JoS JOURNAL OF STROKE

Journal of Stroke 2024;26(2):312-320
https://doi.org/10.5853/jos.2023.03426

Original Article

A Multimodal Ensemble Deep Learning Model for Functional Outcome Prognosis of Stroke Patients

Hye-Soo Jung,[1] Eun-Jae Lee,[1] Dae-Il Chang,[2] Han Jin Cho,[3] Jun Lee,[4] Jae-Kwan Cha,[5] Man-Seok Park,[6] Kyung Ho Yu,[7] Jin-Man Jung,[8] Seong Hwan Ahn,[9] Dong-Eog Kim,[10] Ju Hun Lee,[11] Keun-Sik Hong,[12] Sung-Il Sohn,[13] Kyung-Pil Park,[14] Sun U. Kwon,[1] Jong S. Kim,[1] Jun Young Chang,[1] Bum Joon Kim,[1] Dong-Wha Kang[1]; KOSNI Investigators

[1]Department of Neurology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea
[2]Department of Neurology, Kyung Hee University Medical Center, Seoul, Korea
[3]Department of Neurology, Pusan National University Hospital, Busan, Korea
[4]Department of Neurology, Yeungnam University Medical Center, Daegu, Korea
[5]Department of Neurology, Dong-A University Hospital, Busan, Korea
[6]Department of Neurology, Chonnam National University Hospital, Gwangju, Korea
[7]Department of Neurology, Hallym University Sacred Heart Hospital, Anyang, Korea
[8]Department of Neurology, Korea University Ansan Hospital, Ansan, Korea
[9]Department of Neurology, Chosun University Hospital, Gwangju, Korea
[10]Department of Neurology, Dongguk University Ilsan Hospital, Ilsan, Korea
[11]Department of Neurology, Hallym University Kangdong Sacred Heart Hospital, Seoul, Korea
[12]Department of Neurology, Inje University Ilsan Paik Hospital, Ilsan, Korea
[13]Department of Neurology, Keimyung University Medical Center, Daegu, Korea
[14]Department of Neurology, Pusan National University Yangsan Hospital, Yangsan, Korea

Background and Purpose The accurate prediction of functional outcomes in patients with acute ischemic stroke (AIS) is crucial for informed clinical decision-making and optimal resource utilization. As such, this study aimed to construct an ensemble deep learning model that integrates multimodal imaging and clinical data to predict the 90-day functional outcomes after AIS.

Methods We used data from the Korean Stroke Neuroimaging Initiative database, a prospective multicenter stroke registry to construct an ensemble model integrated individual 3D convolutional neural networks for diffusion-weighted imaging and fluid-attenuated inversion recovery (FLAIR), along with a deep neural network for clinical data, to predict 90-day functional independence after AIS using a modified Rankin Scale (mRS) of 3–6. To evaluate the performance of the ensemble model, we compared the area under the curve (AUC) of the proposed method with that of individual models trained on each modality to identify patients with AIS with an mRS score of 3–6.

Results Of the 2,606 patients with AIS, 993 (38.1%) achieved an mRS score of 3–6 at 90 days post-stroke. Our model achieved AUC values of 0.830 (standard cross-validation [CV]) and 0.779 (time-based CV), which significantly outperformed the other models relying on single modalities: b-value of 1,000 s/mm^2 (P<0.001), apparent diffusion coefficient map (P<0.001), FLAIR (P<0.001), and clinical data (P=0.004).

Conclusion The integration of multimodal imaging and clinical data resulted in superior prediction of the 90-day functional outcomes in AIS patients compared to the use of a single data modality.

Keywords Modified Rankin Scale; Stroke; Prognosis; Deep learning

Correspondence: Dong-Wha Kang
Department of Neurology, Asan Medical Center, University of Ulsan College of Medicine, 88 Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Korea
Tel: +82-2-3010-3440
E-mail: dwkang@amc.seoul.kr
https://orcid.org/0000-0002-2999-485X

Received: October 13, 2023
Revised: April 12, 2024
Accepted: April 23, 2024

Introduction

Acute ischemic stroke (AIS) results in long-term functional disability, inflicting significant social and economic burdens.[1] Accurate prediction of stroke functional outcomes is important to achieve informed clinical decision-making and improve patients' quality of life.[2] However, this prediction is difficult because of the heterogeneous nature of post-stroke disability.[3] Stroke functional outcome is influenced by various factors, encompassing clinical aspects such as age,[3] patient characteristics,[4] cognition,[5] treatment,[6] comorbidities,[7] stroke severity,[8] and even imaging biomarkers.[9,10] Widely used prognostic systems, including Ischemic Stroke Predictive Risk Score[11] or Acute Stroke Registry and Analysis of Lausanne,[12] incorporate some of these clinical factors to predict stroke functional outcomes. However, the selective inclusion of clinical factors may lead to missing important information and failing to consider patient-specific details.

Machine learning (ML) and deep learning (DL) have achieved significant success in the field of medicine, offering a broad range of variables and algorithms. ML methods, such as support vector machines, decision trees, random forests, and deep neural networks, have been used to predict stroke functional outcomes, demonstrating improved performance compared to traditional risk scores.[13,14] However, these models mainly rely on clinical data, and do not incorporate imaging data, which contain valuable information for predicting stroke outcomes, such as the extent of tissue damage, penumbra, and collateral circulation. Recent studies have demonstrated the feasibility of incorporating imaging data into ML/DL models to predict stroke outcomes; however, most focused on selective populations undergoing reperfusion treatment, thus limiting their generalizability.[15-19] Furthermore, studies focusing on the general AIS population have predominantly focused on single-modality data from a single center.[20,21]

This study aimed to predict long-term functional outcomes in AIS patients using a comprehensive model. The proposed model combines multiple magnetic resonance (MR) scans and clinical data from multiple centers. This ensemble model enhanced the performance, minimized biases, and reduced variations in the prediction results. Interpretability methods were employed to visualize the decision-making process based on unique patterns observed in the input data.

Methods

Data collection

Data were obtained from the Korean Stroke Neuroimaging Initiative (KOSNI) Registry, a prospective observational study conducted at 18 tertiary stroke centers in South Korea over an 8-year period (2011–2018). The study protocol was approved by the Institutional Review Board of Asan Medical Center (IRB number: 2013-0162), and informed consent was obtained from all participants. The inclusion criteria for enrollment in the KOSNI registry were as follows: (1) individuals aged >20 years and (2) those presenting with neurological symptoms indicative of stroke, including transient ischemic attacks (TIAs).

Of the 5,018 patients, 2,606 were eligible after applying the following exclusion criteria: (1) missing the 90-day modified Rankin Scale (mRS),[22] (2) presentation >24 hours after stroke onset, (3) absence of stroke lesions in baseline images, and (4) poor image quality or image preprocessing failure (Figure 1). Descriptive statistics comparing baseline characteristics between the study population and excluded participants are provided in Supplementary Table 1. The study defined binarized 90-day mRS outcomes >2, identifying stroke patients who required assistance for daily activities due to functional limitations.[23] Among the patients, 1,613 (61.90%) exhibited an mRS of 0–2, whereas 993 (38.10%) achieved an mRS of 3–6 at 90 days post-stroke.

Image data preprocessing

Baseline MR encompassed two subtypes of diffusion-weighted imaging (DWI): DWI with a b-value of 1,000 s/mm^2 (b1000) and apparent diffusion coefficient (ADC) map, and fluid-attenuated inversion recovery (FLAIR).

Data pre-processing for raw MR scans involved the following steps: N4 bias field correction[24] was applied to each modality, followed by skull stripping using a brain mask derived by K-means clustering. The DWI images (b1000 and ADC map) were aligned to the Montreal Neurological Institute 152 (MNI 152) space, with 2-mm isotropic voxels using the ANTs SyN registration algorithm.[25] The FLAIR images were subjected to linear coregistration to align each volume with the DWI space within the subjects. The images were then aligned to a standard space using

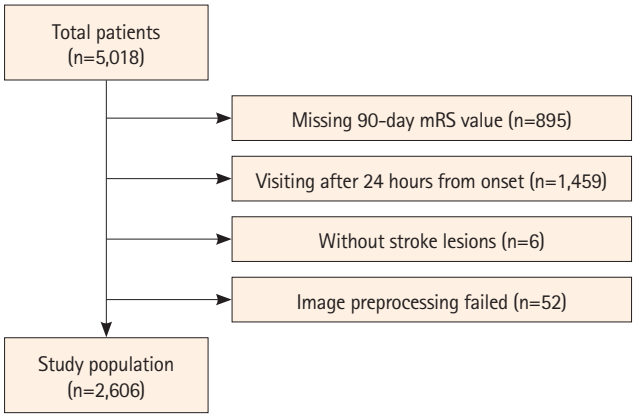


Figure 1. Inclusion and exclusion criteria. Flowchart showing the inclusion and exclusion criteria. mRS, modified Rankin Scale.

DWI deformation. The voxel intensity values in each image were normalized to ensure they fell within the range of 0–1.

Clinical data preprocessing

The clinical variables comprised 22 demographic and clinical features: age; sex; previous history of hypertension, diabetes, hyperlipidemia and stroke (including TIAs); current smoking status; body mass index (BMI); systolic blood pressure; diastolic blood pressure (DBP); hematocrit level; hemoglobin level; blood glucose level; creatinine level; total cholesterol level; high-density lipoprotein cholesterol (HDL-C); low-density lipoprotein cholesterol; total National Institutes of Health Stroke Scale (NIHSS)[8] at admission; duration between stroke onset and admission; reperfusion therapy status; risk status of cardiac embolic sources; and Trial of ORG 10172 in Acute Stroke Treatment (TOAST)[26] subtypes, including large-artery atherosclerosis, cardioembolism, small-vessel occlusion, other determined etiology, and undetermined etiology. Notably, each variable exhibited less than 5% missing data.

In the preprocessing of clinical variables, categorical features such as sex, past medical history, reperfusion therapy status, and TOAST subtypes were subjected to label encoding. Simple mode imputation was further applied to categorical variables with missing values. Conversely, all continuous variables were scaled according to the interquartile range (IQR), without any additional feature engineering, and imputed with the median value in cases where missing values were present.

Proposed approach

Model architecture

The prognostic model framework presented in Figure 2A involved the training of four different models using distinct modalities: clinical data, b1000, ADC map, and FLAIR. Supplementary Table 2 presents the hyperparameter details for each model.

For the clinical data, we employed a simple, fully connected neural network (FCN) consisting of three layers with eight hidden units. This FCN was trained using the Adam optimizer[27] and dropout regularization was applied to prevent overfitting.

By contrast, we used the 3D implementation version of ResNeXt[28] to extract features from the entire MR image. To enhance the performance, we incorporated the Convolutional Block Attention Module (CBAM)[29] after each ResNeXt block (Supplementary Figure 1A). CBAM is a lightweight and versatile attention mechanism that enables the model to focus on both spatial and channel features in the output feature map. It comprises two sequential submodules: a channel attention module and spatial attention module (Supplementary Figure 1B).

The channel attention module filters important information by passing the input feature maps through max-pooling and average-pooling layers, followed by a fully connected layer. The sigmoid function was applied to obtain the channel attention map M_C as follows:

$$M_C = \sigma\left(MLP\left(\mathrm{AvgPool}(F)\right) + MLP\left(\mathrm{MaxPool}(F)\right)\right), \quad (1)$$

where F denotes the input feature map; σ, the sigmoid function; and MLP, the multilayer perceptron in the channel attention module.

The spatial attention module uses the output attention map of the channel attention module to identify locations of meaningful information. The input features sequentially undergo max pooling, average pooling, and convolutional layers to generate the spatial attention map M_S:

$$M_S = \sigma\left(f^{7\times7}([\mathrm{AvgPool}(F); \mathrm{MaxPool}(F)])\right), \quad (2)$$

where $f^{7\times7}$ denotes convolutional operation in spatial attention module.

Following ResNeXt-CBAM, the extracted features comprised 2,048 nodes for the final classification. Class probability was obtained using the sigmoid function for the dichotomized mRS.

To ensure fast convergence and robust training, we used the Rectified Adam[30] optimizer with cosine-annealing learning rate scheduling.[31] To prevent overfitting, additional strategies were employed, including early stopping and RandAugment,[32] a stochastic automated data augmentation method that applies several transformation methods (Supplementary Figure 2).

The dataset used in this study exhibited a class imbalance, which could introduce bias toward the majority class during training. To mitigate this issue, we used focal loss[33] in each single-modality model, which is a modified version of the cross-entropy loss that downweighs the loss assigned to well-classified examples.

Data fusion between baseline models

To improve model performance, we employed a data-fusion technique using a weighted average method. This approach combines probability vectors obtained from each model. The weights for the fusion were determined using the differential evolution method by optimizing the maximum F1 value of the ensemble model. Equation (3) illustrates the computation of the output probability distributions p_i of the single-modality model using the fusion weight w_i, where n denotes the number of models.

$$y = \sum_{i}^{n} w_i p_i. \quad (3)$$

https://doi.org/10.5853/jos.2023.03426

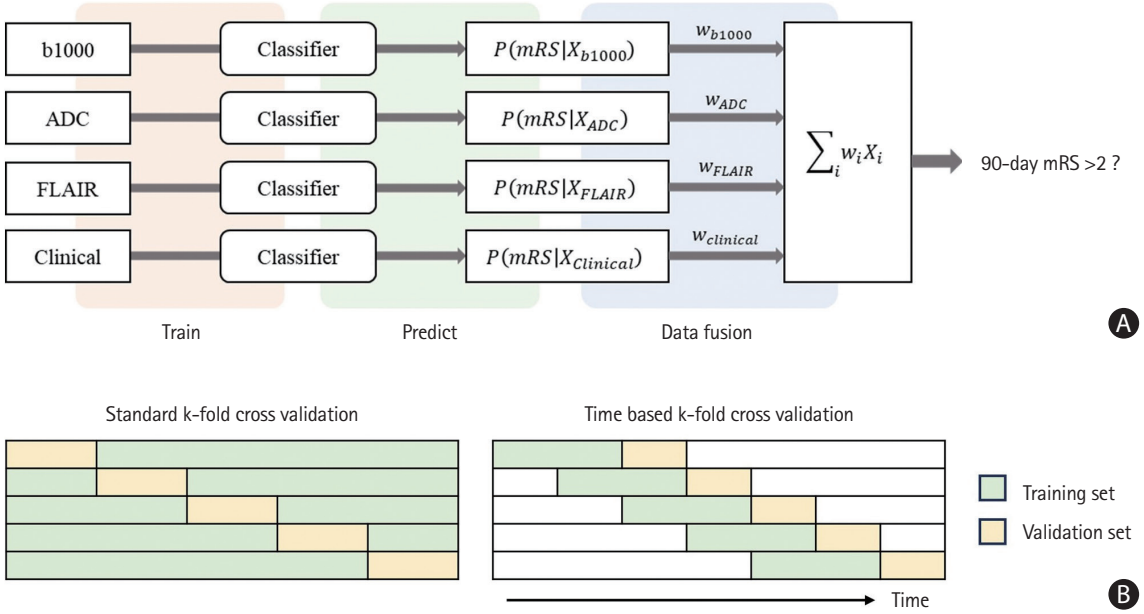


Figure 2. Multimodal classification scheme. (A) Overview of the multimodal classification scheme. Four different modality models (b1000, ADC map, FLAIR, and clinical data) were trained using the training data during the Train phase. These trained models were subsequently applied to the test data in the Predict phase, generating modality-specific outputs referred to as *P*. To obtain the final classification result, the weighted sum of the modality-specific outputs *P* was computed, with the weights (denoted as *w*) optimized during the training process. (B) Training and evaluating the model followed a dual scheme, comprising standard and time-based k-fold CV. b1000, b-value of 1,000 s/mm^2; mRS, modified Rankin Scale; ADC, apparent diffusion coefficient; FLAIR, fluid-attenuated inversion recovery.

Output classes can be effectively determined by generating a hybrid probability distribution with optimized weights (Supplementary Table 3). The final mRS prediction was derived by applying a threshold of 0.45 to calibrate the predicted probabilities on an imbalanced dataset, which was determined by maximizing the F1 score of the results from 5-fold cross validation.

Evaluation

We randomly selected 20% of the entire dataset as the test set to ensure that the distribution of output classes was identical to the remaining data. We employed two distinct approaches for model training and evaluation: standard k-fold cross-validation (CV) and time-based k-fold CV (Figure 2B). In the standard approach, a stratified 5-fold CV was used to maintain consistent proportions of output classes in each fold. In contrast, time-based k-fold CV adopts sliding window CV, a resampling technique used to manage time-series data. After sorting all data by admission date, the training data were split into multiple training and validation subsets. The window size was set to 1,000 instances in each round, and each window was divided into training and validation sets with an 80:20 split. In particular, the validation consistently preceded the training set.

We assessed the performance of the models by measuring sensitivity, specificity, positive predictive value (PPV), negative predictive value, F1 score, and area under the curve (AUC). The AUC was the primary metric. To compare our model with baseline models, we conducted a DeLong test to identify any statistically significant differences based on the models trained on the entire training dataset. The significance level for statistical tests was set at $P<0.05$. We further calculated 95% confidence intervals (CI) using 200 bootstrap samples.

To gain insight into the decision-making processes of each model, we used two explainable AI methods. For the clinical data model, we used the kernel Shapley Additive Explanation (SHAP)[34] to estimate the contribution of each input feature. SHAP calculates global feature importance by averaging the contributions through sample permutation. For the imaging data, we used Grad-CAM[35] to visualize significant brain regions and classify mRS outcomes. Grad-CAM determines the weights of the feature maps based on model information using the global average of the gradient. We obtained voxel-wise average heat maps using Grad-CAM for all AIS patient samples in the test data, and defined the region of interest (ROI) by applying a 50% threshold of voxel intensities. To identify the concentrated areas within the ROI mask, we used automated anatomical labeling to identify specific brain regions.

Experimental setup

All experiments were conducted on a Linux Ubuntu 20.04 LTS workstation with an Intel CPU i9-9940X 3.30 GHz, two NVIDIA

GeForce GTX 2080Ti graphics cards, and 64 GB of RAM. The DL models were implemented and trained in Python 3.8.10 using TensorFlow[36] 2.9.0. For image processing, OpenCV[37] 4.7.0 and scikit-image[38] 0.19.3 were used. The scikit-learn[39] 1.1.3 package was used for model evaluation and training. The interpretability of the clinical data model was visualized using the SHAP[34] 0.41.0 package. Grad-CAM-derived ROI-to-brain anatomy mapping was analyzed using the AtlasQuery tool of the FMRIB Software Library[40] based on the MNI structural atlas, Harvard–Oxford cortical structural atlas, and Harvard–Oxford subcortical structural atlas.

Results

Subjects

The study population comprised 2,606 patients selected from the total registry. Comparison of the baseline characteristics between the study population and excluded participants showed statistically significant differences in 9 features: age, history of diabetes and hyperlipidemia, DBP, Hematocrit, HDL-C, admission NIHSS, reperfusion therapy status, and TOAST subtypes (Supplementary Table 1).

Supplementary Table 4 presents the clinical and demographic characteristics of patients. The median age and baseline NIHSS score were 70 years (IQR, 61–76) and 5 (IQR, 2–10), respectively. After 90 days, 993 (38.1%) patients had poor functional outcomes (mRS score, 3–6), whereas 1,613 (61.9%) did not. Of those with poor functional outcomes, 795 belonged to the training group and 198 to the test group. No clinical inputs exhibited significant differences between the training and test groups (all $P>0.05$).

Prediction performance

Table 1 and Supplementary Table 5 present the average results of the standard 5 and time-based 5-fold CV for the evaluation of the performance of the proposed multimodal model in the prediction of functional outcomes. Compared to models trained with single modalities, our model consistently achieved the highest performance, with an AUC of 0.830 in standard CV, and 0.779 in time-based CV (95% confidence interval [CI]: 0.740, 0.844). All baseline models based on a single MR scan exhibited lower AUC values than our ensemble model.

The receiver operating characteristic curve (ROC) plot illustrated that the ensemble model outperformed those trained using a single modality. The proposed model showed a statistically significant improvement over clinical data ($P=0.004$), b1000 ($P<0.001$), ADC map ($P<0.001$), and FLAIR ($P<0.001$) on comparison of the ROC curves in the DeLong test (Figure 3).

Interpretable model analysis

In the clinical data model, SHAP values quantified the contribution of each feature (Figure 4) to the model results. Analysis

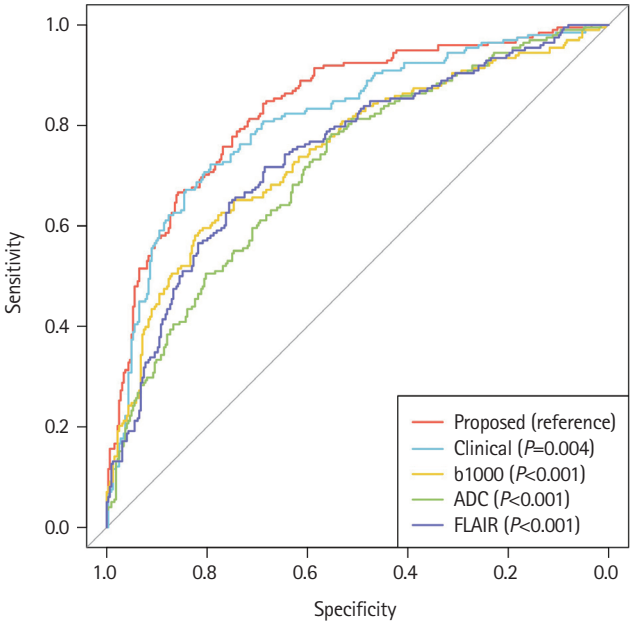


Figure 3. ROC curves for the proposed ensemble and single-modality models. b1000, b-value of 1,000 s/mm^2; ADC, apparent diffusion coefficient; FLAIR, fluid-attenuated inversion recovery; ROC, receiver operating characteristic curve.

Table 1. Performance of standard 5-fold CV with bootstrapped 95% CIs

	Clinical	b1000	ADC	FLAIR	Ensemble
AUC	0.814 (0.662–0.830)	0.748 (0.687–0.763)	0.713 (0.633–0.729)	0.731 (0.610–0.748)	0.830 (0.740–0.844)
F1	0.689 (0.577–0.722)	0.624 (0.568–0.637)	0.596 (0.511–0.620)	0.610 (0.517–0.651)	0.696 (0.619–0.722)
SEN	0.709 (0.661–0.803)	0.643 (0.572–0.776)	0.652 (0.532–0.751)	0.697 (0.557–0.808)	0.759 (0.671–0.818)
SPE	0.787 (0.432–0.794)	0.745 (0.501–0.833)	0.673 (0.492–0.723)	0.643 (0.351–0.716)	0.743 (0.576–0.791)
PPV	0.670 (0.454–0.694)	0.611 (0.491–0.678)	0.550 (0.464–0.574)	0.545 (0.432–0.557)	0.643 (0.527–0.682)
NPV	0.815 (0.764–0.852)	0.775 (0.739–0.799)	0.760 (0.697–0.784)	0.779 (0.699–0.827)	0.834 (0.781–0.858)

Classification performance of standard 5-fold CV experimental results (mean and 95% CI). We reported the following metrics for each model: AUC, F1 score, SEN, SPE, PPV, and NPV. The CIs for the model performance were calculated by bootstrapping samples.
CV, cross-validation; CI, confidence interval; b1000, b-value of 1,000 s/mm^2; ADC, apparent diffusion coefficient; FLAIR, fluid-attenuated inversion recovery; AUC, area under the curve; SEN, sensitivity; SPE, specificity; PPV, positive predictive value; NPV, negative predictive value.

https://doi.org/10.5853/jos.2023.03426

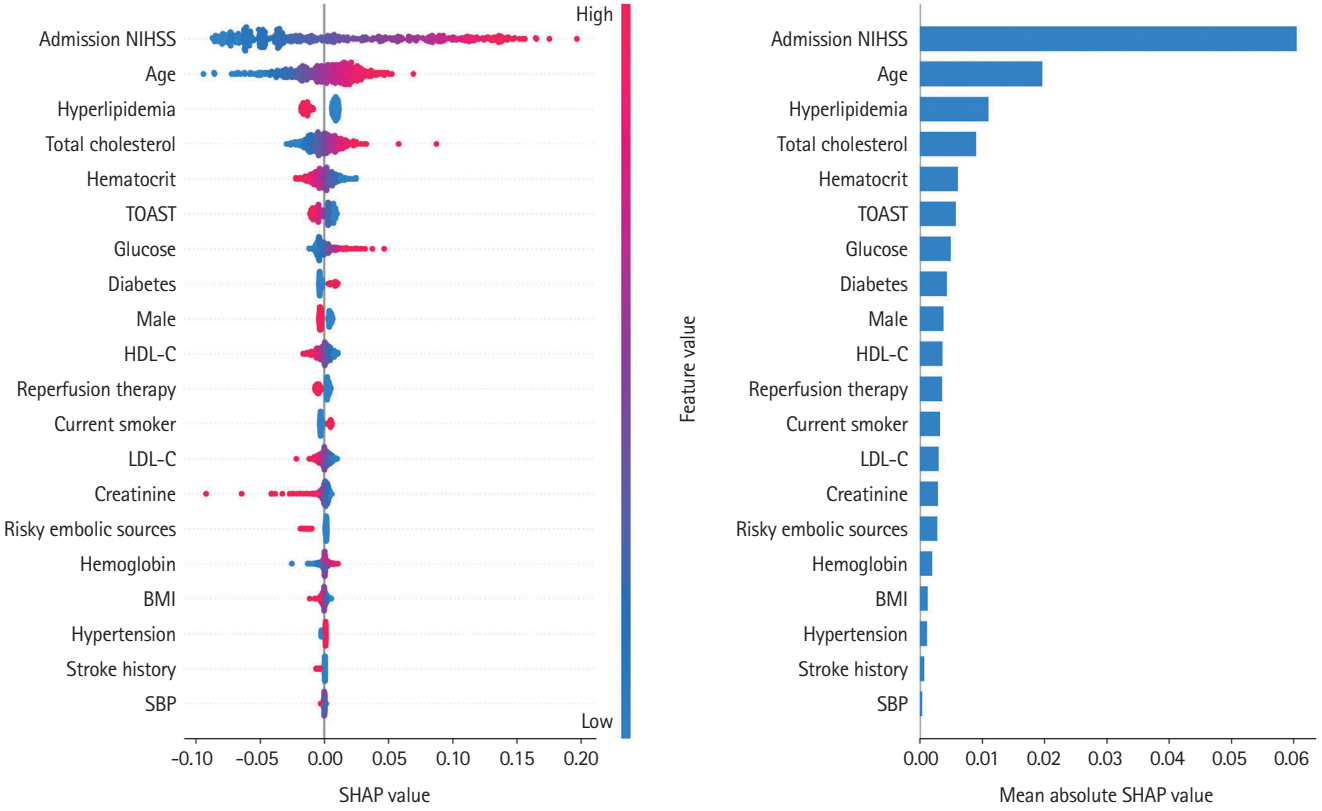


Figure 4. Visualization of SHAP for clinical metadata. The distribution of the SHAP values is presented on the left, whereas the mean absolute SHAP values are presented on the right. All features used in the training were included. The features are presented in order of importance, with the most important features at the top. The color scheme indicates the extent to which the feature values influence the outcome, with high values indicated in red. NIHSS, National Institutes of Health Stroke Scale; TOAST, Trial of ORG 10172 in Acute Stroke Treatment; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; BMI, body mass index; SBP, systolic blood pressure; SHAP, Shapley Additive Explanation.

using SHAP values revealed that age and baseline NIHSS score were the most influential features, whereas other clinical features had relatively minor impacts.

The image models generated average ROI heat maps using Grad-CAM, focusing on the infarcted area in the left hemisphere (Figure 5). Analysis of the different classification groups, true positive (TP), true negative (TN), false positive (FP), and false negative (FN), revealed a consistent ROI, but the intensity in the ROI for TN, FP, and FN was weaker than that for TP (Supplementary Figure 3). The precise anatomical locations of the ROI were consistent with the findings in the left cerebellum and temporo-occipital regions.

Discussion

In this study, we developed an ensemble DL model combining routinely collected multimodal imaging and clinical data to predict the functional outcomes of patients with AIS. Our approach involves the use of 3D CNN models to extract low-level features directly from high-dimensional input images combined with clinical model outputs. This integration led to an improved performance compared to the models trained by each single modality. Techniques such as image augmentation and focal loss were employed to minimize bias derived from data imbalance. Consequently, the final ensemble model achieved an AUC of 0.830 (standard CV) and 0.779 (time-based CV), outperforming the single-modality models.

The key strength of our study was our use of data collected from a multicenter registry, which provided a diverse and representative sample of patients with AIS. This broad coverage enabled training on various stroke types and locations, thus contributing to an improved model performance. Furthermore, the inclusion of diverse imaging protocols from multiple centers adds robustness to the prediction output of the model.

While the ensemble model achieved the highest average performance, the clinical- only model also showed good performance. However, it remains important to acknowledge the vulnerability of clinical data to data drift, as indicated by the wide range of 95% CIs for the AUC (from 0.662 to 0.830), despite the utilization of a multicenter data source. In contrast, our ensemble model exhibited robust performance, providing valuable mitigation against performance fluctuations caused by out-of-data dis-

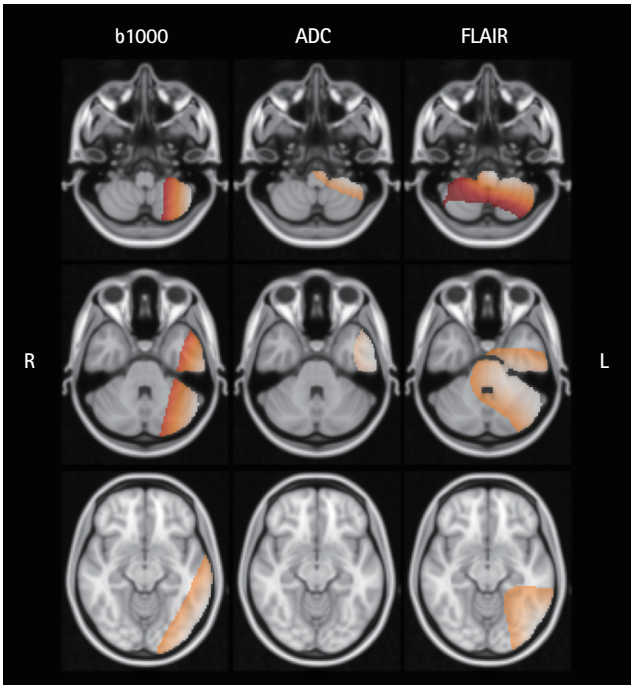


Figure 5. The ROI was determined by applying a 50% intensity threshold to identify TPs. The selected slices are positioned at the following z coordinates in the MNI 152 space in mm: -52, -32, -12. L and R denote left and right sides, respectively. b1000, b-value of 1,000 s/mm^2; ADC, apparent diffusion coefficient; FLAIR, fluid-attenuated inversion recovery; ROI, region of interest; TP, true positive; MNI 152, Montreal Neurological Institute 152.

tributions between the training and test sets. This stability makes the ensemble model particularly advantageous in real-world clinical settings with variations across different hospitals or imaging facilities.

To explore the relationship between the input data and poor functional outcomes, we conducted visual analyses using SHAP and Grad-CAM for each clinical and imaging input. The SHAP plot demonstrated a significant influence of age and NIHSS score on the prediction of functional outcomes based on clinical data, consistent with previous studies.[15] Grad-CAM was used to show which brain regions the model focused on. Interestingly, the average Grad-CAM graphs included distinct brain areas, rather than entire lesions. These findings suggest that poor functional outcomes are correlated with the left cerebellum and temporo–occipital region. Stroke functional outcomes were evaluated using the mRS, which evaluates the level of disability or dependency in daily activities influenced by motor function, balance, and visual function. However, the NIHSS assigns fewer points to ataxia and visual function (two and three points, respectively, out of 42 NIHSS points) than to motor function (19 points, including four points for each limb and three points for facial function). Consequently, given the significant clinical factors of age and the NIHSS score, in the context of imaging factors, the cerebellum, which is responsible for balance control, and the temporo–occipital cortex, which houses the optic pathway, may have been associated with unfavorable outcomes.

Despite these promising results, this study had several limitations. First, our ensemble model was dependent on MR imaging, making it unsuitable for use in many institutions that primarily use CT-based imaging for stroke diagnosis. Thus, the application of our model may be limited to facilities with MRI capabilities. Furthermore, our data may have been subject to several biases. Our study population comprised patients who visited the stroke center within 24 hours of stroke onset, exhibiting higher baseline severity, with an average admission NIHSS score of 5, compared to 3 in the excluded group. This severity mismatch can affect the distribution of each clinical feature between the study population and excluded patients, which could potentially limit the model generalizability.[41] The lack of full lesion growth may also have negatively affected the performance of the model, as most images in the dataset were early baseline images. Similar research has reported improved results using day 1 follow-up images, in which the lesion sizes were clearly seen.[21] Moreover, our model's performance should be further validated in real-world settings as real-world data often show data drift due to variations in distribution over time or across different data sources.[42] Future studies should investigate the effects of these biases to provide more comprehensive insights.

Model training also encountered certain challenges. Training DL models from scratch is inherently difficult as they require extensive data, while small datasets can lead to overfitting. Although we attempted to mitigate overfitting using techniques such as RandAugment and early stopping, the size of the dataset remained limited. To address this, future studies should explore a transfer learning approach with a model pretrained on a larger external dataset, potentially improving the performance and mitigating overfitting. Another challenge during model training is the use of the entire image as an input, which may introduce noise during model training. To overcome these issues, we employed an attention mechanism to enable the model to focus more on the ROI and generate accurate feature extractions. Despite these efforts, some cases still exhibit noise owing to individual differences in lesion size and location. Future research should include strategies to filter out these noises to improve the model performance.

Conclusions

In this study, we constructed a comprehensive model for predicting the 90-day functional outcomes using multiple MR modalities and clinical metadata from a multicenter registry. This

https://doi.org/10.5853/jos.2023.03426

model was superior to other prediction models that rely on a single modality.

Supplementary materials

Supplementary materials related to this article can be found online at https://doi.org/10.5853/jos.2023.03426.

Funding statement

This research was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health and Welfare, Republic of Korea (HR18C0016), and a National IT Industry Promotion Agency (NIPA) grant funded by the Korean government (MSIT) (No. S0252-21-1001; Seoul, Republic of Korea).

Conflicts of interest

The authors have no financial conflicts of interest.

Author contribution

Conceptualization: DWK, HSJ. Study design: HSJ, DWK. Methodology: HSJ. Data collection: all authors. Investigation: HSJ. Statistical analysis: HSJ. Writing—original draft: HSJ. Writing—review & editing: all authors. Funding acquisition: DWK. Approval of final manuscript: all authors.

References

1. James SL, Abate D, Abate KH, Abay SM, Abbafati C, Abbasi N, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. *Lancet* 2018;392:1789-1858.
2. Langhammer B, Sunnerhagen KS, Lundgren-Nilsson Å, Sällström S, Becker F, Stanghelle JK. Factors enhancing activities of daily living after stroke in specialized rehabilitation: an observational multicenter study within the Sunnaas International Network. *Eur J Phys Rehabil Med* 2017;53:725-734.
3. Kugler C, Altenhöner T, Lochner P, Ferbert A. Does age influence early recovery from ischemic stroke? A study from the Hessian Stroke Data Bank. *J Neurol* 2003;250:676-681.
4. Venema E, Mulder MJHL, Roozenbeek B, Broderick JP, Yeatts SD, Khatri P, et al. Selection of patients for intra-arterial treatment for acute ischaemic stroke: development and validation of a clinical decision tool in two randomised trials. *BMJ* 2017;357:j1710.
5. Paker N, Buğdaycı D, Tekdöş D, Kaya B, Dere C. Impact of cognitive impairment on functional outcome in stroke. *Stroke Res Treat* 2010;2010:652612.
6. Molina CA, Alexandrov AV, Demchuk AM, Saqqur M, Uchino K, Alvarez-Sabín J. Improving the predictive accuracy of recanalization on stroke outcome in patients treated with tissue plasminogen activator. *Stroke* 2004;35:151-156.
7. Nichols-Larsen DS, Clark PC, Zeringue A, Greenspan A, Blanton S. Factors influencing stroke survivors' quality of life during subacute recovery. *Stroke* 2005;36:1480-1484.
8. Saposnik G, Guzik AK, Reeves M, Ovbiagele B, Johnston SC. Stroke prognostication using age and NIH stroke scale: SPAN-100. *Neurology* 2013;80:21-28.
9. Boers AMM, Jansen IGH, Beenen LFM, Devlin TG, San Roman L, Heo JH, et al. Association of follow-up infarct volume with functional outcome in acute ischemic stroke: a pooled analysis of seven randomized trials. *J Neurointerv Surg* 2018;10:1137-1142.
10. Laredo C, Zhao Y, Rudilosso S, Renú A, Pariente JC, Chamorro Á, et al. Prognostic significance of infarct size and location: the case of insular stroke. *Sci Rep* 2018;8:9498.
11. Kim YD, Choi HY, Jung YH, Yoo J, Nam HS, Song D, et al. The ischemic stroke predictive risk score predicts early neurological deterioration. *J Stroke Cerebrovasc Dis* 2016;25:819-824.
12. Ntaios G, Faouzi M, Ferrari J, Lang W, Vemmos K, Michel P. An integer-based score to predict functional outcome in acute ischemic stroke: the ASTRAL score. *Neurology* 2012;78:1916-1922.
13. Jang SK, Chang JY, Lee JS, Lee EJ, Kim YH, Han JH, et al. Reliability and clinical utility of machine learning to predict stroke prognosis: comparison with logistic regression. *J Stroke* 2020;22:403-406.
14. Heo J, Yoon JG, Park H, Kim YD, Nam HS, Heo JH. Machine learning-based model for prediction of outcomes in acute stroke. *Stroke* 2019;50:1263-1265.
15. van Os HJA, Ramos LA, Hilbert A, van Leeuwen M, van Walderveen MAA, Kruyt ND, et al. Predicting outcome of endovascular treatment for acute ischemic stroke: potential value of machine learning algorithms. *Front Neurol* 2018;9:784.
16. Ramos LA, Kappelhof M, van Os HJA, Chalos V, Van Kranendonk K, Kruyt ND, et al. Predicting poor outcome before endovascular treatment in patients with acute ischemic stroke. *Front Neurol* 2020;11:580957.
17. Alaka SA, Menon BK, Brobbey A, Williamson T, Goyal M, Demchuk AM, et al. Functional outcome prediction in ischemic stroke: a comparison of machine learning algorithms and re-

gression models. *Front Neurol* 2020;11:889.
18. Hilbert A, Ramos LA, van Os HJA, Olabarriaga SD, Tolhuisen ML, Wermer MJH, et al. Data-efficient deep learning of radiological image data for outcome prediction after endovascular treatment of patients with acute ischemic stroke. *Comput Biol Med* 2019;115:103516.
19. Samak ZA, Clatworthy P, Mirmehdi M. Prediction of thrombectomy functional outcomes using multimodal data. In: Papież B, Namburete A, Yaqub M, Noble J. Medical image understanding and analysis. MIUA 2020. Communications in computer and information science (vol 1248). Cham: Springer, 2020; 267-279.
20. Hatami N, Cho TH, Mechtouff L, Eker OF, Rousseau D, Frindel C. CNN-LSTM based multimodal MRI and clinical data fusion for predicting functional outcome in stroke patients. *Annu Int Conf IEEE Eng Med Biol Soc* 2022;2022:3430-3434.
21. Moulton E, Valabregue R, Piotin M, Marnat G, Saleme S, Lapergue B, et al. Interpretable deep learning for the prognosis of long-term functional outcome post-stroke using acute diffusion weighted imaging. *J Cereb Blood Flow Metab* 2023;43: 198-209.
22. van Swieten JC, Koudstaal PJ, Visser MC, Schouten HJ, van Gijn J. Interobserver agreement for the assessment of handicap in stroke patients. *Stroke* 1988;19:604-607.
23. Sulter G, Steen C, De Keyser J. Use of the Barthel index and modified Rankin scale in acute stroke trials. *Stroke* 1999;30: 1538-1541.
24. Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, et al. N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging* 2010;29:1310-1320.
25. Avants BB, Epstein CL, Grossman M, Gee JC. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med Image Anal* 2008;12:26-41.
26. Adams HP Jr, Bendixen BH, Kappelle LJ, Biller J, Love BB, Gordon DL, et al. Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment. *Stroke* 1993;24:35-41.
27. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv [Preprint]. 2014 [accessed 2023 April 18]. Available from: https://doi.org/10.48550/arXiv.1412.6980.
28. Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated residual transformations for deep neural networks. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21-26; Honolulu, HI, USA. New York: IEEE, 2017. p.5987-5995.
29. Woo S, Park J, Lee JY, Kweon IS. CBAM: convolutional block attention module. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y. Computer vision – ECCV 2018. Lecture notes in computer science (vol 11211). Cham: Springer, 2018;3-19.
30. Liu L, Jiang H, He P, Chen W, Liu X, Gao J, et al. On the variance of the adaptive learning rate and beyond. arXiv [Preprint]. 2019 [accessed 2023 March 7]. Available from: https://doi.org/10.48550/arxiv.1908.03265.
31. Loshchilov I, Hutter F. SGDR: stochastic gradient descent with warm restarts. arXiv [Preprint]. 2016 [accessed 2023 March 7]. Available from: https://doi.org/10.48550/arxiv.1608.03983.
32. Cubuk ED, Zoph B, Shlens J, Le QV. RandAugment: practical automated data augmentation with a reduced search space. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2020 Jun 14-19; Seattle, WA, USA. New York: IEEE, 2020. p.3008-3017.
33. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. *IEEE Trans Pattern Anal Mach Intell* 2020; 42:318-327.
34. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Guyon I, Luxburg U Von, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al. editors, eds. 31st Conference on Neural Information Processing Systems (NIPS 2017); 2017 Dec 4-9; Long Beach, CA, USA. San Diego, CA: Advances in Neural Information Processing Systems; 2017. p. 4765-4774.
35. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int J Comput Vis* 2020;128:336-359.
36. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. Tensorflow: a system for large-scale machine learning. Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation; 2016 Nov 2–4; Savannah, GA, USA. Berkeley: USENIX Association; 2016. p. 265-283.
37. Bradski G. The opencv library. Dr Dobb's Journal: Software Tools for the Professional Programmer 2000;25:120-123.
38. van der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, et al. Scikit-image: image processing in Python. *PeerJ* 2014;2:e453.
39. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011;12:2825-2830.
40. Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM. FSL. *Neuroimage* 2012;62:782-790.
41. Corso G, Bottacchi E, Tosi P, Caligiana L, Lia C, Veronese Morosini M, et al. Outcome predictors in first-ever ischemic stroke patients: a population-based study. *Int Sch Res Notices* 2014; 2014:904647.
42. Turhan B. On the dataset shift problem in software engineering prediction models. *Empir Software Eng* 2012;17:62-74.

https://doi.org/10.5853/jos.2023.03426

Supplementary Table 1. Baseline characteristics of included and excluded individuals in the study

Characteristics	Total (n=5,018)	Included (n=2,606)	Excluded (n=2,412)	*P*	#NA
Age (yr)	69 (60–76)	70 (61–76)	69 (58–76)	0.001	1
Male sex	2,987 (59.53)	1,574 (60.40)	1,413 (58.58)	0.200	0
Hypertension	3,377 (67.30)	1,748 (67.08)	1,629 (67.54)	0.735	1
Diabetes	1,441 (28.72)	735 (28.20)	706 (29.27)	<0.001	270
Hyperlipidemia	2,043 (40.71)	1,001 (38.41)	1,042 (43.20)	<0.001	6
Current smoking status	1,776 (35.39)	906 (34.77)	870 (36.07)	0.354	3
Previous stroke including TIA	419 (8.35)	212 (8.14)	207 (8.58)	0.602	0
BMI (kg/m^2)	23.53 (21.63–25.64)	23.5 (21.48–25.51)	23.61 (21.73–25.78)	0.070	120
SBP (mm Hg)	140 (129–160)	140 (130–160)	140 (128–160)	0.236	10
DBP (mm Hg)	84 (76–96)	86 (79–100)	82.5 (75–92)	<0.001	9
Hematocrit (%)	40.6 (37.3–43.8)	40.4 (37.2–43.7)	40.85 (37.6–43.9)	0.010	23
Hemoglobin (g/dL)	13.8 (12.6–15)	13.8 (12.6–15)	13.9 (12.7–15.1)	0.190	8
Glucose (mg/dL)	126 (107–161)	126 (108–159)	125 (105–163)	0.390	11
Creatinine (mg/dL)	0.84 (0.7–1.01)	0.83 (0.7–1.01)	0.86 (0.7–1.01)	0.794	10
Total cholesterol (mg/dL)	176 (150–205)	176.5 (149–204.25)	176 (150–206)	0.548	25
HDL-C (mg/dL)	46 (38–55)	47 (38–56)	45 (37–54)	<0.001	146
LDL-C (mg/dL)	112 (87–137)	112 (86–137)	111 (87–138)	0.917	118
Admission NIHSS	4 (2–9)	5 (2–10)	3 (1–5)	<0.001	2
Reperfusion therapy	1,361 (27.12)	988 (37.02)	373 (15.88)	<0.001	0
Risk of cardiac embolic sources	465 (9.27)	247 (9.48)	218 (9.04)	0.625	0
TOAST classification				<0.001	28
LAA	1,424 (28.38)	675 (25.90)	749 (31.05)		
CE	1,251 (24.93)	768 (29.47)	483 (20.02)		
SVO	1,080 (21.52)	494 (18.96)	586 (24.30)		
OE	169 (3.37)	63 (2.42)	106 (4.39)		
UE	1,066 (21.24)	598 (22.95)	468 (19.40)		

Baseline characteristics of the included and excluded individuals. Values are expressed as n (%) or medians (interquartile ranges). Statistical tests were conducted using the Wilcoxon signed-rank test for continuous variables, the chi-square test for binary categorical variables, and Fisher's exact test for multiple categorical variables. #NA column provides the actual numbers of each variable for missing data.

TIA, transient ischemic attack; BMI, body mass index; SBP, systolic blood pressure; DBP, diastolic blood pressure; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; NIHSS, National Institutes of Health Stroke Scale; TOAST, Trial of ORG 10172 in Acute Stroke Treatment; LAA, large-artery atherosclerosis; CE, cardioembolism; SVO, small-vessel occlusion; OE, other determined etiology; UE, undetermined etiology.

Supplementary Table 2. Detailed hyperparameters for the imaging and clinical models

Hyperparameter	Value
Imaging data model (CBAM-ResNeXt50)	
Batch size	8
Optimizer	RAdam
Initial learning rate	0.001
Learning rate scheduler	Cosine annealing scheduler with warm restarts
Fraction of initial learning rate	0.5
Steps to decay over	30
Dropout	0.1
Loss function	Binary focal loss
Output activation function	Sigmoid
Image augmentation	RandAugment
Number of operations	1
Magnitude	0.1
Clinical data model (FCN)	
Batch size	32
Hidden layers	8-8-8
Activation function	ELU
Optimizer	Adam
Initial learning rate	0.001
Batch normalization	True
Dropout	0.5
Loss function	Binary focal loss
Output activation function	Sigmoid

List of hyperparameters used in each model.
CBAM, Convolutional Block Attention Module; RAdam, Rectified Adam; FCN, fully connected neural network; ELU, Exponential Linear Unit.

Supplementary Table 3. Optimized weights of each modality

Modality	Weight
Clinical	0.707
b1000	0.055
ADC	0.012
FLAIR	0.226

The weights given to each single modality model.
b1000, b-value of 1,000 s/mm^2; ADC, apparent diffusion coefficient; FLAIR, fluid-attenuated inversion recovery.

https://doi.org/10.5853/jos.2023.03426

Supplementary Table 4. Baseline characteristics of the study subjects

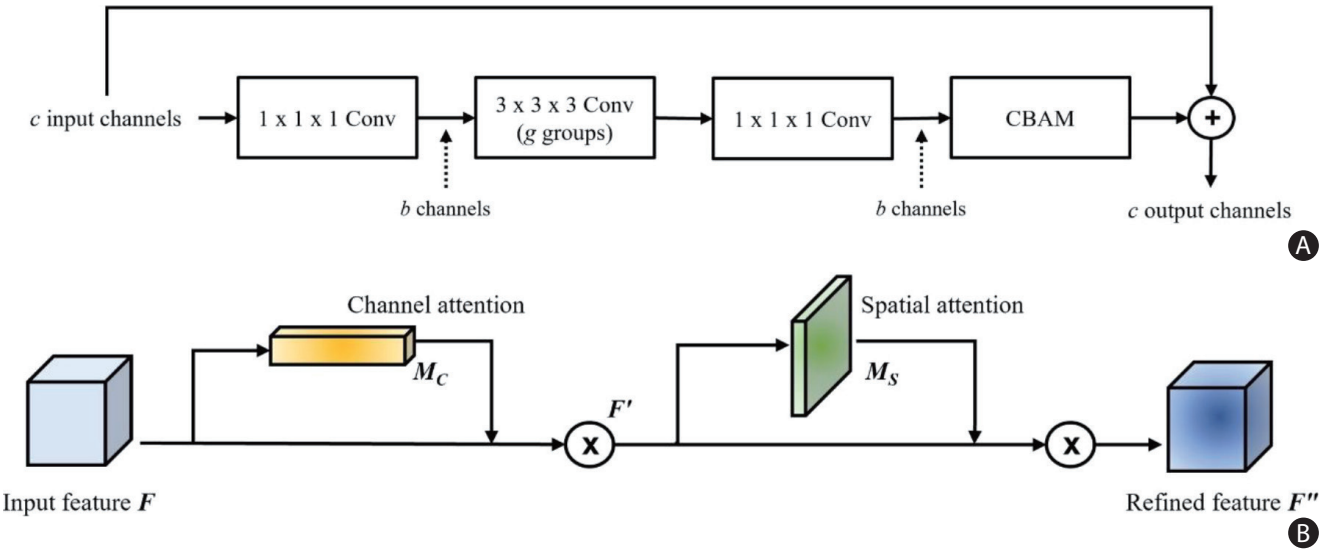
Characteristics	Total (n=2,606)	90-day mRS ≤2 (n=1,613)	90-day mRS >2 (n=993)	*P*	#NA
Age (yr)	70 (61–76)	67 (58–74)	73 (66–79)	<0.001	1
Male sex	1,574 (60.40)	1,045 (64.79)	529 (53.27)	<0.001	0
Hypertension	1,748 (67.08)	1,078 (66.83)	670 (67.47)	0.768	0
Diabetes	735 (28.20)	434 (26.91)	301 (30.31)	0.071	3
Hyperlipidemia	1,001 (38.41)	722 (44.76)	279 (28.10)	<0.001	5
Current smoking status	906 (34.77)	601 (37.26)	305 (30.72)	<0.001	2
Previous stroke including TIA	212 (8.14)	142 (8.80)	70 (7.05)	0.129	0
BMI (kg/m^2)	23.5 (21.48–25.51)	23.71 (21.92–25.78)	23.12 (20.83–24.98)	<0.001	88
SBP (mm Hg)	140 (130–160)	142 (130–160)	140 (120–160)	<0.001	4
DBP (mm Hg)	86 (79–100)	88 (80–100)	83.5 (71–95)	<0.001	3
Hematocrit (%)	40.4 (37.2–43.7)	41.1 (37.8–44.2)	39.4 (36.1–42.6)	<0.001	13
Hemoglobin (g/dL)	13.8 (12.6–15)	14 (12.9–15.2)	13.4 (12.2–14.7)	<0.001	3
Glucose (mg/dL)	126 (108–159)	123 (106–154)	132 (111–166)	<0.001	6
Creatinine (mg/dL)	0.83 (0.7–1.01)	0.86 (0.7–1.02)	0.8 (0.7–1)	0.041	3
Total cholesterol (mg/dL)	176.5 (149–204.25)	177 (149–206)	175 (149–203)	0.266	14
HDL-C (mg/dL)	47 (38–56)	47 (38–56)	47 (38–57)	0.881	86
LDL-C (mg/dL)	112 (86–137)	112 (86–137)	111 (87–135.5)	0.512	71
Admission NIHSS	5 (2–10)	3 (1–6)	10 (5–15)	<0.001	0
Duration between stroke onset and admission	4.04 (1.82–10.6)	4.32 (1.78–11.35)	3.83 (1.85–9.27)	0.089	0
Reperfusion therapy	988 (37.02)	471 (28.32)	517 (51.39)	<0.001	0
Risk of cardiac embolic sources	247 (9.48)	151 (9.36)	96 (9.67)	0.849	0
TOAST classification				<0.001	8
LAA	675 (25.90)	397 (24.61)	278 (28.00)		
CE	768 (29.47)	405 (25.11)	363 (36.56)		
SVO	494 (18.96)	395 (24.49)	99 (9.97)		
OE	63 (2.42)	45 (2.79)	18 (1.81)		
UE	598 (22.95)	365 (22.63)	233 (23.46)		

Baseline characteristics of the study subjects, stratified by 90-day functional status (90-day mRS score >2 or not). Values are expressed as numbers (%) or medians (interquartile ranges). Statistical tests were conducted using the Wilcoxon signed-rank test for continuous variables, the chi-square test for binary categorical variables, and Fisher's exact test for multiple categorical variables. #NA column provides the actual numbers of each variable for missing data.
mRS, modified Rankin Scale; TIA, transient ischemic attack; BMI, body mass index; SBP, systolic blood pressure; DBP, diastolic blood pressure; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; NIHSS, National Institutes of Health Stroke Scale; TOAST, Trial of ORG 10172 in Acute Stroke Treatment; LAA, large-artery atherosclerosis; CE, cardioembolism; SVO, small-vessel occlusion; OE, other determined etiology; UE, undetermined etiology.

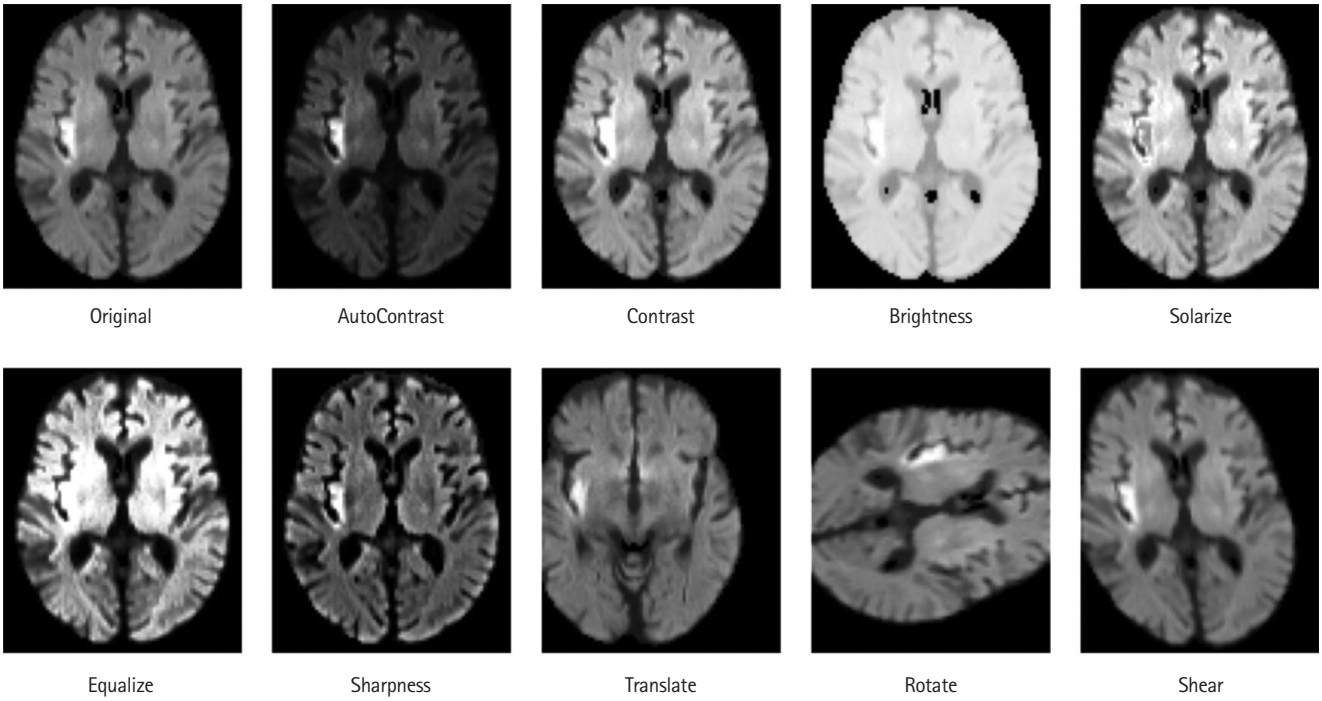
Supplementary Table 5 . Average classification performance of the time-based CV

	Clinical	b1000	ADC	FLAIR	Ensemble
AUC	0.745	0.728	0.694	0.669	0.779
F1	0.609	0.597	0.570	0.559	0.642
SEN	0.683	0.589	0.638	0.637	0.693
SPE	0.658	0.767	0.641	0.614	0.716
PPV	0.564	0.610	0.521	0.506	0.607
NPV	0.777	0.755	0.747	0.740	0.796

Average classification performance of the time-based 5-fold CV. The following metrics were reported for each model: AUC, F1 score, SEN, SPE, PPV, and NPV.
CV, cross-validation; AUC, area under the curve; SEN, sensitivity; SPE, specificity; PPV, positive predictive value; NPV, negative predictive value.

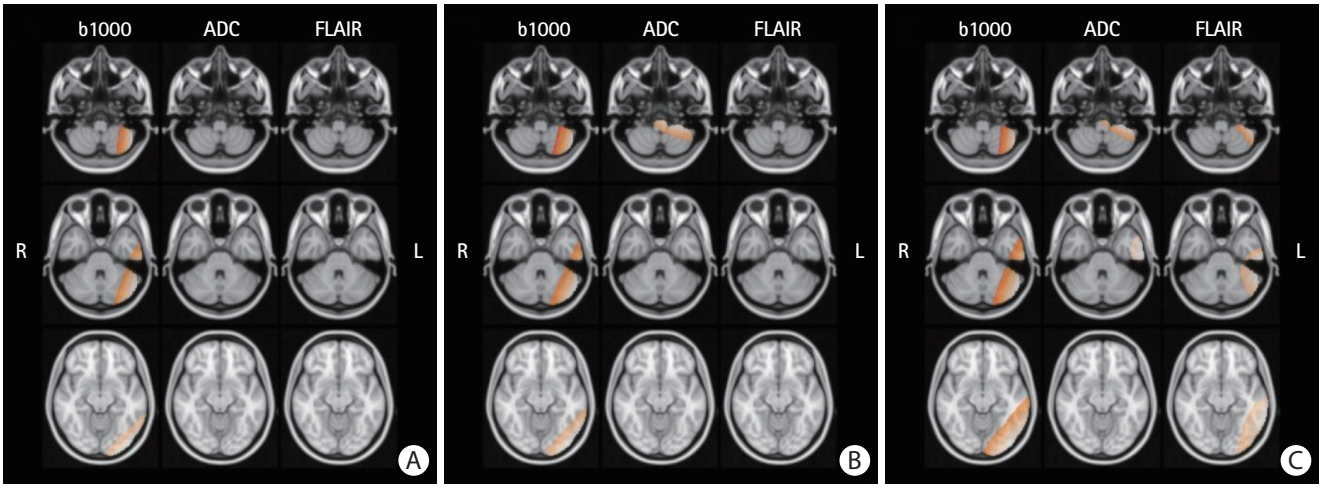


Supplementary Figure 1. Overview of the image model architecture. (A) The Structure of the CBAM-ResNeXt residual block, where *c* denotes the channel size of the input and output features, *b* denotes the number of intermediate channels, and *g* indicates the cardinality size. (B) CBAM attention mechanism. Conv, convloutionaly layer; CBAM, Convolutional Block Attention Module.



Supplementary Figure 2. An example of RandAugment. Sample images augmented by RandAugment with hyperparameters (Number of operations=1 and Magnitude=0.1).

https://doi.org/10.5853/jos.2023.03426



Supplementary Figure 3. The ROI were determined by applying a 50% intensity threshold to identify the following cases: (A) TNs, (B) FPs, and (C) FNs. The selected slices are positioned at the following z coordinates in the MNI 152 space in mm: -52, -32, -12. L and R denote left and right sides, respectively. b1000, b-value of 1,000 s/mm^2; ADC, apparent diffusion coefficient; FLAIR, fluid-attenuated inversion recovery; ROI, region of interest; TN, true negative; FP, false positive; FN, false negative; MNI 152, Montreal Neurological Institute 152.