scientific reports

Check for updates

OPEN

# Advancements in Frank's sign Identification using deep learning on 3D brain MRI

Sungman Jo[1], Jun Sung Kim[2], Min Jeong Kwon[3], Jieun Park[3], Jeong Lan Kim[4], Jin Hyeong Jhoo[5], Eosu Kim[6,7], Leonard Sunwoo[8,9], Jae Hyoung Kim[8,9], Ji Won Han[2] & Ki Woong Kim[1,2,3,10,11]✉

**Frank's sign (FS) is a diagnostic marker associated with aging and various health conditions. Despite its clinical significance, there lacks a standardized method for its identification. This study aimed to develop a deep learning model for automated FS detection in 3D facial images derived from MRI scans. Four deep learning architectures were evaluated for FS segmentation on a dataset of 400 brain MRI scans. The optimal model was subsequently validated on two external datasets, comprising 300 brain MRI scans each with varying FS presence. Dice similarity coefficient (DSC) and receiver operating characteristic (ROC) analysis were employed to assess model performance. The U-net architecture demonstrated superior performance in terms of accuracy and efficiency. On the validation datasets, the model achieved a DSC of 0.734, an intra-class correlation coefficient of 0.865, and an area under the ROC curve greater than 0.9 for FS detection. Additionally, the model identified optimal voxel thresholds for accurate FS classification, resulting in high sensitivity, specificity, and accuracy metrics. This study successfully developed a deep learning model for automated FS segmentation in MRI scans. This tool has the potential to enhance FS identification in clinical practice and contribute to further research on FS and its associated health implications.**

**Keywords** Frank's sign, Deep learning, Segmentation, MRI

Frank's sign (FS) is a diagonal crease in the earlobe extending from the tragus across the lobule to the posterior edge of the auricle[1–3], believed to result from premature aging and loss of dermal and vascular fibers[4–6]. Previous studies have associated FS with an increased risk of cardiovascular disease[1,2], cerebrovascular disease[7], and cognitive impairment[8]. However, the use of FS as a stand-alone biomarker is limited due to inconsistent replication of associations[9–12], unclear underlying mechanisms[13], and other confounding factors such as age and sleep. In addition, not all FSs may be associated with these diseases, and disease-specific characteristics may vary[14,15].

There are no standardized methods for determining the presence and characteristics of FS. Most studies have relied on visual grading based on shape and bilateralism[16–18], but grading methods vary widely, including 'complete/incomplete'[19], 'mild/moderate/severe', 'bilateral/unilateral'[20], 'diagonal /vertical /unclassifiable'[16], or 5-level grading[17,18]. Furthermore, visual rating has limited inter-rater reliability.

3D brain magnetic resonance imaging (MRI) typically provides images of the face, including both ears, before de-facing[21]. Automatic segmentation of FS from these 3D facial images could lead to more accurate, reliable,

[1]Department of Health Science and Technology, Graduate school of convergence science and technology, Seoul National University, Seoul, South Korea. [2]Department of Neuropsychiatry, Seoul National University Bundang Hospital, Seongnam, South Korea. [3]Department of Brain and Cognitive Sciences, Seoul National University of Natural Sciences, Seoul, South Korea. [4]Department of Psychiatry, College of Medicine, Chungnam National University, Daejeon, Republic of Korea. [5]Department of Psychiatry, School of Medicine, Kangwon National University, Chuncheon, Republic of Korea. [6]Graduate School of Medical Science, Brain Korea 21 Project, Yonsei University College of Medicine, Seoul, Korea. [7]Department of Psychiatry, Institute of Behavioral Science in Medicine, Yonsei University College of Medicine, Seoul, Korea. [8]Department of Radiology, Seoul National University Bundang Hospital, Seongnam, Republic of Korea. [9]Department of Radiology, Seoul National University College of Medicine, Seoul, Republic of Korea. [10]Department of Psychiatry, College of Medicine, Seoul National University, Seoul, Korea. [11]Institute of Human Behavioral Medicine, Seoul National University Medical Research Center, Seoul, Korea. ✉email: kwkimmd@snu.ac.kr

and detailed identification and characterization than visual grading of real ears or photographs. However, the automatic identification of FS from the 3D facial image contained in the brain MRI has never been attempted.

In this study, we developed a deep learning model for automatic segmentation of FS from 3D facial images contained in 3D brain MRI and tested its validity using manually segmented FS in 3D facial images contained in 3D brain MRI. In addition, we determined an optimal threshold voxel of automatically segmented FS for detecting the presence of manually segmented FS.

## Methods

### Participants

Three datasets were constructed. Dataset 1 comprised of T1-weighted 3D brain MRIs from 400 older adults with FS. Dataset 2 comprised of T1-weighted 3D brain MRIs from 150 older adults with FS and 150 older adults without FS. Dataset 3 comprised of T1-weighted 3D brain MRIs from 110 older adults with FS and 120 older adults without FS.

Dataset 1 was constructed for the development of a deep learning model for automatic segmentation of FS (DLM4FS). Dataset 2 was created for the external validation of the DLM4FS and the determination of the optimal threshold voxel number of FS automatically segmented by DLM4FS for determining the presence of FS. Dataset 3 was constructed for the cross-center external validation of the DLM4FS.

The brain MRIs included in Dataset 1 and Dataset 2 were acquired at the Seoul National University Bundang Hospital (SNUBH), while those included in Dataset 3 were acquired at one of three university hospitals other than SNUBH (Chungnam National University Hospital [CNUH], Kangwon National University Hospital [KNUH], and Severance Hospital). The participants included in Dataset 1 and Dataset 2 were either participants in the Korean Longitudinal Study on Cognitive Aging and Dementia (KLOSCAD)[22] or visitors to the dementia clingic at Seoul National University Bundang Hospital (SNUBH) from 2013 to 2022. The participants included in Dataset 3 were individuals who visited the dementia clinic at one of three hospitals (CNUH, KNUH, or Severance Hospital) and were participating in the Korean Registry for Activating Trial on Dementia (KREAT-D) from 2021 to 2022.

All participants provided written informed consent, either by themselves or through their legal guardians. This study protocol was approved by the Institutional Review Board of the Seoul National University Bundang Hospital (IRB No. B-2005-615-001, B-0912/089 – 010 and B-2011-651-001), and was performed in accordance with relevant guidelines and regulations.

### Image acquisition and preprocessing

In Dataset 1 and Dataset 2, three-dimensional (3D) T1-weighted magnetic resonance images were acquired in Digital Imaging and Communications in Medicine (DICOM) format using Philips Achieva and Ingenia scanners (Philips Medical Systems; Eindhoven, Netherlands) at SNUBH. The parameters for brain MRI scanning were as follows: voxel size = 1.0 mm × 0.5 mm × 0.5 mm, 1.0 mm sagittal slices with no gap between slices, echo time = 4.6 ms, repetition time = 8.1 ms, field strength = 3.0 Tesla, field-of-view = 240 × 240, acquisition axial plane matrix size = 175 × 480 × 480 in the x-, y-, and z- dimensions, number of excitations = 1, flip angle = 8°, and inversion time = not applied.

Dataset 3 comprises 3D T1-weighted magnetic resonance images acquired in DICOM format using Philips Achieva scanner (Philips Medical Systems; Eindhoven, Netherlands) at KNUH and Philips Ingenia scanner (Philips Medical Systems; Eindhoven, Netherlands) at CNUH and Severance Hospital. The parameters for brain MRI scanning were as follows: voxel size = 1.0 mm × 1.0 mm × 1.0 mm, 1.0 mm sagittal slice with no gap between slices, echo time = 2.9 ms, repetition time = 6.5 ms, field strength = 3.0 Tesla, field-of-view = 256 × 256, acquisition axial plane matrix size = 211 × 256 × 256 in the x-, y-, and z- dimensions, number of excitations = 1, flip angle = 9°, and inversion time = not applied.

All images in DICOM format were converted to Neuroimaging Informatics Technology Initiative (NIFTI) format using MRIcron software (http://www.mricro.com/mricron). The images were then resampled into isotropic voxels ($1 \times 1 \times 1$ mm$^3$) using the *mri_convert* function in FreeSurfer (http://surfer.nmr.mgh.harvard.edu)[23]. Finally, the images were resized to $175 \times 240 \times 240$. Each image consisted of 175 sagittal slices.

We then removed the middle 111 slices (33rd to 143rd ), leaving 64 slices including both ears, 32 slices each on the left (1st to 32nd ) and right (144th to 175th ) sides. Each sagittal slice had a size of $240 \times 240$. Each slice was cropped to a size of $144 \times 144$ pixels (48th to 192nd pixels for width and 24th to 168th pixels for height) by removing the outer pixels. The resulting cropped images had a size of $64 \times 144 \times 144$. Finally, the intensities of the cropped images were normalized. Specifically, voxel-wise intensity value of the cropped images was scaled to a range of [0, 1] by applying the following formula ( $\frac{Intensity-Min}{Max-Min}$), where "Min" and "Max" represent the minimum and maximum intensity values within the cropped image.

### Manual segmentation of FS

We converted the cropped images from grey scale to HSV due to the advantages this conversion offers in the processes of segmentation and feature extraction. The HSV color space separates the intensity (Value) from the color information (Hue and Saturation), enabling intensity-based analyses without interference from color variations. This separation enhances the capture of structural detail, improves the differentiation of features and increases the accuracy of the segmentation. Additionally, HSV minimizes the sensitivity to illumination changes, ensuring robust performance under varying lighting conditions[24–27]. We then segmented FSs primarily in the sagittal plane and confirmed in the axial and coronal planes. For each participant, two trained researchers, both of whom were blinded to demographic and clinical characteristics of the participants, independently segmented FS slice by slice using ITK-SNAP software version 3.8.0 (http://www.itksnap.org)[28]. A diagonal earlobe crease,

extending from the tragus to the posterior edge of the auricle, was identified as FS and carefully distinguished from a wrinkle connecting the inner edge of the earlobe to the cheek. The voxels belonging to the FS were excluded for the following conditions: first, those voxels belonging to the intertragic notch separating the tragus from the antitragus in the outer ear; second, those voxels belonging to the wrinkles connected to the FS but in a different direction; and third, those voxels belonging to the creases connected to the FS but associated with the earrings.

## Development and validation of automatic FS segmentation models

ped to segment FSs from the pre-processed images using four different architectures: U-net[29], U-net++[30], attention U-net[31], and USE-net[32–35]. Figure 1shows the original U-net for 3D FS images, and additional information on the candidate model architectures can be found in Supplementary Fig. 1. The model architecture comprises an encoder, a decoder block, and skip links. The encoder network, a contracting path with convolutional layers, extracts high-level features, thereby decreasing the spatial resolution at each layer. The decoder network, an expanding path, increases the spatial resolution by up-sampling and utilizes feature information to segment the voxels corresponding to the region of interest (ROI). The skip links between the encoder and decoder facilitate the recovery of fine details that might be lost during spatial down-sampling. All architectures employ max pooling layers ($2\times2\times2$) for down-sampling and transposed convolutions for up-sampling ($2\times2\times2$ kernel and strides). Each network comprises four pooling/up-sampling layers with 16 filters for all convolutional blocks in the first layer, which is doubled after each pooling and subsequently halved at each up-sampling layer. The output of each network is a score to classify each voxel as FS or not, generated by a convolutional layer followed by a sigmoid activation. All convolutional blocks included 3D convolutions[36], ReLU activation[37]and batch normalization[38].

The PyTorch libraries[39] were used to implement all candidate architectures on a computer with an Intel Xeon E5-2687 W v4 CPU 3.00 GHz and 126 GB RAM, together with a dedicated GPU (GeForce GTX 1080Ti, 11 GB RAM). The models were trained on the pre-processed images and their corresponding ground truth labels of the development dataset. Randomly selected input images were augmented with horizontal flipping or rotation ($\pm\ \theta\ <10°$) in order to improve the model's capabilities by increasing the level of difficulty. Training was performed with a batch size of 2 for 100 epochs. The network weights were updated using the Adam optimizer with an L2 weight decay of $1\times10^{-5}$ and an initial learning rate of $1\times10^{-3}$. The learning rate was reduced by a factor of 0.5 every 25 epochs. The binary cross entropy (BCE) cube loss function was used to train the model, which combines the cube loss with the standard BCE loss, emphasizing the learning features corresponding to the positive voxels. For purposes of comparison, the same network parameters were employed for all architectures.

## Statistical analysis

Continuous variables were compared between groups using Student t tests and one-way analysis of variance (ANOVA), while categorical variables were evaluated using chi-square tests. To assess the inter-rater reliability of manual FS segmentation between the two researchers, 40 images from Dataset 1 were randomly selected and the two researchers were asked to segment FS independently. Subsequently, the intra-class correlation coefficient (ICC) and the Dice similarity coefficient (DSC) were calculated between the FSs that were manually segmented by two researchers. The DSC was computed to evaluate the spatial overlap between two manually segmented FS, using the formula( $DSC = \frac{2\,|A\cap B|}{|A|+|B|}$), where A represents the one's manually segmented FS and B represents the other's manually segmented FS. This metric provides a measure of overlap, with a value of 1 indicating perfect alignment and 0 indicating no overlap. The ICC, on the other hand, was calculated to assess the consistency of FS segmentations by two researchers. We employed a two-way mixed-effects model (specifically ICC(3,1)) to

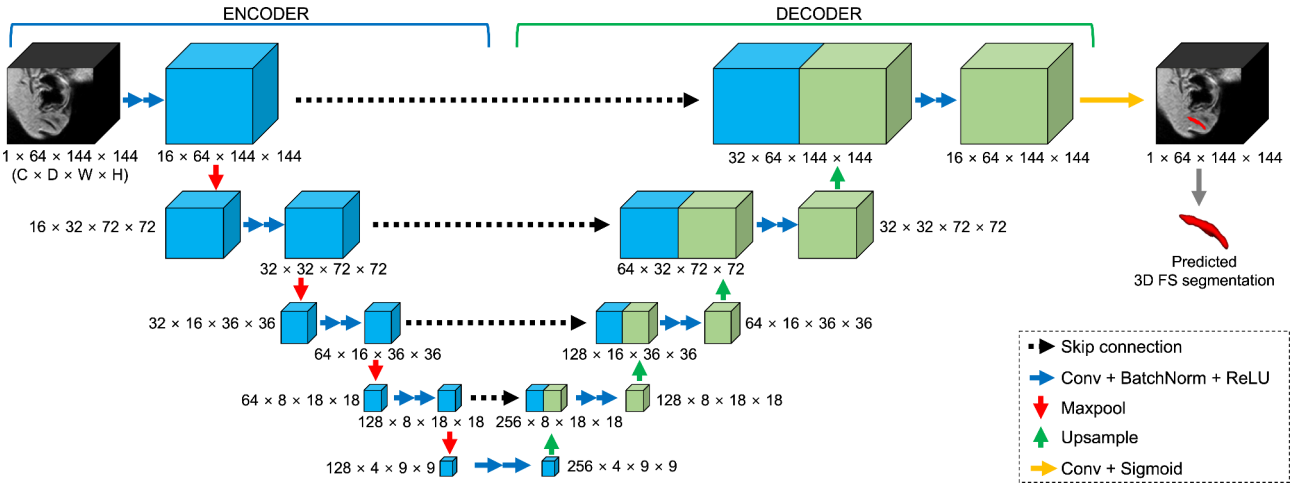


**Fig. 1**. The architecture based on 3D U-net. C, channel; D, depth; H, height; W, width; Conv, convolution $3\times3\times3$ size of the 3D kernels; ReLU, rectified linear unit. Note. The input to the encoder network was pre-processed images and the output of the decoder network was a score to classify each voxel as FS or not.

determine absolute agreement, which is appropriate for comparing measurements from the same subjects and offers a reliability measure for the volumes across segmentation trials. ICC values range from 0 to 1, with higher values indicating stronger agreement.

To identify the optimal automatic FS segmentation model, a 5-fold cross-validation approach was employed to assess the DSC and ICC between manually segmented FSs and automatically segmented FSs in each model with 5-fold cross-validation. Subsequently, the mean DSC and ICC values were compared between models using repeated measures analysis of variance (rmANOVA) with Tukey's honest significant difference (HSD) post-hoc comparisons.

In the two external validation datasets (Dataset 2 and Dataset 3), the accuracy of the DLM4FS for detecting the presence of manually segmented FS was examined using receiver operating characteristic (ROC) curve analysis with a bootstrap sampling estimation (1,000 resamples). In addition, an optimal threshold number of voxels that were automatically segmented as FS for detecting the presence of manually segmented FS was also determined. In each iteration, the accuracy and optimal threshold voxel number were determined in 80% of the ears randomly selected from each dataset using the maximum Youden index (sensitivity + specificity − 1)[40] and validated in the remaining 20% of all ears.

Finally, to evaluate the accuracy of FS volume predictions, we conducted correlation analyses between the model-predicted FS volumes and the ground truth FS volumes across different datasets. And Bland-Altman analysis was conducted across different datasets. For each dataset, Bland-Altman plots were generated, displaying the differences between predicted and ground truth FS volumes.

All statistical analyses were conducted using the most commonly employed Python scientific packages, namely Numpy[41], Pandas[42], statsmodels[43]and SciPy[44].

## Results

The demographic and clinical characteristics were comparable between the three datasets, with the exception of age, MMSE, and the distribution of FS. The participants included in Dataset 3 were younger and performed MMSE better than those included in Dataset 1 and Dataset 2. All participants included in Dataset 1 had manually segmented FS, while approximately half of the participants included in Dataset 2 and Dataset 3 had manually segmented. The volume of manually segmented in Dataset 3 was lower than that in Dataset 1 and Dataset 2 (Table 1). In Dataset 2, the demographic and clinical characteristics were comparable between the participants with manually segmented and those without manually segmented. In Dataset 3, the participants with manually segmented were slightly older and more likely to be male than those without manually segmented (Table 2).

The two researchers demonstrated excellent inter-rater reliability in manual segmentation of FS, as indicated by the ICC (0.949, 95% confidence interval [CI] = 0.904–0.973) and DSC (0.850 ± 0.051).

As demonstrated in Table 3, while the average DSC and ICC between the automatically segmented FSs and the manually segmented FSs were highest in the model employing the U-net architecture, the observed differences were not statistically significant across the architectures utilized in the development of automatic FS segmentation models on the fivefold test sets from dataset 1 ($p > 0.05$). The Supplementary Table 1 presents the DSC and ICC values for the automatically segmented FSs in comparison to the manually segmented FSs across the test sets. However, the mean training time per epoch and the mean inference time per step were found to be significantly shorter for the model utilizing the U-net architecture than for the models using other architectures

| | Dataset 1[a] | Dataset 2[b] | Dataset 3[e] | Statistics[*] | | |
|---|---|---|---|---|---|---|
| | (*n* = 400) | (*n* = 300) | (*n* = 230) | F | *p* | Post hoc |
| Age, years | 76.3 (7.7) | 75.2 (7.1) | 73.4 (7.9) | 7.23 | < 0.001 | c < a, b |
| Women | 250 (62.5) | 192 (64.0) | 147 (63.9) | 0.10 | 0.900 | - |
| Education, years | 11.2 (5.0) | 11.7 (5.0) | 10.5 (5.1) | 2.12 | 0.111 | - |
| MMSE, points | 23.7 (5.5) | 23.3 (5.6) | 24.8 (4.7) | 5.34 | 0.005 | a, b < c |
| CDR | | | | 1.43 | 0.240 | - |
| 0 | 172 (43.0) | 117 (39.0) | 110 (47.8) | | | |
| 0.5 | 207 (51.8) | 156 (52.0) | 94 (40.9) | | | |
| 1 | 21 (5.2) | 27 (9.0) | 26 (11.3) | | | |
| mFS | | | | 10.4 | < 0.001 | b, c < a |
| None | 0 (0) | 150 (50.0) | 120 (52.2) | | | |
| Unilateral | | | | | | |
| Left | 99 (24.8) | 30 (10.0) | 20 (8.7) | | | |
| Right | 104 (26.0) | 51 (17.0) | 33 (14.3) | | | |
| Bilateral | 197 (49.2) | 69 (23.0) | 57 (24.8) | | | |
| $V_{mFS}$, $mm^3$ | 137.4 (77.7) | 136.2 (70.0) | 89.3 (74.9) | 15.3 | < 0.001 | c < a, b |

**Table 1**. Demographic and clinical characteristics of datasets. MMSE, Mini Mental Status Examination; CDR, Clinical Dementia Rating; mFS, manually segmented Frank's sign; VmFS, volume of mFS Note. Continuous variables are presented as mean (standard deviation) and categorical variables as number (percentage). [*]Analysis of variance for variables with Tukey's HSD post-hoc comparison.

| | Dataset 2 | | | Dataset 3 | | |
|---|---|---|---|---|---|---|
| | With mFS ($n=150$) | Without mFS ($n=150$) | $p^*$ | With mFS ($n=110$) | Without mFS ($n=120$) | $p^*$ |
| Age, years | 75.8 (7.5) | 74.8 (6.8) | 0.244 | 74.4 (7.3) | 72.4 (7.9) | 0.021 |
| Women | 100 (66.7) | 92 (61.3) | 0.337 | 62 (56.3) | 79 (65.8) | 0.043 |
| Education, years | 11.3 (5.3) | 12.1 (4.8) | 0.222 | 10.2 (5.2) | 11.0 (5.0) | 0.594 |
| MMSE, points | 24.2 (6.1) | 25.2 (4.6) | 0.237 | 25.1 (4.5) | 24.9 (4.7) | 0.775 |
| CDR | | | 0.073 | | | 0.053 |
| 0 | 66 (44.0) | 53 (35.3) | | 65 (59.1) | 57 (47.5) | |
| 0.5 | 74 (49.3) | 80 (53.3) | | 41 (37.3) | 51 (42.5) | |
| 1 | 10 (6.7) | 17 (11.3) | | 4 (3.6) | 12 (10.0) | |

**Table 2**. Demographic and clinical characteristics of the participants with manually segmented Frank sign and those without manually segmented Frank sign. mFS, manually segmented Frank's sign; MMSE, Mini Mental Status Examination; CDR, Clinical Dementia Rating. Continuous variables are presented as mean (standard deviation) and categorical variables as number (percentage). *Student t tests for continuous variables and chi square tests for categorical variables.

($p<0.001$) (Table 3). Therefore, the U-Net architecture was selected for the development of the automatic FS segmentation model. Figure 2 depicts FSs that have been automatically segmented by this model.

We then validated the performance of this model in Dataset 2 and Dataset 3. There were manually segmented FS in 219 ears out of 600 ears in Dataset 2 and 179 ears out of 460 ears in Dataset 3. In Dataset 2 (an external validation dataset obtained from the same institution as the development dataset), this model demonstrated excellent performance in terms of the area under the ROC curve (AUC), accuracy, sensitivity, and specificity for predicting the presence of manually segmented FS. The mean optimal threshold voxel number for predicting the presence of manually segmented FS was found to be 25/26 (Table 4). In Dataset 3 (an external validation dataset obtained from three institutions different from the institution where the development dataset was obtained), this model also demonstrated excellent AUC, accuracy, sensitivity, and specificity for predicting the presence of manually segmented FS. The mean optimal threshold voxel number for predicting the presence of manually segmented FS was found to be 22/23 (Table 4). The ROC analysis demonstrated high model accuracy across all datasets, with AUC values indicating strong discriminative power. The shaded regions around each ROC curve illustrate the confidence intervals from the 1000 iterations, supporting the model's robustness and reliability in identifying FS across diverse datasets (Fig. 3). These results underscore the model's capability to accurately detect FS across different institutions and patient populations.

The correlation analysis demonstrated strong positive correlations between the predicted and ground truth FS volumes across all datasets (Fig. 4). For Dataset 2, a Pearson correlation coefficient of $r=0.868$ with $p<0.001$ was observed, indicating high prediction accuracy. In Dataset 3, subgroup analyses showed similar trends, with correlation coefficients of $r=0.828$ for CNUH, $r=0.728$ for KNUH, and $r=0.673$ for Severance H, all with $p<0.001$, reflecting robust model performance across different clinical settings. And the Bland-Altman analysis revealed that the mean differences between predicted and ground truth FS volumes varied slightly across datasets (Fig. 5). For Dataset 2, the mean difference was 9.88 $mm^3$, indicating a slight underestimation by the model. In Dataset 3, the mean differences were 19.07 $mm^3$ for CNUH, 7.38 $mm^3$ for KNUH, and 4.63 $mm^3$ for Severance H, with a slight overestimation observed in CNUH and smaller biases in KNUH and Severance H. These results suggest that the model's predictions are closely aligned with the ground truth across various clinical settings, with minor variations in bias reflecting dataset-specific characteristics.

## Discussion

This inaugural study developed a deep learning algorithm to automatically segment FS using face images embedded in 3D T1-weighted brain MRI. The algorithm demonstrated excellent performance in segmenting FS and determining its presence, with comparable results between the development and external validation datasets, suggesting strong potential for new, unseen data. The findings of this study do not indicate that a brain MRI should be conducted for the purpose of detecting Frank's sign. Rather, the results suggest that brain MRIs performed for the diagnosis of numerous other brain disorders can be utilized for the accurate detection of Frank's sign, obviating the necessity for additional testing.

Traditionally, FS detection has relied on semi-quantitative human visual assessment, either by direct observation or examination of photographs. Although brain MRI typically includes facial images, no attempt has been made to automatically segment FS from these images. Using 3D facial images in 3D brain MRI offers significant advantages over 2D ear photography for FS assessment. This study addresses key limitations in prior FS research, which relied heavily on subjective visual assessments or 2D imaging methods[9,12,16]. Unlike these approaches, our study employs a 3D MRI-based deep learning model, leveraging the voxel-level precision of U-Net for FS segmentation. This method not only eliminates grading inconsistencies but also provides higher reliability, as demonstrated by robust performance across multi-center datasets with AUC values exceeding 0.9. Moreover, while previous research often struggled to replicate FS-disease associations due to methodological limitations[7,13], our model offers a standardized framework to explore disease-specific FS characteristics in future

| | Architectures | | | | | Statistics† | | |
|---|---|---|---|---|---|---|---|---|
| | U-net[a] | Attention U-net[b] | U-net + +[c] | USE-net[d] | | F | *p* | *post-hoc* |
| DSC | 0.732 (0.101) | 0.731 (0.171) | 0.731 (0.144) | 0.731 (0.135) | | 0.562 | 0.650 | - |
| ICC | 0.865 (0.006) | 0.850 (0.013) | 0.862 (0.009) | 0.854 (0.010) | | 3.377 | 0.054 | - |
| Training time, sec / epoch | 176.81 (8.40) | 186.50 (7.52) | 347.56 (7.35) | 181.57 (9.39) | | 10142.35 | < 0.001 | a < d < b < c |
| Inference time, ms / step | 44.47 (0.73) | 55.41 (0.90) | 160.64 (1.52) | 45.07 (0.78) | | 14319.83 | < 0.001 | a < d < b < c |

**Table 3**. Performance of deep learning models for automatically segmenting Frank's sign using four different architectures in dataset 1. DSC, dice similarity coefficient between the automatically segmented Frank's sign and the manually segmented Frank's sign; ICC, Intra-class correlation coefficient between the volume of the automatically segmented Frank's sign and that of the manually segmented Frank's sign. Note. All values are presented as a mean (standard deviation) from 5-fold cross-validation. †Repeated measures analysis of variance with Tukey's HSD post-hoc comparison.

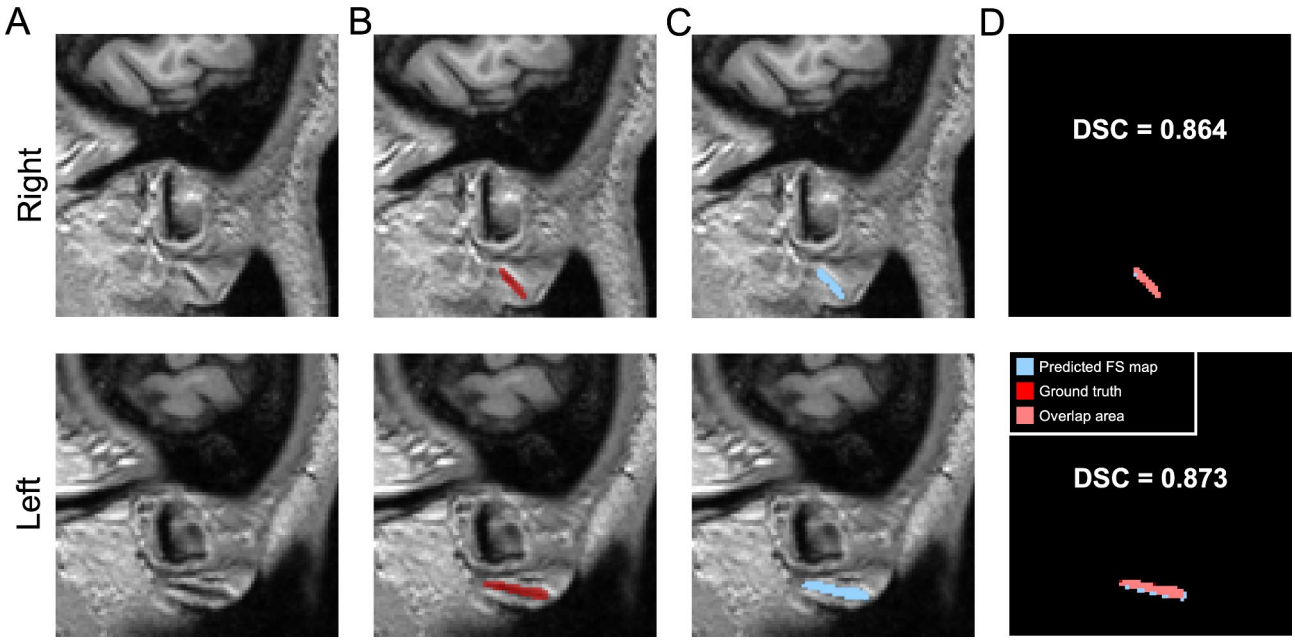


**Fig. 2**. Frank's signs automatically segmented using the U-Net architecture-based model. **(A)** Original three-dimensional ear input images. **(B)** Manually segmented Frank's signs (red) overlaid on the original images. **(C)** Automatically segmented Frank's signs (blue) overlaid on the original images. **(D)** Automatically segmented Frank's signs (blue) overlaid on the manually segmented Frank's signs (red) with the Dice similarity coefficient (DSC) between them.

| | DSC | Optimal cutoff | AUC* | Accuracy* | Sensitivity* | Specificity* |
|---|---|---|---|---|---|---|
| Dataset 2† | | | | | | |
| | 0.734 (0.159) | 25/26 | 0.942 (0.021) | 0.948 (0.019) | 0.918 (0.039) | 0.965 (0.026) |
| Dataset 3‡ | | | | | | |
| All | 0.714 (0.150) | 22/23 | 0.902 (0.046) | 0.911 (0.040) | 0.842 (0.058) | 0.983 (0.031) |
| CNUH | 0.719 (0.128) | 23/24 | 0.919 (0.052) | 0.921 (0.045) | 0.925 (0.041) | 0.913 (0.041) |
| KNUH | 0.712 (0.109) | 23/24 | 0.900 (0.038) | 0.892 (0.029) | 0.800 (0.031) | 1.000 (0.019) |
| Severance Hospital | 0.708 (0.143) | 21/22 | 0.864 (0.043) | 0.923 (0.033) | 0.727 (0.051) | 1.000 (0.014) |

**Table 4**. External validation of the deep learning model for automatically segmenting Frank's sign utilizing U-net architecture in dataset 2 and dataset 3. AUC, area under the receiver operator characteristics curve; CNUH, Chungnam National University Hospital; DSC, dice similarity coefficient; KNUH, Kangwon National University Hospital. All values are presented as a mean (standard deviation) of 1000 iterations. *Receiver operator characteristic analysis with 1000 resamples. †External validation dataset obtained from the same Seoul National University Bundang Hospital as Dataset 1. ‡External validation dataset obtained from hospitals other than Seoul National University Bundang Hospital where Dataset 1 was obtained.

studies. These findings align with advancements in medical imaging[45] and highlight the potential of automated 3D segmentation to overcome limitations in earlier methods, paving the way for broader applications.

The selection of the segmentation algorithm has a considerable impact on the computational efficiency and practical feasibility of the FS segmentation model, as evidenced by our findings. A key advantage is the extraction of intricate 3D FS features unattainable with 2D images, potentially identifying specific FS subgroups associated with cerebrovascular disease, cardiovascular disease, or cognitive disorders. This differentiation is crucial for distinguishing FS from generic earlobe folds. Additionally, FS characteristics may vary across these conditions, facilitating precise identification and understanding of disease-specific manifestations. Moreover, leveraging the typically included but often neglected ear imaging data in brain MRI enhances its utility, expanding clinical applications to provide deeper insights into FS-associated conditions beyond standard brain assessments.

The selection of the segmentation algorithm has a considerable impact on the computational efficiency and practical feasibility of the FS segmentation model, as evidenced by our findings. In this study, we implemented an encoder-decoder deep neural network configuration based on the 3D U-Net architecture, which has been widely recognized for its effectiveness in medical image segmentation[45]. The applicability of traditional segmentation approaches, including intensity thresholding, region growing, and deformable models, was evaluated; however,

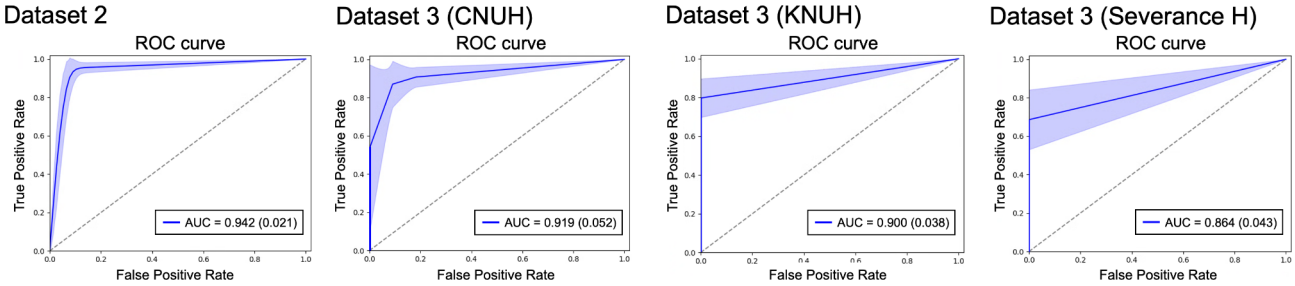


**Fig. 3**. The receiver operating characteristic (ROC) curves for the evaluation of model performance across diverse datasets, based on 1,000 iterations. The shaded regions surrounding each curve represent the confidence intervals derived from the 1,000 iterations. CNUH, Chungnam National University Hospital; KNUH, Kangwon National University Hospital; Severance H, Severance Hospital.

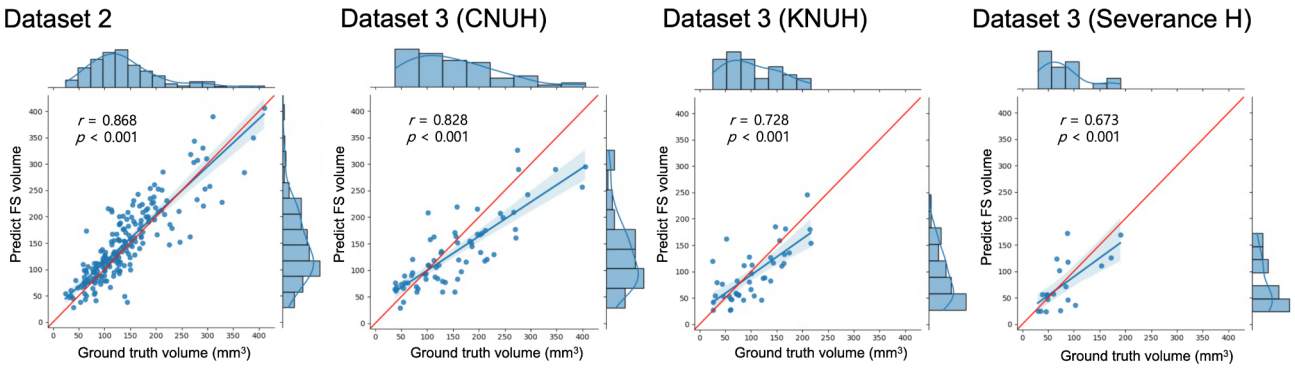


**Fig. 4**. Correlation plots between predicted and ground truth Frank's sign (FS) volumes across different datasets. Each plot shows the predicted FS volume vs. the ground truth volume for Dataset 2 and subgroups within Dataset 3 (CNUH, KNUH, Severance H). The red line represents the line of perfect correlation, while the blue line indicates the regression fit with a 95% confidence interval. The histograms on the x- and y-axes illustrate the distribution of ground truth and predicted volumes, respectively, across each dataset, highlighting the model's predictive accuracy in estimating FS volume. CNUH, Chungnam National University Hospital; KNUH, Kangwon National University Hospital; Severance H, Severance Hospital.

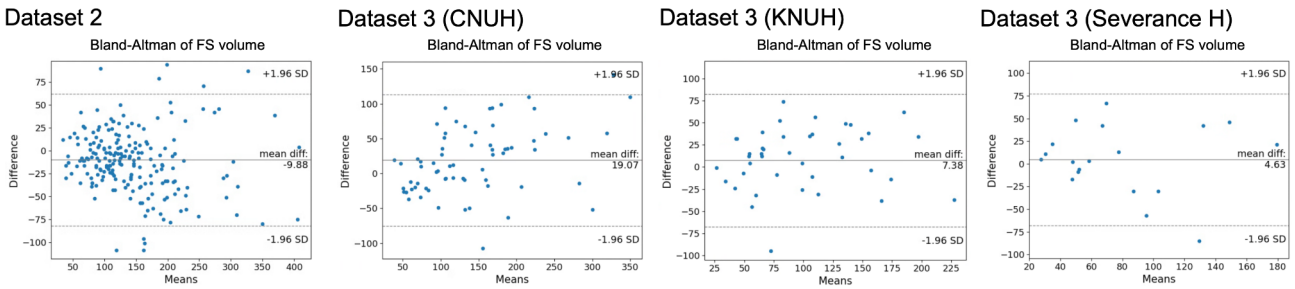


**Fig. 5**. Bland-Altman plots for Frank's sign (FS) volume measurements across different datasets. Each plot shows the difference in FS volume measurements (y-axis) against the mean volume (x-axis) for Dataset 2 and subgroups within Dataset 3 (CNUH, KNUH, Severance H). The solid horizontal line indicates the mean difference (mean diff) between measurements, while the dashed lines represent the limits of agreement at ± 1.96 standard deviations (SD) from the mean difference. CNUH, Chungnam National University Hospital; KNUH, Kangwon National University Hospital; Severance H, Severance Hospital.

these methods were found to have limited utility in FS segmentation. These methods rely heavily on manually defined features, which are less effective for capturing the subtle and complex characteristics of FS. Although these methods are useful for simpler segmentation tasks, they lack the autonomous feature-learning capability that is a hallmark of modern deep learning models[46]. To further enhance the performance of the segmentation, we explored advanced variants of the 3D U-Net model, including 3D U-Net++, Attention 3D U-Net, and 3D

USE-Net. These variants incorporate additional features, such as skip connections and attention mechanisms, which may contribute to improved performance. Attention-based architectures, such as Attention 3D U-Net and 3D USE-Net, could have the potential to enhance the focus on clinically relevant regions by emphasizing small or indistinct features.

However, the more complex models did not yield significantly superior results in FS segmentation when compared to the original 3D U-Net. This finding corroborates the conclusions of previous studies[29], which have demonstrated that the original U-Net model offers an optimal balance between segmentation accuracy and computational efficiency, rendering it a practical choice for applications in clinical settings. The U-Net model exhibited the optimal trade-off between segmentation accuracy and computational efficiency, a finding that aligns with its prevalent use in medical imaging applications. Although the mean DSC and ICC values were highest for the U-Net architecture, the differences in segmentation accuracy across the tested models were not statistically significant ($p > 0.05$). This indicates that all models exhibited comparable accuracy for FS segmentation within the context of this study. However, the U-Net model demonstrated a notable advantage in computational efficiency. The mean training time per epoch and the mean inference time per step were found to be significantly shorter for the U-Net model in comparison to the other architectures ($p < 0.001$), thereby underscoring its practicality for real-world clinical applications, particularly in time-sensitive settings. This efficiency likely contributed to the model's robustness when validated on external datasets, as reduced computational overhead facilitates scalability across diverse clinical settings. Models based on attention mechanisms, such as Attention U-Net and USE-Net, demonstrated the potential to enhance focus on clinically relevant regions. However, the incremental gains in segmentation accuracy were negated by their longer training and inference times, rendering them less suitable for large-scale deployment or scenarios requiring rapid processing. These findings are in accordance with prior studies indicating that simpler architectures, such as U-Net, often provide an optimal balance between accuracy and computational demands, particularly in segmentation tasks involving subtle anatomical structures.

In external validation datasets (Datasets 2 and 3), the U-Net-based model exhibited excellent AUC, accuracy, sensitivity, and specificity, thereby confirming its reliability in diverse clinical settings. While dataset-specific characteristics, such as voxel intensity distributions, influenced optimal threshold values, the strong correlation coefficients and Bland-Altman analyses support the robustness of the model's predictions. It is noteworthy that the observed biases in Dataset 3, particularly in CNUH, reflect the impact of inter-institutional variations, which may also interact with the choice of segmentation algorithm. These minor variations underscore the necessity of validating segmentation models across heterogeneous datasets to guarantee generalizability.

Alternative algorithms, including YOLO[47]and Fast R-CNN[48], were also evaluated for their potential applicability. However, these models are primarily designed for 2D object detection and rely on bounding-box approaches, which are not well-suited for voxel-wise segmentation of subtle structures like FS in 3D MRI data. Although some studies have achieved success in integrating segmentation and classification tasks in two-dimensional contexts, our study opted to separate these processes. This decision allowed us to prioritize segmentation accuracy as a fundamental step for FS analysis. Future research could build on this approach by integrating a classification stage to identify FS subtypes or their clinical implications, thereby enhancing the diagnostic value of the segmentation model[49].

The application of preprocessing techniques is crucial for improving the model and enhancing its interpretability and robustness[50,51]. In this study, we optimized feature dimensions by combining cropped images of bilateral ears across all datasets while maintaining a consistent ratio. This included isotropic voxel resampling and intensity rescaling, potentially enhancing the deep learning models' performance[52]. Furthermore, Additionally, data augmentation was used to refine model performance by effectively increasing the training dataset. Previous research suggests that data augmentation can improve accuracy by up to 5%[53,54].

In the course of our investigation, we noted minor discrepancies in the optimal threshold voxel numbers between Dataset 2 and Dataset 3. These discrepancies are presumably attributable to variations in anatomical, demographic, and clinical characteristics, such as differences in the proportions of participants with varying degrees of cognitive impairment, cerebrovascular disease, and multiple sclerosis. Such variations can influence voxel intensity distributions through disease-related changes in tissue characteristics, as evidenced by previous studies[55,56]. Furthermore, age-related alterations in tissue density and discrepancies in MRI acquisition protocols may also contribute to these discrepancies.

To address these dataset-specific differences, minor adjustments were made to the post-processing thresholds to optimize segmentation accuracy for each dataset while maintaining consistent core model parameters across all analyses. This approach ensured robust and reliable performance, facilitating meaningful comparisons of model performance across datasets despite their inherent variability.

This study has certain limitations. The dataset was composed exclusively of Korean participants, meaning cross-ethnic validation was not performed. Differences in FS characteristics among ethnic groups, influenced by genetic, environmental, and lifestyle factors, suggest that cross-ethnic validation could further generalize the model. Furthermore, while our model effectively segments FS, we did not examine associations between FS and clinical conditions (e.g., cerebrovascular or cardiovascular diseases), which could be explored in future research.

## Conclusion

In conclusion, our study demonstrates the feasibility of 3D U-Net-based segmentation of FS from MRI, offering a more accurate and reliable approach to FS assessment compared to traditional and 2D methods. This tool holds promise for improving FS identification and distinguishing clinically relevant FS features, paving the way for further research and broader applications in health and medical science.

## Data availability

Since the data includes private participant information, it is not available to the public. Upon reasonable request, qualified researchers may be given access to the individual, de-identified participant data that underlie the findings described in this publication. To acquire access, proposals should be sent to K.W.K.

Received: 29 August 2024; Accepted: 9 December 2024
Published online: 18 January 2025

## References

1. Frank, S. T. Aural sign of coronary-artery disease. *N Engl. J. Med.* **289** (6), 327–328 (1973).
2. Mallinson, T. & Brooke, D. Frank's sign as a clinical marker of cardiovascular disease. *J. Paramedic Pract.* **9** (1), 8–10 (2017).
3. Shmilovich, H. et al. Relation of diagonal ear lobe crease to the presence, extent, and severity of coronary artery disease determined by coronary computed tomography angiography. *Am. J. Cardiol.* **109** (9), 1283–1287 (2012).
4. Bouissou, H. et al. Value of skin biopsy in coronary insufficiency. *Arch. Mal. Coeur Vaiss.* **66** (5), 655–660 (1973).
5. Brady, P. M. et al. A new wrinkle to the earlobe crease. *Arch. Intern. Med.* **147** (1), 65–66 (1987).
6. Griffing, G. Frank's sign. *N Engl. J. Med.* **370** (10), e15 (2014).
7. Nazzal, S. et al. Diagonal earlobe crease (Frank's sign): a predictor of cerebral vascular events. *Am. J. Med.* **130** (11), 1324 (2017). e1-1324. e5.
8. Lee, J. S. et al. Diagonal earlobe crease is a visible sign for cerebral small vessel disease and amyloid-β. *Sci. Rep.* **7** (1), 1–7 (2017).
9. Jorde, L. B., Williams, R. R. & Hunt, S. C. Lack of association of diagonal earlobe crease with other cardiovascular risk factors. *West. J. Med.* **140** (2), 220 (1984).
10. Mehta, J. & Hamby, R. Diagonal ear-lobe crease as a coronary risk factor. *N. Engl. J. Med.* **291** (5), 260–260 (1974).
11. Cheng, T. O. More research needed on the association between diagonal earlobe crease and coronary artery disease. *Arch. Intern. Med.* **160** (15), 2393–2393 (2000).
12. Koracevic, G. & Atanaskovic, V. Ear lobe crease: point of disagreement in evidence-based medicine. *Am. J. Forensic. Med. Pathol.* **30** (1), 89 (2009).
13. Overfield, T. & CALL, E. B. Earlobe type, race, and age: effects on earlobe creasing. *J. Am. Geriatr. Soc.* **31** (8), 479–481 (1983).
14. Costa, L. F. D. & Cesar, R. M. Jr *Shape Analysis and Classification: Theory and Practice* (CRC Press, Inc, 2000).
15. Russ, J. C. *The Image Processing Handbook* (CRC, 2006).
16. Rodríguez-López, C. et al. Earlobe crease shapes and cardiovascular events. *Am. J. Cardiol.* **116** (2), 286–293 (2015).
17. Patel, V. et al. Diagonal earlobe creases and atheromatous disease: a postmortem study. *J. R. Coll. Physicians Lond.* **26** (3), 274 (1992).
18. Thilo, C. et al. Diagonal earlobe crease and long-term survival after myocardial infarction. *BMC Cardiovasc. Disord.* **21** (1), 1–8 (2021).
19. Pacei, F. et al. Diagonal earlobe crease (Frank's sign) and increased risk of cerebrovascular diseases: review of the literature and implications for clinical practice. *Neurol. Sci.* **41**, 257–262 (2020).
20. Wang, Y. et al. Relationship between diagonal earlobe creases and coronary artery disease as determined via angiography. *BMJ open.* **6** (2), e008558 (2016).
21. Bischoff-Grethe, A. et al. A technique for the deidentification of structural brain MR images. *Hum. Brain. Mapp.* **28** (9), 892–903 (2007).
22. Han, J. W. et al. Overview of the Korean longitudinal study on cognitive aging and dementia. *Psychiatry Invest.* **15** (8), 767 (2018).
23. Fischl, B. *FreeSurfer Neuroimage*, **62**(2): 774–781. (2012).
24. Gonzalez, R. C. *Digital Image Processing* (Pearson education india, 2009).
25. Stockman, G. & Shapiro, L. G. *Computer Vision* (Prentice Hall PTR, 2001).
26. Gevers, T. & Smeulders, A. W. Color-based object recognition. *Pattern Recogn.* **32** (3), 453–464 (1999).
27. Semary, N. A. *Aproposed hsv-based pseudo-coloring scheme for enhancing medical images. Comput. Sci. Inform. Technol.*, 81. (2018).
28. Yushkevich, P. A. et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage* **31** (3), 1116–1128 (2006).
29. Ronneberger, O., Fischer, P. & Brox, T. *U-net: Convolutional Networks for Biomedical Image Segmentation* (Springer, 2015).
30. Zhou, Z. et al. *Unet++: A Nested u-net Architecture for Medical Image Segmentation, in Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*p. 3–11 (Springer, 2018).
31. Oktay, O. et al. *Attention u-net: Learning where to look for the pancreas.* arXiv preprint arXiv:1804.03999, (2018).
32. Rundo, L. et al. USE-Net: incorporating squeeze-and-excitation blocks into U-Net for prostate zonal segmentation of multi-institutional MRI datasets. *Neurocomputing* **365**, 31–43 (2019).
33. Hu, J., Shen, L. & Sun, G. *Squeeze-and-excitation networks.* (2018).
34. Lee, A. R. et al. Fully automated segmentation on brain ischemic and white matter hyperintensities lesions using semantic segmentation networks with squeeze-and-excitation blocks in MRI. *Inf. Med. Unlocked.* **21**, 100440 (2020).
35. Li, X. et al. White matter hyperintensities segmentation using an ensemble of neural networks. *Hum. Brain. Mapp.* **43** (3), 929–939 (2022).
36. Ji, S. et al. 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **35** (1), 221–231 (2012).
37. Agarap, A. F. *Deep learning using rectified linear units (relu).* arXiv preprint arXiv:1803.08375, (2018).
38. Santurkar, S. et al. *How does batch normalization help optimization?* Advances in neural information processing systems, 31. (2018).
39. Paszke, A. et al. *Pytorch: an imperative style, high-performance deep learning library. Adv. Neural. Inf. Process. Syst.*, 32. (2019).
40. Youden, W. J. Index for rating diagnostic tests. *Cancer* **3** (1), 32–35 (1950).
41. Oliphant, T. E. Python for scientific computing. *Comput. Sci. Eng.* **9** (3), 10–20 (2007).
42. McKinney, W. Pandas: a foundational Python library for data analysis and statistics. *Python high. Perform. Sci. Comput.* **14** (9), 1–9 (2011).
43. Seabold, S. & Perktold, J. *Statsmodels: Econometric and statistical modeling with python*. in *Proceedings of the 9th Python in Science Conference*. Austin, TX. (2010).
44. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods*. **17** (3), 261–272 (2020).
45. Çiçek, Ö. et al. *3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation* (Springer, 2016).
46. O'Mahony, N. et al. *Deep learning vs. traditional computer vision*. in *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1 1*. Springer. (2020).
47. Redmon, J. *You only look once: Unified, real-time object detection*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. (2016).
48. Girshick, R. *Fast r-cnn.* arXiv preprint arXiv:1504.08083, (2015).
49. Almonacid-Uribe, S. L. et al. *Deep Learning for Diagonal Earlobe Crease Detection.* arXiv preprint arXiv:2210.11582, (2022).
50. Alam, S. & Yao, N. The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis. *Comput. Math. Organ. Theory*. **25**, 319–335 (2019).

51. Maharana, K., Mondal, S. & Nemade, B. *A review: Data pre-processing and data augmentation techniques.* Global Transitions Proceedings, 3(1): pp. 91–99. (2022).
52. Chollet, F. *Deep Learning with Python* (Simon and Schuster, 2021).
53. Johnson, J. M. & Khoshgoftaar, T. M. Survey on deep learning with class imbalance. *J. Big Data*. **6** (1), 1–54 (2019).
54. Shorten, C. & Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *J. big data*. **6** (1), 1–48 (2019).
55. Gur, R. C. et al. Sex differences in brain gray and white matter in healthy young adults: correlations with cognitive performance. *J. Neurosci.* **19** (10), 4065–4072 (1999).
56. Han, X. et al. Reliability of MRI-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer. *Neuroimage* **32** (1), 180–194 (2006).

## Acknowledgements
The authors thank all members of the staff and participants at the Seoul National University Bundang Hospital for their important contributions.

## Author contributions
S.J. contributed conceptualization, methodology, formal analysis, investigation, resources, data curation and writing-original draft; J.S.K. contributed methodology, formal analysis and data curation; M.J.K., and J.P. contributed investigation and resources; J.L.K., J.H.J., E.K., L.S., J.H.K., and J.W.H. contributed data collection, resources, and writing-review & editing; K.W.K. contributed conceptualization, methodology, formal analysis, writing-original draft and supervision.

## Funding
This study was supported by a grant from the Korean Health Technology R&D Project, Ministry of Health and Welfare, Republic of Korea [grant no. HI09C1379 (A092077)] and a grant of the Korea Dementia Research Project through the Korea Dementia Research Center (KDRC), funded by the Ministry of Health & Welfare and Ministry of Science and ICT, Republic of Korea (grant number: HU20C0015).

## Declarations

## 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/s41598-024-82756-2.

**Correspondence** and requests for materials should be addressed to K.W.K.