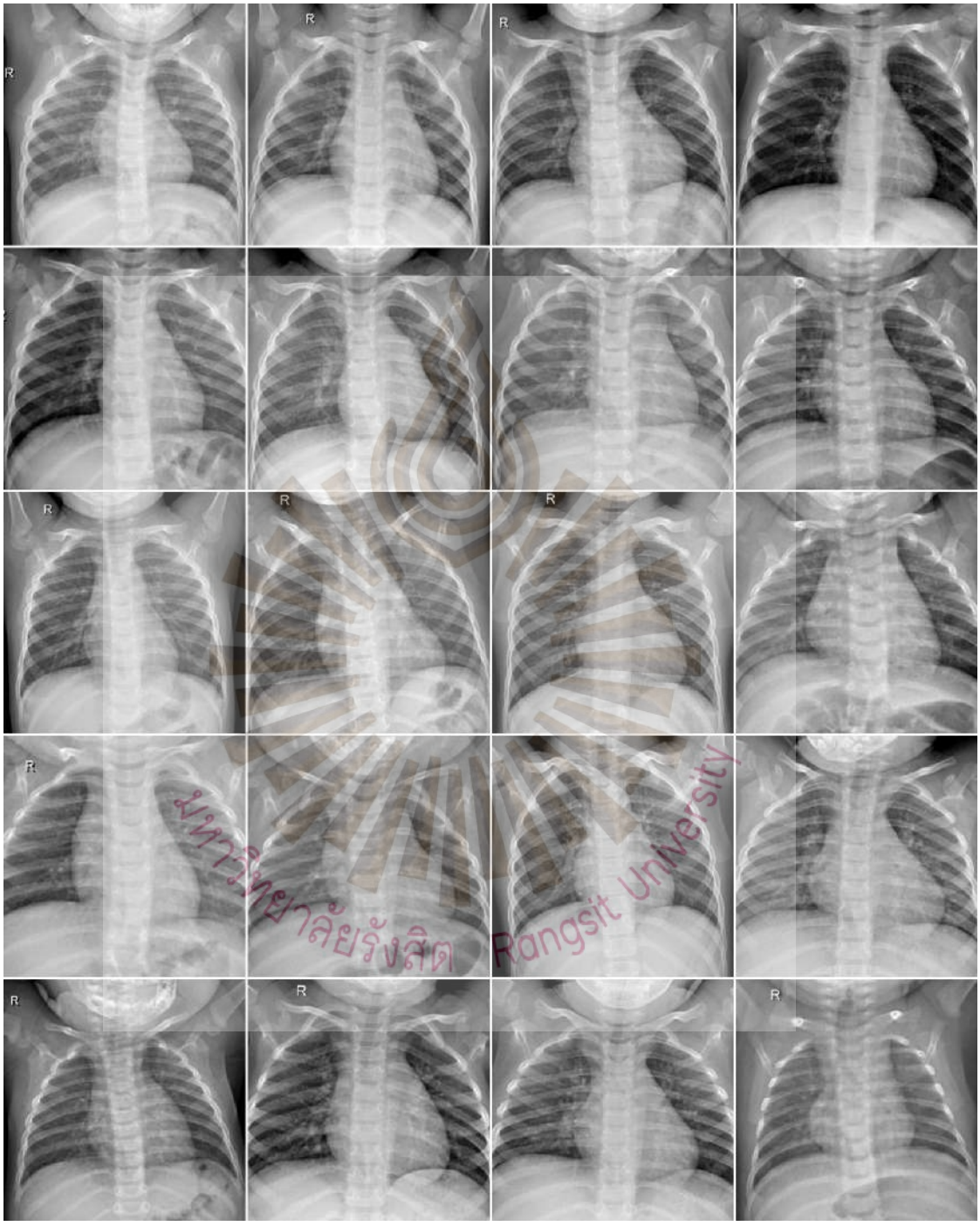
Unrealistic PGGAN-generated synthetic images (batch 2)



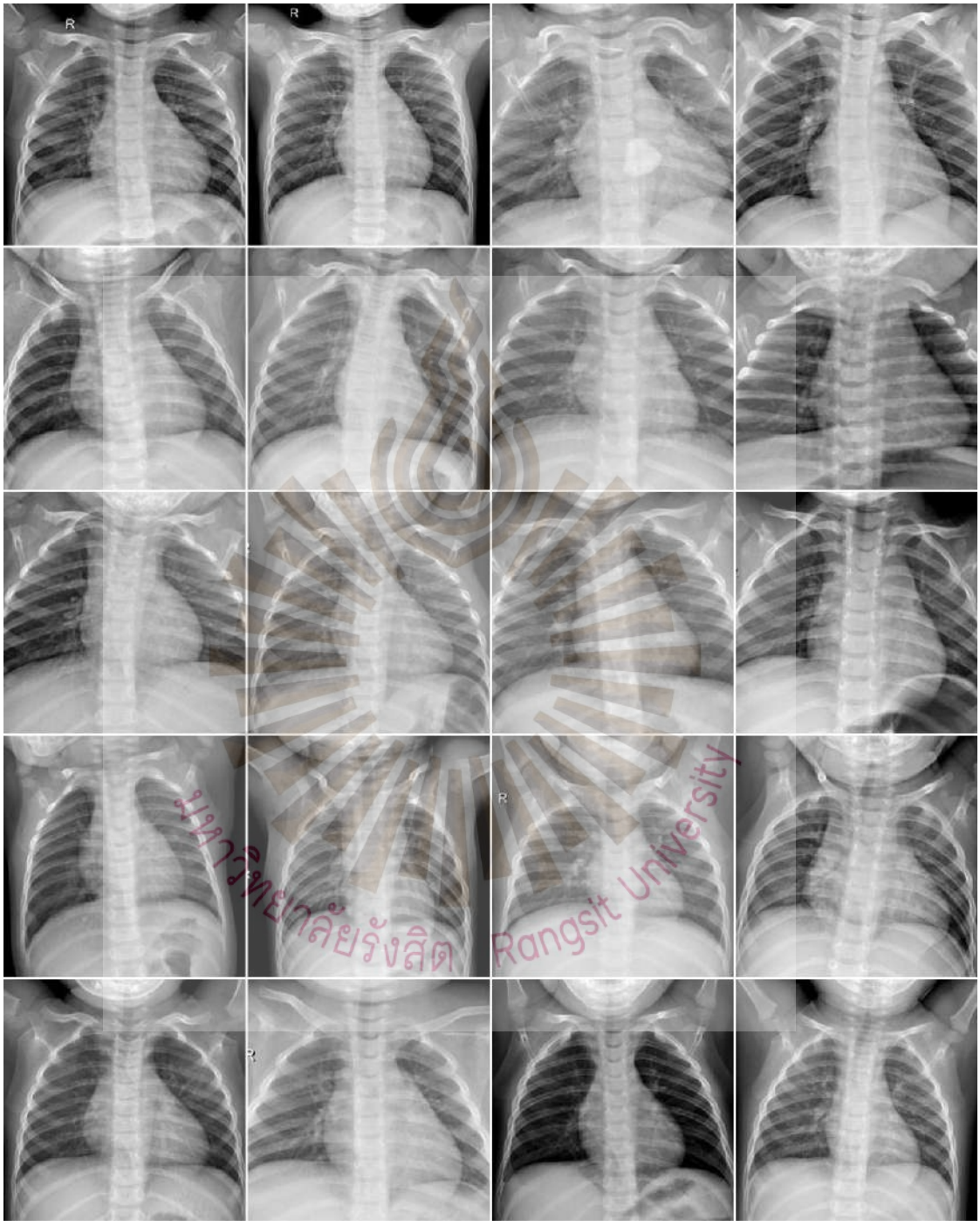
Appendix B

Study 2: Dataset

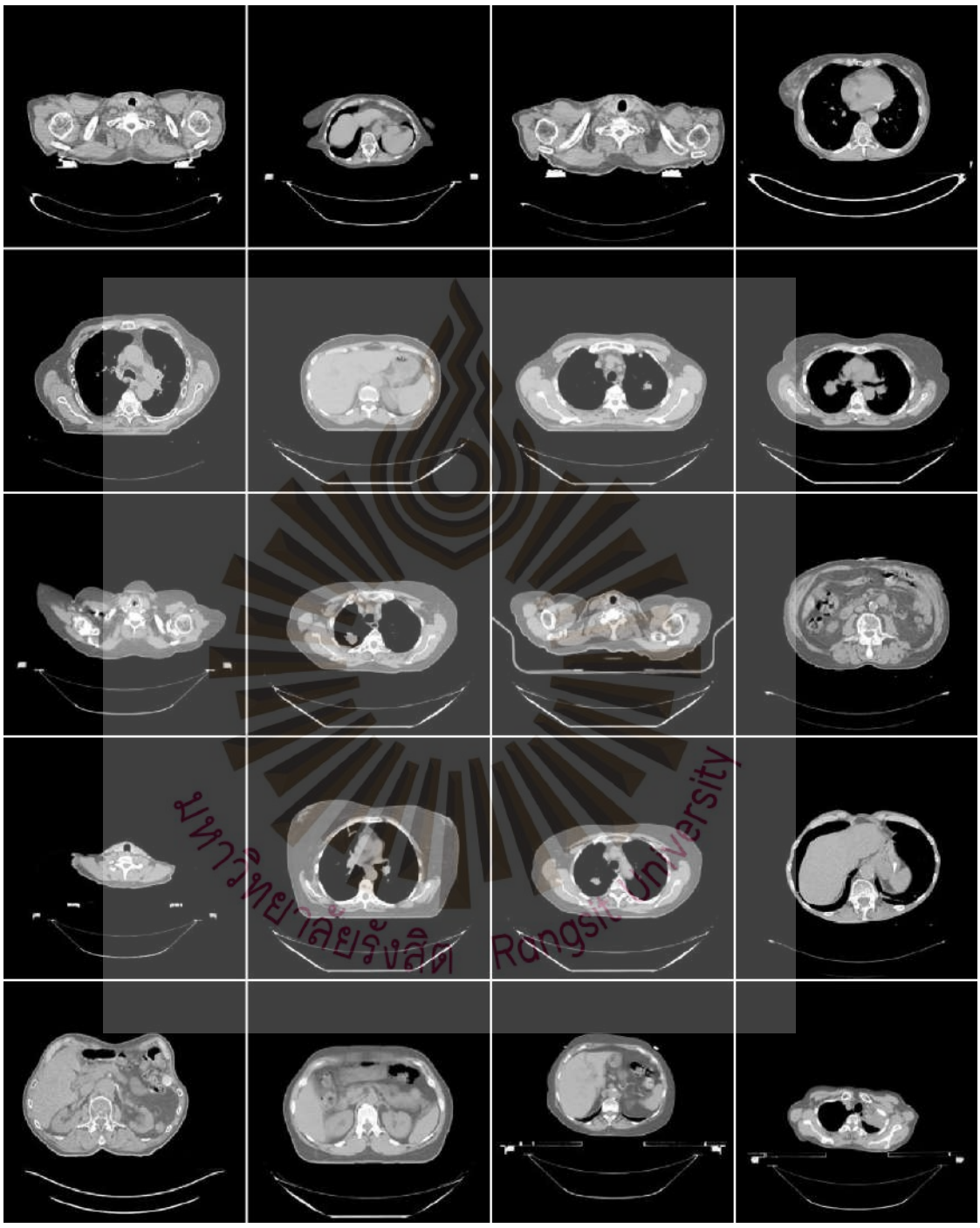
Chest X-ray images (batch 1)



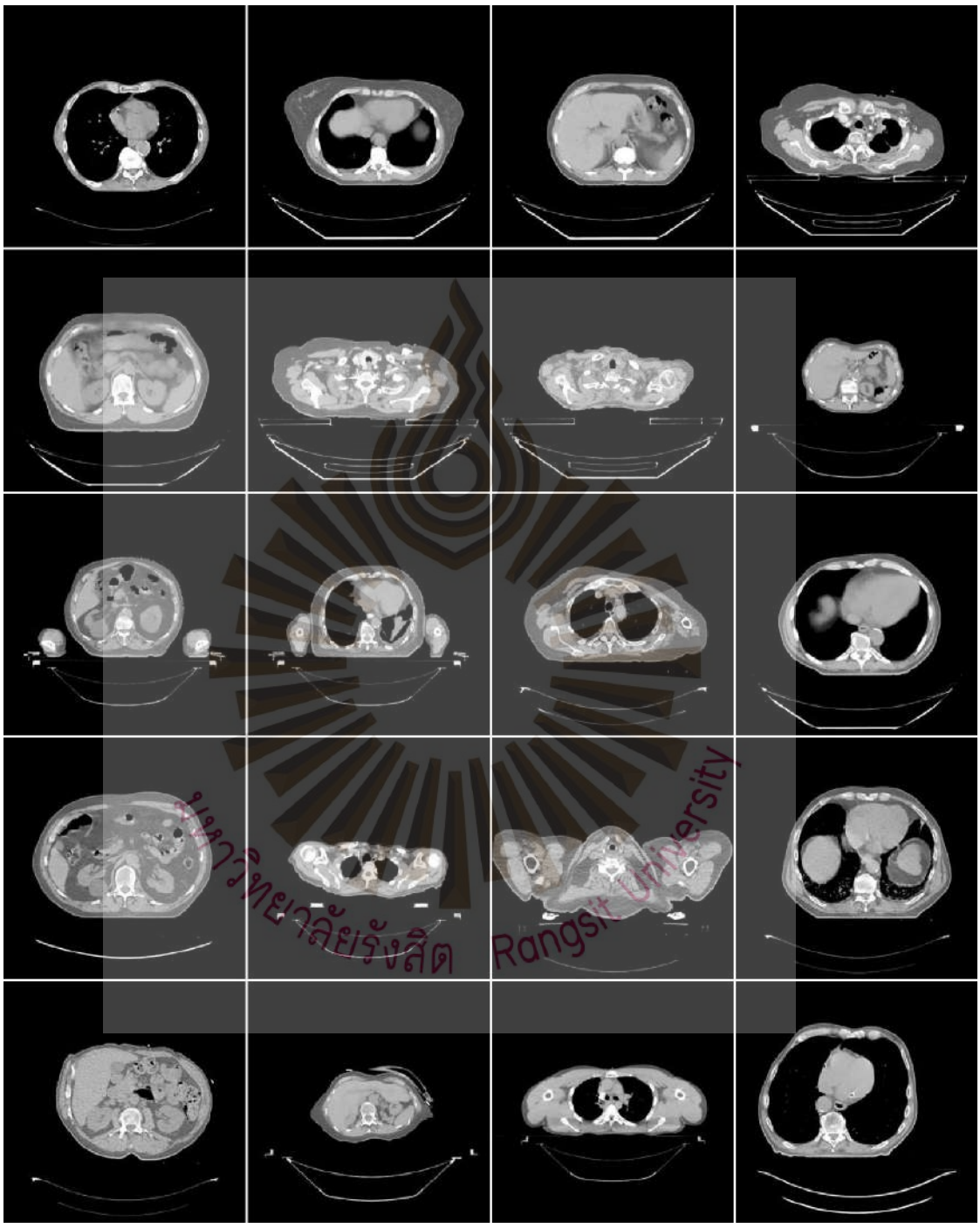
Chest X-ray images (batch 2)



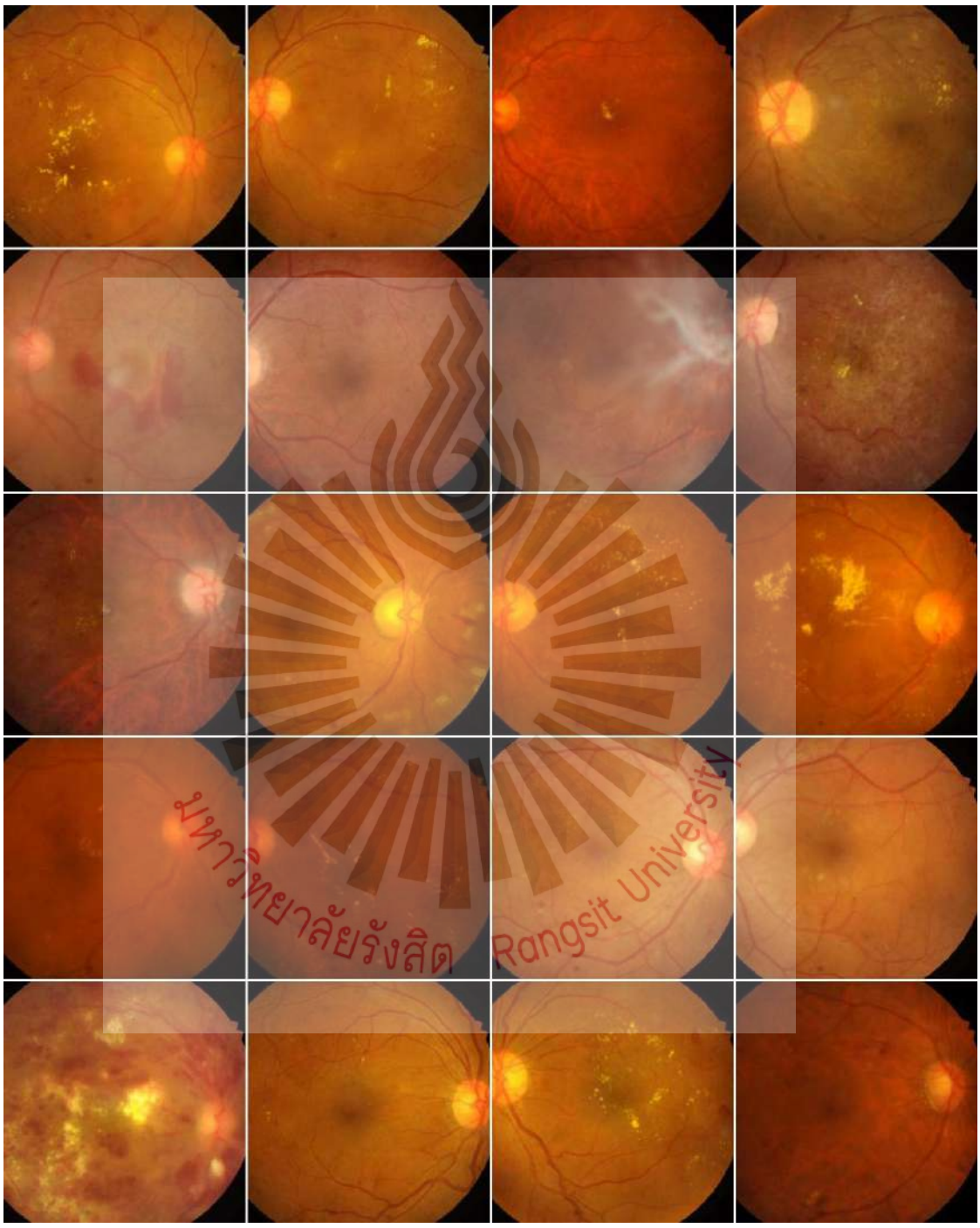
Computed Tomography (CT) scans of the thoracic region (batch 1)



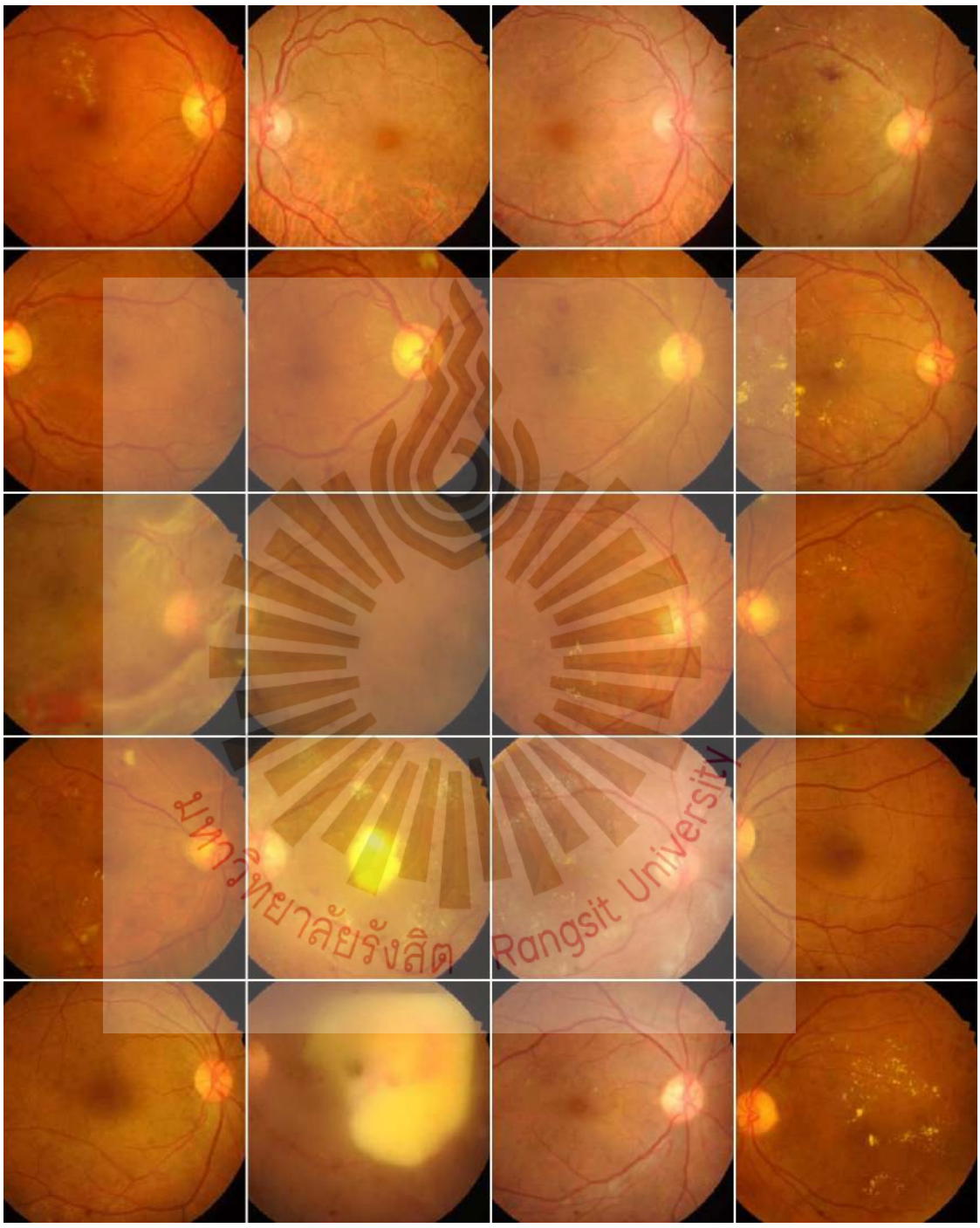
Computed Tomography (CT) scans of the thoracic region (batch 2)



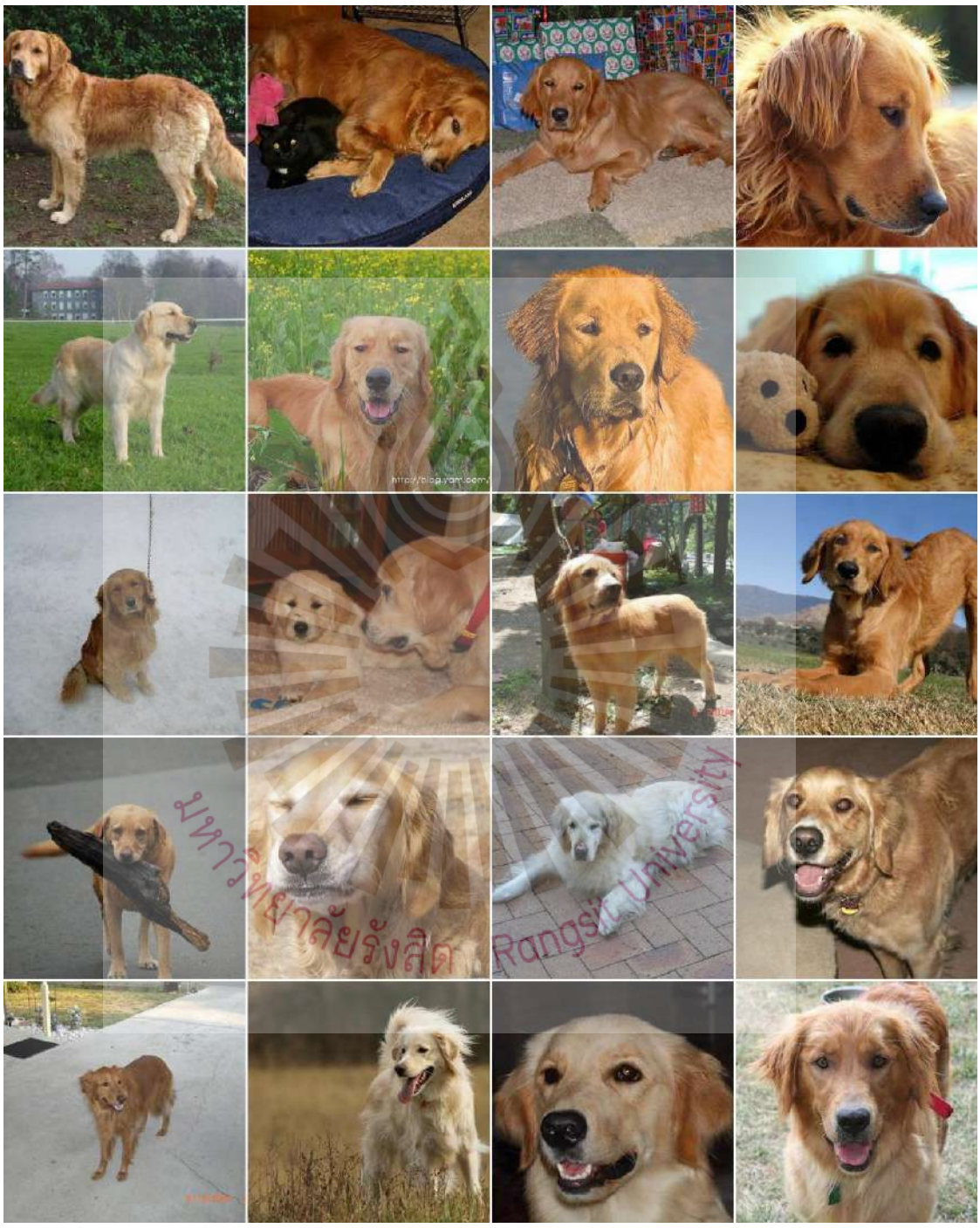
Color fundus photographs (batch 1)



Color fundus photographs (batch 2)



Golden retriever (dog) images (batch 1)

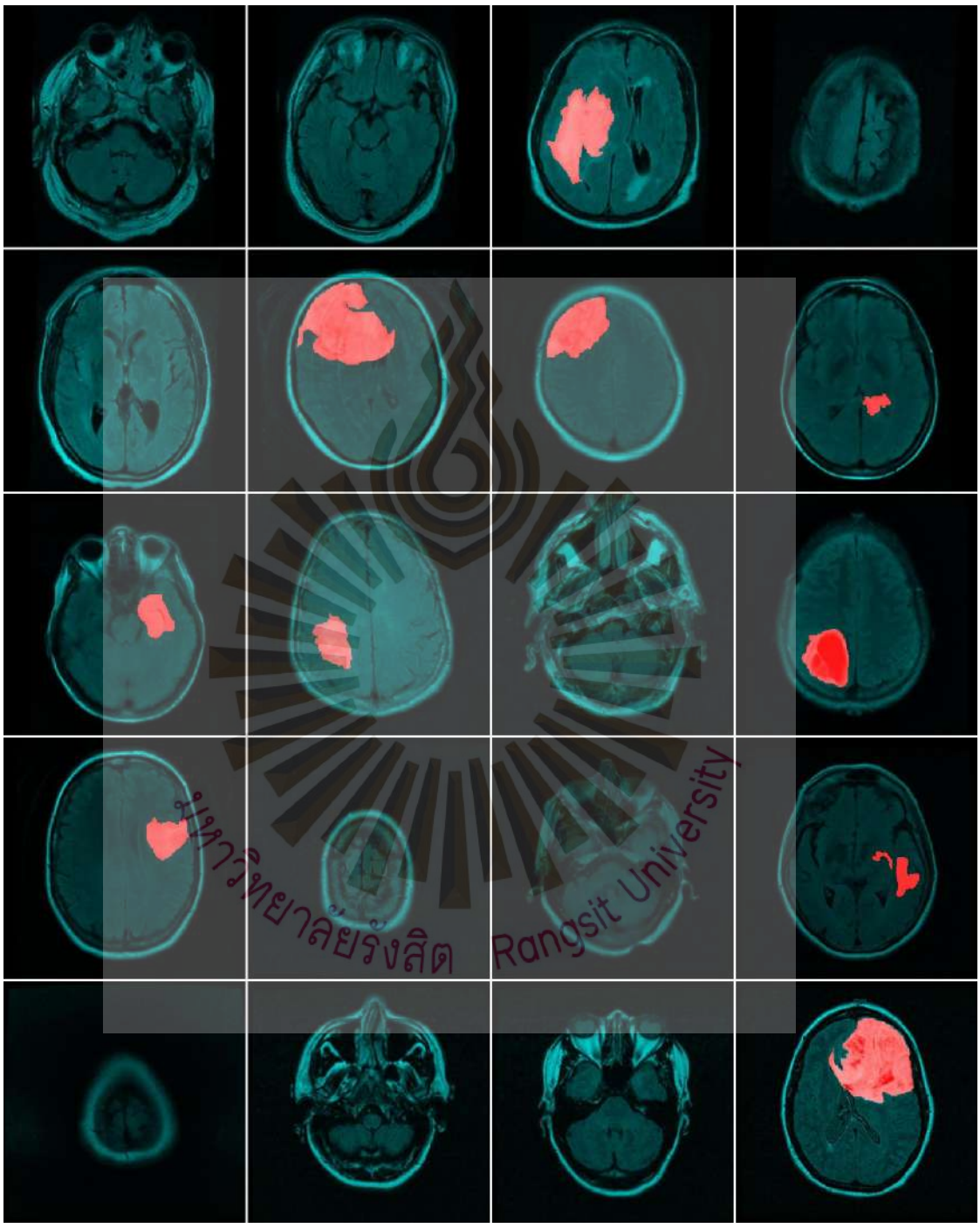


Golden retriever (dog) images (batch 2)

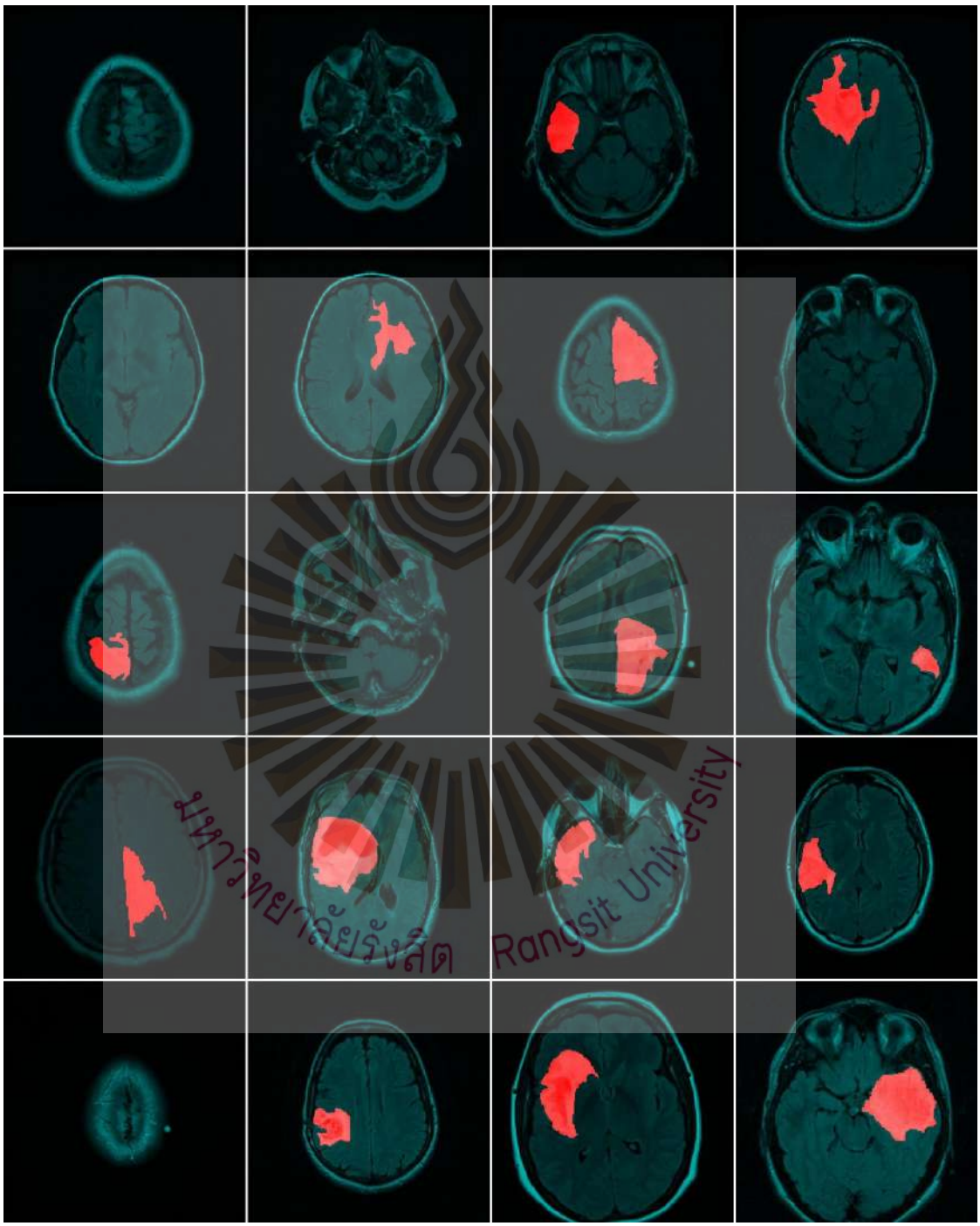
Appendix C

Study 3: Datasets, Realistic, and Unrealistic StyleGAN2-ADA Generated Images

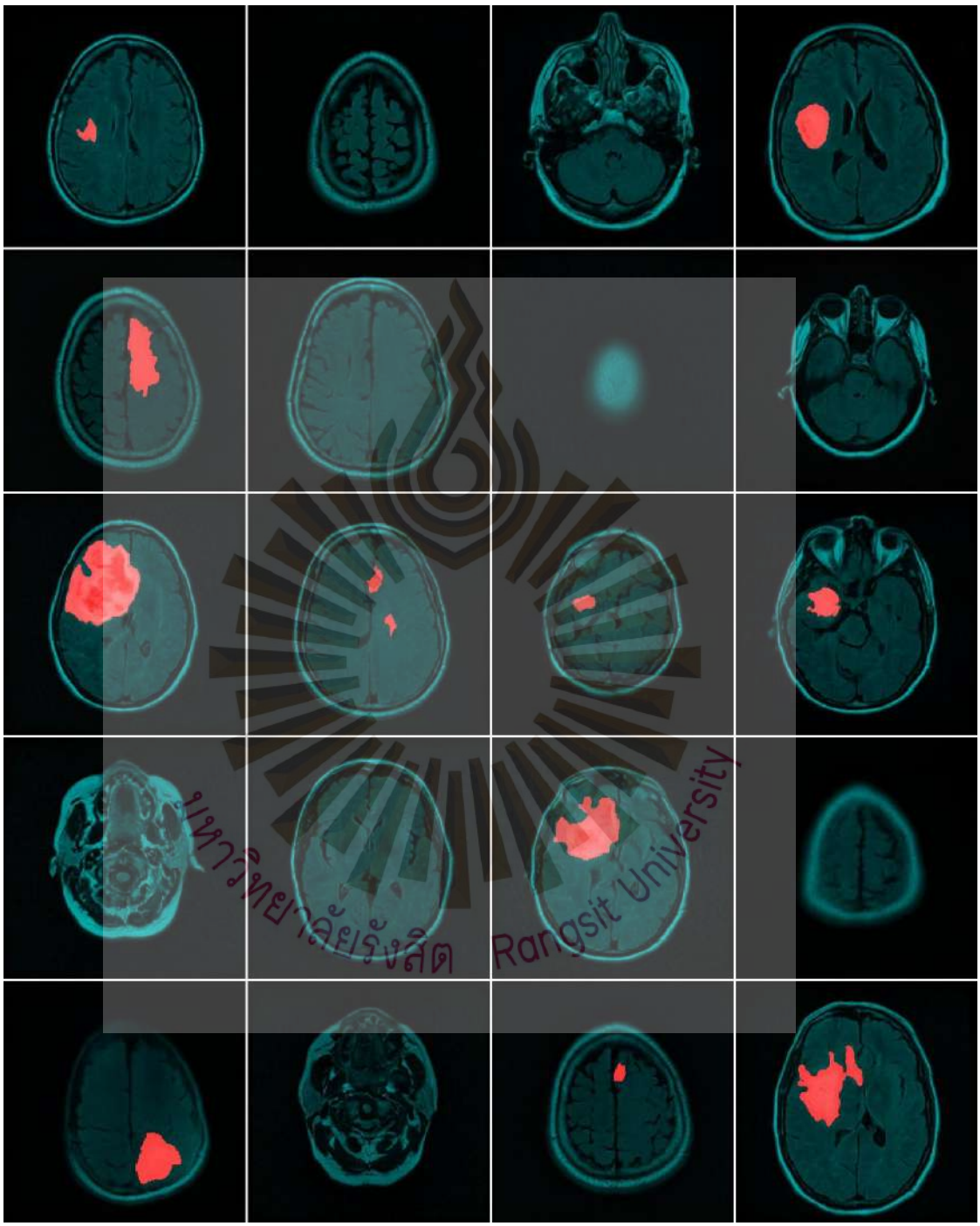
GAN prepared real images (batch 1)



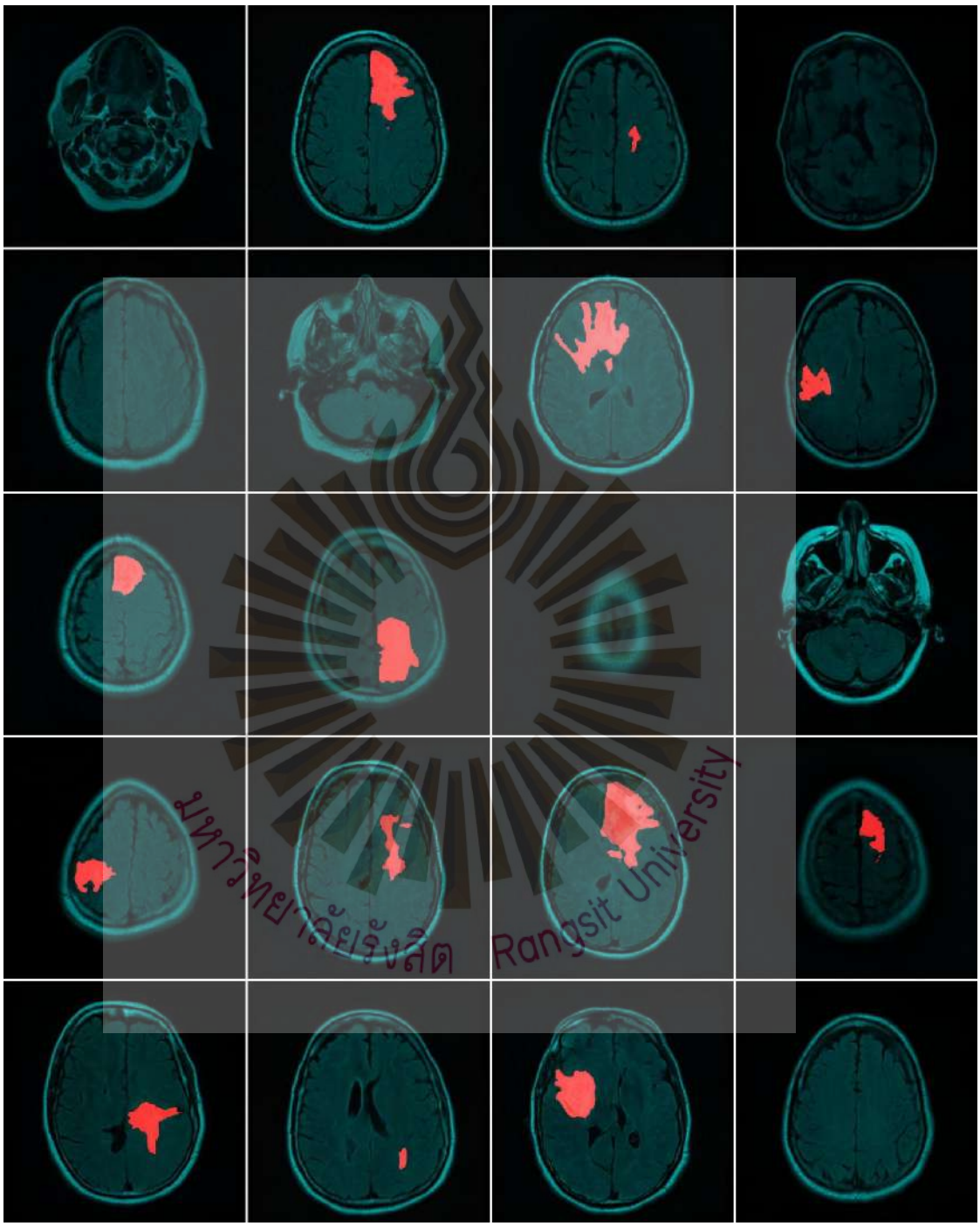
GAN prepared real images (batch 2)



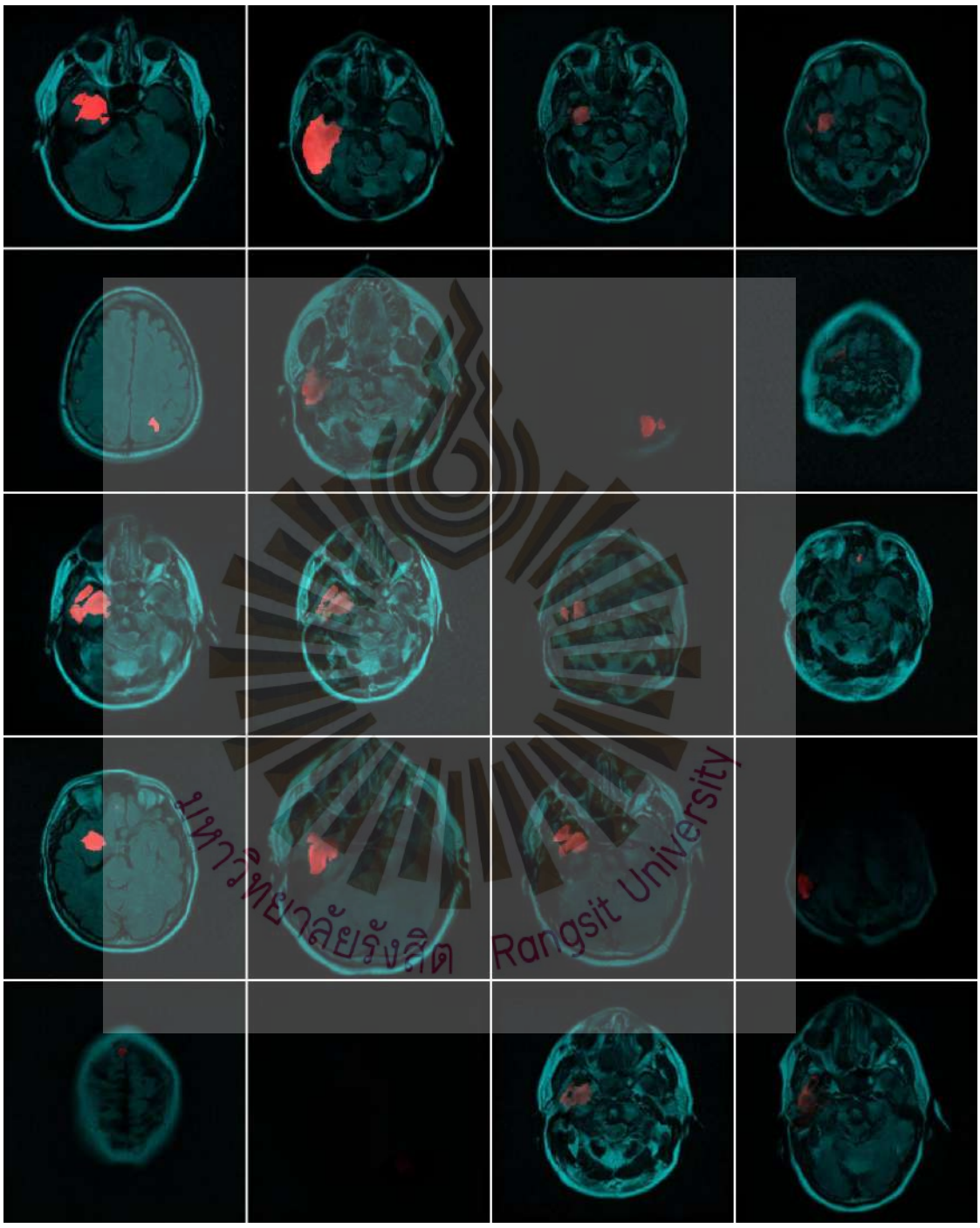
Realistic StyleGAN2, ADA-generated synthetic images (batch 1)



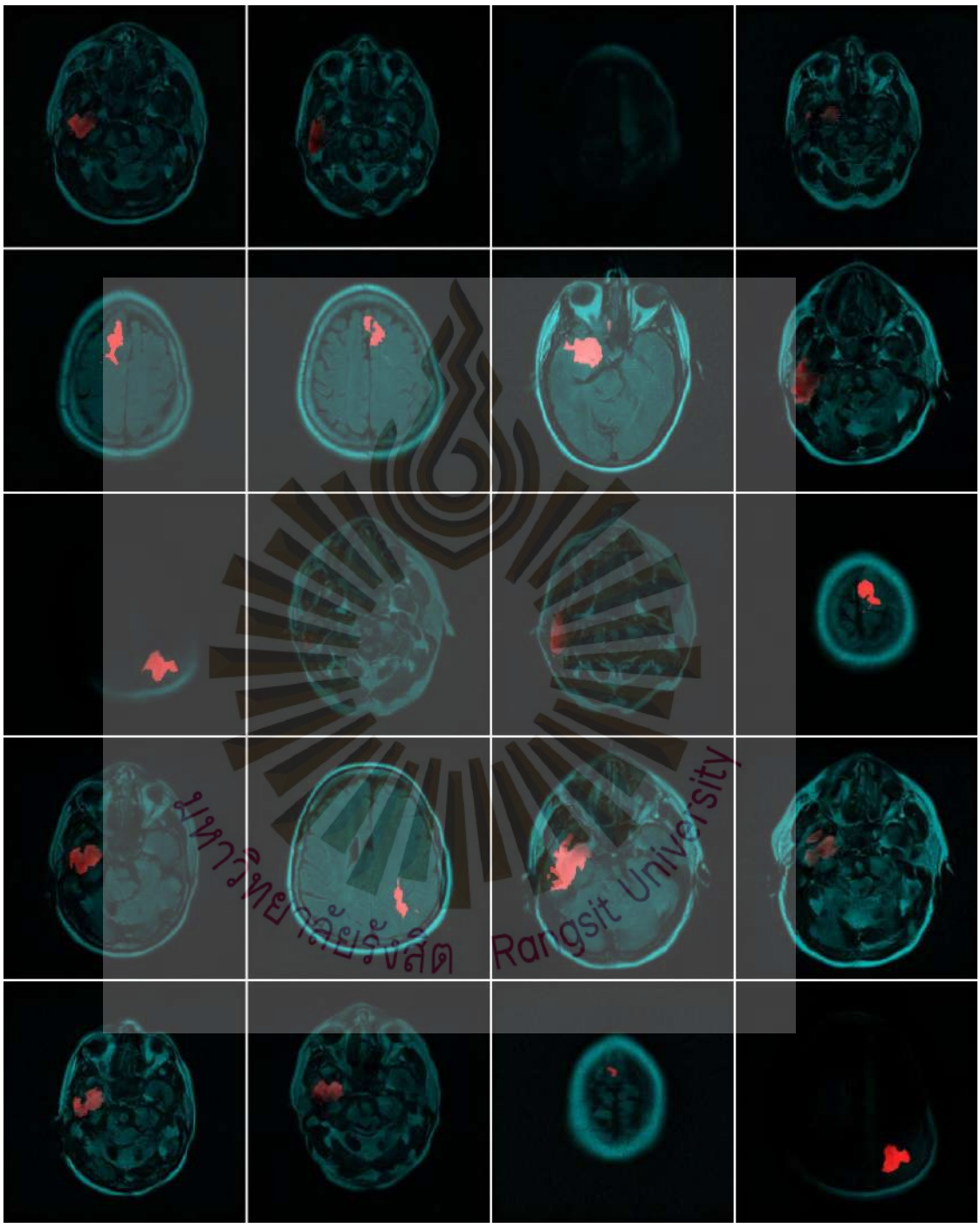
Realistic StyleGAN2, ADA-generated synthetic images (batch 2)



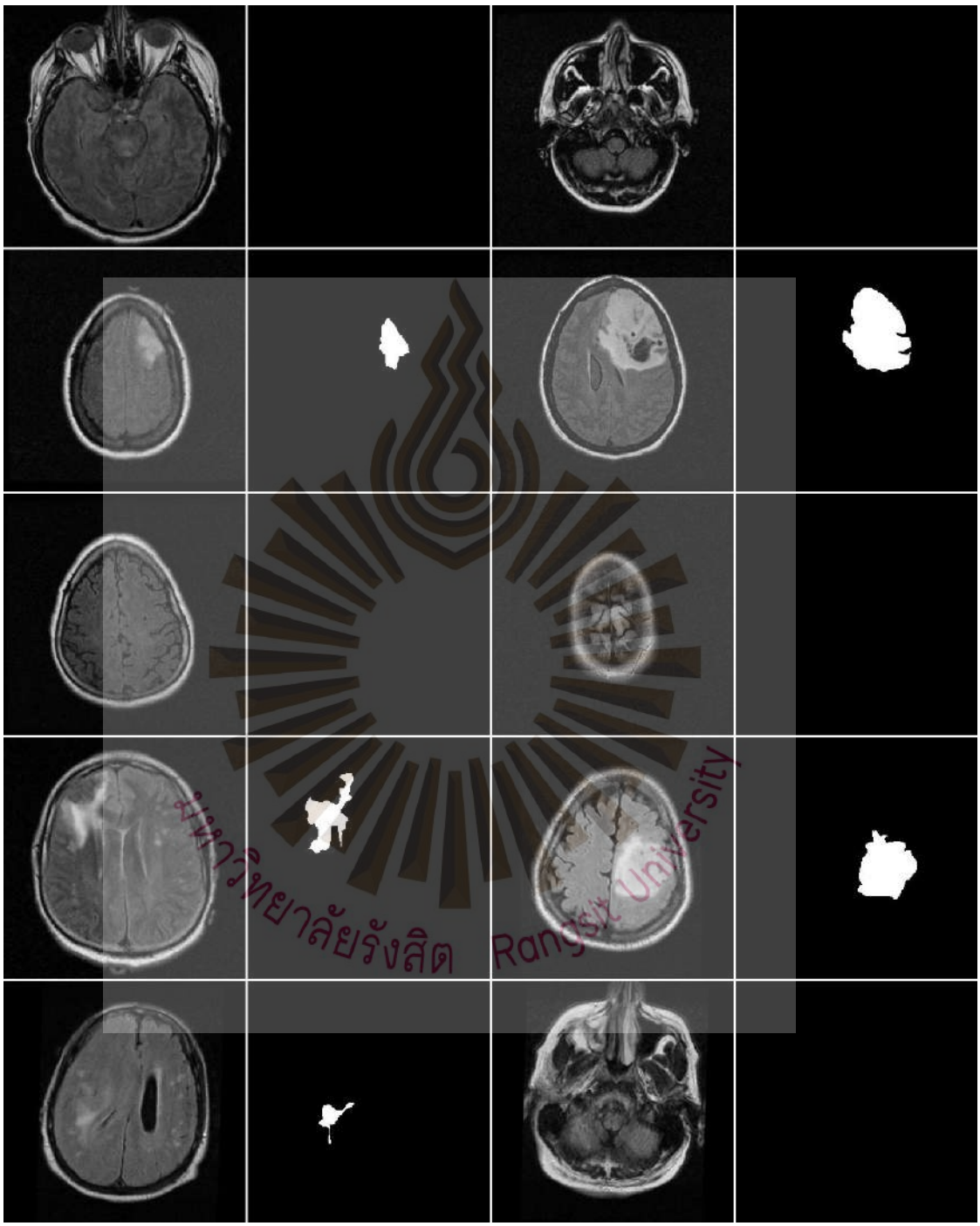
Unrealistic StyleGAN2, ADA-generated synthetic images (batch 1)



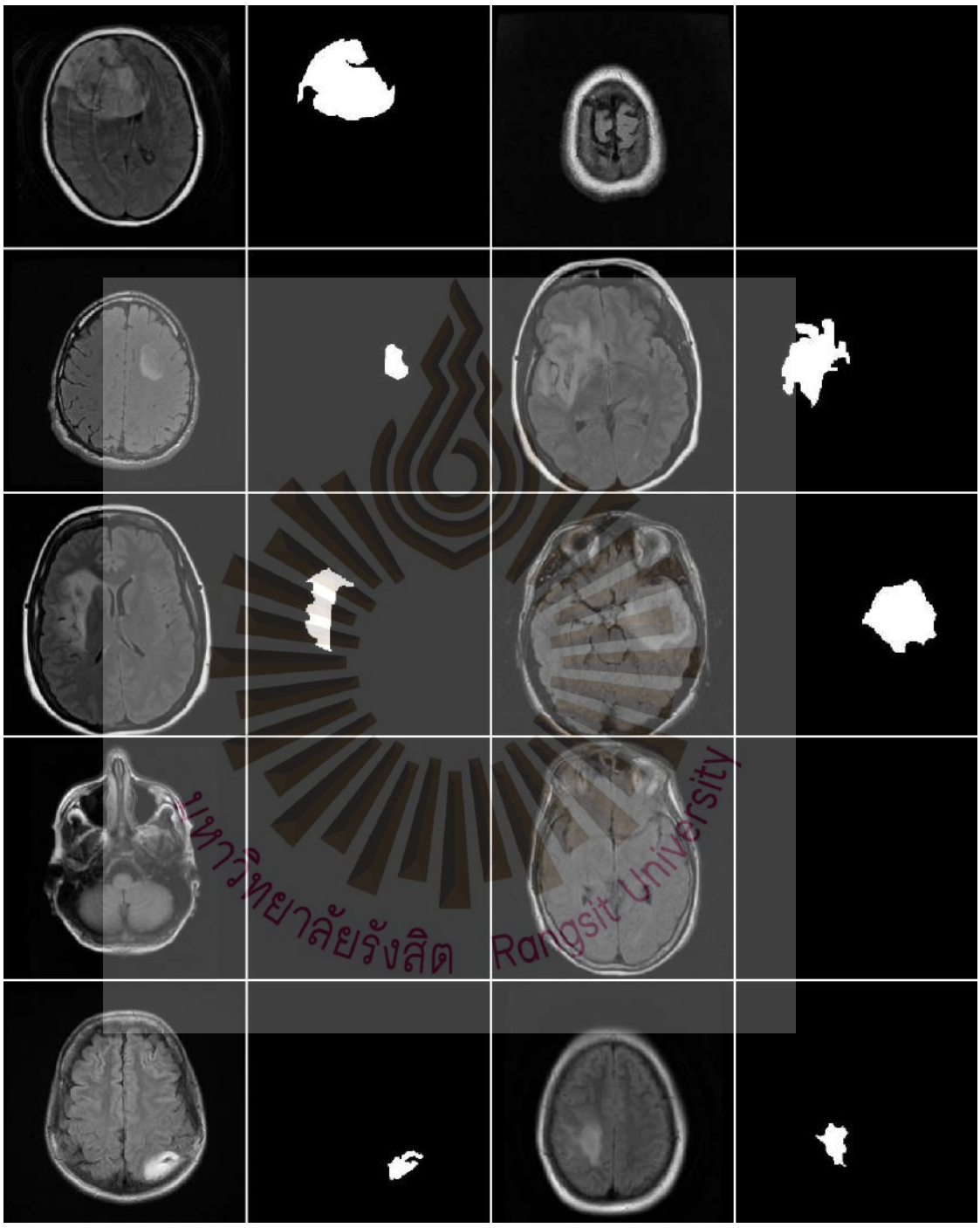
Unrealistic StyleGAN2, ADA-generated synthetic images (batch 2)



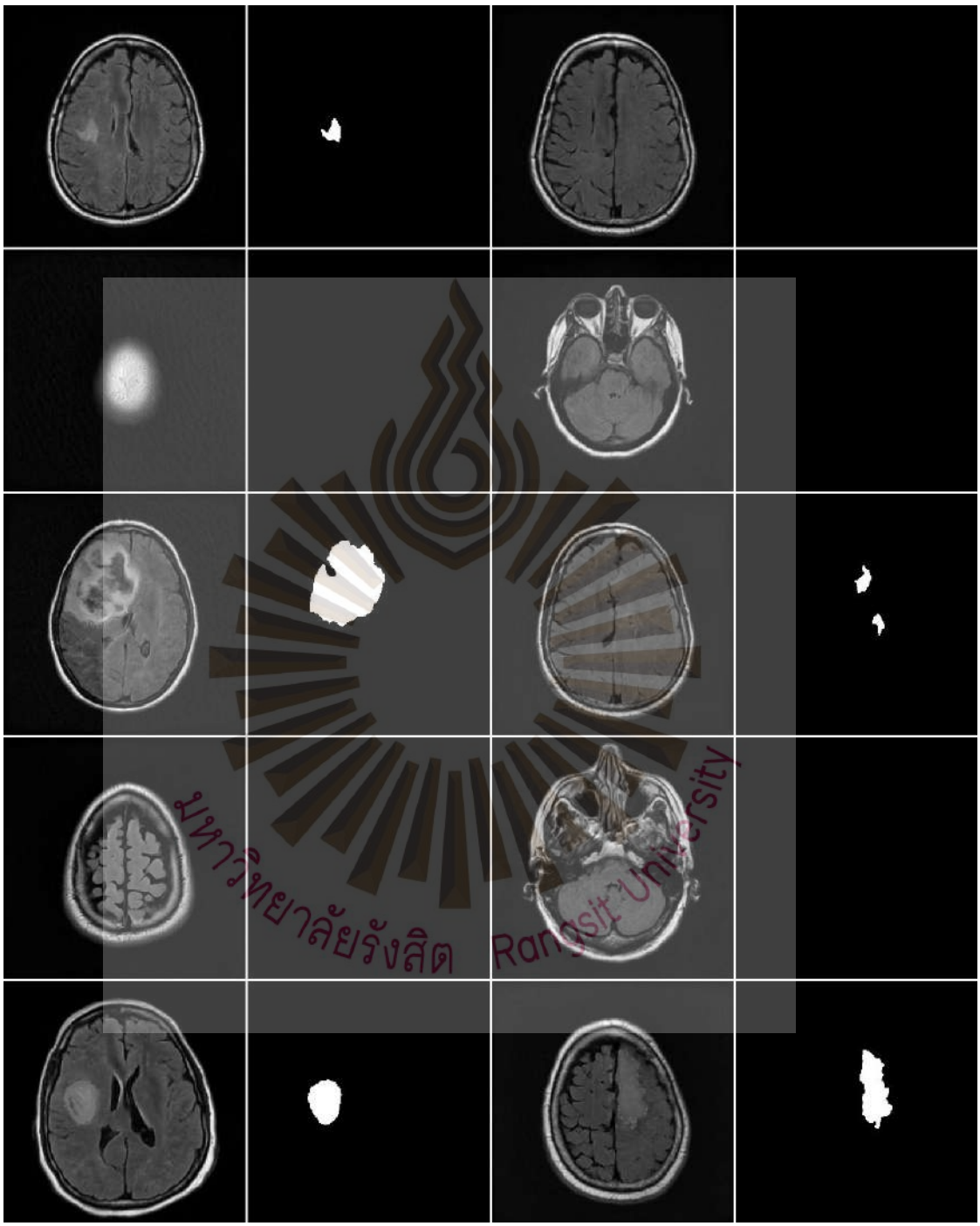
U-nets prepared real images (batch 1)



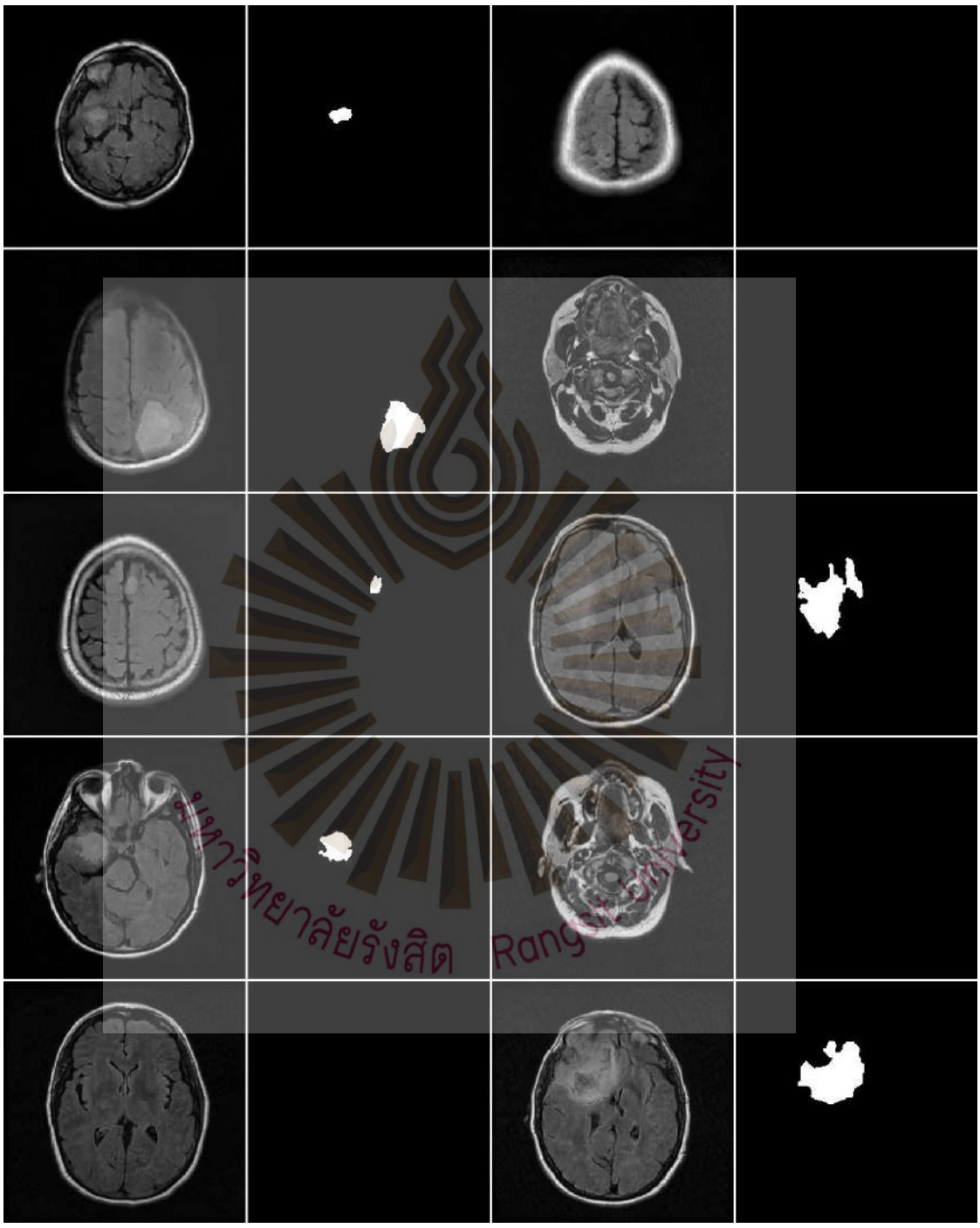
U-nets prepared real images (batch 2)



Generated synthetic data for U-nets (batch 1)



Generated synthetic data for U-nets (batch 2)



Biography

Name	Fawad Asadi
Date of birth	May 19, 1992
Place of birth	Medina, Saudi Arabia
Education background	Taibah University, Saudi Arabia Bachelor in Electrical Engineering, 2018 Rangsit University, Thailand Doctoral in Biomedical Engineering, 2023
Address	Rangsit, Pathum Thani, Thailand
Email Address	eng.fd.ai@gmail.com