scientific reports

Check for updates

OPEN

Comparative validation of AI and non-AI methods in MRI volumetry to diagnose Parkinsonian syndromes

Joomee Song[1,9], Juyoung Hahm[2,3,9], Jisoo Lee[2,4], Chae Yeon Lim[2,5], Myung Jin Chung[2,6,7], Jinyoung Youn[1], Jin Whan Cho[1], Jong Hyeon Ahn[1]✉ & Kyungsu Kim[2,7,8]✉

Automated segmentation and volumetry of brain magnetic resonance imaging (MRI) scans are essential for the diagnosis of Parkinson's disease (PD) and Parkinson's plus syndromes (P-plus). To enhance the diagnostic performance, we adopt deep learning (DL) models in brain MRI segmentation and compared their performance with the gold-standard non-DL method. We collected brain MRI scans of healthy controls ($n = 105$) and patients with PD ($n = 105$), multiple systemic atrophy ($n = 132$), and progressive supranuclear palsy ($n = 69$) at Samsung Medical Center from January 2017 to December 2020. Using the gold-standard non-DL model, FreeSurfer (FS), we segmented six brain structures: midbrain, pons, caudate, putamen, pallidum, and third ventricle, and considered them as annotated data for DL models, the representative convolutional neural network (CNN) and vision transformer (ViT)-based models. Dice scores and the area under the curve (AUC) for differentiating normal, PD, and P-plus cases were calculated to determine the measure to which FS performance can be reproduced as-is while increasing speed by the DL approaches. The segmentation times of CNN and ViT for the six brain structures per patient were 51.26 ± 2.50 and 1101.82 ± 22.31 s, respectively, being 14 to 300 times faster than FS (15,735 ± 1.07 s). Dice scores of both DL models were sufficiently high (> 0.85) so their AUCs for disease classification were not inferior to that of FS. For classification of normal vs. P-plus and PD vs. P-plus (except multiple systemic atrophy - Parkinsonian type) based on all brain parts, the DL models and FS showed AUCs above 0.8, demonstrating the clinical value of DL models in addition to FS. DL significantly reduces the analysis time without compromising the performance of brain segmentation and differential diagnosis. Our findings may contribute to the adoption of DL brain MRI segmentation in clinical settings and advance brain research.

Parkinson's disease (PD) diagnosis is primarily based on clinical presentation. However, for atypical symptoms called red flags[1], brain magnetic resonance imaging (MRI) is essential for diagnosing Parkinson-plus syndromes (P-plus), such as multiple system atrophy (MSA) and progressive supranuclear palsy (PSP). MRI improves the diagnostic accuracy and can be used for monitoring disease progression[2]. Brain MRI can reveal various features that appear in P-plus but not in PD[2–4]. For instance, patients with PSP show marked midbrain atrophy[5], known as the hummingbird sign. In MSA- Parkinsonian type (MSA-P), the putamen is atrophic, with a flattened lateral border, and shows a hypointense signal on T1-weighted gradient-echo images. Patients with MSA-cerebellar type (MSA-C) show predominant atrophy in the pons and middle cerebellar peduncles, resulting in an increased midbrain-to-pons ratio[6] and a decrease in the magnetic resonance Parkinsonism index[7]. Accordingly, quantitative

[1]Department of Neurology and Neuroscience Center, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea. [2]Medical AI Research Center, Research Institute for Future Medicine, Samsung Medical Center, Seoul, Republic of Korea. [3]Department of Biostatistics, Columbia University, New York, NY, USA. [4]Department of Electrical and Computer Engineering, University of Maryland, College Park, MD, USA. [5]Department of Medical Device Management and Research, SAIHST, Sungkyunkwan University, Seoul, Republic of Korea. [6]Department of Radiology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea. [7]Department of Data Convergence and Future Medicine, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea. [8]Department of Radiology, Massachusetts General Brigham and Harvard Medical School, Boston, MA, USA. [9]These authors contributed equally: Joomee Song and Juyoung Hahm. ✉email: jonghyeon.ahn@samsung.com; kskim.doc@gmail.com

measures of the volume of these brain structures have also been assessed, showing high sensitivity and specificity in differentiating PD from P-plus[8].

Although the diagnostic sensitivity and specificity obtained by evaluating the midbrain area are generally high for differentiating between PSP, MSA, and PD[9], the visual assessment of this area is not quantitative, lacks objectivity, and highly dependent on the physician's skills or image acquisition. Consequently, diagnoses based on visual assessments have shown a broad spectrum of accuracy, even falling below 80%[10–12]. To develop a consistent and quantitative analysis of brain MRI, volumetry of the midbrain area has been used as an optimal predictor for accurate diagnosis[6,8,13,14]. Thus, brain image segmentation has become an important stage in most downstream analyses based on prediction models or automated machine-learning (ML) methods for volumetry and diagnosis.

A trained physician's manual segmentation of brain MRI scans is strenuous and time-consuming, and it requires a highly skilled specialist to correctly identify the brain structures. Various automated techniques using atlas-based or deep-learning (DL) techniques have been developed to overcome these problems. Although automated image segmentation models for the brain show limitations[15,16], FreeSurfer (FS)[17] can extract brain structures with relatively high accuracy. Therefore, FS has been widely adopted as a non-DL automated segmentation method[17–21].

Various automated segmentation methods for brain structures have been developed, but their use in clinical practice is limited, being typically used in one-time studies. This is attributable to the time-consuming and complex process of automated segmentation models compared with physicians' simple visual assessments of brain MRI scans. For instance, the automated FS for segmentation takes more than 4.5 h per patient to segment the brain captured in an MRI scan. This complexity problem occurs because existing automated segmentation methods use atlas-based registration[22–25]. In fact, expressing segmentation as an atlas-based registration problem requires considerable time, and FS must be optimized to obtain a coordinate transformation function suitable for the internal atlas model of each test sample.

An automated model for fast segmentation and diagnosis without involving intricate methods should be developed for clinical use. Although DL segmentation has been developed and used in various fields including brain segmentation[26–28], studies for the efficiency and accuracy of segmenting the specific parts of the brain MRI (e.g., separating midbrain and pons for *Brainstem Substructure* pipeline) of specific neurodegenerative diseases is still progressing. Unlike existing non-DL methods, DL may increase the analysis speed by completing segmentation using only forward computations based on learned parameters without requiring optimization processes such as registration. However, it is difficult to predict whether DL shows performance degradation compared with non-DL methods, especially in segmenting brain MRI of neurodegenerative diseases. Our study is significant because it demonstrates the comparative performance of both DL and non-DL methods in segmenting brain MRI and applies them to diagnosing Parkinsonian diseases. In other words, this study *took a further step by showing the differential diagnosis of parkinsonian diseases using brain segmentation by AI and non-AI models*, not merely comparing the performance of segmentation between AI and non-AI models as previous studies[26].

Recent DL segmentation models are classified into convolutional neural network (CNN) and vision transformer (ViT) architectures. Accordingly, a representative model of each framework, V-Net[29] and UNet transformer (UNETR)[30], respectively, were adopted to perform volumetric 3D image segmentation in this study. The DL models were trained to segment brain structures on MRI scans for the diagnosis of neurodegenerative diseases, and their performances were analyzed and compared with an existing non-DL model, FS. Six brain structures that are important in classifying normal, PD, and P-plus cases were segmented: putamen, pallidum, midbrain, pons, caudate, and third ventricle. The volumes of the segmented areas were subsequently used to differentiate between normal, PD, and P-plus cases. As illustrated in Fig. 1, we compared the disease differentiation accuracy and segmentation time of the DL models with those of FS, which were regarded as the reference (i.e., ground truth) for training the DL segmentation models. Therefore, the key contributions of our study from this comparative analysis are as follows: (1) We demonstrate that the gold standard DL models can extensively decrease FS inference time without compromising diagnostic performance (Table 1) and successfully reproduce the brain part segmentation results of FS in neurodegenerative diseases (Fig. 2), (2) Consequently, we show that DL enables a much less complex segmentation and comparable automatic diagnosis of neurodegenerative diseases as the current non-DL approach (Tables 2 and 3), promising the practical usage of DL-based brain MRI segmentation in the diagnosing or studying neurodegenerative diseases (e.g, differential diagnosis between PD, P-plus, and normal cases).

Results

Segmentation time of brain structures.

Table 1 lists the time required to segment the six brain structures per patient. As mentioned in Section "Brain structure segmentation: baseline with FS", the FS sequentially processes the remainder of the *recon-all* pipeline and the complete *Brainstem Substructure* pipeline. In FS segmentation time, we removed the time consumed for preprocessing (i.e., extracting the skull-stripped image from the original MRI) described in Section "Brain structure segmentation: baseline with FS". The resulting time provides a fair comparison of the total times, because FS and DL models use the skull-striped MRI scan as input to derive the final segmentation results, indicated by bold values in Table 1. When including the pre-processing time, CNN-based V-Net and ViT-based UNETR were 14 and 7 times faster than FreeSurfer, respectively.

To compare with FS processing time, we have added time for using CPU in DL models. The CNN-based V-Net and ViT-based UNETR are considerably faster than FS. On average, V-Net took 3.48 s to segment the six brain structures, and UNETR took 48.14 s using GPU and 51.26 s and 1101.82 s using CPU, whereas FS took approximately 15,735 s using CPU, being approximately 307 and 14 times slower than V-Net and UNETR, respectively. Despite calculating the time using CPU, the DL models were faster than FS by 14 to 300 times. Not

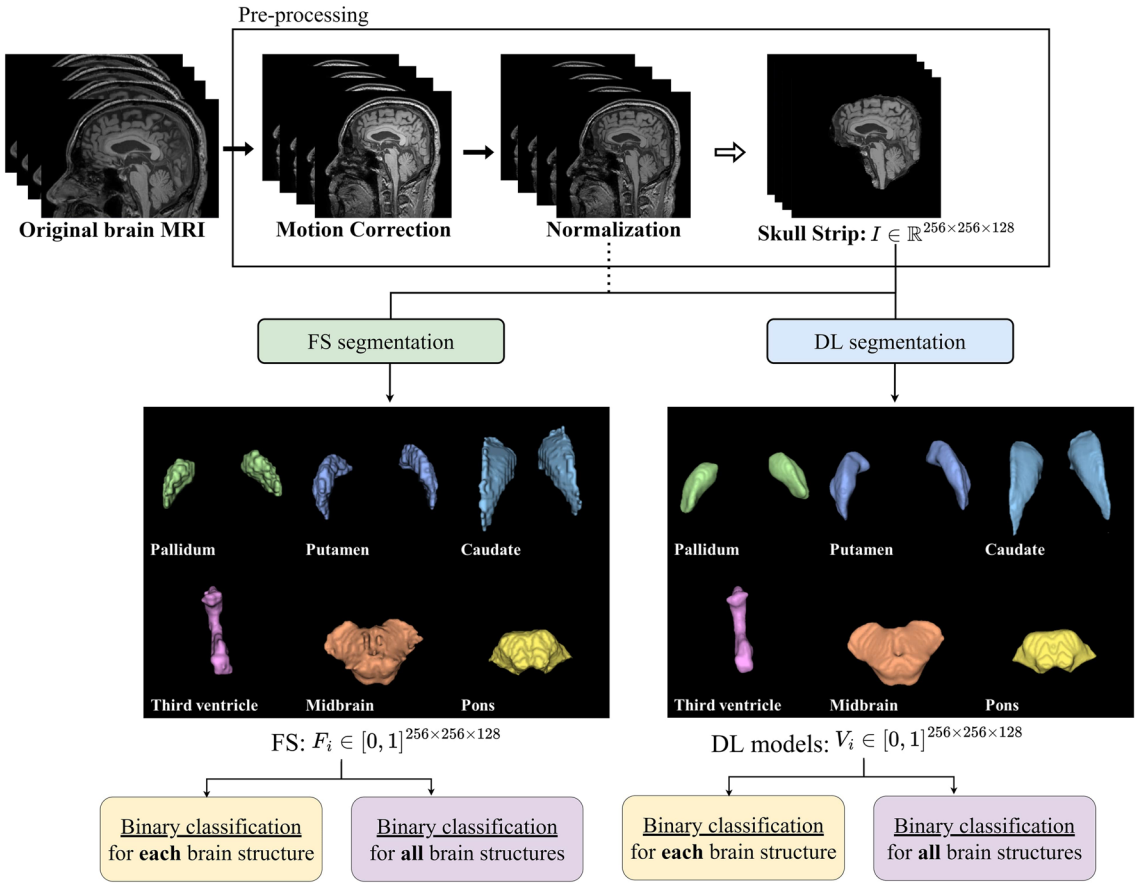

Figure 1. Overview of the study. The diagnostic performance of Parkinsonian syndrome regarding analysis time and accuracy for extracting and segmenting brain structures were compared between DL models and FS. Disease diagnosis was performed using the extracted structures individually or comprehensively.

	CNN (s)	ViT (s)	FS (s)
Midbrain	8.5739 ± 2.50 (0.5827 ± 0.17)	174.37 ± 10.81 (7.5817 ± 0.47)	1698 ± 0.144
Pons	8.5385 ± 2.35 (0.5803 ± 0.16)	207.05 ± 46.46 (9.2242 ± 2.02)	
V3	8.5341 ± 2.35 (0.5800 ± 0.16)	178.25 ± 10.35 (7.7525 ± 0.45)	14,037 ± 1.5
Caudate	8.4590 ± 2.35 (0.5749 ± 0.16)	176.20 ± 10.35 (7.6610 ± 0.23)	
Putamen	8.5561 ± 2.50 (0.5815 ± 0.17)	179.63 ± 10.81 (7.8112 ± 0.47)	
Pallidum	8.6032 ± 2.65 (0.5847 ± 0.18)	186.32 ± 21.39 (8.1019 ± 0.93)	
Total	51.26 ± 2.50 (3.48 ± 0.17)	1101.82 ± 22.31 (48.14 ± 0.97)	15,735 ± 1.07

Table 1. Measured segmentation time per patient obtained by using CNN-based V-Net, ViT-based UNETR, and FS using CPU. GPU running time is shown in the (). The time was calculated after the skull-stripped image was obtained. Data are shown as mean ± standard deviation. (V3, third ventricle).

only GPU-based DL models, but CPU-utilized DL models also demonstrated to have significant performance compared to the non-AI method (i.e., FS).

Dice score of brain structure segmentation using DL models. Segmentation and prediction results of V-Net, UNETR and FS are illustrated in Fig. 2. The Dice score was obtained (Supplementary Table S1) to evaluate the performance of 3D image segmentation. The CNN- and ViT-based models showed high Dice scores above 0.85 for all the brain structures. The Dice scores were higher for the midbrain and pons than for the basal ganglia (i.e., caudate, putamen, pallidum), possibly because the brainstems are surrounded by cerebrospinal fluid and provide a stronger contrast for accurate segmentation. The ViT-based model showed a higher Dice score than the CNN-based model, which in turn showed a much shorter segmentation time than the ViT-based model (e.g., 51.26 s for V-Net and 1101.82 s for UNETR, as shown in Table 1). Although we evaluated V-Net and UNETR in different development environments of TensorFlow and PyTorch, respectively, we expect the CNN-based V-Net to be competitive in speed with the ViT-based UNETR given the segmentation speed difference of

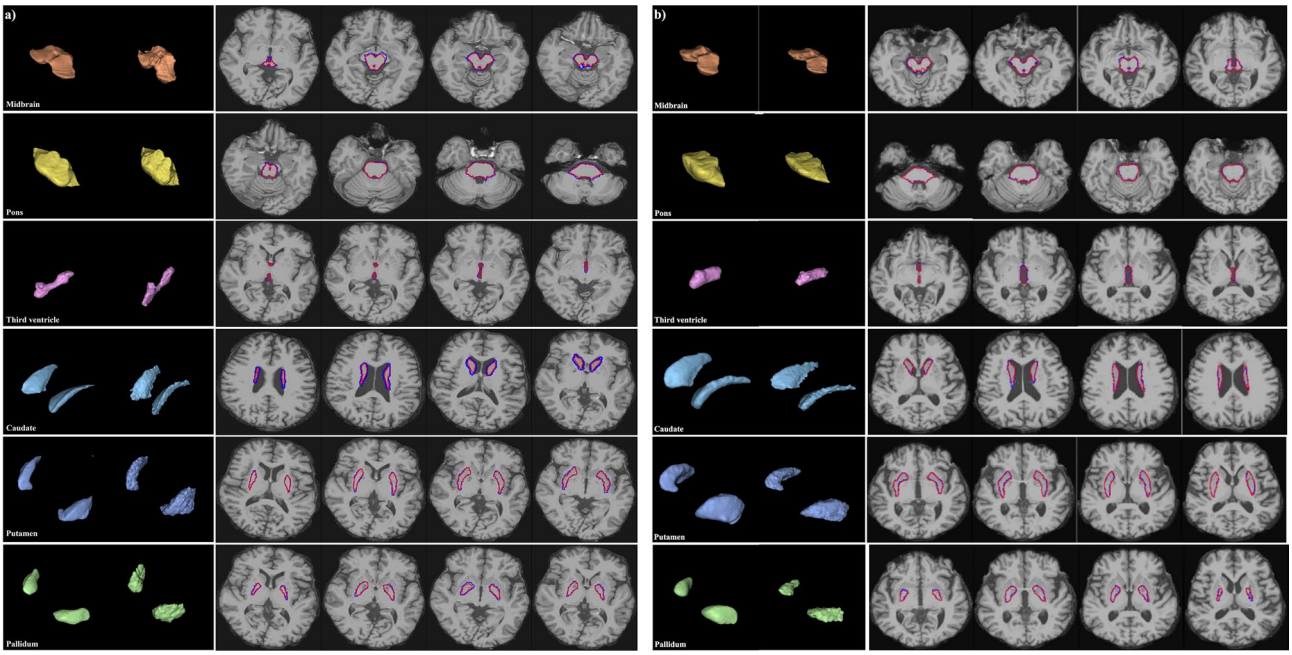

Figure 2. (**a**) Segmentation results of CNN-based V-Net (left 3D images in first column and red-highlighted areas in second column) and FS (right 3D images in first column and blue-highlighted areas in second column) for each brain structure. (**b**) Segmentation results of ViT-based UNETR (left 3D images in first column and red-highlighted areas in second column) and FS (right 3D images in first column and blue-highlighted areas in second column) for each brain structure.

Case		Midbrain	Pons	Midbrain/pons	V3	Caudate	Putamen	Pallidum
Normal vs. PSP	V-Net	0.73 ± 0.06	0.69 ± 0.03*	0.65 ± 0.08	0.83 ± 0.09*	0.54 ± 0.04	0.74 ± 0.01*	0.78 ± 0.05*
	UNETR	0.69 ± 0.06	0.64 ± 0.05*	0.60 ± 0.07	0.84 ± 0.08*	0.57 ± 0.03	0.69 ± 0.03*	0.76 ± 0.05
	FS	0.70 ± 0.06	**0.89 ± 0.05**	0.65 ± 0.08	0.82 ± 0.02	0.56 ± 0.06	0.62 ± 0.09	0.72 ± 0.11
Normal vs. MSA-P	V-Net	0.63 ± 0.05	0.60 ± 0.05	**0.73 ± 0.03**	**0.73 ± 0.03**	0.61 ± 0.02	**0.73 ± 0.03**	0.67 ± 0.10
	UNETR	0.61 ± 0.05	0.67 ± 0.06	0.70 ± 0.05	0.70 ± 0.03	0.59 ± 0.03	**0.73 ± 0.03**	0.65 ± 0.09
	FS	0.64 ± 0.04	0.65 ± 0.04	0.70 ± 0.05	**0.73 ± 0.03**	0.60 ± 0.06	0.70 ± 0.03	0.66 ± 0.08
Normal vs. MSA-C	V-Net	0.76 ± 0.11	0.90 ± 0.04	**0.91 ± 0.02**	0.56 ± 0.02	0.61 ± 0.01	0.65 ± 0.11	0.66 ± 0.09
	UNETR	0.73 ± 0.08	0.86 ± 0.06	0.81 ± 0.15	0.58 ± 0.01*	0.57 ± 0.04	0.62 ± 0.10	0.66 ± 0.07
	FS	0.76 ± 0.10	0.90 ± 0.04	**0.91 ± 0.02**	0.56 ± 0.01	0.62 ± 0.10	0.65 ± 0.10	0.71 ± 0.09
Normal vs. PD	V-Net	0.55 ± 0.02	0.53 ± 0.02	0.57 ± 0.04	0.61 ± 0.07	0.57 ± 0.02	0.55 ± 0.03	0.56 ± 0.02
	UNETR	0.58 ± 0.03	0.55 ± 0.03	0.53 ± 0.04	**0.63 ± 0.06**	0.55 ± 0.04	0.54 ± 0.05	0.54 ± 0.03
	FS	0.56 ± 0.02	0.52 ± 0.01	0.57 ± 0.04	0.61 ± 0.08	0.54 ± 0.02	0.57 ± 0.01	0.54 ± 0.02
PD vs. PSP	V-Net	0.71 ± 0.07	0.67 ± 0.03	0.58 ± 0.07	0.74 ± 0.09	0.54 ± 0.02	0.67 ± 0.03	**0.75 ± 0.03**
	UNETR	0.67 ± 0.09	0.63 ± 0.07	0.57 ± 0.08	0.72 ± 0.10	0.52 ± 0.01	0.63 ± 0.01	0.72 ± 0.05
	FS	0.69 ± 0.06	0.65 ± 0.04	0.58 ± 0.07	0.74 ± 0.11	0.52 ± 0.01	0.62 ± 0.03	0.71 ± 0.08
PD vs. MSA-P	V-Net	0.59 ± 0.07	0.67 ± 0.02	0.58 ± 0.06	0.65 ± 0.08	0.58 ± 0.02	0.67 ± 0.01*	0.65 ± 0.07
	UNETR	0.56 ± 0.02	0.64 ± 0.02	0.67 ± 0.04*	0.59 ± 0.05	0.61 ± 0.03	0.66 ± 0.01*	0.63 ± 0.05
	FS	0.59 ± 0.03	0.66 ± 0.02	0.58 ± 0.06	0.65 ± 0.08	0.59 ± 0.04	**0.69 ± 0.04**	0.66 ± 0.07
PD vs. MSA-C	V-Net	0.74 ± 0.06	0.90 ± 0.03	**0.94 ± 0.02**	0.56 ± 0.08	0.59 ± 0.06	0.50 ± 0.07	0.64 ± 0.06
	UNETR	0.69 ± 0.02	0.82 ± 0.11	0.81 ± 0.16	0.56 ± 0.08	0.58 ± 0.01	0.57 ± 0.05	0.62 ± 0.05
	FS	0.71 ± 0.06	0.90 ± 0.03	**0.94 ± 0.02**	0.57 ± 0.08	0.59 ± 0.06	0.59 ± 0.08	0.69 ± 0.10

Table 2. Disease binary classification based on individual brain structures. Segmentation AUC of CNN-based V-Net, ViT-based UNETR, and FS. Mean ± standard deviation for threefold cross-validation and midbrain-to-pons ratio segmentation are listed. $^*p < 0.05$ indicates a significant difference in AUC between the DL models and FS. The best result for each volume segmentation method based on FS and DL in binary classification is shown in bold.

Case	V-Net		UNETR		FS	
	LR	XGBoost	LR	XGBoost	LR	XGBoost
Normal vs. PSP	**0.89 ± 0.07**	0.86 ± 0.05	0.89 ± 0.08	0.84 ± 0.04	**0.89 ± 0.07**	0.87 ± 0.06
Normal vs. MSA-P	0.78 ± 0.04	0.73 ± 0.003*	**0.81 ± 0.03**	0.77 ± 0.04	0.79 ± 0.001	**0.82 ± 0.01**
Normal vs. MSA-C	0.90 ± 0.03	**0.93 ± 0.04**	0.85 ± 0.12	0.90 ± 0.10	0.88 ± 0.04	**0.95 ± 0.03**
Normal vs. PD	0.60 ± 0.07	**0.66 ± 0.02***	0.60 ± 0.04	0.60 ± 0.03	0.65 ± 0.07	**0.70 ± 0.05**
PD vs. PSP	**0.80 ± 0.08**	0.78 ± 0.001*	0.77 ± 0.13	0.75 ± 0.01	**0.77 ± 0.10**	0.76 ± 0.03
PD vs. MSA-P	**0.76 ± 0.07**	0.66 ± 0.03	0.71 ± 0.02*	0.68 ± 0.07*	**0.79 ± 0.08**	0.71 ± 0.02
PD vs. MSA-C	**0.91 ± 0.04**	0.87 ± 0.03	0.80 ± 0.19	0.80 ± 0.12	0.89 ± 0.07	**0.91 ± 0.05**

Table 3. Binary classification of diseases based on all the brain structures. AUC in LR and XGBoost of CNN-based V-Net, ViT-based UNETR, and FS. The AUC is expressed as the mean from threefold cross-validation. LR; logistic regression, XGBoost; eXtreme Gradient Boosting. The best result for each volume segmentation method based on FS and DL in each binary classification is shown in bold. *$p < 0.05$ indicates a significant difference in AUC between the DL models and FS.

at least 10 times in our experiments. In addition, the CNN-based V-Net had a similar performance to the ViT-based UNETR in actual disease classification, as listed in Table 2.

Binary classification based on individual brain structures. Using the estimated volumes, we performed binary classification for cases normal vs. P-plus, normal vs. PD, and PD vs. P-plus, where P-plus comprised PSP, MSA-P, and MSA-C cases. The AUCs of the brain structures for each model were compared, as listed in Table 2, which also presents the AUC of the midbrain-to-pons ratio[31].

Among the 98 cases (7 cases of binary classification × 2 DL models × 7 cases of brain structures), there was no significant difference in AUC between the DL models and FS, except for 11 cases (i.e., cases where the p-value is less than 0.05). In over half of the 11 cases (i.e., 7 cases), AUCs of the DL models (i.e., CNN-based V-Net and ViT-based UNETR) were also no less than those of FS. This result demonstrated that the DL model reproduces the performance of the FS model successfully (i.e., obtains a performance similar to that of the FS). Furthermore, most of the cases for the CNN-based V-Net showed no lower AUC for disease classification than the cases for the ViT-based UNETR.

The highest AUCs in the comparison between the methods were higher in normal or PD vs. MSA-C (0.91–0.94) than in normal or PD vs. PSP (0.75–0.89). Among the brain structures, the midbrain-to-pons ratio showed the best performance in normal vs. MSA-C and PD vs. MSA-C, while the third ventricle and pallidum showed the best performance in normal vs. PSP and PD vs. PSP. The highest AUCs were not significantly different in the classification of normal or PD vs. MSA-P (0.69–0.73) or PD (0.63).

Binary classification based on complete brain structures. Most AUCs of the DL models were not significantly different from those of FS, as listed in Table 3, although a considerable difference existed in the segmentation speed between the models and FS, as listed in Table 1. In Table 3, the highest AUC of FS and DL models for each binary classification are indicated in bold. The highest AUCs of classification between PD vs. P-plus and normal vs. P-plus were higher than 0.8 in both DL models, except for PD vs. MSA-P (AUC > 0.76). There was no significant difference between FS and the DL models (p-value of 0.05 or higher) in all highest AUCs.

Table 3 shows that of the 28 cases (2 ML models × 2 DL models × 7 binary classifications), most cases (i.e., 24 cases) had no significant differences with FS (i.e., with p-values above 0.05), proving the successful reproducability of the performance of FS by DL models. Like listed in Table 2, the CNN-based V-Net achieved a better AUC than the ViT-based UNETR; in 9 of the 14 pairs of cases, the CNN-based V-Net outperformed the ViT-based UNETR. From the results of both LR and XGBoost, we confirm that considering all six brain structures (Table 3) resulted in a significantly higher AUC than when considering the individual structures (Table 2).

Discussion

We developed two DL models, V-Net and UNETR, which showed significantly faster brain segmentation than FS and a comparable accuracy. Our DL models shortened the segmentation time by 14 to 300 times compared with FS. Moreover, they showed robust high performance in differential diagnosis between PD and P-plus cases using the volume of segmented brain structures. The DL models were efficient (i.e., analysis speed 14 to 300 times faster than FS) and effective (i.e., comparable to FS in Dice score and AUC) in automated brain segmentation and disease diagnosis, even for simultaneous analysis of all brain structures and their individual analyses. Thus, the proposed DL models may promote the application of automated brain segmentation in clinical practice and facilitate efficient and accurate brain research in medicine.

Automated tools have scarcely been adopted for brain segmentation in clinical practice despite their high accuracy in the differential diagnosis of patients with Parkinsonism[13,16]. This is mainly attributable to the complicated and time-consuming process of automated brain segmentation compared with physicians' qualitative visual assessment of brain MRI scans. Consequently, automated segmentation models have mainly been used in research settings that require quantitative brain measurements. Nevertheless, their application in clinical settings may increase with our DL models, which have shown much faster segmentation than FS with a similar accuracy.

The DL models may contribute to improve the accuracy of clinical diagnosis of PD or P-plus cases by providing precise brain image analysis. In addition, clinical trials that require quantitative brain measurement from a large population may be conveniently conducted using our fast and accurate DL models. In the past, methods for brain image analysis were time- and resource-consuming, even with an automated segmentation tool such as FS.

While V-net and UNETR showed significantly faster segmentation in both CPU and GPU, with satisfactory accuracy, the CNN-based V-Net may be more suitable in clinical settings for diagnosis based on volumetry of brain MRI scans. Note that the time was computed without the pre-processing time for the fairness of measuring the time. FreeSurfer's segmentation time corresponds to the time taken in registration-based segmentation for the *recon-all* pipeline and *brainstem substructures* pipeline (Supplementary Fig. S1). Even if the pre-processing time is calculated, CNN-based V-Net and ViT-based UNETR were 14 and 7 times faster than FreeSurfer, respectively. Although the ViT-based UNETR is the most recent DL model and shows a high Dice score, the number of training parameters is approximately 46 times larger than that of V-Net. As presented in Table 1, using CPU may take longer by 14 to 22 times. However, it is evident that compared to FS segmentation time, DL models' processing time is quicker and have equivalent performance with FS. As the number of calculations increases with the number of trainable parameters, the hardware requirements increase in terms of graphics processing unit (GPU) memory and processing power. Consequently, the ViT-based UNETR may be considerably demanding for training and evaluation, requiring high specification GPU. The CNN-based V-Net showed an AUC generally higher than that of UNETR and lower Dice scores. Until the ViT performance is further improved, the CNN-based V-Net, which uses fewer GPU resources, seems to be the best option for clinical practice.

Using gold standard machine learning based approaches, we showed AUC of diagnosis based on FreeSurfer's segmentation and DL method's segmentation to show that there is no significant difference between FreeSurfer's and DL models' segmentation results. Since our DL models are 14 to 300 times faster than FS without sacrificing diagnostic performance, they are superior to FS in terms of clinical efficacy. In binary classification using individual brain structures, the relative order of the AUC of each brain structure was consistent with previously reported results[10,32]. For instance, the pons and midbrain-to-pons ratio showed the highest AUC in classification of normal vs. MSA-C and PD vs. MSA-C cases. The third ventricle and pallidum showed the highest AUC in classification of normal vs. PSP and PD vs. PSP cases. The putamen showed the highest AUC in classification of PD and MSA-P cases. In the classification of PD vs. PSP cases, the third ventricle showed a higher AUC, whereas the midbrain showed a relatively lower AUC. Single measurements of the midbrain have failed to differentiate PSP from PD or MSA[33–35], despite classic MRI studies showing atrophic midbrain in PSP[7,11]. On the other hand, the third ventricle has been shown to be a reliable marker for diagnosing early stage PSP from PD and late-stage PSP[36], and it has been added to a new version of the magnetic resonance Parkinsonism index[37].

For binary classification based on all six brain structures, significant improvements in the AUC were achieved in all models. In both DL models, the highest AUC of classification of PD vs. P-plus and normal vs. P-plus cases was above 0.8, except for PD vs. MSA-P cases. The relatively low AUC of classification between PD and MSA-P cases based on brain MRI cases has also been reported in previous studies[10,32]. The limitation of clinical diagnosis may have contributed to the relatively low AUCs in these studies owing to the overlapping manifestations between PD and MSA-P cases. Clinical diagnosis of PSP and MSA-P has been reported to have the most frequent discrepancy from autopsy-proven diagnosis, even when considering diagnostic criteria[38]. No significant difference in brain MRI scans has been found between normal and PD cases, resulting in no significant AUC differences for classification between these cases.

Our study has some limitations. First, the diagnoses of PD, PSP, and MSA-C were not pathologically verified. Instead, movement specialists provided clinical diagnoses based on validated clinical consensus, providing only probable diagnosis. Second, we segmented six brain structures, namely, midbrain, pons, medulla, putamen, pallidum, and third ventricle, but disregarded other brain structures that may reflect different pathologic characteristics between PD and P-plus (e.g., cerebellum, middle cerebellar peduncle). We excluded those structures owing to the low segmentation accuracy achieved by FS. Also, DL methods learn the coarse features in priority because they are the common region of training data, which are low frequency regions. This results in smoother image than the that of FS, mitigating the images' minor artifacts on the outer rims. However, in the case of cerebellum where smaller changes are essential, more specific study is needed to know whether our DL methods will be applicable to the segmentation of cerebellum's small gyri, comparing them to manual segmentation of the cerebellum. Nevertheless, the differential diagnosis of P-plus using only the brain structures included in this study has been reported as reliable[31]. Third, given memory limitations, we downscaled the output shape from $256 \times 256 \times 256$ to $256 \times 256 \times 128$, which may have caused an information loss. Nevertheless, the Dice scores suggest a negligible impact of information loss, whereas using a downscaled input accelerates training and inference in DL models. Fourth, FreeSurfer does not support GPU (i.e., CUDA) for segmentation, which makes it difficult to compare the time between DL models. We have calculated the segmentation time using CPU and still concluded that DL models are faster by 14 to 300 times.

Automated segmentation of brain MRI scans has become an influential method for diagnosing neurodegenerative diseases, including movement disorders. Using the high-performance CNN- and ViT-based models, we significantly shortened the segmentation time of deep brain structures while obtaining comparable accuracy to the conventional FS segmentation. Despite the superior DL performance, no quantitative results of the comparative analysis and evaluation of the performance of DL have been reported to date for the differential diagnosis of neurodegenerative diseases, including PD and P-plus. We found that the cost-effective CNN-based model achieves satisfactory performance in both segmentation and differential diagnosis compared with the most recent ViT-based model. Our DL models may contribute to the development of patient- and clinician-friendly segmentation methods that enable fast and accurate diagnosis and may provide a meaningful reference for hospitals planning to introduce DL brain segmentation and diagnosis for neurodegenerative diseases.

This study focuses on comparing whether AI is more effective in diagnostic performance than the existing representative non-AI method. Therefore, as the subject of this study is to compare techniques, comparison with clinicians was not performed in this study. It would be a promising future study to compare the accuracy of diagnosis between the machine learning methods and clinicians.

Methods

In this section, we describe the brain MRI data (Section "Data preparation"), FS implementation (Section "Brain structure segmentation: baseline with FS"), and DL method implementation (Section "DL models for brain structure segmentation") for the volumetric analysis of key brain structures to diagnose neurodegenerative diseases. Figure 1 shows an overview of the study process considering the evaluation and comparisons between FS and DL models (i.e., modified V-Net and UNETR representing CNN and ViT DL architectures, respectively). Supplementary Fig. S1 shows a diagram of the overall performance comparison. We developed DL models with faster processing but similar segmentation performance to FS. The DL models were trained to reproduce and segment the results of FS for each brain structure $F_i \in [0,1]^{256\times256\times128}$ as model output $V_i \in [0,1]^{256\times256\times128}$ by taking skull-stripped brain image $I \in \mathbb{R}^{256\times256\times128}$ as input ($i \in \{pallidum, putamen, caudate, third\ ventricle, midbrain, pons\}$), with resolution ($h$, w, d) (height $h = 256$, width $w = 256$, depth $d = 128$). The DL segmentation results for the six brain structures were stored as 3D binary masks (F_i and V_i indicate the FS and DL-model masks for brain structure i, respectively), where each mask output contained intensities between 0 and 1 (area outside and inside the target brain structure, respectively). By calculating the absolute volume of each or all the brain structures predicted by FS or DL models, we performed binary classification of PD, MSA-C, MSA-P, PSP, and normal cases, and calculated the area under the curve (AUC) of segmentation.

Ethical approval. All authors of this study confirm that all methods or experiments were performed in accordance with the Declaration of Helsinki and the relevant guidelines and regulations provided by the policies of the Nature Portfolio journals. This study was approved by the Institutional Review Board of the Samsung Medical Center (IRB number: SMC 2021-07-026). The written informed consent of the patients was waived by the Institutional Review Board of Samsung Medical Center because we used deidentified and retrospective data.

Data preparation. *Study population and clinical assessments.*
We retrospectively screened patients from the Neurology Department of Samsung Medical Center between January 2017 and December 2020. Patients diagnosed with PD, probable MSA, or probable PSP were included in this study. The diagnosis for each patient was determined by movement disorder specialists based on the following criteria: PD was determined according to the United Kingdom PD Society Brain Bank criteria[39] using [18F] N-(3-fluoropropyl)-2β-carbon ethoxy-3β-(4-iodophenyl) nortropane positron emission tomography, while probable MSA and PSP were diagnosed according to the second consensus diagnosis of MSA[40] and movement disorder society clinical diagnostic criteria for PSP[41], respectively. MSA cases were further classified as either MSA-P or MSA-C after reaching consensus[40]. Patients with concomitant or structural brain lesions, including stroke and tumors, which may affect brain MRI scans, were excluded from the study. An age-matched healthy elderly population was included as the control group. Demographic information on age, sex, and disease duration until the brain MRI examination was collected, as listed in Table 4. We analyzed the data from 411 individuals and performed threefold cross-validation to train and evaluate the DL models. Each group consisted of 105 healthy controls and 105 PD, 69 PSP, 69 MSA-C, and 63 MSA-P cases.

We applied cross-validation with three outer folds for evaluation to mitigate bias in the validation and test sets and analyze the effect of set composition (combinations of cases in groups). The data were randomly divided into three sections, one for testing and two for training. Each group comprised 35 normal, 35 PD, 23 PSP, 23 MSA-C, and 21 MSA-P cases.

Data acquisition and standardization. Axial brain MRI scans were acquired using a standard protocol for T1-magnetization-prepared rapid acquisition of gradient echo, with repetition/echo time of 11,000/125 ms, inversion time of 2800 ms, field of view of 240 mm, acquisition matrix size of 320 × 249, echo train length of 27, 1 signal average, slice thickness of 5 mm, interslice gap of 1.5 mm, and scanning time of 198 s.

We included six brain structures that are involved in Parkinsonian syndromes in the gray matter, namely, the midbrain, pons, putamen, pallidum, caudate, and third ventricle. These areas are reported to have the

	PD ($n = 105$)	PSP ($n = 69$)	MSA-P ($n = 63$)	MSA-C ($n = 69$)	Normal ($n = 105$)
Age (years)	69.83 ± 10.14	73.86 ± 7.85	71.58 ± 9.30	64.6 ± 9.04	68.29 ± 9.69
Sex (male)	56 (53.33)	45 (65.22)	46 (66.67)	31 (49.21)	52 (49.52)
Onset to MRI (years)	6.02 ± 6.09	4.68 ± 3.34	4.27 ± 2.82	3.01 ± 2.63	–

Table 4. Demographic and clinical characteristics of patients enrolled in this study. Data are shown as mean ± standard deviation or *n* (%). *PD* Parkinson's disease, *PSP* progressive supranuclear palsy, *MSA-P* multiple systemic atrophy-Parkinsonian type, *MSA-C* multiple systemic atrophy-cerebellar type.

highest sensitivity and specificity for differentiating Parkinsonian syndromes[13,16]. The MRI scans were resized to 256 × 256 × 128 (i.e., number of slices in the coronal/sagittal/axial planes) to segment each structure.

The FS accepts Digital Imaging and Communications in Medicine (DICOM) or Neuroimaging Informatics Technology Initiative (NIfTI) files as inputs. DICOM is a compelling and flexible but complex format that provides interoperability between several hardware and software tools. Given its complexity, DICOM format was converted to NIfTI format[42]. NIfTI is a more straightforward format than DICOM and preserves the essential metadata. In addition, it maintains the volume as a single file and uses raw data after a simple header, and NIfTI files can be loaded and processed faster than DICOM files for whole brain images. Therefore, we converted files in the brain MRI DICOM format into files in the NIfTI format using MRIcroGL.

Brain structure segmentation: baseline with FS. The extraction of brain structures obtained using atlas-based automated segmentation are necessary for training and validation before establishing an automated DL segmentation model. In this study, we used these results as DL ground-truth labels and evaluated the validity of DL model for generating the same label. As a representative technology for atlas-based automated segmentation (see details in Supplementary Section A.2), we selected FS (version 7.2), which is publicly available for neuroscience research and provides high segmentation performance[18–21,43,44]. Additionally, FS no longer supports CUDA, thus unable to calculate the time using GPU.

To segment and extract the six brain structures using FS, it sequentially executes the *recon-all*[45] pipeline and *Brainstem Substructure* pipeline[46]. We used both pipelines because the *recon-all* pipeline does not support segmentation of brainstem structures (e.g., pons and midbrain). However, because the *Brainstem Substructure* pipeline receives pre-processed inputs from the *recon-all* pipeline, both pipelines should be executed. Therefore, the extraction of the six brain structures through FS can be divided into MRI scan pre-processing in the *recon-all* pipeline and the remaining segmentation of the *recon-all* pipeline along with segmentation in the *Brainstem Substructure* pipeline. These processes are explained in Section "Data preparation" and Section "Brain structure segmentation: baseline with FS".

MRI scan pre-processing for FS: motion correction and skull removal. The MRI scan pre-processing in the *recon-all* pipeline of FS mainly consists of (1) motion correction, (2) normalization, and (3) skull stripping. Motion correction is conducted before averaging when various source volumes are used, compensating for small motion variations between volumes. FS constructs cortical surface models and the boundary between white matter and cortical gray matter to automatically match the brain images of patients, using software[17]. In addition, intensity normalization is applied to the original volume. However, adjusting for intensity fluctuations may hinder intensity-based segmentation. Instead, we scale the intensities of all voxels to the mean value (110) of white matter.

After correcting for motions and normalizing the data, FS removes the skull and provides the skull-stripped brain MRI scan. Removing intracranial brain cavities (e.g., skin, fat, muscle, neck, and eyeballs) may reduce human rater variability[47] and promote automated brain image segmentation and improve analysis quality. Therefore, brain MRI scans should be pre-processed to isolate the brain from extracranial or nonbrain tissues in a process known as skull stripping[48]. FS developers devised and applied in-house automated skull-stripping algorithms to isolate intracranial cavities by default.

In this study, the steps of brain MRI scan pre-processing (i.e., skull stripping with motion correction and normalization of a brain MRI scan) took approximately 20 min. We converted the final skull-stripped images to NIfTI files with size of 256 × 256 × 128, while the original brain MRI scan had a size of 256 × 256 × 256, which was adjusted for efficient comparison with the DL models.

FS for brain structure segmentation. After pre-processing (Section "Brain structure segmentation: baseline with FS"), FS segments the six brain structures by applying the remaining processes of the *recon-all* pipeline and the complete *Brainstem Substructure* pipeline. After skull stripping, registration-based segmentation proceeds as follows. FS determines and refines the white and gray matter interfaces for both hemispheres. Then, FS searches for the edge of the gray matter, which represents the pial surface. With pial surfaces, FS expands and inflates sulci banks and gyri ridges. Subsequently, it extends again into a sphere and parcellates the cortex. After applying these processes, FS segments the brain. The *recon-all* pipeline encompasses some brain structures (i.e., putamen, caudate, pallidum, and third ventricle), while the *Brainstem Substructure* pipeline segments the midbrain and pons.

In this study, the final segmentation result was assessed with the same input size of 256 × 256 × 128. The original size of the segmentation result was 256 × 256 × 256, but it was adjusted to 256 × 256 × 128 for comparison with the DL models. In addition, we replaced FS with a DL model applied to the skull-stripped MRI scan (i.e., pre-processing result of the *recon-all* pipeline) to perform segmentation. For the replacement, we evaluated whether the DL analysis is faster than FS analysis and whether the segmentation result of DL is sufficiently reproducible compared with that of FS. The differences between FS and DL segmentation are illustrated in Fig. 2.

DL models for brain structure segmentation. In this study, we used DL models and FS to segment the same skull-stripped images (i.e., images pre-processed by the FS *recon-all* pipeline, as described in Section "Brain structure segmentation: baseline with FS"). The original size of the skull-stripped image generated by FS was 256 × 256 × 256, which was adjusted to 256 × 256 × 128 for DL segmentation owing to the limited GPU memory. We evaluated and compared the performance and analysis time of the DL models by replacing the segmentation process of FS after skull stripping with DL. FS may be inefficient because it segments the entire brain image, requiring many hours of processing. In fact, FS takes at least 4.5 h to segment the six brain structures considered in this study because it requires atlas-based registration to transform the coordinates of the entire

MRI scan to segment specific brain structures. Consequently, FS cannot notably reduce the processing time even if only six brain structures were to be segmented. On the other hand, we verified that DL segmentation (e.g., using V-Net or UNETR) takes less than 1 min to 18 min per case to segment the six target brain structures. As DL models do not require complex registration, unlike non-artificial-intelligence methods (e.g., FS), they can substantially increase the processing efficiency. The implementation details of the DL models are described herein. As DL models, we adopted the CNN-based V-Net[29] and ViT-based UNETR[30] using the segmentation results provided by FS as labels (Section "Brain structure segmentation: baseline with FS"). The two models were trained to reproduce FS segmentation.

CNN-based V-Net. V-Net has been used to segment an entire volume after training an end-to-end CNN on MRI volumes for revealing the prostate[29,49,50] The architecture of V-Net is V-shaped, where the left part of the network is a compression path, whereas the right part decompresses the features until the original input size is recovered. The left part of the network is separated into stages that operate at varying resolutions.

In this study, one to three convolutional layers were used in each step. A residual function was learned at each level. The input of the residual part was used in the convolutional layers and nonlinear operations. This output was added to the last convolutional layer of the stage. The rectified linear unit (ReLU) was used as the nonlinear activation function. Convolutions were applied throughout the compression path. The right part of the network learned a residual function similar to that of the left part. V-Net has shown promising segmentation results, and using this model in our application improved performance. The model was adjusted according to the available memory. The proposed architecture is illustrated in Fig. 3. The left part used a residual block (ResBlock) and maximum pooling (MaxPooling). ResBlock was applied to all the blocks with an input size of $256 \times 256 \times 128$. On the other hand, 3D MaxPooling reduced the depth, height, and width of the feature maps to reduce their resolution. The right part also used ResBlock but replaced MaxPooling with UpConvolution, which consisted of 3D upsampling, batch normalization, ReLU activation, and convolutional layers ($5 \times 5 \times 5$ filter, same padding, and stride of 1). Upsampling increased the resolution of the feature maps, and batch normalization improved convergence throughout the network[51].

ViT-based UNETR. UNETR[30] is a transformer architecture for 3D medical-image segmentation. There is a study that used UNETR as brain tumor segmentation[52], but no study was held for brain parts segmentation. It uses a transformer as the encoder to learn the sequence representations of the input volume and capture global multi-scale information while adopting U-shaped architectures for the encoder and decoder. The proposed architecture is illustrated in Fig. 4. UNETR followed a contracting–expanding path with an encoder comprising a stack of transformers connected to a decoder through skip connections. The encoder directly used 3D patches

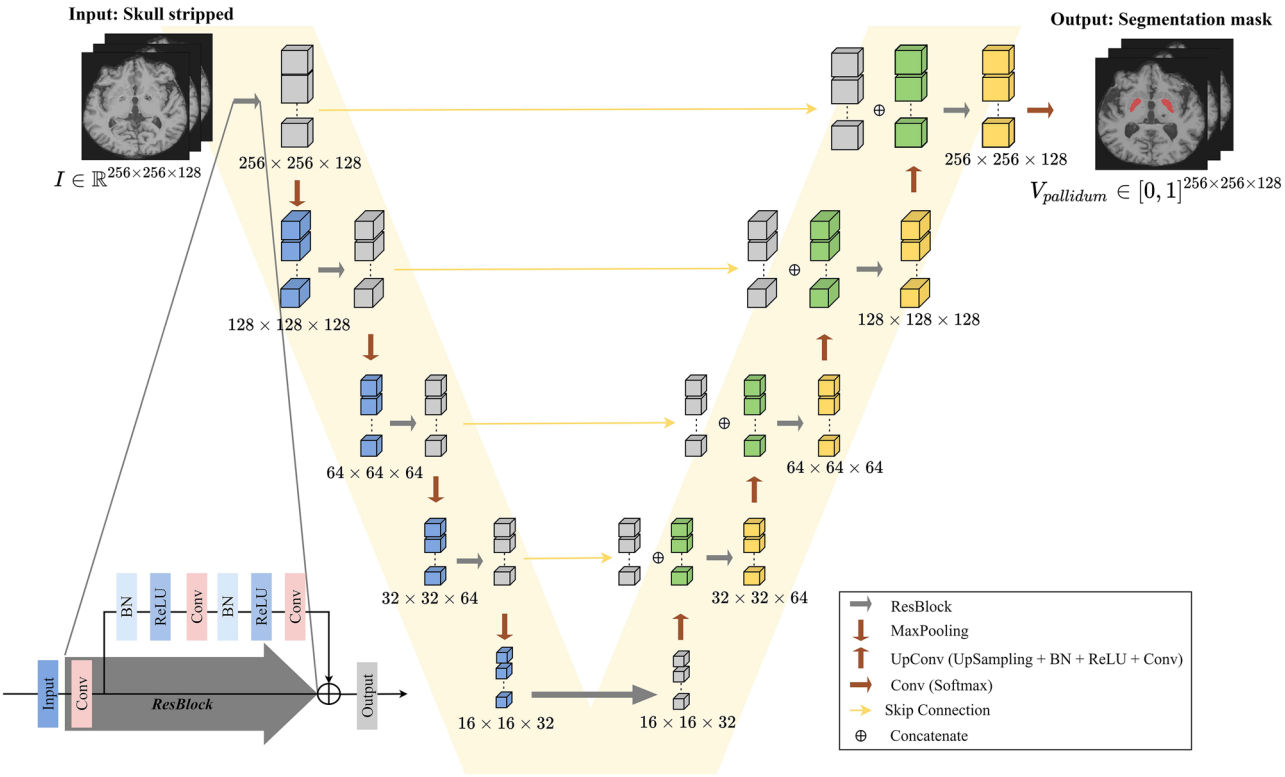

Figure 3. Architecture of CNN-based 3D segmentation using V-Net. ResBlock, MaxPooling, and UpConvolution were used to reduce the depth, height, and width. The output shown in the figure is the segmentation of pallidum. (*Conv* convolution layer, *BN* batch normalization).

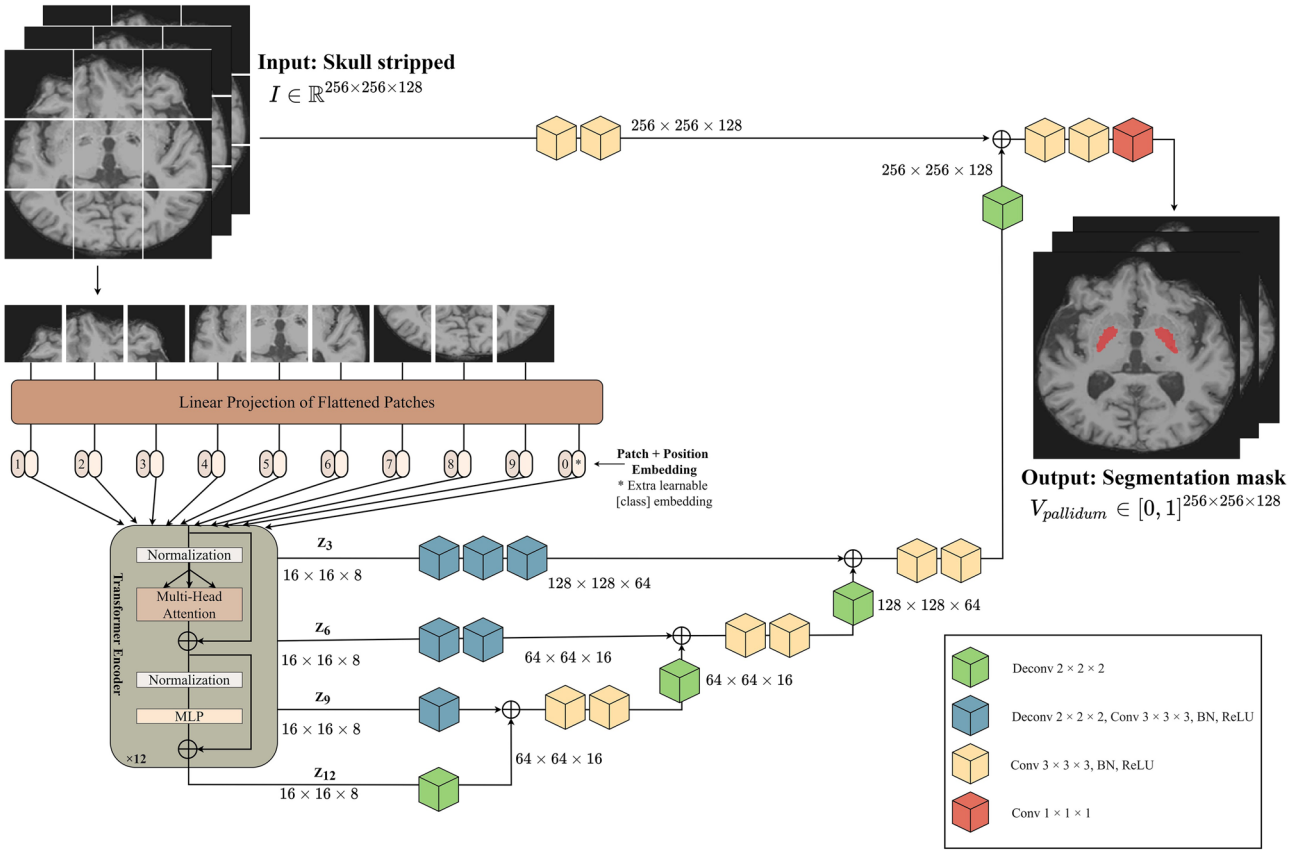

Figure 4. Architecture of ViT-based UNETR directly connected to a CNN-based decoder via skip connections at different resolutions for segmentation. (*Deconv* deconvolution layer, *Conv* convolution layer, *BN* batch normalization, *MLP* multilayer perceptron).

and was connected to a CNN-based decoder via a skip connection. A 3D input volume was split into homogeneous nonoverlapping patches and projected onto a subspace using a linear layer. Position embedding was applied to the sequence and then used as input to the transformer. The encoded representations at different levels in the transformer were retrieved and sent to a decoder via skip connections to obtain the segmentation results.

Implementation details of DL models: training and inference. For the DL models, the input comprised a brain mask and the corresponding patient's segmented brain structures in the MRI scans, which were merged into an array of dimension 256 × 256 × 128. The ground truth of each brain structure was segmented using FS. For evaluation, threefold cross-validation of the test data was applied to calculate the Dice score and Dice loss. We implemented V-Net in TensorFlow and Keras and trained it for 100 epochs. For UNETR, PyTorch and MONAI[53] were applied, and the model was trained for 20,000 iterations. Both models used Python language and were trained using an NVIDIA Tesla V100 DGXS GPU with a batch size of 1 and an initial learning rate of 0.0001. For CPU, Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20Ghz was used.

We evaluated the accuracy of the evaluated models using the Dice score by comparing the expected segmentation with V-Net (or UNETR) and FS outputs. The Dice score measures the overlap between the reference and predicted segmentation masks. A Dice score of 1 indicates perfect spatial correspondence between the two binary pictures, whereas a score of 0 indicates no correlation. We used the Dice loss to determine the performance of the three outer cross-validations on their test sets for the corresponding structures. If F_i and V_i are the ground-truth mask and its prediction for each brain structure, respectively (i.e., FS segmentation mask F_i and its DL prediction mask V_i, respectively, as shown in Fig. 1), the Dice score[54] for each brain structure $i \in$ {pallidum, putamen, caudate, third ventricle, midbrain, pons} is derived as

$$Dice = \frac{2||V_i \circ F_i||_1}{||V_i||_1 + ||F_i||_1}, \tag{1}$$

where $\circ$ denotes the Hadamard product (i.e., component-wise multiplication) and $||\cdot||_1$ is the L1-norm (i.e., sum of absolute values of all components). Moreover, we measured the segmentation time for evaluation.

Statistical analysis for binary classification of cases. We obtained the absolute volumes from the six segmented brain structures (i.e., pons, putamen, pallidum, midbrain, caudate, and third ventricle) predicted by the DL models (i.e., CNN-based V-Net or ViT-based UNETR) or FS. Based on the absolute volume of the individual

brain structures, we calculated the AUC of the binary classification of diseases, normal vs. P-plus, normal vs. PD, and PD vs. P-plus cases. The AUC was computed based on the receiver operating characteristic curve produced by the correlation between the predicted absolute volume of each brain structure and each case.

Disease binary classification was conducted using the six segmented brain structures individually or collectively. For individual analysis, the AUC was derived through thresholding-based binary classification by obtaining the absolute volume of the individual structures. For a comprehensive analysis of all structures, we additionally considered an ML classification algorithm to perform disease binary classification with the six volumes as inputs. For the classification algorithm, binomial logistic regression (LR) and extreme gradient boosting (XGBoost) were used. LR is a statistical model widely used in ML classification[55–57]. XGBoost is a well-established method that produces advanced results among gradient-boosting-based techniques[58] (e.g., XGBoost successfully won 17 out of the 29 ML tasks posted on Kaggle by 2015[59]). In both methods, we evaluated the AUC obtained by the DL model and FS through threefold cross-validation.

Data availability

The authors declare that the main data supporting the results of this study are available within the paper. The raw datasets from Samsung Medical Center are protected to preserve patient privacy but can be made available upon reasonable request provided that approval is obtained from the corresponding Institutional Review Board. For the request for data availability, please contact Jong Hyeon Ahn at jonghyeon.ahn@samsung.com.

Code availability

The code that used for the DL models is available on GitHub: https://github.com/kskim-phd/AI_vs_FS.

Received: 15 October 2022; Accepted: 21 February 2023
Published online: 01 March 2023

References

1. Berg, D. *et al.* MDS research criteria for prodromal Parkinson's disease. *Movem. Disord.* **30**, 1600–1611. https://doi.org/10.1002/mds.26431 (2015).
2. Meijer, F. J. A., Goraj, B., Bloem, B. R. & Esselink, R. A. J. Clinical application of brain MRI in the diagnostic work-up of Parkinsonism. *J. Park. Dis.* **7**, 211–217. https://doi.org/10.3233/JPD-150733 (2017).
3. Watanabe, H. *et al.* Clinical and imaging features of multiple system atrophy: Challenges for an early and clinically definitive diagnosis. *J. Movem. Disord.* **11**, 107. https://doi.org/10.14802/jmd.1802 (2018).
4. Whitwell, J. L. *et al.* Radiological biomarkers for diagnosis in PSP: Where are we and where do we need to be?. *Movem. Disord.* **32**, 955–971. https://doi.org/10.1002/mds.27038 (2017).
5. Jankovic, J., Hallett, M., Okun, M. S., Comella, C. L. & Fahn, S. *Principles and Practice of Movement Disorders E-Book* (Elsevier Health Sciences, Amsterdam, 2021).
6. Hussl, A. *et al.* Diagnostic accuracy of the magnetic resonance Parkinsonism index and the midbrain-to-pontine area ratio to differentiate progressive supranuclear palsy from Parkinson's disease and the parkinson variant of multiple system atrophy. *Movem. Disord.* **25**, 2444–2449. https://doi.org/10.1002/mds.23351 (2010).
7. Quattrone, A. *et al.* MR imaging index for differentiation of progressive supranuclear palsy from Parkinson disease and the Parkinson variant of multiple system atrophy. *Radiology* **246**, 214–221. https://doi.org/10.1148/radiol.2453061703 (2008).
8. Paviour, D. C., Price, S. L., Jahanshahi, M., Lees, A. J. & Fox, N. C. Regional brain volumes distinguish PSP, MSA-P, and PD: MRI-based clinico-radiological correlations. *Movem. Disord.* **21**, 989–996. https://doi.org/10.1002/mds.20877 (2006).
9. Zanigni, S. *et al.* Accuracy of MR markers for differentiating progressive supranuclear palsy from Parkinson's disease. *NeuroImage Clin.* **11**, 736–742. https://doi.org/10.1016/j.nicl.2016.05.016 (2016).
10. Massey, L. A. *et al.* Conventional magnetic resonance imaging in confirmed progressive supranuclear palsy and multiple system atrophy. *Movem. Disord.* **27**, 1754–1762. https://doi.org/10.1002/mds.24968 (2012).
11. Schrag, A. *et al.* Differentiation of atypical Parkinsonian syndromes with routine MRI. *Neurology* **54**, 697–697. https://doi.org/10.1212/WNL.54.3.697 (2000).
12. Kim, Y. E., Kang, S. Y., Ma, H.-I., Ju, Y.-S. & Kim, Y. J. A visual rating scale for the hummingbird sign with adjustable diagnostic validity. *J. Park. Dis.* **5**, 605–612. https://doi.org/10.3233/JPD-150537 (2015).
13. Saeed, U., Lang, A. E. & Masellis, M. Neuroimaging advances in Parkinson's disease and atypical Parkinsonian syndromes. *Front. Neurol.* **1189**, 572976. https://doi.org/10.3389/fneur.2020.572976 (2020).
14. Möller, L. *et al.* Manual MRI morphometry in Parkinsonian syndromes. *Movem. Disord.* **32**, 778–782. https://doi.org/10.1002/mds.26921 (2017).
15. Despotović, I., Goossens, B. & Philips, W. MRI segmentation of the human brain: Challenges, methods, and applications. Comput. Math. Methods Med. https://doi.org/10.1155/2015/450341 (2015).
16. Fawzi, A., Achuthan, A. & Belaton, B. Brain image segmentation in recent years: A narrative review. *Brain Sci.*https://doi.org/10.3390/brainsci11081055 *(2021).*
17. Fischl, B. Freesurfer. *NeuroImage* **62**, 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021 (2012).
18. Dewey, J. *et al.* Reliability and validity of MRI-based automated volumetry software relative to auto-assisted manual measurement of subcortical structures in HIV-infected patients from a multisite study. *NeuroImage* **51**, 1334–1344. https://doi.org/10.1016/j.neuroimage.2010.03.033 (2010).
19. Eggert, L. D., Sommer, J., Jansen, A., Kircher, T. & Konrad, C. Accuracy and reliability of automated gray matter segmentation pathways on real and simulated structural magnetic resonance images of the human brain. *PLOS ONE*https://doi.org/10.1371/journal.pone.0045081 *(2012).*
20. Mayer, K. N. *et al.* Comparison of automated brain volumetry methods with stereology in children aged 2 to 3 years. *Neuroradiology* **58**, 901–910. https://doi.org/10.1007/s00234-016-1714-x (2016).
21. Klauschen, F., Goldman, A., Barra, V., Meyer-Lindenberg, A. & Lundervold, A. Evaluation of automated brain MR image segmentation and volumetry methods. *Hum. Brain Mapp.* **30**, 1310–1327. https://doi.org/10.1002/hbm.20599 (2009).
22. Pham, D. L., Xu, C. & Prince, J. L. Current methods in medical image segmentation. *Annu. Rev. Biomed. Eng.* **2**, 315–337. https://doi.org/10.1146/annurev.bioeng.2.1.315 (2000).
23. Christensen, G. E., Joshi, S. C. & Miller, M. I. Volumetric transformation of brain anatomy. *IEEE Trans. Med. Imaging* **16**, 864–877. https://doi.org/10.1109/42.650882 (1997).
24. Collins, D. L., Holmes, C. J., Peters, T. M. & Evans, A. C. Automatic 3-D model-based neuroanatomical segmentation. *Hum. Brain Mapp.* **3**, 190–208. https://doi.org/10.1002/hbm.460030304 (1995).

25. Iosifescu, D. V. *et al.* An automated registration algorithm for measuring MRI subcortical brain structures. *NeuroImage* **6**, 13–25. https://doi.org/10.1006/nimg.1997.0274 (1997).
26. McClure, P. *et al.* Knowing what you know in brain segmentation using Bayesian deep neural networks. *Front. Neuroinform.*https://doi.org/10.3389/fninf.2019.00067 *(2019).*
27. Rastogi, D., Johri, P. & Tiwari, V. Brain tumor segmentation and tumor prediction using 2D-Vnet deep learning architecture. In *2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART)*https://doi.org/10.1109/smart52563.2021.9676317 *(2021).*
28. Hatamizadeh, A. *et al.* Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In *International MICCAI Brainlesion Workshop*, 272–284 (Springer, 2022).
29. Milletari, F., Navab, N. & Ahmadi, S.-A. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, 565–571, https://doi.org/10.1109/3DV.2016.79 (2016).
30. Hatamizadeh, A. *et al.* Unetr: Transformers for 3D medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 574–584 (2022).
31. Chougar, L. *et al.* Automated categorization of Parkinsonian syndromes using magnetic resonance imaging in a clinical setting. *Movem. Disord.* **36**, 460–470. https://doi.org/10.1002/mds.28348 (2021).
32. Sjöström, H., Granberg, T., Hashim, F., Westman, E. & Svenningsson, P. Automated brainstem volumetry can aid in the diagnostics of parkinsonian disorders. *Park. Relat. Disord.* **79**, 18–25 (2020).
33. Brooks, D. J. & Seppi, K. Proposed neuroimaging criteria for the diagnosis of multiple system atrophy. *Movem. Disord.* **24**, 949–964. https://doi.org/10.1002/mds.22413 (2009).
34. Hotter, A., Esterhammer, R., Schocke, M. F. & Seppi, K. Potential of advanced MR imaging techniques in the differential diagnosis of parkinsonism. *Movem. Disord.* **24**, S711–S720. https://doi.org/10.1002/mds.22648 (2009).
35. Oba, H. *et al.* New and reliable MRI diagnosis for progressive supranuclear palsy. *Neurology* **64**, 2050–2055. https://doi.org/10.1212/01.WNL.0000165960.04422.D0 (2005).
36. Quattrone, A. *et al.* A new MRI measure to early differentiate progressive supranuclear palsy from de novo Parkinson's disease in clinical practice: an international study. *Movem. Disord.* **36**, 681–689. https://doi.org/10.1002/mds.28364 (2021).
37. Quattrone, A. *et al.* A new MR imaging index for differentiation of progressive supranuclear palsy-parkinsonism from Parkinson's disease. *Park. Relat. Disord.* **54**, 3–8. https://doi.org/10.1016/j.parkreldis.2018.07.016 (2018).
38. Rizzo, G. *et al.* Accuracy of clinical diagnosis of Parkinson's disease: A systematic review and meta-analysis. *Neurology* **86**, 566–576. https://doi.org/10.1212/WNL.0000000000002350 (2016).
39. Hughes, A. J., Daniel, S. E., Kilford, L. & Lees, A. J. Accuracy of clinical diagnosis of idiopathic Parkinson's disease: A clinico-pathological study of 100 cases. *J. Neurol. Neurosurg. Psychiatry* **55**, 181–184. https://doi.org/10.1136/jnnp.55.3.181 (1992) https://jnnp.bmj.com/content/55/3/181.full.pdf.
40. Gilman, S. *et al.* Second consensus statement on the diagnosis of multiple system atrophy. *Neurology* **71**, 670–676. https://doi.org/10.1212/01.wnl.0000324625.00404.15 (2008) https://n.neurology.org/content/71/9/670.full.pdf.
41. Höglinger, G. U. *et al.* Clinical diagnosis of progressive supranuclear palsy: The movement disorder society criteria. *Movem. Disord.* **32**, 853–864. https://doi.org/10.1002/mds.26987 (2017).
42. Whitcher, B., Schmid, V. J. & Thorton, A. Working with the DICOM and NIfTI data standards in R. *J. Stat. Softw.* **44**, 1–29. https://doi.org/10.18637/jss.v044.i06 (2011).
43. Heinen, R. *et al.* Robustness of automated methods for brain volume measurements across different MRI field strengths. *PLoS ONE*https://doi.org/10.1371/journal.pone.0165719 *(2016).*
44. Velasco-Annis, C., Akhondi-Asl, A., Stamm, A. & Warfield, S. K. Reproducibility of brain MRI segmentation algorithms: Empirical comparison of local map PSTAPLE, FreeSurfer, and FSL-first. *J. Neuroimaging* **28**, 162–172. https://doi.org/10.1111/jon.12483 (2017).
45. recon-all.
46. Iglesias, J. E. *et al.* Bayesian segmentation of brainstem structures in MRI. *NeuroImage* **113**, 184–195. https://doi.org/10.1016/j.neuroimage.2015.02.065 (2015).
47. Kleesiek, J. *et al.* Deep MRI brain extraction: A 3D convolutional neural network for skull stripping. *NeuroImage* **129**, 460–469. https://doi.org/10.1016/j.neuroimage.2016.01.024 (2016).
48. Kalavathi, P. & Prasath, V. B. S. Methods on skull stripping of MRI head scan images: A review. *J. Dig. Imaging* **29**, 365–379. https://doi.org/10.1007/s10278-015-9847-8 (2015).
49. Bocchetta, M. *et al.* Automated brainstem segmentation detects differential involvement in atypical Parkinsonian syndromes. *J. Movem. Disord.* **13**, 39–46. https://doi.org/10.14802/jmd.19030 (2020).
50. Manjón, J. V. *et al.* pBrain: A novel pipeline for Parkinson related brain structure segmentation. *NeuroImage Clin.* **25**, 102184. https://doi.org/10.1016/j.nicl.2020.102184 (2020).
51. Ioffe, S. & Szegedy, C. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. https://doi.org/10.48550/ARXIV.1502.03167 (2015).
52. Hatamizadeh, A. *et al.* Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In BrainLes@MICCAI (2022).
53. Consortium, M. Monai: Medical open network for ai. *Tech. Rep.*https://doi.org/10.5281/zenodo.6903385 *(2022).*
54. Sheller, M. J. *et al.* Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. *Sci. Rep.*https://doi.org/10.1038/s41598-020-69250-1 *(2020).*
55. Austin, P. C., Tu, J. V., Ho, J. E., Levy, D. & Lee, D. S. Using methods from the data-mining and machine-learning literature for disease classification and prediction: A case study examining classification of heart failure subtypes. *J. Clin. Epidemiol.* **66**, 398–407. https://doi.org/10.1016/j.jclinepi.2012.11.008 (2013).
56. Thabtah, F., Abdelhamid, N. & Peebles, D. A machine learning autism classification based on logistic regression analysis. *Health Inf. Sci. Syst.*https://doi.org/10.1007/s13755-019-0073-5 *(2019).*
57. Nusinovici, S. *et al.* Logistic regression was as good as machine learning for predicting major chronic diseases. *J. Clin. Epidemiol.* **122**, 56–69. https://doi.org/10.1016/j.jclinepi.2020.03.002 (2020).
58. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **29**, 1189–1232. https://doi.org/10.1214/aos/1013203451 (2001).
59. Ogunleye, A. & Wang, Q.-G. XGBoost model for chronic kidney disease diagnosis. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **17**, 2131–2140. https://doi.org/10.1109/TCBB.2019.2911071 (2020).

Acknowledgements

This study was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (2021R1F1A106153511), by the Korea Medical Device Development Fund grant funded by the Korean government (Ministry of Science and ICT, Ministry of Trade, Industry and Energy, Ministry of Health & Welfare, Ministry of Food and Drug Safety) (202011B08-02, KMDF_PR_20200901_0014-2021-02), by the Technology Innovation Program (20014111) funded by the Ministry of Trade, Industry & Energy (MOTIE,

Korea), and by the Future Medicine 20*30 Project of the Samsung Medical Center (SMX1210791). The funders provided support in the form of salaries for some authors but did not have any additional involvement in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of the authors are reported in the corresponding section.

Author contributions

(1) Research Project: A. Conception and design, B. Acquisition of data, C. Analysis and interpretation of data. (2) Manuscript: A. Writing of the first draft, B. Review and critique. (3) Others: A. Statistical analysis, B. Obtaining funding, C. Technical support, D. Study supervision, E. Supervision of data collection. J.S.: 1A, 1B, 1C, 2A, 2B, J.H.: 1C, 2A, 2B, 3A, 3C, J.L.: 2B, 3A, 3C, C.Y.L.: 3C, M.J.C.: 3B, 3C, J.Y.: 1B, 2B, J.W.C.:1B, 2C, 3E, J.H.A.: : 1B, 1C, 2B, 3D, 3E, K.K.: 1A, 1C, 2A, 2B, 3A, 3C, 3D.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-30381-w.

Correspondence and requests for materials should be addressed to J.H.A. or K.K.