공학석사 학위논문

3 차원 확산 확률론적 생성 모델을 이용한 파킨슨병 환자들의 18F-FP-CIT PET 영상 분석 및 신경학적 응급 환자들의 뇌 CT 영상의 비정상 영역 탐지

Study on Image Analysis of 18F-FP-CIT PET in Parkinson's Disease Patients and Anomaly Detection of Brain CT in Neurological Emergency Patients using 3D Diffusion Probabilistic Models

울 산 대 학 교 대 학 원

의 과 학 과

원종준

3 차원 확산 확률론적 생성 모델을 이용한 파킨슨병 환자들의 18F-FP-CIT PET 영상 분석 및 신경학적 응급 환자들의 뇌 CT 영상의 비정상 영역 탐지

지도교수 김남국

이 논문을 공학석사 학위논문으로 제출함

2024 년 8 월

울 산 대 학 교 대 학 원

의 과 학 과

원종준

원종준의 공학석사 학위논문을 인준함

심사위원 홍길선 인

심사위원 김남국 인

심사위원 정선주 인

울 산 대 학 교 대 학 원

2024 년 8 월

Contents

Contents of Tables

Contents of Figures

1. Abstract

Purpose This work tried to verify two abilities of deep generative models using several tasks. The first task is the performance to distinguish features in the image analysis task of 18F-FP-CIT PET in Parkinson's Disease patients, and the second is the data distribution learning (generator) performance in unsupervised anomaly detection of non-contrast 3D brain CT.

Methods The first study involved 2,672 18F-FP-CIT PET scans retrospectively collected. We build a 3D pre-training model with the generative method and evaluate the model's performance to discriminate features using linear probing and fine-tuning. We did binary classification of Essential Tremor (ET) / early onset Parkinson's disease and multi-class classification of Parkinson's disease (PD), multiple system atrophy (MSA) and progressive supranuclear palsy (PSP). Also, the pre-trained model was used in the motor-symptom onset year regression task, and the model's performance was evaluated similarly to classifications. The second study involved 34,085 non-contrast brain CT scans retrospectively collected with healthy subjects. We trained a 3D generative model using only the normal CT scans. After learning the normal data distribution, this model could detect abnormal scans that deviated from the normal distribution.

Results In the first task, the proposed network achieved the area under the receiver operating characteristic curve (AUROC) of 0.997(internal), 0.994 (external) with the area under the precision-recall curve (AUPRC) of 0.998(internal), 0.991(external) in the cross-validation of ET/PD classification and AUROC of 0.920(internal), 0.919(external) with AUPRC of 0.881(internal), 0.670(external) in the cross-validation of PD/MSA/PSP classifications. In the regression task, the model achieved the Mean Absolute Error (MAE) of 2.013(internal) and 1.965(external) with the concordance correlation coefficient (CCC) of 0.701(internal) and 0.733(external). In the second task, our network detected normal and abnormal images with0.91 0.93, 0.91, and 0.91 of accuracy, precision, recall, and F1-score in the internal test set and 82, 0.82, 0.82, and 0.82 in the external test set, respectively.

Conclusion This study suggested that the deep generative models could become a discriminant of the functional brain images' clinical features. Also, this could become a distribution learner of the structural brain images, which detect whether an image is in the distribution or deviated.

2. Introduction

The landscape of deep learning has seen remarkable advancements with the development of generative models to capture the underlying data distribution, allowing them to create new, realistic samples. These models have become integral to various applications, demonstrating significant potential in generating new samples and understanding existing data.[1][2] Recent models like GPT (Generative Pre-trained Transformer) and T5 (Text-To-Text Transfer Transformer) have garnered significant attention by unifying generative and discriminative tasks. Unlike traditional models focusing solely on classification or generation, these models leverage a generative pre-training approach to enhance their performance on discriminative tasks.[3][4] Generative models excel in creating new samples by approximating the data distribution. This capability implies that they also achieve a semantic

understanding of the raw visual data, which is crucial for recognition tasks. By learning to generate data, these models inherently learn to discriminate between different data classes.[5] Diffusion Probabilistic Models have recently dominated the field of image generation. [6][7] These models generate images through a gradual denoising process, starting from pure noise and progressively refining the image. This approach has proven effective in producing high-quality, realistic images. Based on DPMs, some models emerged to solve the problem of DPMs having no practical and meaningful latent space. Among them, Hierarchical Diffusion Autoencoders (HDAE) perform well in extracting the fine-to-coarse-level feature for the latent space of diffusion models. In this sense, this model could simultaneously become a great discriminant and data distribution learner (generator). So, we did two tasks using HDAE, especially 3D inflated HDAE. One was the image analysis of 18F-FP-CIT PET in patients with Parkinson's disease, mainly using the faculty in view of classifier and regressor. Another is the unsupervised anomaly detection in 3D non-contrast brain CT, which employs the ability of data distribution understanding.

The First task is the image analysis of 3D 18F-FP-CIT PET in Parkinson's Disease patients. Parkinson's Disease (PD) is a chronic and progressive neurodegenerative disorder that primarily affects the central nervous system, leading to significant motor system impairments. The hallmark motor symptoms of PD include tremors, rigidity, bradykinesia (slowness of movement), and postural instability, collectively known as parkinsonism. These symptoms typically emerge gradually and worsen over time, severely impacting affected individuals' quality of life and functional abilities. [1] FP-CIT (Fluoropropyl-carbomethoxyiodophenyl-tropane) is a radioligand used in neuroimaging to visualize dopamine transporters in the brain. It binds specifically to the dopamine transporter (DAT) sites in the striatum, enabling detailed imaging of dopaminergic neuron integrity. FP-CIT imaging is critical in the evaluation of diseases that affect dopaminergic pathways, such as PD. FP-CIT Single Photon Emission Computed Tomography (SPECT) is a widely used imaging technique that employs FP-CIT as a radiotracer. This method is beneficial for diagnosing PD and differentiating it from other conditions that affect motor function. However, FP-CIT Positron Emission Tomography (PET) is a more advanced imaging technique that also uses FP-CIT as a radiotracer. Compared to SPECT, PET imaging offers higher resolution and greater sensitivity, providing more detailed and accurate images of dopamine transporter distribution.[8]

Diagnosing PD can be challenging due to its symptom overlap with other neurodegenerative disorders such as Essential Tremor (ET), Multiple System Atrophy (MSA), and Progressive Supranuclear Palsy (PSP). These conditions share similar motor symptoms, making accurate diagnosis difficult. However, accurate differentiation between these disorders is crucial for developing effective therapeutic strategies.[9] Without a characteristic, PD has similar diseases; one of the other features of PD is progressive neurodegeneration. Progressive loss of the ascending dopaminergic projection in the basal ganglia is a fundamental pathological feature of Parkinson's disease. Imaging biomarkers in PD are increasingly important for monitoring progression in clinical trials and also have the potential to improve clinical care and management. [10]

For this reason, it is meaningful to build a reliable computer-aided diagnosis and imaging biomarker that reflects progressive neurodegeneration that could assist neurologists in making precise diagnoses.[11] Conventional quantitative analyses of FP-CIT imaging typically focus on specific brain region uptake values, such as the striatal binding ratio (SBR) or Standardized uptake value ratio (SUVR). While these measures are helpful, conventional analyses often focus on predefined regions of interest (ROIs), potentially overlooking relevant information in other areas of the brain. Deep learning has demonstrated exceptional performance in various domains, including medical image analysis. It can automatically learn and extract relevant features from raw images without the need for manual ROI delineation. However, FP-CIT PET is a sophisticated imaging modality that is difficult to acquire, expensive, and challenging to image. Consequently, there have been relatively few studies utilizing this modality despite its potential for providing detailed insights into dopaminergic dysfunctions in PD. By building deep learning models trained on a large dataset of FP-CIT PET images, this study aims to support neurologists in making accurate diagnoses and monitoring disease progression.

This study compares several generative pre-training techniques to determine their efficacy in downstream classification and regression tasks. By pre-training on a generative task, models can learn robust feature representations that are beneficial for subsequent discriminative tasks such as classification and regression. Various methods for generative pre-training exist, each with its unique advantages. After generative pre-training, the encoder is subjected to linear probing and fine-tuning. Linear probing involves training only a trainable simple linear classifier on the pre-trained features while fine-tuning adjusts all the pre-trained weights on the specific downstream tasks. This study evaluates the encoder's performance in two classification tasks: ET vs. early PD and PD vs. MSA vs. PSP. Early PD patients include patients whose FP-CIT PET scans were acquired undergone years of motor symptom onset. In addition, we performed a regression task to predict the onset year of motor symptoms. Also, using this regression network, we calculated the correlation coefficient between the actual duration of PD progression and the predicted duration of PD. See Figure 1 for overall workflow.

We adopted the Hierarchical Diffusion Autoencoder (HDAE) [12], Denoising Diffusion Autoencoder (DDAE) [13], Simple masked image modeling (SimMIM) [10], Disruptive AE (DisAE) [14], and Pixel2Style2Pixel (PSP) [15]. Both DDAE and HDAE are based on diffusion pre-training, but HDAE includes a semantic encoder, while DDAE does not. We also compared the model using Discrete Wavelet Transforms (DWT) as preprocessing to enhance the performance of HDAE and DDAE, naming the Hierarchical Wavelet Diffusion Autoencoder (HWDAE) and Wavelet Denoising Diffusion Autoencoder (WDDAE), respectively. We also trained 3D stylegan2 [16] with cascaded training methods and built a generative adversarial network inversion network PSP based on the encoder method to find W+ space in stylegan2. In addition, we adopt two models SimMIM and DisAE which have swin transformer encoder but have different pre-training strategy. SimMIM uses masking, and DisAE uses super-resolution and denoising pre-training strategy. Lastly, we compared classification and regression performances using these pre-trained models to those of the supervised network, 3D ResNet34. Among various

diffusion models and tasks, HWDAE demonstrated generally superior linear probing and fine-tuning performance among the various models tested. This model's ability to capture intricate details during the generative pre-training phase translates into enhanced discriminative capabilities, making it particularly effective for medical image analysis.

In summary, our contributions to this first study are as follows:

- We compared the semantic encoder of generative models and demonstrated that pre-trained generative self-supervised models achieve comparable classification and regression performance to supervised models in differential diagnosis and symptom onset prediction of PD, using a large dataset of FP-CIT PET scans.
- We suggested the predicted symptom onset years could be a reasonable imaging biomarker to monitor the progression of PD.
- We reconfirmed the correlation between the actual disease progression interval and the changes in presynaptic dopaminergic neuron degradation observed in FP-CIT PET scans.

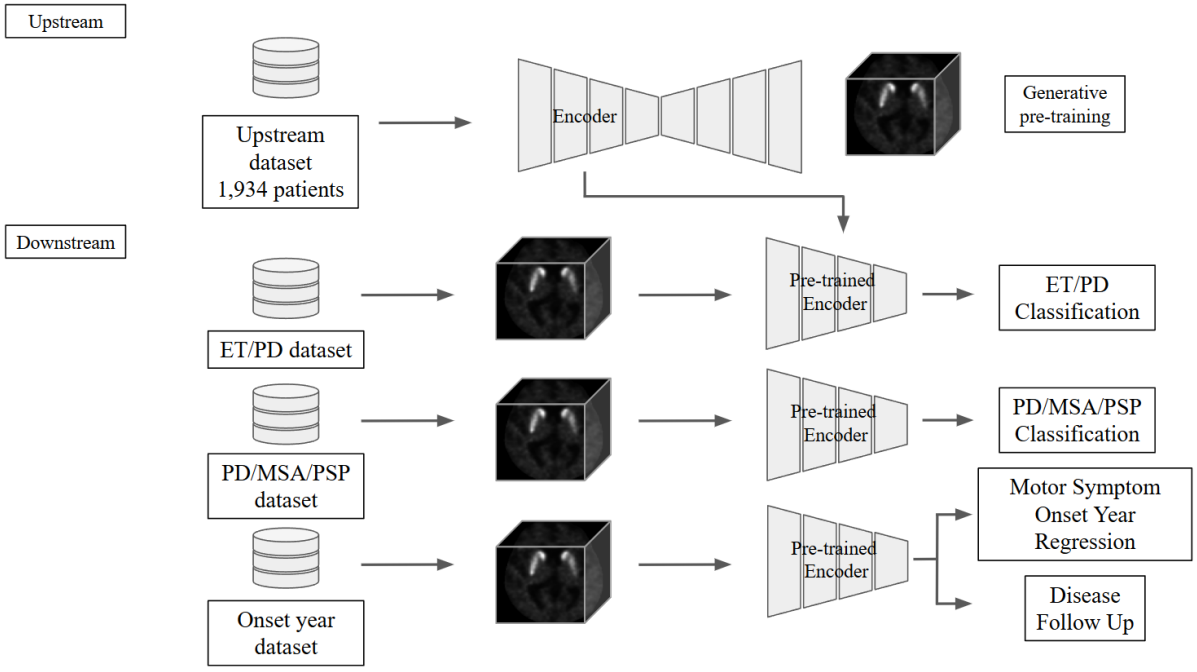

Figure 1. The overall workflow of upstream and downstream tasks. In the upstream process, a generative self-supervised model was trained. In the downstream process, the pre-trained encoder weights from the generative model were used for binary and three-class classification, symptom onset regression, and follow-up image analysis.

Our second task is unsupervised anomaly detection (UAD) in 3D non-contrast brain CT. Neurological emergencies require diagnosing and treating a wide variety of lesions they may encounter at any time. Non-contrast brain computed tomography (NCCT) is the current standard imaging modality for the initial screening

and diagnosis of neurological conditions because it can detect a broad spectrum of disorders and has a fast-scanning time. With the development of deep learning, several studies have demonstrated its potential to help with various radiology tasks. These studies show that deep learning-based radiology classification can improve workflow efficiency, speed up radiology reporting, and timely manage patients with critical findings. However, building large-scale annotated training datasets remains a major hurdle in developing deep learning systems in medicine, especially in emergencies where they must deal with various lesions with heterogeneous morphology, sizes, locations, and intensities.

Different from supervised deep learning with a narrow clinical focus, which has limited usage in this condition, unsupervised anomaly detection, recently DDPMs, has been studied to detect diverse lesions that deviate from the normal distribution. Deep generative models learn to capture normal data distribution; hence, they can detect anomalous data that deviate from the normal distribution without prior knowledge of anomalies. These methods are categorized as reconstruction-based methods, in which the raw pixel difference between the source image and its reconstruction by generative model indicates the degree of anomaly. Normal patients do not differ much from the training distribution, so the reconstruction has a small residual error from the original image. In contrast, the model maps abnormal patients who deviate from the training distribution to the normal distribution. The normal-like reconstruction image has a relatively large residual error from the original image. The overall workflow is in Figure 2. However, although previous studies using deep generative models have attracted considerable attention, we suspected that some conceptual rethinking and expanding of anomaly detection were needed. (1) The anomaly lesion segmentation perspective was the focus, and the Dice Coefficient Score was one of the main evaluation metrics; (2) Many studies set further study as the expansion model to 3D.

We first started with doubts about the evaluation metric and raised the need for a combination of the structure-conscious and the texture-focusing anomaly detection model that approaches anomalies from a different perspective. The evaluation metric used in most existing papers is the Dice Coefficient Score. This is a quantitative result of how well the anomaly map matches the existing lesion. We questioned this approach as we observed the results in Figure 3, which displays three different cases of brain anomalies. Column (b) shows the reconstruction results of one of our models, the texture-focusing anomaly detection model, and column (c) shows the reconstruction results of the other model, both the texture and structure-conscious anomaly detection model. If we let the model evaluate the Dice Coefficient Score as the anomaly segmentation point, the results in column (b), which only considers texture, would be good normal distribution predictors. But even without the quantitative results, good or not, a casual observer might think that the images in column (b) are some kind of anomalies out of normal distribution. Unlike networks that only consider texture, performing texture and structure-conscious anomaly detection methods yields the results in column (c), which a casual observer might consider more normal than column (b). For brevity, we will refer to the texture and structure-conscious anomaly detection model as the anatomical model and the texture-focusing anomaly detection model as the texture model.

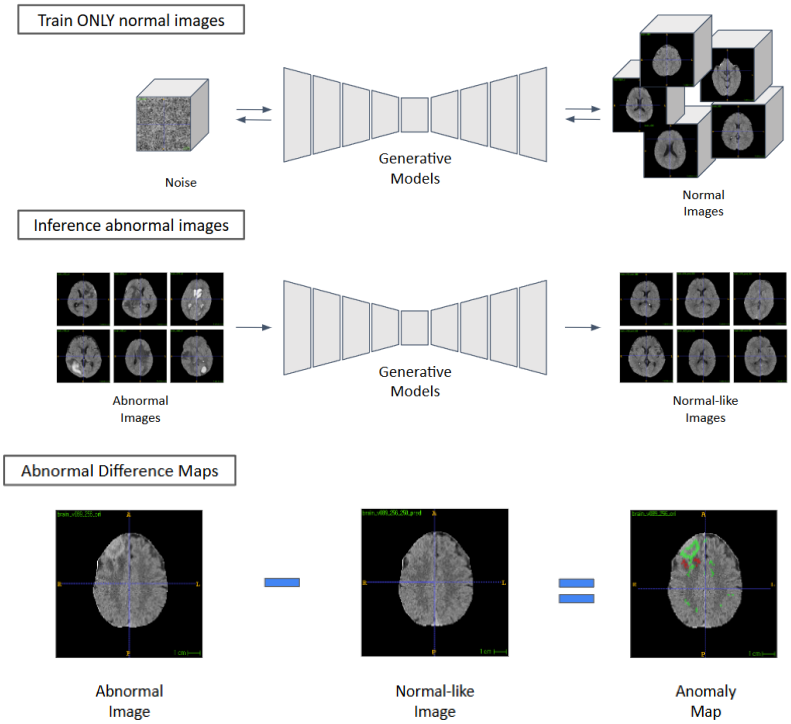

Figure 2. The overall workflow of the unsupervised anomaly detection tasks. In the training process, a generative model was trained only using normal brain CT images. In the inference process, when abnormal images were input, the generative model predicted normal-like images, which have similarities to the original images. After that, we could yield the anomaly map by using a difference map of the abnormal image and the normal-like image. The display windowing is [0, 80] HU

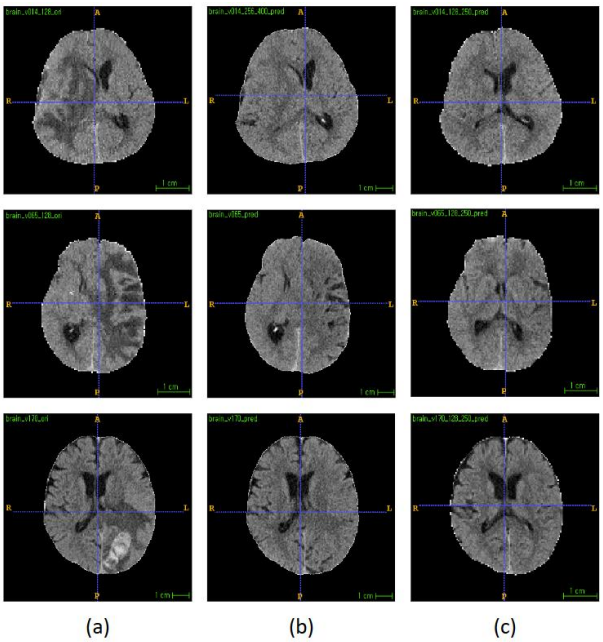

Figure 3. Scans of three cases of brain anomalies, each row corresponding to a different case. Within each case, Column (a) presents the original scans exhibiting anomalies, Column (b) shows the predictions only consider texture, and Column (c) displays the predictions that consider texture and structure. The display windowing is [0, 80] HU

So, is the texture model a bad thing? It should not be. Lesions in the brain have varying intensities, morphologies, locations, and sizes. However, even without lesions, the normal distribution of the gyrus, sulcus, and ventricle is highly randomized across individuals due to the nature of the brain. Look at Figure 4, which displays three other different cases of brain anomalies. Even if the anatomical model predicts the normal distribution, the brain structure of the prediction could be different from the original patient's uniquely characterized brain structure, like gyrus or sulcus, resulting in false positives in residual error as anomalies. These false positives can be measured with larger anomaly scores than the lesions, especially with small intensity differences from surroundings or small size. Therefore, preserving the structure and selecting only the anomalies as the view of the texture model is also important.

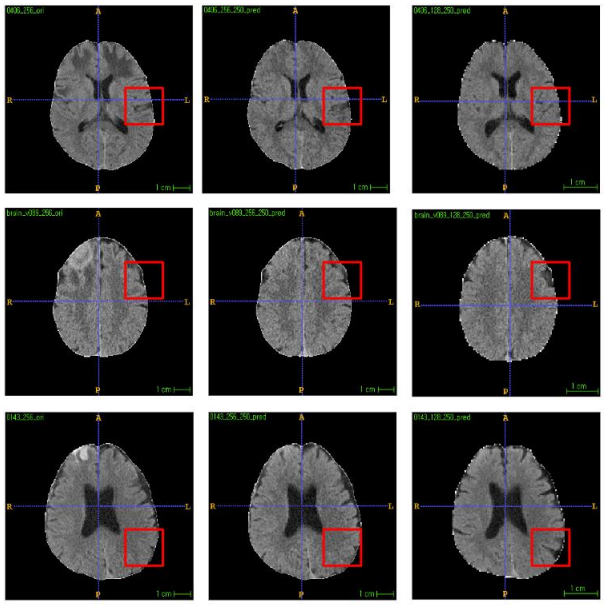

Figure 4. Scans of three cases of brain anomalies, each row corresponding to a different case. Within each case, Column (a) presents the original scans exhibiting anomalies, Column (b) shows the predictions only consider texture, and Column (c) displays the predictions that consider texture and structure. The display windowing is [0, 80] HU

Given the above, we used to combine the anatomical and texture models to derive accurate abnormality scores. With these anomaly scores, we solved the anomaly detection problem from the perspective of binary classification (normal and emergency) based on anomaly scores rather than segmentation labels. The radiologist in the emergency room assigned the labels normal and abnormal.

Unlike most previous work, which mainly considered the anomaly segmentation perspective, our study gave new insight into anomalies based on anatomical and texture views. We also explored the UAD as a 3D model and tried to find the gain of scaling up. We use real-world datasets; there are abnormal patients in about 10 percent of the whole dataset. The model we adopted is the 3D inflated Hierarchical Diffusion autoencoder [15], which model is

composed of DDPM and a semantic encoder.

Our contribution can be summarized as follows.

- We proposed decision-making methods for classifying normal and abnormal patients, which consider anatomical and texture, using real-world datasets with internal and external validation.
- We explored 3D UAD.

3. Materials and Method

3.1 Datasets, Image Acquisition, and Preprocessing of the 18F-FP-CIT PET Image Analysis

We retrospectively collected patients who visited the movement disorder clinic at the AMC from January 2005 to March 2022. Patients diagnosed with Parkinson's disease (PD), multiple system atrophy (MSA), progressive supranuclear palsy (PSP), essential tremor (ET), and other types of secondary Parkinsonism were included. The United Kingdom PD Society Brain Bank criteria1, probable MSA criteria from the second consensus statement on the diagnosis of MSA, probable PSP criteria from the Movement Disorder Society criteria for PSP, and the diagnostic criteria for essential tremor were utilized. The clinical diagnoses of secondary Parkinsonism were based on the clinical presentation and imaging findings as evaluated by movement specialists. Only patients who undertook both 18F-FP-CIT PET and brain magnetic resonance imaging (MRI) scans at the AMC and who were assessed by the two designated movement specialists (SJC and SYJ) were included. Exclusions were made for patients with PET and MRI scans over five years apart, significant PET image artifacts, ischemic striatum lesions on PET, scans post-deep brain stimulation surgery, or PET images with a slice thickness of 3 mm. See Figure 5. for details.

18F-FP-CIT was synthesized according to the method previously described5. 18F-FP-CIT PET images were acquired 180 minutes after administration of 185 MBq 18F-FP-CIT intravenously, utilizing two scanners: the Biograph TruePoint 40 (Bio40) and Vision 600 (Vision) (Siemens, Knoxville, TN, USA). These scanners achieve an in-plane spatial resolution of 2.0 mm full width at half maximum at the center of the field of view. A low-dose brain computed tomography (CT) scan, typically at 120 kVp and 20 mAs with a slice thickness of 1.5 mm, was performed immediately prior to the PET imaging to assist in image fusion and attenuation correction. The PET scanning duration was 10 minutes for Bio40 and seven minutes for Vision, both conducted in three-dimensional mode. For image reconstruction, we employed the TrueX algorithm for Bio40 and also time-of-flight implementation for Vision, using all-pass filters across matrices of 336 × 336 (Bio40) and 440 × 440 (Vision). All brain MRI T1 images were acquired in the axial orientation with parameters presented as median [interquartile ranges] values, reflecting parameter variations resulting from the study's retrospective design: TR 450.0 ms [6.5, 450.0], TE 8.0 ms [3.0, 10.8], flip angle 69.0° [9.0, 70.0], x, y-voxel spacing 0.5 × 0.5 mm [0.4, 0.5], slice thickness 5.0 mm [2.0, 5.0], and spacing between slices 7.0 mm [2.0, 7.5].

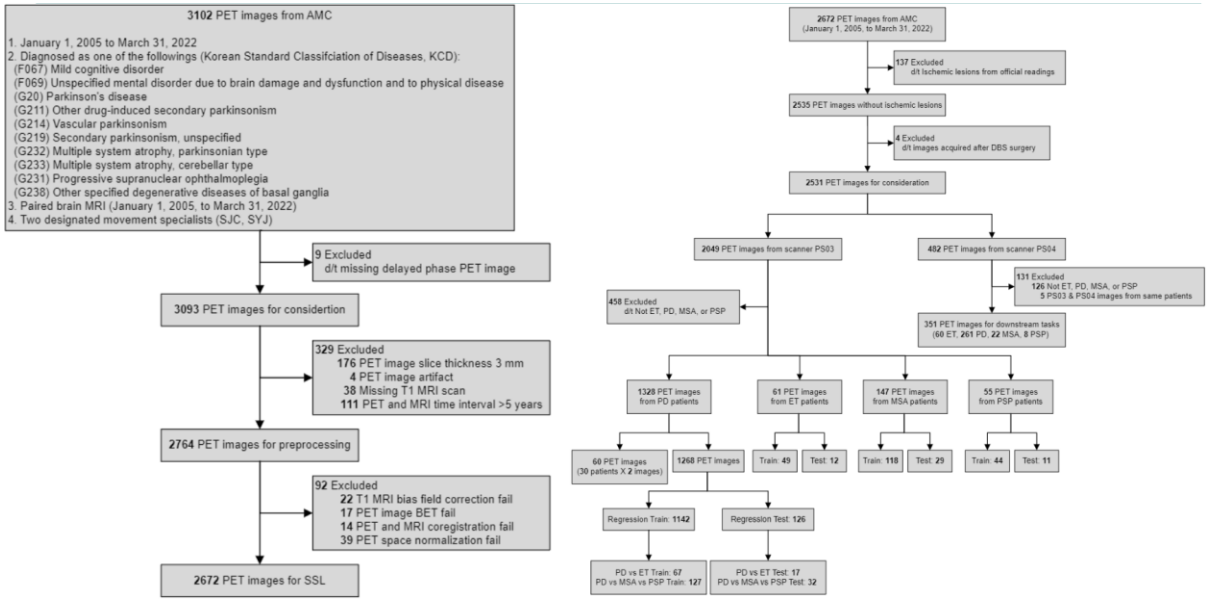

Figure 5. Schema of data collection and the labeling criteria used in the 18F-FP-CIT PET image analysis.

We first converted PET and MRI images from the DICOM to the Nifti format and removed skulls using SynthStrip6 for PET and HD-BET [17] for MRI images. After bias field correction, T1 images were co-registered to the corresponding PET images, and then T1 images were normalized to the Montreal Neurological Institute (MNI) template using SPM12 software (Statistical Parametric Mapping, the Wellcome Trust Centre for Neuroimaging). Subsequently, we applied the inverse deformation map obtained from the previous normalization step to the following region-of-interests (ROIs) from the brain atlases defined in the same MNI space: ventral striatum, caudate, putamen, limbic, executive, and sensorimotor striatum from the Oxford-GSK-Imanova structural8 and connectivity9 striatal atlases; calcarine cortex from the Automated Anatomical Labelling Atlas 310. Accordingly, we gained an ROI mask on the native space of the co-registered T1 and PET images. The standardized uptake values (SUVs) of the ROIs were calculated using these masks. For model training and testing, PET images were intensity normalized by dividing by the average SUV of the bilateral calcarine cortices of each image, followed by min-max normalization. Finally, we cropped the PET images to a uniform size of 192 × 192 × 96, centered on the nonzero region of each image.

3.2 Datasets, Image Acquisition, and Preprocessing of Unsupervised Anomaly Detection

34,085 healthy non-contrast brain CT scans were retrospectively collected from Jan. 2000 to Aug. 2018 in a tertiary academic hospital to create the training dataset. Additionally, brain CT scans were from individuals undergoing emergency screenings for suspected neurological issues in the emergency departments (EDs). These 34,085 scans have a mean age of 42.9 years, a standard deviation of 19.6, and a gender ratio of 46.5:53.5. Internal (N=80) and external (N=160) test sets were enrolled with normal and abnormal (1:1 ratio) patients comprising various brain abnormalities such as acute infarctions, brain mass-like lesions, hydrocephalus, and intracranial hemorrhages.

As the pre-processing, we first performed skull stripping using CT BET [46], and the CT volume was resized into 256*256*32 for efficient training. Input images are intensity normalized [-10, 90] HU to [0, 1].

3.3 Diffusion Probabilistic Models

Drawing Inspiration from non-equilibrium statistical physics, Sohl-Dickstein et al. [6] introduced Diffusion Probabilistic Models (DPMs), tractable and flexible generative models capable of matching a data distribution by training to reverse a gradual, multi-step noising process. They derive bounds on the entropy. More recently, Ho et al. [7] showed Denoising Diffusion Probabilistic Models (DDPM), establishing a connection between DPMs and score matching [18][19] with Langevin dynamics. A predefined diffusion process progressively adds random noise in these models, erasing information. The denoising process, which works in the opposite direction, is approximated with a neural network. These models generate samples by gradually removing noise from a signal, and their training objective is formulated as a reweighted variational lower bound.

In the medical domain, there are various applications of DPMs, including image generation, anomaly detection, segmentation, and others.[2] 3D imaging generation and its applications are also beginning to emerge. [20] The progression of PD usually brings volumetric degeneration. Because this property is challenging to observe in 2D imaging, we did generative pre-training in 3D PET images. Also, for the lesions of brain CT that have volumetric properties, we did generative normal training using 3D brain CT images.

3.4 Hierarchical Diffusion Autoencoder

While DPMs beat other generative models with superior image sample quality, unlike GANs or VAEs, their latent variables lack semantic meaning and are useless for other applications. Preechakul et al. [21] searched to extract a linear, semantically meaningful, and decodable representation of an input image via autoencoding, which uses DPMs and proposed Diffusion Autoencoders (DAE). The DAE consists of a semantic encoder and a conditional DDIM. A learnable semantic encoder captures the semantics of input images, and a conditional DDIM is the image decoder that reconstructs the original input images. In the backward process, the conditional DDIM is conditioned on an additional latent vector derived from the semantic encoder. In the medical image domain, DAE is sometimes used as a diagnosis model with good accuracy, interpretability, and meaningful latent space to reflect the progression or disease classes. [22][23].

However, Zeyu Lu et al. [12] found that a semantic latent code of DAE cannot fully reflect the information of hierarchical feature representation from low level to high level, resulting in insufficient image reconstruction. To mitigate those limitations, they propose Hierarchical Diffusion Autoencoders (HDAE) that exploit the fine-grained-to-abstract and low-level-to-high-level feature hierarchy for the latent space of diffusion models. While the DAE puts a single 512-dimensional vector drawn from the last part of the encoder as the condition for every

encoder-decoder block of the DPMs, the HDAE extracts different vectors depending on the resolution of the feature maps of the semantic encoder and then provides them as conditions for every block of the encoder and decoder of the DPMs with the exact resolution. The structure of the semantic encoder is the same as the encoder of a DDPM UNet.

In our 3D inflated DAE, we also experimentally observed that DAE cannot reconstruct the original FP-CIT PET images or brain CT well and even distorted the crucial parts. So, we adopt HDAE as the 3D inflated model with a semantic encoder. For PET datasets, we reconstructed the input data regardless of what type of disease they had, while for brain CT, we only reconstructed normal data as input data.

Like previous works [22] [23], for PD has progressive degenerative properties and other similar but different diseases, HDAE could be a compelling analytical tool for the diagnosis of PD. Also, in the unsupervised anomaly detection task, although DDPMs generate high-fidelity images, the backward reconstruction process could accumulate errors and misalignment, and sometimes, it does not exactly reconstruct the original structure of the scans. To address this issue, conditioning through semantic encoders was devised and applied to restore the reconstruction image close to the original one. [21] [24] [25] In this view, a Hierarchical Diffusion Autoencoder as the UAD model is a fascinating reconstruction tool.

3.5 Architecture Improvements with Discrete Wavelet Transform in the 18F-FP-CIT PET Image Analysis

Brain imaging often has a three-dimensional tensor, unlike X-ray or fundus images. Three-dimensional images consume a relatively large amount of GPU memory during the training process, limiting the model's size. In the case of PET image resolution 192*192*96, we observed that the initial channel of HDAE is only 8 channels and DDAE is 16 channels based on 2 batch sizes on a 48GB GPU. In [26] [27] [28], they used Discrete Wavelet Transform in DPMs for fast inference conditional probabilities as the method to generate realistic high-resolution images and to avoid high-resolution feature maps. We replaced the learnable first layer of DDAE and HDAE with preprocessing using the discrete wavelet transform to increase the model's size and enhance image recognition performance, which was named WDDAE and HWDAE. As a result, we could increase the number of initial channels to 128. We decompose a 3D PET scan $y \in R^{192*192*96}$ into 8 wavelet coefficients at half-resolution resolution. See Figure 6 for the HWDAE model architecture.

3.6 Methodological Improvements with Anatomical & Texture Ensemble in Unsupervised Anomaly Detection

For the trait of unsupervised anomaly detection, we need to set the threshold as post-processing. However, there is a trade-off. A high threshold alleviates false positives but passes over small anomalies as false negatives, and a low threshold could detect small anomalies but yield false positives notoriously. We tried to solve this trade-off

problem by combining models.

We built two models independently to cover the anatomical and texture views. Before our study, several studies argued that structure was essential and envisioned SSIM loss in their networks. [29] [30] However, our approach is to make separate and combined models of both perspectives. The anatomical model is an HDAE network trained with a resolution of 128*128*32 to see the coarse feature, the anatomical structure. The anomaly score is then obtained after up-sampling to the original resolution of 256*256*32. The model mainly focuses on coarse anatomical anomalies to detect large anomalies like hydrocephalus or extensive hemorrhages. We applied a relatively high threshold in this model to relieve the false positives.

Unlike large volumetric or intensity difference anomalies, there are anomaly types that have small differences. Sometimes, these anomalies can easily be confused with normal tissues or passed under a threshold that is handled by post-processing. To see these types, we built another texture-focusing network trained with a resolution of 256*256*32. Unlike an anatomical model, which could easily create false positives in brain structures as normal variation, this texture model needs to restrict false positives, which is equal to preserving the unique brain structure of the original patients as much as possible. So, it was necessary for hard conditioning. We determined that the gyrus, sulcus, and ventricle are the most distorted parts during reconstruction. Therefore, we put the above segmentation map [31] as a condition to see the texture while preserving the structure. This model could possibly detect anomalies that have less dense difference HU values, like acute infarctions and intracranial hemorrhage. We applied a relatively low threshold in this model to detect the small volumetric or intensity difference anomalies. Lastly, we combined models as a multi-stage sequence where the texture model is first performed to distinguish normal from abnormal; then, the abnormal group is inferred once more by the anatomical model to add the abnormality scores. See Figure 7 for the architectures of models.

3.7 Implementation Details of the 18F-FP-CIT PET Image Analysis

All experiments, including upstream, downstream, and other ablations, were conducted using a single NVIDIA RTXA6000 48GB GPU. The optimizer for all models is AdamW [32], with optimizer momentum parameters set at 0.9 and 0.999 and a weight decay of 0.05. A cosine decay learning rate scheduler [33]is used. Data augmentation techniques applied include rotation, flip, and zoom-in-out.

For upstream, the learning rate starts at a base value of 4e-5 and reduces to an end value of 2e-5. The loss function across all models is L2 loss. Training parameters show a batch size of 2 for all models, with a warm-up period of 5 epochs on the whole dataset. The initial channels for the models are specified as follows: HWDAE at 128, HDAE at 8, WDDAE at 128, DDAE at 16, SimMIM & DisAE at 48, and PSP at 64. Channel multiplication factors vary across models, with HWDAE using 1, 1.2, 2, and 4; HDAE employing 1, 4, 8, 16, 32, and 64; WDDAE using 1, 1, 2, and 4; DDAE using 1, 2, 4, and 8; SimMIM & DisAE using 1, 2, 4, and 8; and PSP also using 1, 2,

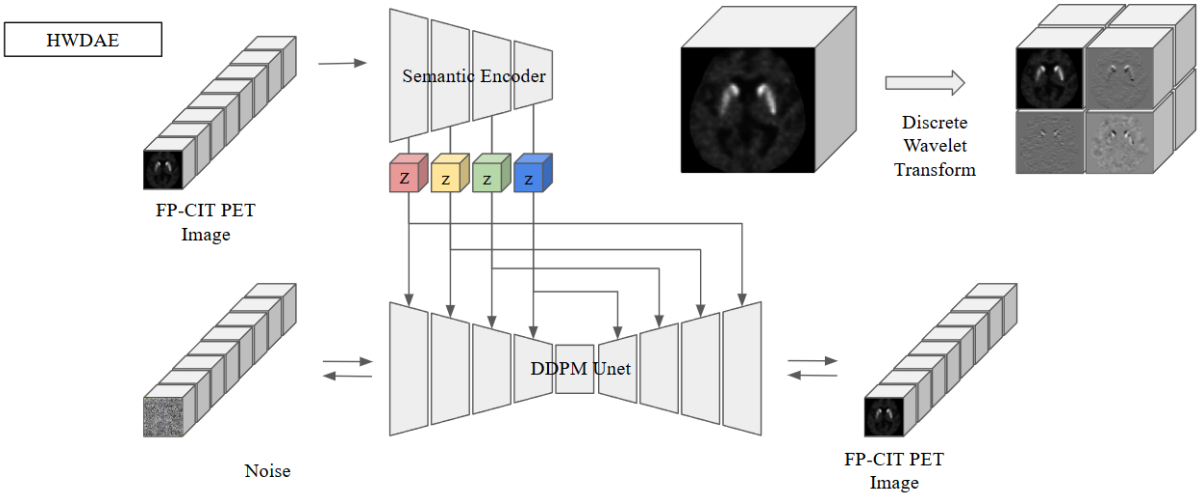

Figure 6. The model architecture of HWDAE in the 18F-FP-CIT PET image analysis. The FP-CIT pet scans decompose and concatenate as channel dimensions first.

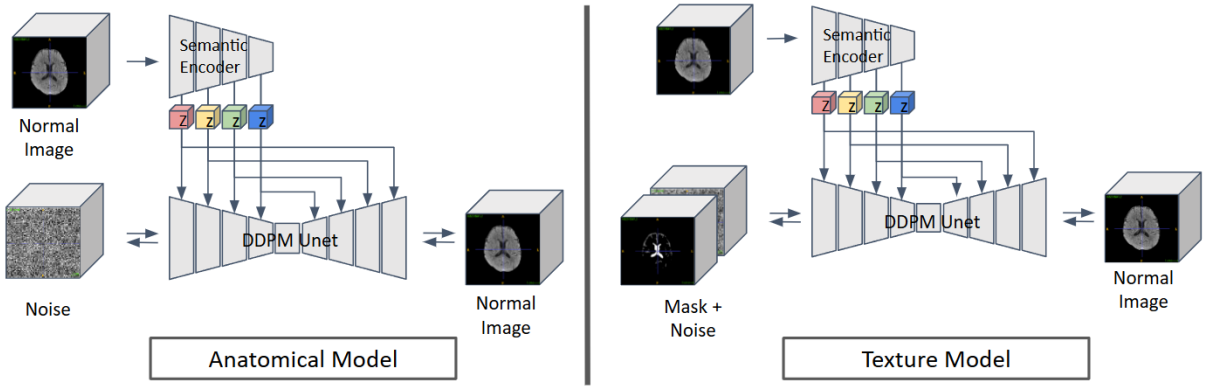

Figure 7. The model architecture of anatomical and texture HDAE in unsupervised anomaly detection. The FP-CIT pet scans decompose and concatenate as channel dimensions first.

4, and 8. The training epochs are set to 400 for HWDAE, HDAE, WDDAE, DDAE, SimMIM, and DisAE, and PSP is trained for 30 epochs. Correspondingly, the training durations are roughly 14 days for HWDAE, 9 days for HDAE, 12 days for WDDAE, 7 days each for DDAE and SimMIM & DisAE, and 7 days for PSP.

Downstream tasks are divided into Classification and Regression, with each having three different setups: Linear Probe, Scratch, and Fine-tuning. In the Classification section, the loss function used is cross-entropy, with a label smoothing value of 0.1. For Regression, the loss function is Huber loss. The selection criteria for weight are micro AUROC for Classification and loss for Regression. The batch size for Linear Probe in both Classification and Regression is set to 6, while for Scratch and Fine-tuning, it is set to the maximum that can fit in a GPU. The learning rate (lr) parameters are specified as follows: end lr is 1e-6 for all setups, base lr varies with 1e-3 for Linear Probe, 5e-5 for Scratch, and 1e-3 for Fine-tuning in Classification, and 1e-4 for all setups in Regression. The warm-up epochs differ, with Classification Linear Probe having 10, Scratch 20, and Fine-tuning 20, while

Regression has 30 for Linear Probe, 60 for Scratch, and 60 for Fine-tuning. The total epochs for training are uniformly set to 50 for Classification and 30 for Regression across all setups.

3.8 Implementation Details of the Unsupervised Anomaly Detection

All experiments were performed on a single NVIDIA RTX A6000 48GB GPU. The optimizer is AdamW [33], with optimizer momentum parameters set at 0.9 and 0.999 and a weight decay of 0.05. A cosine decay learning rate scheduler [34]is used. The base learning rate is 4e-5, and the end learning rate is 2e-5. Total epochs are 10 with 1 warm-up epoch. The texture-HDAE has the initial channels as 16 and channel multiplication factors as 1,2,4,8 and 16. The anatomical-HDAE has the initial channels as 64 and channel multiplication factors as 1,2,4 and 8. Only the block before the bottleneck has self-attention. The batch size is 2 of the texture-HDAE and 5 of the anatomical-HDAE, respectively.

4. Experiments and Results of the 18F-FP-CIT PET Image Analysis

We mainly selected ablation models with pretraining and fine-tuning phases with 3D generative or image reconstruction methods. We compared HWDAE and WDDAE with HDAE, and DDAE does not have a DWT preprocessing step, and SimMIM, Disruptive Autoencoder both have the same swin-transformer-based encoder, but the pretraining method is different. SimMIM has image masking of 0.6 ratio, and DisAE has channel masking of 0.6 ratio with denoising and super-resolution tasks. In addition, we had a question about the encoder-based generative adversarial network(GAN) inversion model, which would be a good classifier. So, we first trained 3D stylegan2 using cascaded methods and second trained 3D Pixel2Style2Pixel, one of the encoder-based GAN inversion models. Lastly, we set the Resnet scratch model as the anchor.

After the pretraining phase on unconditional image generation or reconstruction, we perform three downstream tasks: ET / (early onset) PD Classification, PD/MSA/PSP Classification, and PD patients' motor symptom onset year regression. Lastly, we selected 30 patients who took pet images twice and analyzed the correlation between real-time PD progress and prediction duration. For classification, we use the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC) as evaluation metrics. We use the mean absolute error (MAE) and concordance correlation coefficient (CCC) for regression. We use the Pearson correlation coefficient as an evaluation metric for the follow-up patients.

We conduct linear probing and fine-tuning evaluations with training from scratch to evaluate the feature extract performance. We evaluated all the linear probing, fine-tuning, and scratch with 5-fold cross-validation with the hold-out test set PS03, which is the unseen internal test dataset scanned from the same scanner as the upstream training dataset. PS04 is another unseen test dataset scanned by different scanners in the same clinical center and considered an external set.

4.1 ET/PD Classification

Essential Tremor (ET) and Parkinson's Disease (PD) are both common movement disorders and have tremor and gait difficulty. However, accurate diagnosis is important to treat properly because they require different management and treatment strategies. Several studies classify ET/PD using conventional quantitative analyses, machine learning, and deep learning with the data of UPDRS or medical imaging (FP-CIT SPECT, FP-CIT PET). [34] [35] [36] [37]

We showed the quantitative ET / (early onset) PD binary classification results for pre-trained models. As previous studies usually show more than 90% accuracy, our experiments showed high performance; even linear probing achieved near 1.0 AUROC and AUPRC in both PS03 and PS04. We only perform linear probing except for resnet34 because it already yielded good performance in linear probing. HWDAE achieved the best performance in the AUROC AUPRC. WDDAE generally ranked second. In this task, we observed that DWT could enhance model performance compared to HWDE vs HDAE and WDDAE vs DDAE. See Table 1 and Figure 8 for the results.

Table 1. Mean AUROCs, AUPRCs, and SDs of ET/PD classification. All values are the results of linear probing, except Resnet(scratch). Bold text represents the best linear probing performance. Underlined text is the second-best linear probing performance.

Mean (SD)	ET / PD classification			
Models (Linear Probe)	AUROC		AUPRC	
	PS03	PS04	PS03	PS04
Resnset (**scratch**)	1.000 (0.000)	0.999 (0.001)	1.000 (0.000)	0.999 (0.001)
HWDAE	**0.997** (0.004)	**0.994** (0.001)	**0.998** (0.003)	**0.991** (0.002)
HDAE	0.928 (0.009)	0.949 (0.004)	0.957 (0.994)	0.925 (0.008)
WDDAE	<u>0.987</u> (0.011)	0.979 (0.006)	<u>0.991</u> (0.009)	<u>0.967</u> (0.008)
DDAE	0.969 (0.002)	0.953 (0.010)	0.975 (0.002)	0.936 (0.012)
SimMIM	0.827 (0.025)	0.813 (0.030)	0.842 (0.047)	0.718 (0.038)
DisAE	0.946 (0.004)	<u>0.980</u> (0.002)	0.963 (0.003)	0.966 (0.002)
PSP	0.966 (0.006)	0.966 (0.008)	0.976 (0.005)	0.949 (0.012)

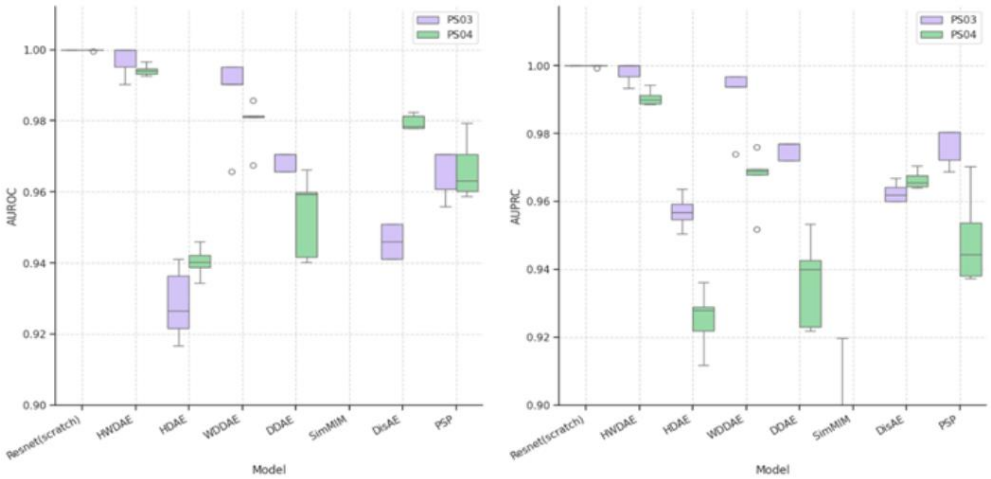

Figure 8. AUROCs, AUPRCs box plots of ET/PD linear probing results on PS03, PS04

4.2 PD/MSA/PSP Classification

In several studies that validated the diagnosis of PD using the pathologic examination at autopsy, they concluded that the general accuracy of diagnosis of Parkinsonian disorders is not satisfying and challenging. About 80% of patients have accurate diagnoses, even movement disorders experts have decided. [38] [39]. The most misdiagnosed atypical parkinsonian syndromes (APS) are multiple system atrophy (MSA) and progressive supranuclear palsy (PSP). Reducing misdiagnosis could avoid inappropriate medicines and treatments. Several studies tried to discriminate PD and APS using conventional methods, machine learning, and deep learning. Zhao et al. introduced DAT-Net for the diagnosis of PD and APS, which showed 0.934 as the average AUROC of PD, MSA, and PSP. They proved that the deep-learning-based method could achieve a more accurate diagnosis than conventional binding ratio quantification. [40]

Our study also confirmed that deep learning classification methods showed high AUROCs and AUPRCs overall. In addition, we proved that a model trained only using a scanner could accurately diagnose PD and APS. We organized the quantitative results of PD/MSA/PSP Classification for pre-trained models. We conduct linear probing except for resnet34. In the PS03, DDAE achieved the best linear probing performance, and HWDAE and WDDAE achieved a comparable performance in the AUROC AUPRC. In the PS04, HWDAE outperformed all other models. In these results, DWT selectively enhances model performance only for HWDAE. See Table 2 and Figure 9 for the results.

In the results of the fine-tuning, HWDAE achieved the best fine-tuning performance in the AUROC and AUPRC except for the AUPRC of the PS04. Although HWDAE and HDAE have similar scratch performance, DWT enhances fine-tuning performance in this case. In other cases, WDDAE showed fine-tuning performance similar to that of DDAE. (Table 2, Figure 9) Like linear probing results, we observed DWT selectively enhances model performance only for HWDAE.

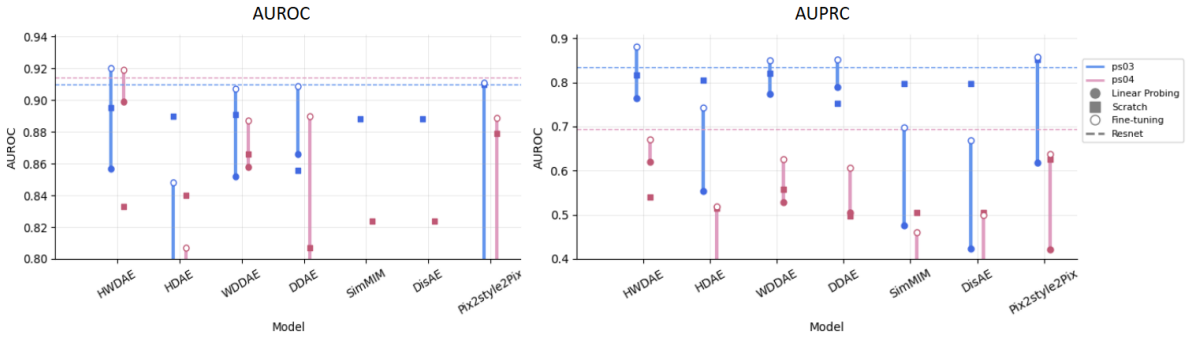

Figure 9. The average AUROCs and AUPRCs bar plots of the results of PD/MSA/PSP. The results of linear probing, scratch, and fine-tuning were evaluated on PS03 and PS04 test sets.

Table 2. Mean AUROCs, AUPRCs, and SDs of the average of PD/MSA/PSP on PS03 and PS04 test sets. All Resnet values are the results of training from scratch. Other models' values have the results of linear probing, training from scratch, and fine-tuning. Bold text represents the best performance, and the underlined text represents the second-best performance.

	Linear Probing				Scratch				Fine tuning			
Models	AUROC		AUPRC		AUROC		AUROC		AUPRC		AUROC	
	PS03	PS04	PS03	PS03	PS04	PS03	PS03	PS04	PS03	PS03	PS04	PS03
Resnset	0.910 (0.013)	0.914 (0.014)	0.834 (0.027)	0.694 (0.057)	0.910 (0.013)	0.914 (0.014)	0.834 (0.027)	0.694 (0.057)	0.910 (0.013)	0.914 (0.014)	0.834 (0.027)	**0.694** (0.057)
HWDAE	0.857 (0.008)	**0.899** (0.015)	0.765 (0.022)	**0.620** (0.015)	0.895 (0.008)	0.833 (0.053)	0.816 (0.040)	0.540 (0.063)	**0.920** (0.017)	**0.919** (0.025)	**0.881** (0.013)	0.670 (0.029)
HDAE	0.722 (0.012)	0.632 (0.024)	0.553 (0.019)	0.370 (0.005)	0.890 (0.020)	0.840 (0.023)	0.805 (0.028)	0.514 (0.022)	0.848 (0.020)	0.807 (0.050)	0.742 (0.037)	0.518 (0.052)
WDDAE	0.852 (0.012)	0.858 (0.013)	0.774 (0.028)	0.528 (0.044)	0.891 (0.022)	0.866 (0.054)	0.820 (0.023)	0.557 (0.056)	0.907 (0.031)	0.887 (0.035)	0.849 (0.030)	0.626 (0.039)
DDAE	**0.866** (0.006)	0.781 (0.010)	**0.789** (0.014)	0.506 (0.015)	0.856 (0.017)	0.807 (0.048)	0.752 (0.030)	0.498 (0.035)	0.909 (0.027)	0.890 (0.042)	0.852 (0.026)	0.606 (0.033)
SimMIM	0.617 (0.024)	0.498 (0.012)	0.476 (0.023)	0.347 (0.007)	0.888 (0.020)	0.824 (0.048)	0.797 (0.029)	0.506 (0.032)	0.798 (0.042)	0.761 (0.062)	0.698 (0.074)	0.460 (0.041)
DisAE	0.576 (0.013)	0.615 (0.011)	0.424 (0.020)	0.364 (0.004)	0.888 (0.020)	0.824 (0.048)	0.797 (0.029)	0.506 (0.032)	0.772 (0.025)	0.798 (0.050)	0.669 (0.054)	0.500 (0.052)
PSP	0.778 (0.017)	0.684 (0.014)	0.619 (0.038)	0.422 (0.013)	0.910 (0.016)	0.879 (0.031)	0.851 (0.025)	0.626 (0.053)	0.911 (0.013)	0.889 (0.010)	0.858 (0.029)	0.637 (0.038)

We also present the result of fine-tuning each for PD/MSA/PSP. HWDAE achieved almost the best fine-tuning performance in the AUPRCs for each class and comparable results in the AUROCs. See Table 3 and Figure 10 for the results.

Table 3. Mean AUROCs, AUPRCs, and SDs of PD/MSA/PSP for each class. The results of fine-tuning were evaluated on PS03 and PS04. Bold text represents the best performance, and underlined text represents the second-best performance.

Mean (std)	PD/MSA/PSP			
	AUROC		AUROC	
Models	Fine-Tuning		Fine-Tuning	
	PS03	PS03	PS03	PS03
Resnset (scratch)	0.902/**0.918**/0.911 (0.021/0.007/0.018)	0.925/0.895/**0.921** (0.02/0.049/0.038)	0.874/0.908/0.721 (0.027/0.009/0.073)	**0.991**/0.722/**0.371** (0.002/0.154/0.07)
HWDAE	0.911/0.909/**0.939** (0.005/0.013/0.011)	0.928/**0.932**/0.898 (0.007/0.004/0.035)	**0.878**/**0.909**/**0.854** (0.013/0.015/0.031)	**0.991**/**0.776**/0.242 (0.001/0.043/0.09)
HDAE	0.802/0.869/0.872 (0.02/0.029/0.038)	0.807/0.838/0.777 (0.076/0.046/0.092)	0.782/0.839/0.605 (0.016/0.036/0.098)	0.971/0.485/0.099 (0.014/0.129/0.038)
WDDAE	0.9/0.899/0.922 (0.023/0.023/0.038)	0.897/0.903/0.861 (0.027/0.012/0.042)	0.847/0.89/0.811 (0.042/0.033/0.032)	0.986/0.616/0.276 (0.005/0.049/0.094)
DDAE	**0.916**/0.901/0.91 (0.014/0.012/0.043)	**0.93**/0.845/0.893 (0.006/0.029/0.03)	0.872/0.896/0.786 (0.018/0.036/0.048)	**0.991**/0.506/0.32 (0.001/0.116/0.077)
SimMIM	0.741/0.784/0.87 (0.049/0.043/0.041)	0.761/0.789/0.732 (0.062/0.067/0.09)	0.722/0.75/0.623 (0.067/0.036/0.132)	0.962/0.34/0.077 (0.013/0.086/0.032)
DisAE	0.747/0.728/0.841 (0.02/0.043/0.044)	0.804/0.806/0.784 (0.048/0.093/0.075)	0.673/0.699/0.635 (0.044/0.067/0.125)	0.969/0.414/0.117 (0.01/0.169/0.049)
PSP	0.905/0.912/0.917 (0.015/0.008/0.028)	0.89/0.857/0.921 (0.013/0.016/0.02)	0.872/0.904/0.799 (0.021/0.025/0.065)	0.984/0.615/0.313 (0.003/0.071/0.086)

4.3 Motor Symptom Onset Year Regression

Information on the disease progression rate is useful in planning clinical trials. However, determining the progression of PD is difficult. PD may not progress constantly because each stages have a different rate. Also, sometimes, they rely on demographic factors that could be different from each other. [41] Despite these challenges, many studies quantitatively revealed the disease progression associated knowledge by using several imaging modalities like MRI, PET, and Free-water imaging. They tried to show the clear temporal relevance of diagnostic and progression imaging biomarkers to be used by clinicians and researchers. [42] Morrish et al. [43] found that putamen influx and clinical rating correlate. Huang et al. [44] found that PD progression was associated with increasing metabolism in the subthalamic nucleus (STN) and internal globus pallidus, as well as in the dorsal pons and primary motor cortex. Gaurav et al. [45] observed a progressive and measurable decrease in neuromelanin-based substantia nigra (SN) signal and volume in PD. They showed the correlation between neuromelanin SN changes and disease severity and duration. We aim to build a deep-learning model of the onset year prediction and use the prediction as a biomarker that could monitor progression. Because the predicted onset year is just from the FP-CIT PET images, the model's onset year output equals degradation on the imaging level. If a PET image has a small predicted onset year, this image has a relatively small degradation. If a PET image has a large predicted onset year, this image relatively has a large degradation.

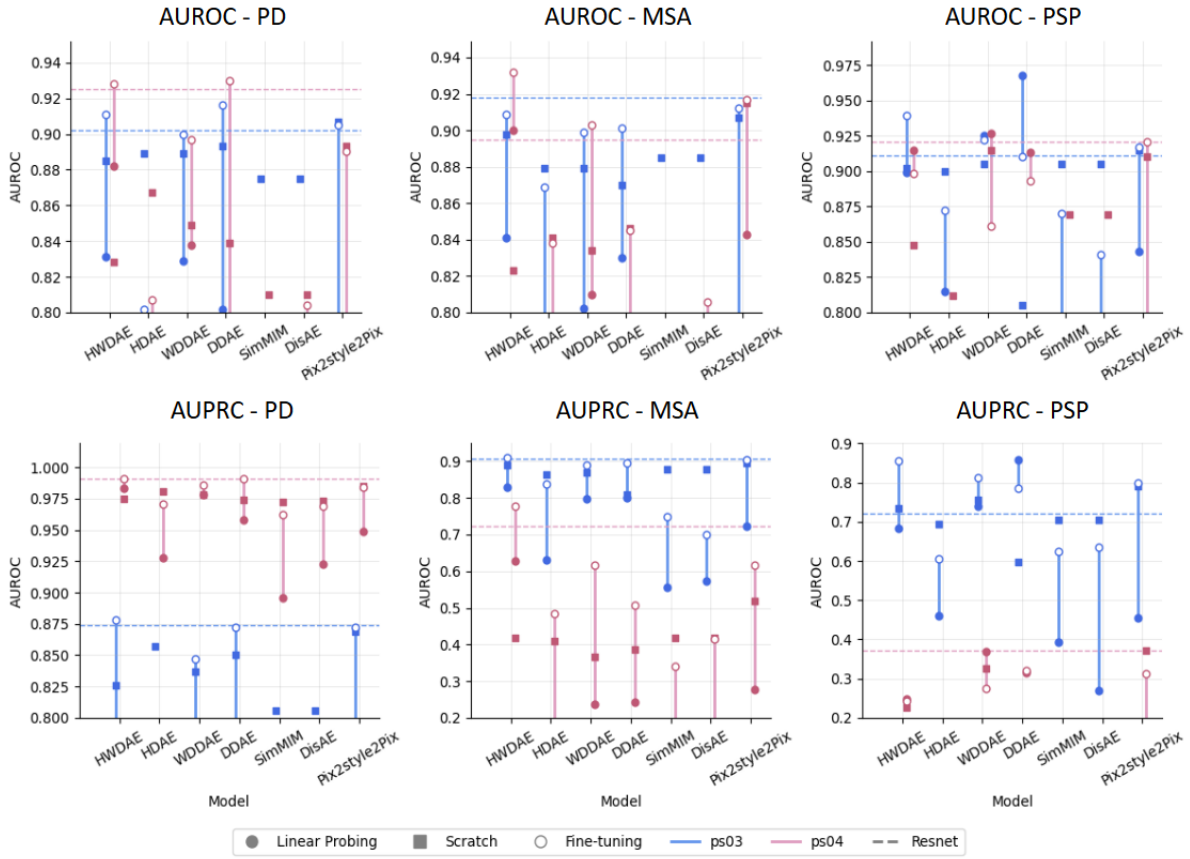

Figure 10. AUROCs and AUPRCs bar plots of the results of PD/MSA/PSP for each. The results of linear probing, scratch, and fine-tuning were evaluated on PS03 and PS04.

Previous studies using dopaminergic PET usually assert three opinions. 1) symptom duration in preclinical and prodromal PD cases has a poor relationship with the change in PET imaging 2) There are large changes in 1 year, with little change between 2 and 4 years 3) striatal dopaminergic markers follow an exponential decline and largely plateau within 5 years of diagnosis.[43] Despite these studies, there are fewer observations about the progression of moderate to late-stage PD. In our case, we observed not only the opinions that are usually asserted in previous papers but also those of late-onset PD patients.

We first showed the qualitative results of onset year regression using scatter plots of Resnet, HWDAE, and WDDAE which showed relatively reasonable results than other ablation models. Both PS03 and PS04, our scatter plots have inclinations 1) From onset 0 years to 4 years, there is little correlation between the actual and predicted onset years. 2) After the onset of 5 years, the distribution of the scatter plot has a slope of less than 45 degrees and is located under the diagonal line. These attributes mean that there is a plateau of decrease of striatal dopaminergic markers. 3) Although the plateau decreases, moderate to late-stage PD decreases steadily. See Figure 11 for the results.

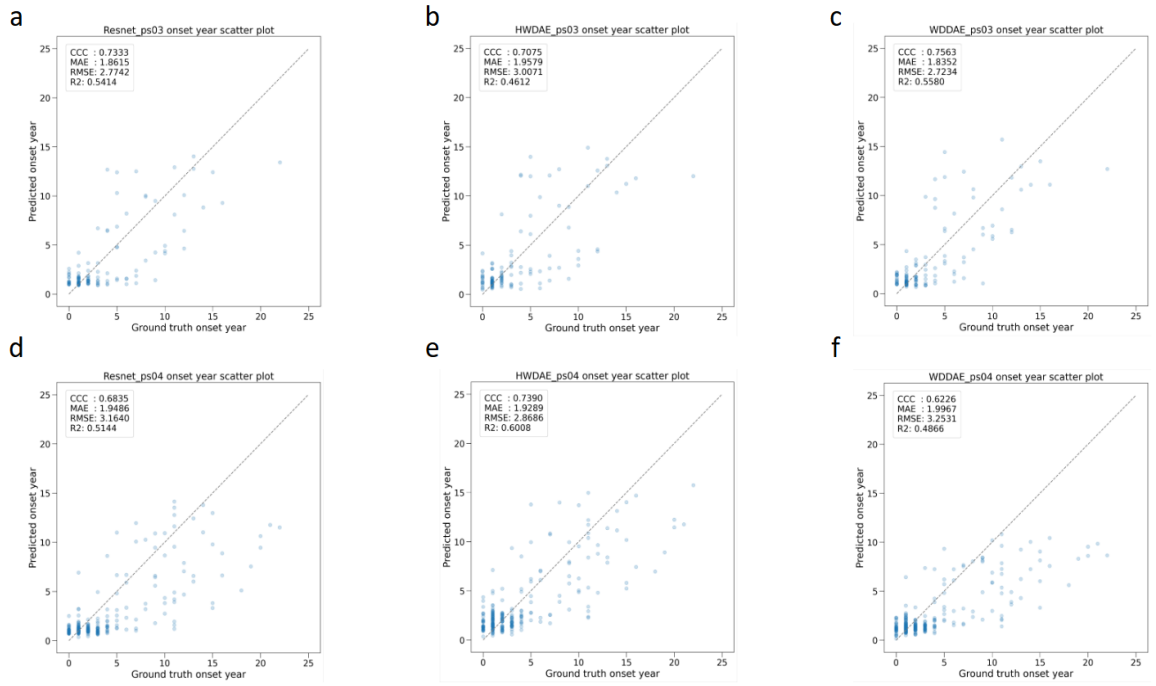

Figure 11. Scatter plots of regression. a, b, and c are Resnet, HWDAE, and WDDAE of PS03, respectively. Scatter plots d, e, and f Resnet, HWDAE, and WDDAE of PS04, respectively.

Table 4. Mean MAEs, CCCs, and SDs on PS03 and PS04 test sets. All values are the results of linear probing. Bold text represents the best performance, excluding Resnet(scratch). Underlined text represents the second-best performance.

Models (Linear Probe) Mean(std)	MAE		CCC	
	PS03	PS04	PS03	PS04
Resnset (scratch)	1.966 (0.027)	2.029 (0.089)	0.711 (0.021)	0.668 (0.051)
HWDAE	**2.294** (0.016)	**2.510** (0.009)	**0.414** (0.015)	**0.397** (0.012)
HDAE	2.509 (0.021)	2.868 (0.028)	0.250 (0.010)	0.182 (0.005)
WDDAE	2.499 (0.004)	2.799 (0.004)	0.180 (0.005)	0.110 (0.004)
DDAE	<u>2.411</u> (0.007)	2.895 (0.041)	<u>0.316</u> (0.007)	0.180 (0.006)
SimMIM	2.766 (0.000)	2.970 (0.001)	0.001 (0.000)	-0.002 (0.000)
DisAE	2.659 (0.004)	2.888 (0.003)	0.055 (0.002)	0.023 (0.001)
PSP	2.416 (0.0220)	<u>2.776</u> (0.089)	0.312 (0.011)	<u>0.358</u> (0.015)

In DDAE, WDAE, we showed the quantitative results of onset regression for pre-trained models. We conducted linear probing except for resnet34. HWDAE achieved the best performance in all the evaluation metrics and in the internal and external datasets. However, in these linear probing results, unlike the classification tasks, WDDAE and DDAE didn't show comparable performance to HWDAE. See Table 4 for the results.

In the results of the fine-tuning (Table 5, Figure 12**)**, WDDAE achieved the best fine-tuning performance in all the evaluation metrics of the PS03, and HWDAE achieved the best of the PS04. DWT enhances fine-tuning performance in both WDDAE and HWDAE. In the linear regression of onset year using SUVR, the results are 2.592(2.084) MAE, 0.490 CCC, and 2.760(2.564) MAE, 0.458 CCC respectively PS03, PS04. The deep learning regression models showed better performance than the classical analytic method using SUVR.

Table 5. Mean MAEs, CCCs, and SDs on PS03 and PS04 test sets from the results of scratch and fine-tuning. Bold text represents the best performance, and underlined text represents the second-best performance.

Mean(std)	MAE				CCC			
Models	Scratch		Fine-Tuning		Scratch		Fine-Tuning	
	PS03	PS04	PS03	PS04	PS03	PS04	PS03	PS04
Resnset (scratch)	1.966 (0.027)	2.029 (0.089)	1.966 (0.027)	<u>2.029</u> (0.089)	0.711 (0.021)	0.668 (0.051)	0.711 (0.021)	<u>0.668</u> (0.051)
HWDAE	2.036 (0.065)	2.06 (0.04)	2.013 (0.039)	**1.965** (0.026)	0.686 (0.017)	0.62 3(0.026)	0.701 (0.013)	**0.733** (0.006)
HDAE	2.075 (0.029)	2.331 (0.105)	2.061 (0.019)	2.188 (0.053)	0.678 (0.011)	0.493 (0.082)	0.661 (0.012)	0.545 (0.024)
WDDAE	2.073 (0.033)	2.412 (0.063)	**1.923** (0.068)	2.055 (0.062)	0.686 (0.024)	0.424 (0.035)	**0.736** (0.01)	0.607 (0.033)
DDAE	2.111 (0.063)	2.412 (0.048)	<u>1.961</u> (0.039)	2.233 (0.04)	0.671 (0.018)	0.41 (0.038)	<u>0.716</u> (0.021)	0.506 (0.036)
SimMIM	2.111 (0.076)	2.532 (0.04)	2.029 (0.071)	2.249 (0.068)	0.650 (0.031)	0.378 (0.033)	0.682 (0.024)	0.54 (0.03)
DisAE	2.111 (0.076)	2.532 (0.04)	2.187 (0.08)	2.336 (0.121)	0.650 (0.031)	0.378 (0.033)	0.608 (0.036)	0.485 (0.052)
PSP	2.078 (0.07)	2.199 (0.142)	2.143 (0.049)	2.258 (0.098)	0.656 (0.026)	0.536 (0.082)	0.662 (0.04)	0.485 (0.065)

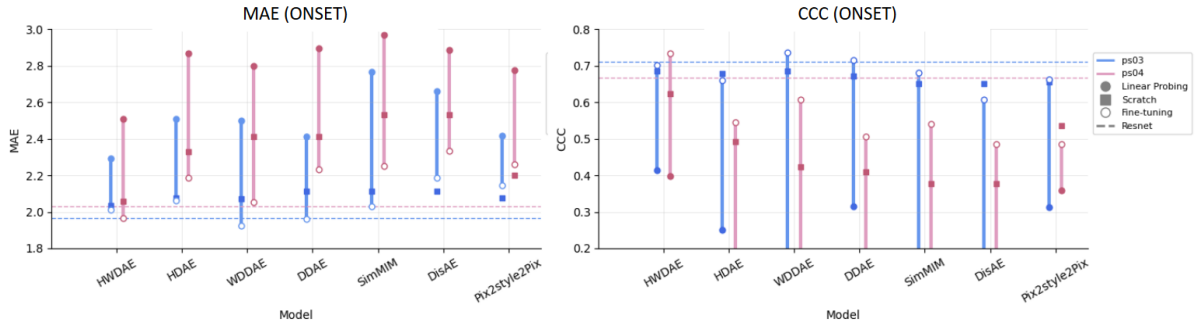

Figure 12. MAEs and CCCs bar plots of the results of onset year regression. This was evaluated on PS03 and PS04 based on the results of linear probing, scratch, and fine-tuning.

4.4 Follow-Up Patients Analysis

Using the regression models, we evaluate the actual duration of PD progression and the predicted duration of PD [43]. For the patients in 30 pairs, 60 images are composed of an initial FP-CIT PET image and a follow-up FP-CIT PET. Those images are scanned from PS03. We calculate the actual duration of PD (onset year of the followed image - onset year of the initial image). Also, we calculate the deep-learning-based predicted duration of PD (onset year of the followed image- predicted onset year of the initial image). We observe there is a mild correlation between actual progress in PD and predicted (Pearson Correlation Coefficient 0.600, p-value<0.001). This means the actual duration of progression and the degradation of PET imaging have a mild correlation. See Figure 13 for the results.

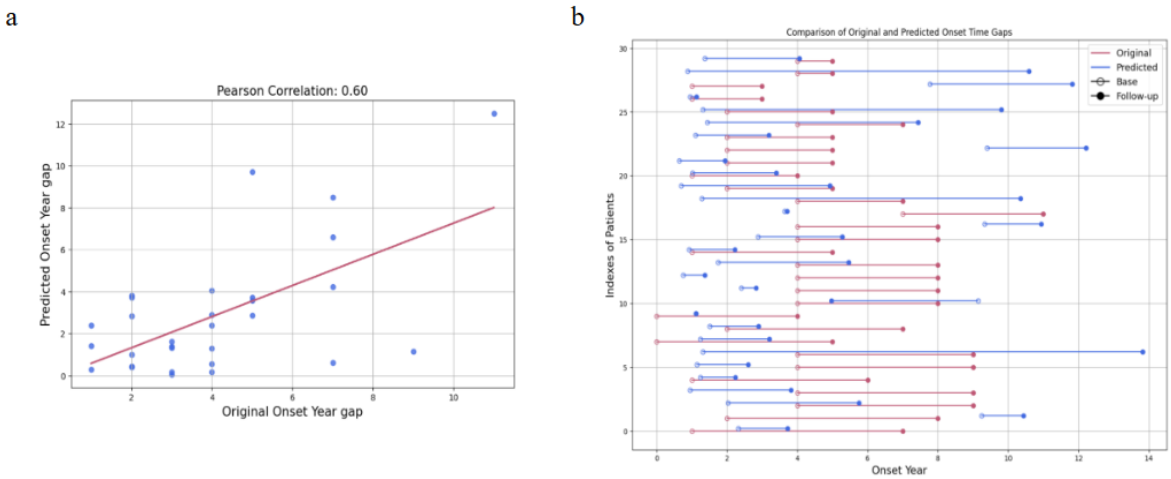

Figure 13. a) A scatter plot of the onset year gap between original and predicted onset years with the Pearson correlation coefficient. b) A plot that indicates longitudinal information has the value of the original, predicted onset year of initial and follow-up images. The length of the line is equal to the duration.

5. Experiments and Results of the Unsupervised Anomaly Detection

After we trained normal data only, we inferred that internal and external test sets compose normal and abnormal images at a ratio of 1:1. After processing these maps, normalization, erosion, median filter, and removing small

objects were performed sequentially. We show some qualitative results of abnormal images (see Figure 14) and normal images (see Figure 15).

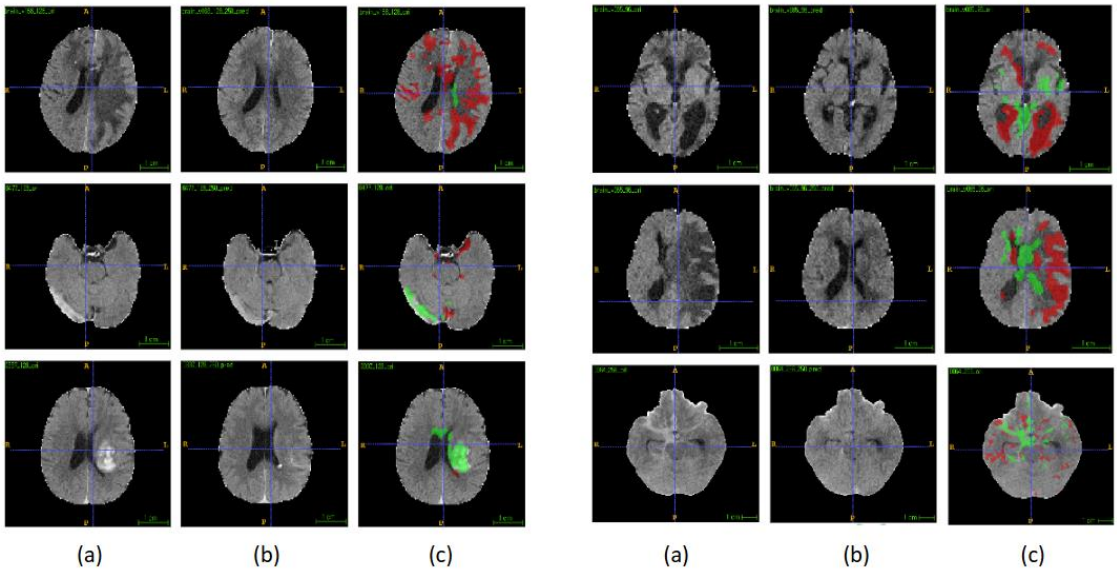

Figure 14. Some cases of abnormal images and their inference results. The first column (a) is of the axial images of original abnormal CT scans as the inputs, the second column (b) is their output normal-like images, and the third column overlaps images of the anomaly map on the original image. The green means the original image has a higher value than the normal-like image, and the red area the original image has a lower value than the normal-like image.

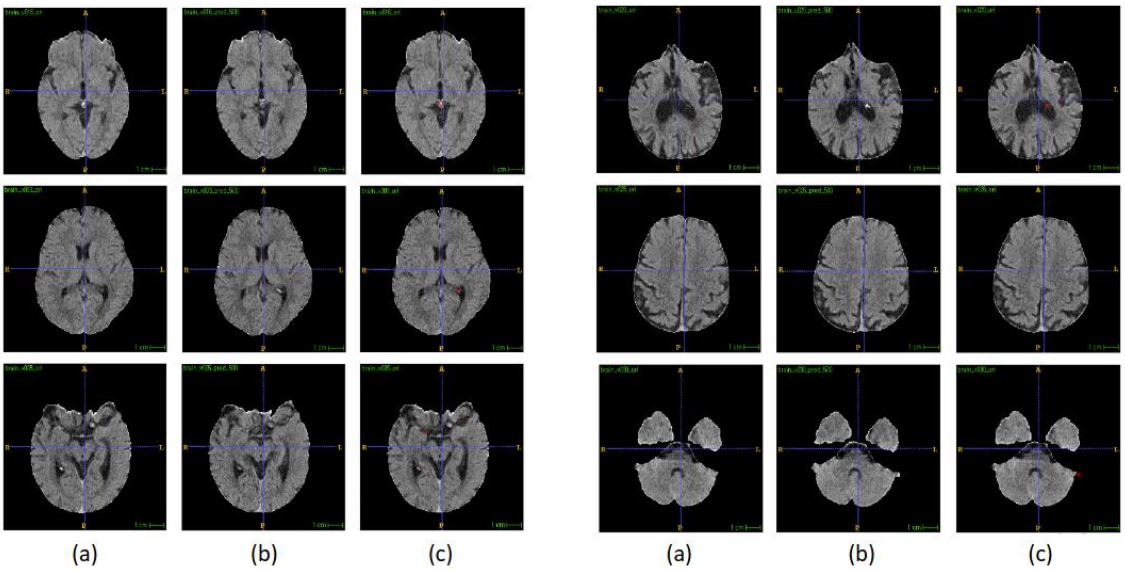

Figure 15. Some cases of normal images and their inference results. The first column (a) is of the axial images of original normal CT scans as the inputs, the second column (b) is their output reconstructed images, and the third column overlaps images of the difference map on the original image. The red area in the original image has a lower value than the normal-like image

In the qualitative inference results of abnormal images (see Figure 13), we could observe the reconstruction images somewhat convert abnormal tissues to normal-like tissues. As can be seen in the various cases, our model was able to detect lesions of various sizes and intensities.

In the qualitative inference results of normal images (see Figure 14), we could observe the reconstruction images with hard conditioning (segmentation mask) preserve normal tissues well. However, some false positives were mainly present in the calcified parts of the ventricles and the outer surface of the brain.

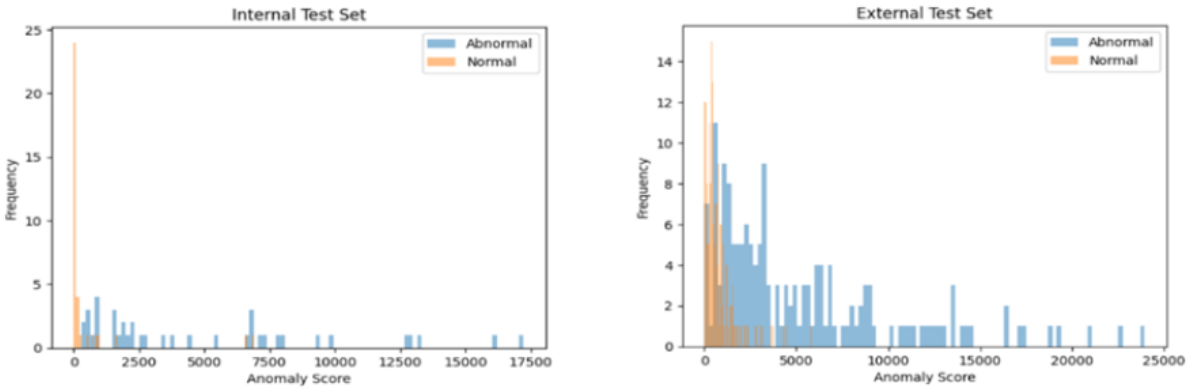

Figure 16. The histograms of anomaly scores in the internal and external test sets. The orange parts are normal anomaly scores, and the blue parts are abnormal anomaly scores.

Table 6. The quantitative results of unsupervised anomaly detection in internal test set and external test set.

	Precision	Recall	F1-score	Accuracy
Internal Test Set	0.93	0.91	0.91	0.91
External Test Set	0.82	0.82	0.82	0.82

In the quantitative inference results of internal and external test sets (see Table 5 and Figure 16), using an optimal threshold, we parted the normal and abnormal images and calculated the precision, recall, f1-score, and accuracy like the binary classification. Unlike the test results of the internal test set, which achieved an accuracy of 0.91, the test results of the external test set relatively achieved a low accuracy of 0.82.

6. Discussion of the 18F-FP-CIT PET Image Analysis

We evaluated diffusion models as generative pre-training methods for FP-CIT PET image recognition tasks. Competent performance with linear probing showed that these models learned good feature representations during the upstream training phase. Our proposed model, HWDAE, demonstrated stable and good performance on both internal and external test sets for classification and regression tasks. Discrete Wavelet Transform preprocessing enhanced feature representation learning by providing multiple channels while conserving memory through down sampling. This preprocessing achieved superior performance in some settings, though occasionally without meaningful gains. Our study has several limitations.

- Consistency of network backbone: Due to limited computational resources, we could not unify the network backbone of the generative models to identical structures. However, we followed guidelines from renowned academic papers, such as those on guided diffusion and simple diffusion, for model structure modifications.
- Computational efficiency: While generative pre-trained models show superior performance in several tasks, the improvements can be relatively small and may even fall short of the supervised model ResNet34, despite their high training costs.
- Class imbalance: Reflecting the real-world prevalence gap among PD, MSA, and PSP, our dataset had considerable class imbalance. Nevertheless, our model, HWDAE, demonstrated relatively uniform performance across all three classes. We also calculated both macro and micro AUC and PRC, with the micro versions accounting for class weights.
- Relationship of symptom onset and FP-CIT PET findings: We recognize that dopaminergic cell loss observed in FP-CIT PET scans does not always correlate with disease progression measured by symptom duration. Nonetheless, our study showed that image features extracted from deep learning models can surpass the conventional SUVR values in predicting disease duration. This result suggests that these features may serve as future markers for disease progression.
- Disease subtypes and unexplored diagnoses of parkinsonism: Our dataset did not include other parkinsonism diagnoses such as corticobasal syndrome and Lewy body dementia. Additionally, disease subtypes of PD, MSA, and PSP, which could present different FP-CIT PET image findings, were not considered in disease classification due to a limited number of data.

7. Discussion of the Unsupervised Anomaly Detection

We built diffusion models (anatomical and texture models) for non-contrast 3D brain CT as normal distribution learners. Using these models, we detected abnormal areas that deviated from the normal distribution. We confirm that our models detected various lesions in the real data sets and even preserved normal tissues well in the reconstruction process. Our proposed models and combining method demonstrated good performance on internal and external test sets for binary classification of normal and abnormal. However, our study has several limitations and needs further study.

- The lack of proper evaluation metrics: we started our suggestions with doubts about the segmentation view and the dice coefficient score. However, even we thought our doubts and suggestions are reasonable, we encountered difficulties in finding proper evaluation metrics that accurately assessed anomalies. Our binary classification is one of the evaluation methods, but it has the limitation of not being able to assess whether the location of the lesion is correct.
- Difficulty in real-world clinical usage: Although we designed our study to closely match real-world

data and settings, we still suffered from the inherent problem of slow inference time, which is a major drawback of DDPM. In addition, we recognize the cumbersome process of adding BET and gyrus sulcus segmentation to be used in actual clinical practice.

8. Conclusion

In this study, we build a generative model HDAE and perform several tasks. We use this model in the 18F-FP-CIT PET image analysis task as a discriminant of image features and in another task of unsupervised anomaly detection of non-contrast 3D brain CT as a data distribution learner (generator). As we observed the results of these tasks, we confirmed that the generative model could achieve good performances even though the tasks have different characteristics. In addition, the fact that brain images have three-dimensional characteristics led us to explore three-dimensional generation models. Despite being less studied than traditional two-dimensional generative models, we observed that the three-dimensional generative model performed well.

References

1. Bond-Taylor, Sam, et al. "Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models." IEEE transactions on pattern analysis and machine intelligence 44.11 (2021): 7327-7347.
2. Kazerouni, Amirhossein, et al. "Diffusion models in medical imaging: A comprehensive survey." Medical Image Analysis 88 (2023): 102846.
3. Radford, Alec, et al. "Language models are unsupervised multitask learners." OpenAI blog 1.8 (2019): 9.
4. Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." Journal of machine learning research 21.140 (2020): 1-67.
5. Huang, Shih-Cheng, et al. "Self-supervised learning for medical image classification: a systematic review and implementation guidelines." NPJ Digital Medicine 6.1 (2023): 74.
6. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015, June). Deep unsupervised learning using nonequilibrium thermodynamics. In International conference on machine learning (pp. 2256-2265). PMLR.
7. Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." Advances in neural information processing systems 33 (2020): 6840-6851.
8. Kong, Yanyan, et al. "Imaging of dopamine transporters in Parkinson disease: a meta-analysis of 18F/123I-FP-CIT studies." Annals of Clinical and Translational Neurology 7.9 (2020): 1524-1534.
9. McFarland, Nikolaus R. "Diagnostic approach to atypical parkinsonian syndromes." CONTINUUM: Lifelong Learning in Neurology 22.4 (2016): 1117-1142.
10. Redgrave, Peter, et al. "Goal-directed and habitual control in the basal ganglia: implications for Parkinson's disease." Nature Reviews Neuroscience 11.11 (2010): 760-772.

11. Mitchell, Trina, et al. "Emerging neuroimaging biomarkers across disease stage in Parkinson disease: a review." JAMA neurology 78.10 (2021): 1262-1272.

12. Lu, Zeyu, et al. "Hierarchical diffusion autoencoders and disentangled image manipulation." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024.

13. Xiang, Weilai, et al. "Denoising diffusion autoencoders are unified self-supervised learners." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

14. Valanarasu, Jeya Maria Jose, et al. "Disruptive Autoencoders: Leveraging Low-level features for 3D Medical Image Pre-training." arXiv preprint arXiv:2307.16896 (2023).

15. Richardson, Elad, et al. "Encoding in style: a stylegan encoder for image-to-image translation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.

16. Karras, Tero, et al. "Analyzing and improving the image quality of stylegan." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

17. Isensee, Fabian, et al. "Automated brain extraction of multisequence MRI using artificial neural networks." Human brain mapping 40.17 (2019): 4952-4964.

18. Song, Yang, and Stefano Ermon. "Generative modeling by estimating gradients of the data distribution." Advances in neural information processing systems 32 (2019).

19. Song, Yang, et al. "Score-based generative modeling through stochastic differential equations." arXiv preprint arXiv:2011.13456 (2020).

20. Pinaya, Walter HL, et al. "Generative ai for medical imaging: extending the monai framework." arXiv preprint arXiv:2307.15208 (2023).

21. Preechakul, Konpat, et al. "Diffusion autoencoders: Toward a meaningful and decodable representation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

22. Keicher, Matthias, et al. "Semantic latent space regression of diffusion autoencoders for vertebral fracture grading." arXiv preprint arXiv:2303.12031 (2023).

23. Ijishakin, Ayodeji, et al. "Interpretable Alzheimer's Disease Classification Via a Contrastive Diffusion Autoencoder." arXiv preprint arXiv:2306.03022 (2023).

24. Behrendt, Finn, et al. "Guided Reconstruction with Conditioned Diffusion Models for Unsupervised Anomaly Detection in Brain MRIs." arXiv preprint arXiv:2312.04215 (2023).

25. Gao, Qi, et al. "CoreDiff: Contextual error-modulated generalized diffusion model for low-dose CT denoising and generalization." IEEE Transactions on Medical Imaging (2023).

26. Guth, Florentin, et al. "Wavelet score-based generative modeling." Advances in Neural Information Processing Systems 35 (2022): 478-491.

27. Phung, Hao, Quan Dao, and Anh Tran. "Wavelet diffusion models are fast and scalable image generators." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

28. Friedrich, Paul, et al. "WDM: 3D Wavelet Diffusion Models for High-Resolution Medical Image Synthesis." arXiv preprint arXiv:2402.19043 (2024).

29. Behrendt, Finn, et al. "Capturing inter-slice dependencies of 3D brain MRI-scans for unsupervised anomaly detection." *Medical Imaging with Deep Learning*. 2022.

30. Behrendt, Finn, et al. "Diffusion Models with Ensembled Structure-Based Anomaly Scoring for Unsupervised Anomaly Detection." arXiv preprint arXiv:2403.14262 (2024).

31. Cai, Jason C., et al. "Fully automated segmentation of head CT neuroanatomy using deep learning." Radiology: Artificial Intelligence 2.5 (2020): e190183.

32. Loshchilov, Ilya, and Frank Hutter. "Decoupled weight decay regularization." arXiv preprint arXiv:1711.05101 (2017).

33. Loshchilov, Ilya, and Frank Hutter. "Sgdr: Stochastic gradient descent with warm restarts." arXiv preprint arXiv:1608.03983 (2016).

34. Moon, Sanghee, et al. "Classification of Parkinson's disease and essential tremor based on balance and gait characteristics from wearable motion sensors via machine learning techniques: a data-driven approach." Journal of neuroengineering and rehabilitation 17 (2020): 1-8.

35. Cheng, FuChao, et al. "Identifying and distinguishing of essential tremor and Parkinson's disease with grouped stability analysis based on searchlight-based MVPA." BioMedical Engineering OnLine 21.1 (2022): 81.

36. Xiao, Pan, et al. "Combined brain topological metrics with machine learning to distinguish essential tremor and tremor-dominant Parkinson's disease." Neurological Sciences (2024): 1-12.

37. Bajaj, Nin, Robert A. Hauser, and Igor D. Grachev. "Clinical utility of dopamine transporter single photon emission CT (DaT-SPECT) with (123I) ioflupane in diagnosis of parkinsonian syndromes." Journal of Neurology, Neurosurgery & Psychiatry 84.11 (2013): 1288-1295.

38. Bajaj, Nin, Robert A. Hauser, and Igor D. Grachev. "Clinical utility of dopamine transporter single photon emission CT (DaT-SPECT) with (123I) ioflupane in diagnosis of parkinsonian syndromes." Journal of Neurology, Neurosurgery & Psychiatry 84.11 (2013): 1288-1295.

39. Rizzo, Giovanni, et al. "Accuracy of clinical diagnosis of Parkinson disease: a systematic review and meta-analysis." Neurology 86.6 (2016): 566-576.

40. Zhao, Yu, et al. "Decoding the dopamine transporter imaging for the differential diagnosis of parkinsonism using deep learning." European journal of nuclear medicine and molecular imaging 49.8 (2022): 2798-2811.

41. Alves, Guido, et al. "Progression of motor impairment and disability in Parkinson disease: a population-based study." Neurology 65.9 (2005): 1436-1441.

42. Mitchell, Trina, et al. "Emerging neuroimaging biomarkers across disease stage in Parkinson disease: a review." JAMA neurology 78.10 (2021): 1262-1272.

43. Morrish, P. K., G. V. Sawle, and D. J. Brooks. "Clinical and [18F] dopa PET findings in early Parkinson's disease." Journal of Neurology, Neurosurgery & Psychiatry 59.6 (1995): 597-600.

44. Huang, Chaorui, et al. "Changes in network activity with the progression of Parkinson's disease." Brain 130.7 (2007): 1834-1846.

45. Gaurav, Rahul, et al. "Longitudinal changes in neuromelanin MRI signal in Parkinson's disease: a progression marker." Movement Disorders 36.7 (2021): 1592-1602.

46. Akkus, Zeynettin, et al. "Robust brain extraction tool for CT head images." Neurocomputing 392 (2020): 189-195.

Abstract (with Korean)

목적 본 연구에서는 몇 가지 작업을 통해 심층 생성 모델의 두 가지 능력을 검증하고자 했다. 첫 번째는 파킨슨병 환자의 18F-FP-CIT PET 영상 분석에서 영상의 특징을 추출 및 구별하는 능력이고, 두 번째는 3D 뇌 CT 의 비지도 이상 검출에서 데이터 분포를 학습(생성자)하는 능력이다.

방법 첫 번째 연구에는 후향적으로 수집된 2,672 개의 18F-FP-CIT PET 스캔이 포함된다. 생성 방법으로 3D 사전 훈련 모델을 구축하고 선형 검증 및 미세 조정을 사용하여 특징을 구별하는 모델의 성능을 평가했다. 본태 떨림 / 파킨슨병의 이진 분류와 파킨슨병, 다계통 위축증, 진행성 핵상 마비의 다중 클래스 분류를 수행했다. 또한 사전 훈련된 모델을 운동 증상 발병 연도 회귀 과제에 사용했으며, 모델의 성능을 분류와 유사하게 평가했다. 두 번째 연구는 건강한 피험자를 대상으로 후향적으로 수집한 34,085 개의 뇌 CT 스캔을 대상으로 했다. 정상 CT 스캔만을 사용하여 3D 생성 모델을 학습시켰다. 정상 데이터 분포를 학습한 후 이 모델은 정상 분포에서 벗어난 비정상적인 스캔을 감지할 수 있었다.

결과 첫 번째 과제에서 제안된 모델은 본태 떨림 / 파킨슨병의 분류의 교차 검증에서는 수신자 조작 특성 곡선(AUROC) 0.997(내부 검증) 0.994(외부 검증)의 정량결과와 정밀도-재현율 곡선(AUPRC) 0.998(내부), 0.991(외부) 의 정량 결과를 보였다. 파킨슨병 / 다계통 위축증 / 진행성 핵상 마비의 다중 클래스 분류에서는 수신자 조작 특성 곡선에서 0.920(내부 검증), 0.919(외부 검증) 정밀도-재현율 곡선에서 0.881(내부 검증), 0.670(외부 검증)의 정량 결과를 보였다. 회귀 작업에서 모델은 평균 절대 오차(MAE)가 2.013(내부 검증), 1.965(외부 검증) 일치 상관 계수(CCC)는 0.701(내부 검증), 0.733(외부 검증)의 정량 결과를 보였다. 두 번째 과제에서 네트워크는 내부 검증 세트에서 정확도, 정밀도, 재현율, F1 점수가 각각 0.91, 0.93, 0.91, 0.91, 외부 검증 세트에서 82, 0.82, 0.82 로 정상 및 비정상 이미지를 탐지했다.

결론 이 연구는 심층 생성 모델이 기능적 뇌 이미지의 임상적 특징을 판별할 수 있는 분류기가 될 수 있음을 시사한다. 동시에 심층 생성 모델이 구조적 뇌 이미지의 정상 분포 학습자가 되어 이미지가 분포 내에 있는지(정상) 또는 이탈했는지(비정상)를 감지할 수 있음을 확인하였다.