scientific reports

Check for updates

OPEN

Deep learning-based classification of diffusion-weighted imaging-fluid-attenuated inversion recovery mismatch

Pum Jun Kim[1,9], Dongyoung Kim[1,9], Joonwon Lee[2], Hyung Chan Kim[3], Jung Hwa Seo[4], Suk Yoon Lee[5], Doo Hyuk Kwon[6], Hyungjong Park[7], Jaejun Yoo[1]✉ & Seongho Park[2,8]✉

The presence of a diffusion-weighted imaging (DWI)–fluid-attenuated inversion recovery (FLAIR) mismatch holds potential value in identifying candidates for recanalization treatment. However, the visual assessment of DWI–FLAIR mismatch is subject to limitations due to variability among raters, which affects accuracy and consistency. To overcome these challenges, we aimed to develop and validate a deep learning-based classifier to categorize the mismatch. We screened consecutive acute ischemic stroke patients who underwent DWI and FLAIR imaging from a four stroke centers. Two centers were used for model development and internal testing (derivation cohort), while two independent centers served as external validation cohorts. We developed Convolutional Neural Network-based classifiers for two binary classifications: DWI–FLAIR match versus non-match (Label Set I) and match versus mismatch (Label Set II). A total of 2369 patients from the derivation set and 679 patients from two external validation sets (350 and 329 patients) were included in the analysis. For Label Set I, the internal test set AUC was 0.862 (95% CI 0.841–0.884, with external validation AUCs of 0.829 (0.785–0.873) and 0.835 (0.790–0.879). Label Set II showed higher performance with internal test AUC of 0.934 (0.911–0.957) and external validation AUCs of 0.883 (0.829–0.938) and 0.913 (0.876–0.951). A deep learning-based classifier for the DWI–FLAIR mismatch can be used to diminish subjectivity and support targeted decision-making in the treatment of acute stroke patients.

Keywords Diffusion-FLAIR mismatch, Ischemic stroke, Cerebral infarction, Deep learning, Machine learning

The presence of a diffusion-weighted imaging (DWI)–fluid-attenuated inversion recovery (FLAIR) mismatch can be utilized to identify patients who may benefit from intravenous thrombolytic treatment in an uncertain time window[1–5]. Additionally, it may hold value in identifying candidates for recanalization treatment due to its potential to reflect salvageable tissue post-reperfusion[6–8].

During the progression of ischemic stroke, the temporal changes captured in DWI and FLAIR images likely reflect different stages of tissue injury. After cerebral artery occlusion, diffusion changes, indicating cytotoxic edema, appear before FLAIR changes, which reflect vasogenic edema following blood-brain barrier disruption (Figure S1)[9–13]. Therefore, the DWI-FLAIR mismatch may serve as a marker indicating the occurrence of cytotoxic edema, but before irreversible changes have set in.

Interestingly, recent randomized controlled trials have suggested that the traditional concept of irreversible tissue damage might need reconsideration, as even patients with large core infarcts showed benefits from

[1]Graduate School of Artificial Intelligence, Ulsan National Institute of Science and Technology, Ulsan, Republic of Korea. [2]Department of Neurology, Inje University College of Medicine, Inje University Haeundae Paik Hospital, Busan, Republic of Korea. [3]Department of Neurology, Ulsan Hospital, Ulsan, Republic of Korea. [4]Department of Neurology, Dong-A University Hospital, Dong-A University College of Medicine, Busan, Republic of Korea. [5]Department of Neurology, Busan Paik Hospital, Inje University College of Medicine, Busan, Republic of Korea. [6]Department of Neurology, Yeungnam University College of Medicine, Daegu, Republic of Korea. [7]Department of Neurology, School of Medicine, Keimyung University, Dageu, Republic of Korea. [8]Department of Neurology, College of Medicine, Hanyang University, Seoul, Korea. [9]Pum Jun Kim and Dongyoung Kim contributed equally to this work. ✉email: jaejun.yoo@unist.ac.kr; risepsh@gmail.com

recanalization therapy[8,14–16]. In this evolving landscape of stroke treatment, the mismatch could potentially offer additional insights into tissue viability assessment.

Despite its potential clinical utility, a major limitation in the assessment of the mismatch is its dependence on human visual interpretation to determine the presence of FLAIR hyperintense lesions[17]. This method is susceptible to interrater variability, which can impact both the accuracy and consistency of assessments.

To address this challenge, our study proposes the use of a deep learning-based classifier for measuring the mismatch. This approach could provide a more objective tool for clinicians, particularly when there's uncertainty in assessing the mismatch, aiding them in selecting the right patients for recanalization therapy.

Furthermore, unlike previous studies that have used DWI and FLAIR imaging to predict the time of stroke onset[18], we have focused on developing and validating a deep learning classifier that distinguishes between the presence or absence of the mismatch. This approach carries distinct clinical implications, introducing a novel and necessary perspective to the field.

Methods

Study design and data source

This diagnostic test accuracy study for a deep learning model classifying diffusion-FLAIR mismatch and match included four historical cohorts based on a prospectively collected stroke registries. Two cohorts, Haeundae Paik Hospital and Busan Paik Hospital, were used as derivation cohorts, while the remaining two cohorts, Yeungnam University Hospital (YN) and Keimyung University Dongsan Hospital (KM), served as external validation cohorts.

For this study, we selected patients over 18 years of age who, within a single study session, underwent imaging that included all four modalities: FLAIR, DWI (b = 1000 and b = 0 s/mm2), and Apparent Diffusion Coefficient (ADC) sequences, and were diagnosed with acute ischemic stroke.

Patients with a diffusion restriction volume exceeding 5 mL were included in the study. Lesions smaller than 5 mL, including scattered lesions too small to determine whether Diffusion-FLAIR match or mismatch, were excluded due to the difficulty in establishing a ground truth and the premise that the DWI-FLAIR mismatch classifier is intended to identify patients who could benefit from recanalization therapy. It was determined that lesions too small may not derive significant advantage from reperfusion, rendering them less relevant for the purposes of this mismatch classification. To select volumes over 5 mL, we utilized a segmentation model developed from previous research[19].

The study adhered to the STARD reporting guidelines[20]. Data was anonymized using de-identification methods provided by each data provider (Figure S2).

This retrospective study was conducted in accordance with the principles of the Declaration of Helsinki and was approved by the Institutional Review Board (IRB No. 2021-09-025-006) of Haeundae Paik Hospital. The requirement for informed consent was waived by the Institutional Review Board of Haeundae Paik Hospital due to the retrospective nature of the study.

Labeling process for imaging analysis

The labeling process involved the manual classification of lesions observed in simultaneously acquired DWI and FLAIR images from the same patient, based on signal intensity discrepancies, into five categories: (a) Diffusion-FLAIR (DF) match, (b) DF mismatch, (c) subtle FLAIR changes, (d) regional heterogeneous FLAIR changes, and (e) not indicative of ischemic stroke or not suitable for analysis (Fig. 1).

To reflect both clinical reality and ensure rigorous model evaluation, we designed two binary classification schemes. Binary Label Set I was designed to address real-world clinical scenarios by combining categories b (DF mismatch), c (subtle FLAIR changes), and d (regional heterogeneous FLAIR changes) as non-matches, contrasting them with category a (DF Match). However, recognizing that categories c and d represent ambiguous cases that could potentially affect the assessment of our model's core capability, we also defined Binary Label Set II. This second set focuses solely on contrasting category a (DF Match) with category b (DF Mismatch), allowing us to evaluate the model's fundamental ability to distinguish between clear-cut cases of match and mismatch patterns. This dual approach aims to enable both rigorous technical validation and assessment of real-world clinical utility.

To establish the ground truth for the mismatch, the manual labeling work was carried out by five stroke specialists, including four neuro-interventionists and one neurologist specializing in stroke. To enhance the accuracy and reliability of the labeling, serial independent review of a common dataset was performed, followed by an assessment of interobserver agreement. Multiple collaborative reviews were conducted until a unanimous decision was reached among the raters. After a consensus was reached among the raters, the remainder of the dataset was individually assigned to each of the five raters for final independent review. The raters were provided only with the neuroimaging data for evaluation and were not informed about the clinical information.

Data preparation and model development

All DWI, ADC, and FLAIR images of the subjects underwent the same image preprocessing steps, which included normalization, interpolation, and registration (Method S1). Based on Label Set I, we developed our binary classification algorithm using a 3D Convolutional Neural Network architecture (Method S2). For this purpose, 90% of the derivation set was utilized for training and tuning to determine the hyperparameters, while the remaining 10% was allocated for evaluation. The development of this algorithm was carried out using Python (version 3.8.17) and Pytorch (version 1.13.1 + cu177).

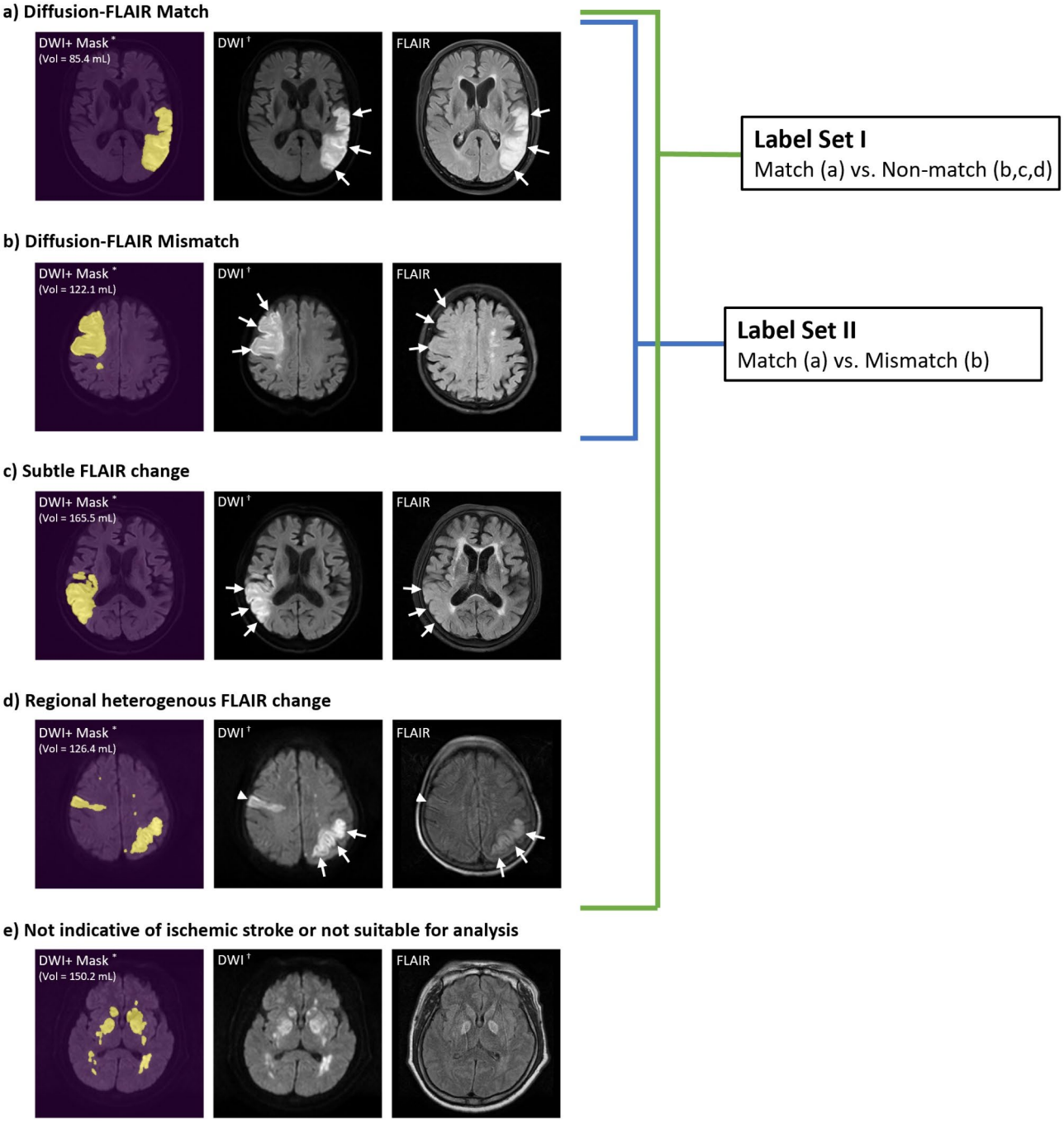

Fig. 1. Image Labeling Framework. Representative MRI scans illustrating the five categorization classes used by all raters for manual labeling of brain images. (**a**) Diffusion-FLAIR (DF) match, identified by diffusion restriction with corresponding FLAIR hyperintensity. (**b**) DF mismatch, characterized by the presence of diffusion restriction without corresponding FLAIR hyperintensity. (**c**) Subtle FLAIR change, where only faint FLAIR signal alterations are present. (**d**) Heterogeneous FLAIR change, assigned when DF mismatch is detected in motor eloquent areas. Here, the white arrow indicates a DF match, while the arrowhead indicates a DF mismatch. (**e**) Non-acute cerebral infarction cases, such as those found in hypoxic-ischemic encephalopathy. To simplify the classification process for binary decision-making, categorizations have been organized into two sets of binary labels. The first set, Binary Label Set I, groups cases that are DF Mismatches (**b**), exhibit subtle FLAIR changes (**c**), or show regional heterogeneous FLAIR changes (**d**) into a 'Non-match' category for comparison with Match cases. This set is designed to include the radiologic complexities encountered in clinical practice by considering a broader range of diagnostic scenarios as non-matches. The second set, Binary Label Set II, directly contrasts clear cases of DF Match (**a**) against Mismatch (**b**), thus facilitating a straightforward radiological assessment. * Images overlayed with masks predicted automatically using a segmentation model on DWI. Individual infarcted core volume measured based on the predicted masks. † Diffusion-weighted imaging with b = 1000. DWI, diffusion weighted imaging; FLAIR, fluid attenuated inversion recovery.

Algorithm validation and statistical analysis

To validate the developed algorithm, we conducted tests on both derivation Sets and two external validation datasets. The performance of the derivation dataset was assessed using a cross-validation method, as illustrated in Figure S3. For the external datasets, we evaluated the performance by using the entirety of each data set as a test set.

For each dataset, we evaluated the performance of our algorithm in discriminating the mismatch by examining the Receiver Operating Characteristic (ROC) curves and 95% Confidence Interval (CI) for both Label Set I and Label Set II. We compared differences between the AUCs using DeLong's test[16]. For secondary outcome metrics, we calculated positive predictive value, and negative predictive value. The sensitivity and specificity values at the threshold defined by Youden's index J (J = sensitivity + specificity − 1) were also determined.

To evaluate the accuracy and reliability of the predictive model, we calculated the Brier Score and generated Calibration plots[22,23]. The Brier Score quantitatively assesses the model's predicted probabilities against actual outcomes, and Calibration Plots visually demonstrate the agreement between the model's predictions and observed results across different probability percentiles. These methods offer a comprehensive evaluation of the model's performance in predicting actual outcomes.

To enhance the interpretability of the imaging model, saliency maps were generated using Guided Grad-CAM[24]. This technique combines gradient-based localization with class activation mappings to visually emphasize the critical regions within the input images that significantly contribute to the model's predictions.

We used the Fleiss' Kappa method to measure the agreement among multiple raters.

All statistical analyses were conducted using R version 4.1.3. We considered a 2-tailed P-value < 0.05 to be statistically significant.

Results

The analysis included 3,022 patients from the derivation cohort and 753 patients from the external validation cohorts, KM and YN (Figure S4). Manual labeling was performed on these individuals, resulting in a Fleiss' Kappa of 0.91 for the final labels.

After excluding cases not indicative of ischemic stroke or unsuitable for analysis, 2,369 patients from the derivation cohort were used for model development and validation. For the external validation, 350 patients from KM and 329 from YN were used. Within these groups, 1,443 patients (60.9%) in the derivation cohort, 213 (60.8%) in KM, and 160 (48.6%) in YN were identified as having a Diffusion-FLAIR match (Table 1).

Performances of the model

For Label Set I, the area under the curve (AUC) for the internal test set was 0.862 (95% CI : 0.841–0.884). For the external validation sets, KM and YN, the AUCs were 0.829 (95% CI: 0.785–0.873) and 0.835 (95% CI: 0.790–0.879), respectively. For Label Set II, the AUC for the internal test set was 0.934 (95% CI: 0.911–0.957), while for the external validation sets, KM and YN, the AUCs were 0.883 (95% CI: 0.829–0.938) and 0.913 (95% CI: 0.876–0.951), respectively (Table 2; Fig. 2). No significant differences were observed in the performance across the data sets.

Model calibration was undertaken to evaluate the probability that a new observation falls into each of the predefined categories. The calibration slopes demonstrated a negligible difference between the predicted and observed probabilities of DWI-FLAIR mismatch, suggesting an excellent fit of the model (Figure S5). The Brier scores for the internal test set and the external validation sets (KM and YN) were 0.02, 0.16, and 0.23, respectively, indicating the reliability of the model predictions across these cohorts.

Model interpretability analysis

To provide insights into the model's decision-making process, we generated saliency maps using Guided Grad-CAM (Fig. 3). The visualization revealed distinct activation patterns between match and mismatch cases. In match cases, the saliency maps showed focused activation primarily within the lesion areas where both DWI and FLAIR signals were positive. In contrast, for mismatch cases, the activation patterns extended beyond the core DWI lesion areas, suggesting the model's consideration of broader regional characteristics in its classification process. Representative cases demonstrating these distinctive activation patterns are shown in Fig. 3A (match cases) and 3B (mismatch cases).

	Derivation cohort ($n = 2,369$)	External cohort (KM) ($n = 350$)	External cohort (YN) ($n = 329$)
Age in years, mean (SD)	72.9 (13.5)	71.8 (13.3)	71.3 (12.4)
Male (%)	1,340 (56.6)	193 (55.1)	200 (60.8)
Core volume in mL*, median (IQR)	19.7 (9.4–54.0)	19.6 (9.7–55.5)	19.6 (8.9–49.2)
Labels			
Diffusion-FLAIR match (%)	1,443 (60.9)	213 (60.8)	160 (48.6)
Diffusion-FLAIR mismatch (%)	289 (12.2)	23 (6.6)	69 (21.0)
Subtle FLAIR change (%)	410 (17.3)	63 (18.0)	54 (16.4)
Regional heterogenous FLAIR change (%)	227 (9.6)	51 (14.6)	46 (14.0)

Table 1. Internal and External Data Sets in the algorithm development. *Estimated core volume of cerebral infarction based on segmentation model[14].

	Label set I*		
	Internal test set‡ (*n* = 1,185)	External test set (KM) (*n* = 350)	External test set (YN) (*n* = 329)
AUROC (95% CI)	0.862 (0.841–0.884)	0.829 (0.785–0.873)	0.835 (0.790–0.879)
Sensitivity (95% CI)	0.833 (0.805–0.861)	0.685 (0.623–0.748)	0.706 (0.636–0.777)
Specificity (95% CI)	0.754 (0.715–0.792)	0.839 (0.778–0.901)	0.870 (0.819–0.921)
PPV (95% CI)	0.826 (0.798–0.854)	0.869 (0.818–0.920)	0.837 (0.775–0.899)
NPV (95% CI)	0.763 (0.725–0.801)	0.632 (0.562–0.702)	0.758 (0.697–0.818)
	Label set II†		
	Internal test set‡ (*n* = 842)	External test set (KM) (*n* = 236)	External test set (YN) (*n* = 229)
AUROC (95% CI)	0.934 (0.911–0.957)	0.883 (0.829–0.938)	0.913 (0.876–0.951)
Sensitivity (95% CI)	0.872 (0.847–0.897)	0.779 (0.724–0.835)	0.844 (0.787–0.900)
Specificity (95% CI)	0.869 (0.816–0.923)	0.870 (0.732–1.000)	0.913 (0.847–0.980)
PPV (95% CI)	0.968 (0.954–0.982)	0.982 (0.962–1.002)	0.957 (0.924–0.991)
NPV (95% CI)	0.602 (0.537–0.666)	0.299 (0.189–0.408)	0.716 (0.622–0.810)

Table 2. Model performance metrics on internal and external test sets for label sets I and II. *Label Set I categorizes images as 'Diffusion-FLAIR Match' for a binary comparison against a combined 'Non-match' group, which encompasses 'Diffusion-FLAIR Mismatch,' 'Subtle FLAIR Change,' and 'Regional Heterogeneous FLAIR Change'.. †Label Set II represents a binary categorization distinguishing only between 'Diffusion-FLAIR Match' and ' Diffusion-FLAIR Mismatch,' excluding all other categories. ‡This number for internal test set represents the total number of patients aggregated from the results after performing a 5-fold validation on the 10% of patients allocated to the test set in the derivation cohort (Figure S2). AUROC, Area Under the Receiver Operating Characteristic; PPV, Positive Predictive Value; NPV, Negative Predictive Value.

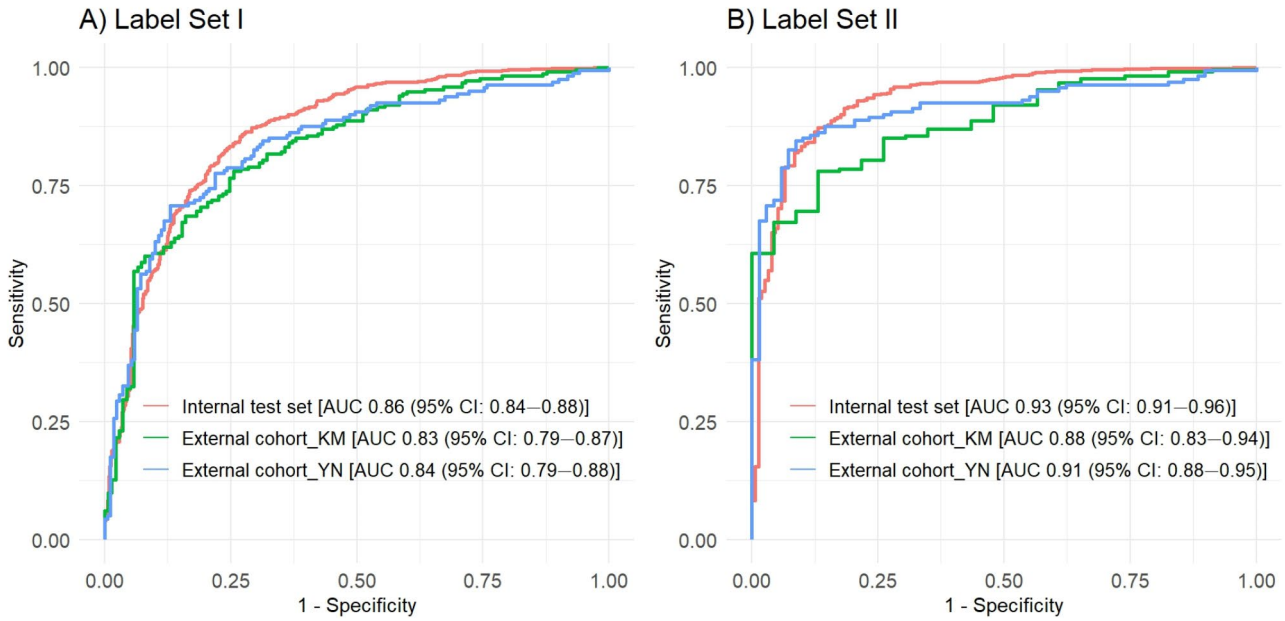

Fig. 2. Receiver operating characteristic curve comparisons for internal and external test sets (KM, YN) across label sets I (A) and II (B).

Discussions

In this study, we developed and evaluated a deep learning-based classifier that leverages routinely acquired MR images from patients in the early acute phase of stroke to assess Diffusion-FLAIR (DF) match status. This novel approach addresses the inherent limitations associated with the subjective visual assessment of the mismatch[17], a critical diagnostic imaging marker in the management of acute ischemic stroke. Our classifier represents the first of its kind to tackle the challenges of interrater variability that arise from traditional visual evaluations in clinical settings, providing a standardized and objective tool for mismatch assessment.

The model demonstrated robust performance in distinguishing between matches and non-matches within DF comparisons, with particularly outstanding capabilities in differentiating clear matches from definitive

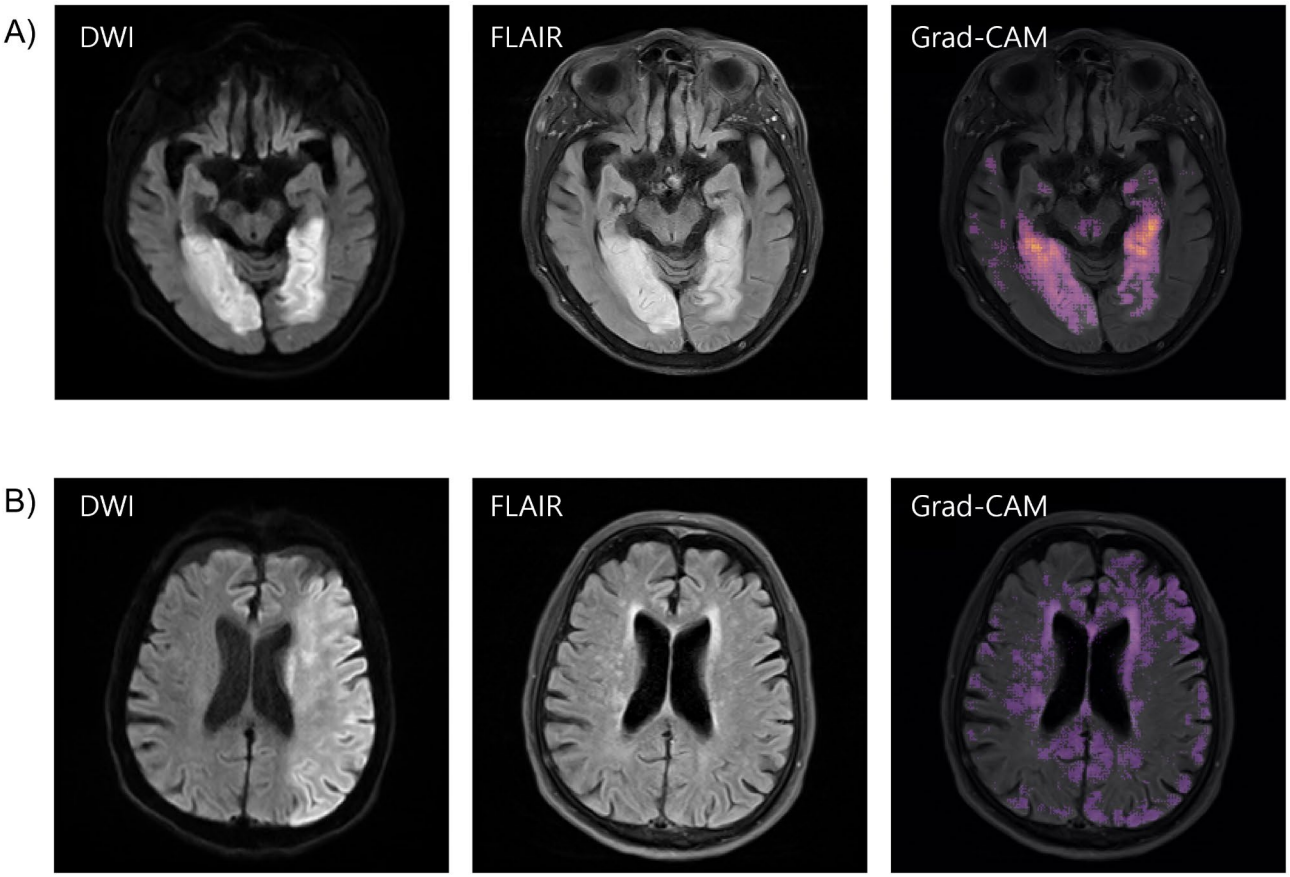

Fig. 3. Guided Grad-CAM visualization of lesion in Diffusion-FLAIR match (**A**) and mismatch (**B**) Cases. Panel A illustrates a Diffusion-FLAIR match case where the Guided Grad-CAM highlights the lesion, presenting a focused pattern of attention on the abnormality. Conversely, panel B displays a Diffusion-FLAIR mismatch case, where the Guided Grad-CAM extends its attention beyond the lesion, indicating a less targeted pattern.

mismatches. These findings have shown to be generalizable across both an internal multicenter test set and two external single-center registries, underscoring the potential of our approach to enhance the precision of stroke imaging diagnostics in clinical settings.

To rigorously evaluate the model's performance and clinical utility, we employed two distinct binary classification schemes: Label Set I and Label Set II. Label Set II, which focused exclusively on distinguishing clear matches from definitive mismatches, demonstrated superior performance across all cohorts. This enhanced performance was expected, as Label Set II excluded ambiguous cases with subtle or heterogeneous FLAIR changes. However, the clinical value of Label Set I should not be underestimated, as it was specifically designed to reflect real-world scenarios by incorporating these challenging cases. In routine clinical practice, encountering patients with ambiguous FLAIR changes is not uncommon. Therefore, the assistance of an automated classification model trained on Label Set I can be particularly crucial for clinicians when dealing with these challenging cases, enhancing the model's practical applicability in everyday clinical settings.

Previous studies have developed machine learning models using DWI and FLAIR imaging to predict the time since stroke (TSS)[18]. The ultimate goal of estimating time in these studies was to use the estimated time as an indication for reperfusion therapy if it falls within a specific time window. However, this approach raises several concerns. Firstly, the accuracy of predicting time based on DWI and FLAIR images is questionable[4,25]. For example, in patients with good collateral blood flow, FLAIR positive lesions may be less dependent on the elapsed time since stroke onset[6]. Secondly, the mere presence within a specific time window does not directly reflect the reversibility of brain tissue. It is known that some patients benefit from reperfusion therapy even beyond the traditional time windows, suggesting that the presence within a specific window is not an absolute indicator of salvageable tissue. Lastly, models that provide only the TSS without information about the intermediary processes limit their interpretability. Given the critical nature of healthcare, it is safer for AI to function as a supportive tool that aids clinicians in making important medical decisions. When checking for the presence or absence of the mismatch, clinicians can visually verify the AI's predictions by reviewing the images themselves. However, in the case of TSS predictions, clinicians do not have the option to visually confirm the AI-generated values. If it is difficult to explain or verify the AI's reasoning, this could limit its practical applicability in clinical settings.

In contrast, our study established the mismatch as the outcome instead of using TSS as an outcome. This approach could potentially enhance the usability of clinical decisions by physicians, as the various radiological changes over the progression of cerebral tissue ischemia are likely to reflect the tissue's state more precisely and directly than that of time[26].

The DWI-FLAIR mismatch has the potential to extend beyond the traditional scope of perfusion-diffusion mismatch (PDM) and clinical-diffusion mismatch (CDM), offering broader applicability in scenarios with extended time windows. (Figure S1)[9–13]While conventional methods often classify diffusion-restricted tissue as non-salvageable, the mismatch can refine the selection process by identifying regions with heterogeneous viability within tissue that has traditionally been considered ischemic core[27–30]. The possibility of recovery within such core tissue has become particularly significant in light of recent randomized trials, which have shown that even in large core strokes with diffusion volumes exceeding 50 mL, endovascular thrombectomy can provide superior outcomes compared to medical treatment[8–11]. The mismatch's ability to distinguish varying tissue states within large core infarctions suggests why reperfusion therapy may be effective in these cases, further emphasizing its value in identifying candidates who could benefit from such interventions, even in traditionally challenging scenarios.

To enhance the interpretability of our model's decisions and understand factors influencing its performance, we conducted comprehensive analyses of various clinical and imaging features. Our volume-dependent performance analysis, detailed in Supplementary Materials (Table S2), specifically compared model performance between lesions of 5-30mL and those exceeding 30mL. While lesions larger than 30mL showed slightly better performance, the differences were not statistically significant. Similarly, in our temporal analysis, we observed that Label Set I showed minimal accuracy improvements with shorter onset-to-imaging times, while Label Set II displayed more uniform performance across time intervals, though these temporal patterns also lacked statistical significance. Although not statistically significant, these subtle patterns might suggest that conditions that minimize labeling ambiguity - such as larger lesion volumes, shorter onset-to-imaging times, and more definitive label criteria (as in Label Set II) - could potentially contribute to more reliable model performance through clearer ground truth establishment.

Additionally, the saliency maps generated using Guided Grad-CAM provided valuable insights into regions the model considers important for classification. The distinct activation patterns observed between match and mismatch cases offer valuable visual guidance for clinicians in their decision-making, although it's important to note that while these visualizations highlight regions the model considers important, they do not explicitly explain the underlying significance of these regions.

Our analysis revealed interesting patterns in how the model learned from human annotations. While human annotators primarily focused on comparing FLAIR signals to DWI lesions during the labeling process, the varying complexity of DWI signal patterns - from extensive diffusion changes to complex cortical restricted diffusion patterns - presented challenges in maintaining consistent evaluation criteria. This variability suggests that human annotators may have struggled to consistently exclude unaffected areas and accurately assess the corresponding FLAIR signals, potentially introducing clinician-intended biases into the annotations. As a result, the AI model may have learned patterns that go beyond simple signal differences between DWI and FLAIR, incorporating these inherent biases from clinical expertise. Since the AI model was trained under human supervision based on these potentially biased annotations, it may have internalized more complex relationships than initially expected. While this AI model has demonstrated good performance, often exceeding human expectations by recognizing such complex patterns, identifying the exact mechanisms behind these capabilities remains challenging. To determine which features the model has captured and to understand its limitations, ongoing human intervention—through clinical application, validation, feedback, and updates—is required.

We observed slight variations in accuracy between the derivation dataset's test set and the two external test sets. This discrepancy could stem from the process of preparing brain imaging data uniformly across 36 slices, which sometimes required the interpolation of images when the original number was fewer, potentially distorting the original signal patterns. This interpolation process, necessary to meet the input requirements of the deep learning model, might have introduced artifacts, affecting the model's performance across different datasets. Further investigation is needed to fully understand the impact of this and other factors on the model's generalizability.

This study has several limitations. First, despite collecting and validating training samples across multiple cohorts, concerns regarding the generalizability of our findings persist. The heterogeneity of imaging protocols across institutions introduces variability that could impact the algorithm's ability to accurately classify the mismatch in various clinical settings. For instance, the potential differences in classification performance due to varied MR parameter settings and MRI manufacturers warrant further investigation.

Secondly, while we dedicated substantial effort to establishing precise ground truth labels through a meticulous manual labeling process by stroke experts, the inherent subjectivity in visual interpretation remains a challenge. The initial moderate inter-rater agreement (Fleiss' Kappa of 0.57) primarily stemmed from disagreements in assessing subtle FLAIR changes and evaluating regional heterogeneous FLAIR patterns, particularly in determining thresholds for subtle changes and their involvement in eloquent areas. Through multiple consensus meetings focused on establishing consistent criteria for identifying potential thrombectomy candidates, we achieved progressive improvement in agreement metrics (Fleiss' Kappa values of 0.57, 0.86, and 0.91). However, these labeling uncertainties and their potential impact on model performance warrant careful consideration in future implementations.

Thirdly, while the performance metrics of the developed classifier are promising, assessing the clinical benefits derived from model utilization necessitates separate consideration. We hypothesize that the ultimate clinical utility of this model lies in its potential to identify salvageable tissue using the presence or absence of the mismatch, thereby facilitating decision-making for recanalization therapy in patients likely to benefit

from improved outcomes. However, it remains unclear whether our defined labels are effective in distinguishing salvageable tissue within the context of recanalization therapy. Therefore, to validate the efficacy and applicability of this model, prospective studies are required to evaluate the diagnostic yield or comparative effectiveness on clinical outcomes in settings with and without the use of the model.

In conclusion, our study enhances the use of deep learning for assessing Diffusion-FLAIR mismatch in stroke patients, highlighting its potential to reduce subjectivity and support targeted decision-making in acute ischemic stroke treatment.

Data availability

This study involves data from human participants, and thus, there are ethical concerns related to privacy protection. However, the data can be made available upon reasonable request, with approval from both the corresponding author and the Ethics Committee.

Received: 15 September 2024; Accepted: 11 February 2025
Published online: 18 February 2025

References

1. Aoki, J. et al. FLAIR can estimate the onset time in acute ischemic stroke patients. *J. Neurol. Sci.* **293**, 39–44 (2010).
2. Ebinger, M. et al. MRI-based intravenous thrombolysis in stroke patients with unknown time of symptom onset. *Eur. J. Neurol.* **19**, 348–350 (2012).
3. Petkova, M. et al. MR imaging helps predict time from symptom onset in patients with acute stroke: Implications for patients with unknown onset time. *Radiology* **257**, 782–792 (2010).
4. Thomalla, G. et al. DWI-FLAIR mismatch for the identification of patients with acute ischaemic stroke within 4· 5 h of symptom onset (PRE-FLAIR): A multicentre observational study. *Lancet Neurol.* **10**, 978–986 (2011).
5. Thomalla, G. et al. MRI-guided thrombolysis for stroke with unknown time of onset. *N. Engl. J. Med.* **379**, 611–622 (2018).
6. Wouters, A. et al. Association between time from stroke onset and fluid-attenuated inversion recovery lesion intensity is modified by status of collateral circulation. *Stroke* **47**, 1018–1022 (2016).
7. Rocha, M. & Jovin, T. G. Fast versus slow progressors of infarct growth in large vessel occlusion stroke: Clinical and research implications. *Stroke* **48**, 2621–2627 (2017).
8. Yoshimura, S. et al. Endovascular therapy for acute stroke with a large ischemic region. *N. Engl. J. Med.* **386**, 1303–1313 (2022).
9. Nagaraja, N., Forder, J. R., Warach, S. & Merino, J. G. Reversible diffusion-weighted imaging lesions in acute ischemic stroke: A systematic review. *Neurology* **94**, 571–587 (2020).
10. Xu, X. et al. Comparative study of the relative signal intensity on DWI, FLAIR, and T2 images in identifying the onset time of stroke in an embolic canine model. *Neurol. Sci.* **35**, 1059–1065 (2014).
11. Burdette, J. H., Ricci, P. E., Petitti, N. & Elster, A. D. Cerebral infarction: Time course of signal intensity changes on diffusion-weighted MR images. *AJR Am. J. Roentgenol.* **171**, 791–795 (1998).
12. Ayata, C. & Ropper, A. H. Ischaemic brain oedema. *J. Clin. Neurosci.* **9**, 113–124 (2002).
13. Simard, J. M., Kent, T. A., Chen, M., Tarasov, K. V. & Gerzanich, V. Brain oedema in focal ischaemia: Molecular pathophysiology and theoretical implications. *Lancet Neurol.* **6**, 258–268 (2007).
14. Sarraj, A. et al. Trial of endovascular thrombectomy for large ischemic strokes. *N. Engl. J. Med.* **388**, 1259–1271 (2023).
15. Huo, X. et al. Trial of endovascular therapy for acute ischemic stroke with large infarct. *N. Engl. J. Med.* **388**, 1272–1283 (2023).
16. Bendszus, M. et al. Endovascular thrombectomy for acute ischaemic stroke with established large infarct: Multicentre, open-label, randomised trial. *Lancet* **402**, 1753–1763 (2023).
17. Scheldeman, L. et al. Diffusion-weighted imaging and fluid-attenuated inversion recovery quantification to predict diffusion-weighted imaging-fluid-attenuated inversion recovery mismatch status in ischemic stroke with unknown onset. *Stroke* **53**, 1665–1673 (2022).
18. Offersen, C. M. et al. Artificial intelligence for automated DWI/FLAIR mismatch assessment on magnetic resonance imaging in stroke: A systematic review. *Diagnostics* **13**, 2111 (2023).
19. Jo, H. et al. Combining clinical and imaging data for predicting functional outcomes after acute ischemic stroke: An automated machine learning approach. *Sci. Rep.* **13**, 16926 (2023).
20. Cohen, J. F. et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: Explanation and elaboration. *BMJ open.* **6**, e012799 (2016).
21. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* 837–845 (1988).
22. Rufibach, K. Use of Brier score to assess binary predictions. *J. Clin. Epidemiol.* **63**, 938–939 (2010).
23. Park, S. H. & Han, K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology* **286**, 800–809 (2018).
24. Selvaraju, R. R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* 618–626 (2017).
25. Emeriau, S. et al. Can diffusion-weighted imaging–fluid-attenuated inversion recovery mismatch (positive diffusion-weighted imaging/negative fluid-attenuated inversion recovery) at 3 Tesla identify patients with stroke at < 4.5 Hours? *Stroke* **44**, 1647–1651 (2013).
26. Bivard, A., Spratt, N., Miteff, F., Levi, C. & Parsons, M. W. Tissue is more important than time in stroke patients being assessed for thrombolysis. *Front. Neurol.* **9**, 41 (2018).
27. Nogueira, R. G. et al. Thrombectomy 6 to 24 hours after stroke with a mismatch between deficit and infarct. *N. Engl. J. Med.* **378**, 11–21 (2018).
28. Albers, G. W. et al. Thrombectomy for stroke at 6 to 16 hours with selection by perfusion imaging. *N. Engl. J. Med.* **378**, 708–718 (2018).
29. Marcoux, F., Morawetz, R., Crowell, R., DeGirolami, U. & Halsey, J. Jr. Differential regional vulnerability in transient focal cerebral ischemia. *Stroke* **13**, 339–346 (1982).
30. Labeyrie, M-A. et al. Diffusion lesion reversal after thrombolysis: A MR correlate of early neurological improvement. *Stroke* **43**, 2986–2991 (2012).

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2022R1G1A1005686). This research was supported by the K-Brain Project of the National Research Foundation (NRF) funded by the Korean government (MSIT) (No.

RS-2023-00265393). This study was supported by Inje University Haeundae Paik Hospital.

Author contributions

P.K., D.K., and J.Y. contributed to the development of the prediction model. J.L., H.K., J.S., S.L., and S.P. were responsible for data labeling. D.K. and H.P. provided and curated external data. J.S., S.L., D.K., H.P., and J.Y. contributed to data review. S.P. contributed to the development of the research idea, study design, interpretation of results, drafting of the manuscript, and final decision-making.

Funding

No funding was received for conducting this study.

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/s41598-025-90214-w.

Correspondence and requests for materials should be addressed to J.Y. or S.P.