Title

Functional Outcome Prediction in Acute Ischemic Stroke Using a Fused Imaging and Clinical Deep Learning Model.

Journal

Stroke, 54(9)

Authors

Liu, Yongkai
Yu, Yannan
Ouyang, Jiahong
et al.

Publication Date

2023-09-01

DOI

10.1161/STROKEAHA.123.044072

Peer reviewed

HHS Public Access

Author manuscript

Stroke. Author manuscript; available in PMC 2024 September 01.

Published in final edited form as:

Stroke. 2023 September ; 54(9): 2316–2327. doi:10.1161/STROKEAHA.123.044072.

Functional Outcome Prediction in Acute Ischemic Stroke using a Fused Imaging and Clinical Deep Learning Model

Yongkai Liu, PhD[1], **Yannan Yu, MD**[1], **Jiahong Ouyang, MS**[1,2], **Bin Jiang, MD**[1], **Guang Yang, PhD**[3], **Sophie Ostmeier, MD**[1], **Max Wintermark, MD**[4], **Patrik Michel, MD**[5], **David S. Liebeskind, MD**[6], **Maarten Lansberg, MD, PhD**[7], **Gregory Albers, MD**[7], **Greg Zaharchuk, MD, PhD**[1]

[1]Department of Radiology, Stanford University, Stanford, CA, USA

[2]Department of Electrical Engineering, Stanford University, Stanford, CA, USA

[3]National Heart and Lung Institute, Imperial College London, London, UK

[4]Department of Neuroradiology, University of Texas MD Anderson Center, Houston, TX, USA

[5]Neurology Service, Department of Clinical Neurosciences, Lausanne University Hospital and University of Lausanne, Switzerland

[6]Department of Neurology, UCLA, Los Angeles, CA, USA

[7]Department of Neurology, Stanford, Stanford, CA, USA

Abstract

Background: Predicting long-term clinical outcome based on early acute ischemic stroke (AIS) information is valuable for prognostication, resource management, clinical trials, and patient expectations. Current methods require subjective decisions about which imaging features to assess and may require time-consuming post-processing. This study's goal was to predict ordinal 90-day modified Rankin Scale (mRS) in AIS patients by fusing a deep learning model of DWI images and clinical information from the acute period.

Methods: 640 AIS patients who underwent MRI within 1–7 days post-stroke and had 90-day mRS follow-up data were randomly divided into 70% (n=448) for model training, 15% (n=96) for validation, and 15% (n=96) for internal testing. Additionally, external testing on a cohort from Lausanne University Hospital (LUH) (n=280) was performed to further evaluate model generalization. Accuracy for ordinal mRS, accuracy within ±1 mRS category, mean absolute prediction error, and determination of unfavorable outcome (mRS>2) were evaluated for clinical only, imaging only, and 2 fused clinical-imaging models.

Results: The fused models demonstrated superior performance in predicting ordinal mRS score and unfavorable outcome in both internal and external test cohorts when compared to the clinical

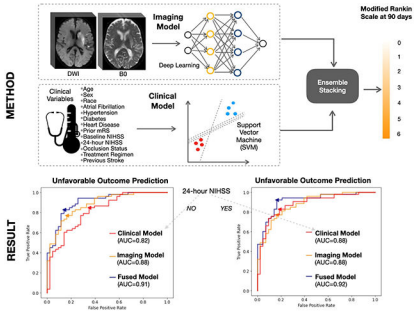
[*]**Correspondence to:** Yongkai Liu, Ph.D., Department of Radiology, 1201 Welch Rd., Stanford, California 94305-5488, yongkliu@stanford.edu.

Supplemental Material
Supplemental Figures S1–S3
Supplemental Tables S1–S6

and imaging models. For the internal test cohort, the top fused model had the highest AUC of 0.92 for unfavorable outcome prediction and the lowest mean absolute error (MAE: 0.96; 95% CI: 0.77–1.16), with the highest proportion of mRS score predictions within ±1 category (79%; 95% CI: 71–88%). On the external LUH cohort, the best fused model had an AUC of 0.90 for unfavorable outcome prediction and outperformed other models with an MAE of 0.90 (95% CI: 0.79–1.01), and the highest percentage of mRS score predictions within ±1 category (83%; 95% CI: 78–87%).

Conclusions: A deep learning-based imaging model fused with clinical variables can be used to predict 90-day stroke outcome with reduced subjectivity and user burden.

Graphical Abstract



INTRODUCTION

With around 795,000 strokes occurring yearly, stroke is a leading cause of disability and death worldwide[1]. Stroke survivors commonly suffer from disability and function loss that impacts their quality of life significantly[2]. Predicting the long-term degree of clinical impairment based on information available early in the course of acute ischemic stroke (AIS) would be valuable for optimizing rehabilitation strategies, prognostication, clinical trials, resource management, and patient expectations[1,3]. However, long-term outcome prediction is challenging because various factors may directly or indirectly impact how disabled a patient will be in the long run[1,4,5]. For example, many studies have shown that the size of the initial infarct is a relatively weak predictor of outcome[6–9].

Some studies[1,4,10] have attempted to predict the long-term degree of outcome by estimating the modified Rankin Scale obtained at three months (90-day mRS) following hospital discharge. For example, Zhang et al.[10] showed that a model relying only on age and NIHSS score at the discharge time could be used to predict the 90-day mRS score. Xie et al.[1] and Heo et al.[4] predicted the 90-day mRS score using machine learning models based on clinical and imaging variables. Brugnara et al.[5] introduced a multimodal machine learning model of clinical, multimodal imaging, and angiographic characteristics to predict clinical outcome after endovascular treatment. However, these approaches may be subject to inconsistent and sub-optimal performance for several reasons: firstly, they all rely on human-crafted imaging features which may not be optimal for prediction and secondly, they may require clinical measurements that are either not routinely acquired in non-specialist centers or may be subjective. In particular, the choice and extraction of imaging features add additional

layers of subjectivity and typically require time-consuming, often manual post-processing resources.

Deep learning (DL), using convolutional neural networks, has recently demonstrated remarkable capabilities in comprehending radiologic images and enhancing medical imaging diagnosis and prognostication[11,12]. DL can adaptively learn representative information from raw medical imaging without any preconceptions related to the human-involved feature extraction process. To the best of our knowledge, no studies have leveraged deep learning to identify optimal features from medical imaging for predicting long-term disability outcomes, particularly the specific score on the 90-day mRS, a challenging but valuable task.

This study aimed to develop and evaluate a DL-based fused predictive model incorporating a DL model of diffusion-weighted imaging (DWI) and routinely obtained clinical variables from the acute stroke period for predicting the exact 90-day mRS as well as favorable outcome (mRS<=2). The fused models were evaluated using two separate test cohorts and compared with the models only using imaging or clinical information.

METHODS

Patients and MRI datasets

Data underpinning the findings of this study can be obtained from the corresponding author upon a reasonable request. This study was carried out according to the United States Health Insurance Portability and Accountability Act (HIPAA) of 1996 with institutional review board (IRB) approval. All subjects provided written informed consent or the need for consent was waived by the local IRB. This study included AIS patients from four prospective multicenter trials and two single-center registries (multicenter trials: Imaging Collaterals in Acute Stroke (iCAS) (April 2014 to June 2019; n = 188)[13], Diffusion Weighted Imaging Evaluation for Understanding Stroke Evolution Study–2 (DEFUSE-2) (July 2008 to October 2012; n = 140)[14], Endovascular Therapy Following Imaging Evaluation for Ischemic Stroke 3 (DEFUSE 3) (May 2016 to May 2017; n=182)[15], Computed Tomography Perfusion to Predict Response to Recanalization in Ischemic Stroke Project (CRISP) (August 2008 to June 2012; n=201)[16]; single-center registries: University of California, Los Angeles (UCLA) stroke registry (2012 to 2016; n = 196), and Lausanne University Hospital (LUH) stroke registry[1] (January 2008 to December 2017; n = 1723). The enrollment details of the clinical trials can be found in the publications cited above; the enrollment criteria for the registries can be found at the studies[1,17,18].

The long-term clinical outcome measure was mRS, which ranged between 0 (no disability) and 6 (death)[19,20] and was assessed at a median of 90 days (range: 60–120 days) following discharge. We excluded patients without a 90-day mRS, and those without day 1–7 DWI or B0 images. Clinical variables included in the model are listed in Table 1. If the clinical variable was missing, the mean value of this variable in the training dataset was used. Rates of missing data situation are shown in Table 1 and were generally low. We built the models using a random subset of patients from iCAS, DEFUSE2, DEFUSE3, CRISP, and the UCLA registry for training and validation. The remaining patients from these datasets were used as an internal test cohort. In addition, we included LUH patients as an external generalization

cohort to further test the model's performance. The LUH patient cohort differs from the internal cohort in several respects, including fewer patients with hypertension and diabetes, and lower stroke severity as measured by NIHSS. A flow chart of the subjects included in the study can be found in Figure 1.

Data Preprocessing

We co-registered and normalized all MRI images to the Montreal Neurological Institute template using SPM12 software (Statistical Parametric Mapping, The Wellcome Trust Centre for Neuroimaging). DWI and B0 images were normalized by their means before being fed into the deep learning model. Categorical variables (such as history of diabetes or hypertension) were transformed into dummy variables before being fed into the clinical model.

Model Structure, Training, and Testing

We are testing 5 different machine learning models in this study. This includes two clinical models, an imaging-only model, and two DL fused models that combine the imaging model with one or the other clinical model. These are described in more detail below.

Clinical models ––We employed a Support Vector Regression (SVR) model as the basis for our clinical model. Commonly collected clinical variables, including age, sex, prior mRS, baseline NIHSS, large vessel occlusion status (yes/no), and history of hypertension, diabetes, atrial fibrillation, heart diseases, and previous stroke (yes/no), as well as the treatment regimen (no treatment, intravenous tissue plasminogen activator (IV tPA) only, endovascular therapy (EVT) only, or both IV tPA and EVT), were input into the clinical model to generate a continuous prediction of a 90-day mRS (Clinical Model I). Clinical Model II expands upon Clinical Model I by also including the 24-hour NIHSS score. These were evaluated separately since many sites do not routinely collect 24-hour NIHSS scores. SVR was selected for its proven robustness, effective handling of high-dimensional data, and strong generalization performance, which has yielded accurate predictions across diverse applications[21].

Imaging model ––A customized 3D CNN with a 3D ResNet[22] as the backbone was adopted as the imaging model. Details of the model can be found in Supplemental Figure S1. We reduced the number of neurons in the last fully connected layer to one, which enables the CNN to perform a regression task. The imaging model takes the DWI and B0 images (obtained 1 to 7 days after the stroke onset) as the input and outputs a continuous prediction of 90-day mRS.

Fused models -––The deep learning fused models consist of a DL-based imaging model combined with a clinical model, which are subsequently integrated through an ensemble stacking technique[23] to produce the mRS prediction, which is again a continuous variable. Predictions larger than 6 were adjusted to 6, and less than 0 to 0. Ensemble stacking is a well-defined technique that involves training multiple base models independently and then aggregating their individual predictions through a higher-level, or meta-model[24]. This approach capitalizes on the unique strengths and advantages of each base model, resulting

in enhanced prediction accuracy and minimal overfitting. Two fused models were developed: Fused Model I, which combines Clinical Model I with the Imaging Model, and Fused Model II, which integrates Clinical Model II and the Imaging Model. More details can be found in Figure 2.

We employed stochastic gradient descent (SGD) as the optimizer and smooth L1 loss[25] as the loss function during deep learning model training. To augment the imaging data sample size during model training, we randomly introduced blurring, noise, and changes in image contrast. The primary dataset was randomly divided into a development cohort (n=544) for model development and an internal testing cohort (IT Cohort, n=96) for model evaluation. During each training run, the development cohort was further split into training (n=448) and validation (n=96) subsets. We set the maximum number of training epochs to 100 and selected the optimal model for each run based on the lowest validation loss after 30 training epochs. The IT Cohort remained separate from the model development process. Each model was trained and developed using the development cohort 50 times with random initialization weights. We averaged predictions across these multiple models for the final predictions on the test sets, where the variance in predictions also yields an estimate of the uncertainty of the model. For patients with multiple MRI scans between days 1–7 post-stroke, we used all acquired MRI images for model training, while only the latest MRI scans within that time period were used for model validation and testing. To assess this latter decision, we also examined model results when using only the latest MRI scan as part of the training process, which can be found in Supplemental Materials.

Performance Evaluation

Mean absolute error (MAE), accuracy for a specific mRS (ACC), and mRS accuracy within ±1 score (±1ACC) were used to measure the model's ordinal outcome prediction performance. MAE evaluates the average absolute discrepancy between the predicted score and the ground truth of the 90-day mRS, with a smaller MAE signifying superior model performance. The predicted score, as a continuous variable, was rounded to the nearest integer (0–6) to enable the calculation of ACC and ±1ACC for the ordinal prediction of each patient's mRS score. ACC measures the percentage of correct predictions over all the predictions. ±1ACC measures the accuracy of prediction within ±1 mRS category. We have additionally calculated a tertile outcome metric, assessing accuracy as binned into the following mRS categories: 0–2, 3–4, 5–6[26]. These results can be found in Supplemental Materials. Finally, we evaluated the performance of the model in subgroups of patients based on type of intervention (none, IV tPA only, EVT only, and combined IV tPA and EVT), which is also included in Supplemental Materials.

The area under the receiver operating characteristic curve (AUC), sensitivity, and specificity were used to measure the model's ability to distinguish favorable from unfavorable outcome (90-day mRS 0–2 vs. 3–6, respectively). AUC was calculated by thresholding at a predicted mRS score of 2.5. Predicted scores ranging from 0–2.5 and 2.5–6 were linearly rescaled to 0–0.5 and 0.5–1, respectively. This rescaling facilitated the calculation of AUC, effectively representing the model's capacity to differentiate between favorable and unfavorable outcomes. For obtaining sensitivity and specificity, the Youden index operating point was

used (solid triangles on the ROC curves). Similar analyses were performed for determining excellent outcome (mRS 0–1) and a large core trial metric (mRS 0–3) from other classes, and these can be found in the Supplemental Materials.

For interpreting the imaging model, saliency maps were created using channel activation maps for the regression task[27,28], visually highlighting the significant regions within the input images that contribute to the model's predictions. To assess how important the saliency maps were for prediction, we evaluated the quality of our model's lesion detection by dividing the results into three categories: 1) no detection of the lesion; 2) moderate detection, the model partially captured the stroke lesion, often touching its boundary; and 3) excellent detection, accurately including the entirety of the ground-truth lesion. Lesion detection was performed by an independent reader who was blinded to both the 90-day outcome prediction and the reference standard. The results can be found in Supplemental Materials.

Statistical Analysis

The Diebold-Mariano test was performed to assess for significant differences between two MAEs. The significance of differences between AUC was obtained using the two-sided Delong test. The McNemar test was used to compare sensitivity and specificity. The permutation test was used to compare ACC and ±1ACC. P values less than 0.05 were considered statistically significant. To address multiple testing concerns, we used the Benjamini-Hochberg procedure, keeping the false discovery rate at a 0.05 level[29]. 95% confidence interval (CI) for sensitivity and specificity was calculated using Wilson's method.

RESULTS

Patient Characteristics

A total of 1,028 patients were initially considered for the training, validation, and internal test sets, sourced from four major clinical trials (DEFUSE2, DEFUSE3, iCAS, CRISP) and the UCLA registry. After applying the inclusion and exclusion criteria, 640 patients were selected and randomly divided into 70% (n=448) for model training, 15% (n=96) for validation, and 15% (n=96) for internal testing. The LUH generalization cohort included 280 patients, who were selected based on the same criteria.

Performance Analysis on the Internal Test Cohort

Figure 3 presents three representative examples of outcome prediction by the fused, imaging, and clinical models. In Case A, both the fused (type I) and imaging models accurately predict the patient's 90-day mRS as 4, whereas the clinical model I incorrectly predicts it as 3, potentially influenced by the patient's young age. In Case B, the fused and clinical models (type I) predict the 90-day mRS as 2, while the imaging model estimates it as 3. The saliency map indicates that the model may not have learned to focus on the infarct's location. In Case C, all three models (type I for fused and clinical models) inaccurately predict an mRS of 3 for a patient who was deceased (mRS 6) at 90 days.

Table 2 and Figure 4(A) show the quantitative comparisons between clinical models, imaging model, and fused models on the internal test cohort. Table 2 (upper part) describes the performance of clinical, imaging, and fused models in predicting specific mRS outcomes in the internal test set. The Fused Model II had the lowest MAE (0.96; 95% CI: 0.77–1.16) among all the models, significantly better than the clinical model I (MAE: 1.23; 95% CI: 1.03–1.45; p<0.001) and Clinical Model II (MAE: 1.03; 95% CI: 0.83–1.23; p=0.002), but not significantly different from the other models. Fused Model II also demonstrated a higher proportion of mRS score predictions within ±1 category (79%; 95% CI: 70–85%) compared to Clinical Model I (65%; 95% CI: 48–69%; p=0.007), Clinical Model II (74%; 95% CI: 64–80%; p=0.13), Imaging Model (75%; 95% CI: 67–83%; p=0.48), and Fused Model I (74%; 95% CI: 66–82%; p=0.33). Fused Model II also exhibited the highest exact mRS accuracy of 35% (95% CI: 26–46%), though this did not also significantly differ from the other models.

To predict unfavorable outcome (Figure 4), the clinical model I had an AUC of 0.82 (95% CI: 0.73–0.90). Clinical Model II, which integrated the 24-hour NIHSS score into the model, demonstrated a superior AUC of 0.88 (95% CI: 0.81–0.94), although the difference did not reach statistical significance (p=0.14). The imaging model had an AUC of 0.88 (95% CI: 0.82–0.94; p=0.16 and 98 when compared to both clinical models). Fusing the imaging model with clinical model I (Fused Model I) resulted in an AUC of 0.91 (95% CI: 0.84–0.96; p=0.01 and 0.15 in comparison to clinical model I and the imaging model, respectively). Fused Model II demonstrated the best performance, achieving an AUC of 0.92 (95% CI: 0.86–0.97, p=0.02 compared to the clinical model I; p=0.05 compared to the clinical model II; p=0.19 compared to the imaging model; p=0.68 compared to Fused Model I).

Moreover, at the Youden index point, Fused Model II attained the highest sensitivity (0.91; 95% CI: 0.80–0.96) among all models, while maintaining a relatively high specificity (0.84; 95% CI: 0.70–0.92). Fused Model I demonstrated the highest specificity (both 0.86; 95% CI: 0.73–0.93). More details can be found in Supplemental Table S1.

Supplemental Figure S2(A) shows the median and range of the fused model predictions compared with the ground truth for each mRS level, showing better performance over the mRS 0–4 interval compared with the mRS 5–6 interval. Supplemental Figures S3 (A&C) display the ROC curve performance for excellent outcomes (mRS 0–1) and (mRS 0–3), respectively, both demonstrating similar trends.

Performance Analysis on the External LUH Cohort

Table 2(lower part) and Figure 4(B) show the quantitative comparisons between clinical models, imaging model, and fused models on the external LUH test cohort. To predict mRS, Fused Model II achieved the lowest MAE (MAE: 0.90; 95% CI: 0.79–1.01). This performance was comparable to that of Clinical Model II (MAE: 0.91; 95% CI: 0.80–1.03). Moreover, it significantly outperformed the Imaging Model (MAE: 1.03; 95% CI: 0.91–1.15; p=0.005), Clinical Model I (MAE: 1.15; 95% CI: 1.03–1.27; p<0.001) and Fused Model I (MAE: 0.99; 95% CI: 0.87–1.11; p=0.02). In terms of ±1 Accuracy (±1ACC), Fused Model II and Clinical Model II were the top performers, with similar outcomes

(Fused Model II: 83%, 95% CI: 78–87%; Clinical Model II: 84%, 95% CI: 79–88%; p=0.66). Both of these models outperformed Clinical Model I (71%; 95% CI: 65–76; p<0.001) and also exceeded the performance of Imaging Model (77%; 95% CI: 72–82; p=0.17 and 0.10) and Fused Model I (81%; 95% CI: 76–85; p=0.19 and 0.13). The best exact mRS accuracy was achieved by Fused Model II, with 36% (95% CI: 30–41%), followed closely by the Imaging Model and Fused Model I, both at 35% (95% CI: 29–40%). Clinical Model II recorded an accuracy of 34% (95% CI: 29–40%), while Clinical Model I lagged behind at 28% (95% CI: 22–34%). However, there was no significant difference in accuracy between Fused Model II and the other models.

For predicting unfavorable outcomes, Clinical Model I achieved an AUC of 0.77 (95% CI: 0.71–0.82). Clinical Model II performed significantly better with an AUC of 0.88 (95% CI: 0.84–0.92; p<0.001). The Imaging Model attained an AUC of 0.85 (95% CI: 0.80–0.90), significantly surpassing Clinical Model I (p<0.02), yet trailing behind Clinical Model II (p=0.18). Fused Model I exhibited an AUC of 0.87 (95% CI: 0.83–0.90), outperforming Clinical Model I significantly (p<0.001), but not significantly different from the Imaging Model (p=0.49). The best performance was achieved by Fused Model II, recording an AUC of 0.90 (95% CI: 0.86–0.93; p<0.001 compared to Clinical Model I; p=0.03 compared to Clinical Model II; p=0.004 compared to the Imaging Model; p=0.02 compared to Fused Model I). Furthermore, at the Youden index, Fused Model II achieved the highest sensitivity among all models at 0.94 (95% CI: 0.87–0.97). It significantly outperformed Clinical Model I (0.76; 95% CI: 0.66–0.83; p<0.001), Clinical Model II (0.84; 95% CI: 0.76–0.90; p=0.007), and the Imaging Model (0.79; 95% CI: 0.69–0.86; p=0.002), but did not differ significantly from Fused Model I (0.94; 95% CI: 0.87–0.97; p=0.74). The Imaging Model (0.76; 95% CI: 0.73–0.93) and Clinical Model II (0.76; 95% CI: 0.70–0.82) had the highest specificity, differing significantly from both fused models. Further details are provided in Supplementary Table S1.

Supplemental Figure S2(B) shows the median and range of the fused model predictions compared with the ground truth for each mRS level, with similar performance as in the internal test set. Supplemental Figures S3 (B&D) display the ROC curve performance for excellent outcomes (mRS 0–1) and (mRS 0–3), respectively, both showing similar trends.

Interestingly, there is less differentiation of the different models for excellent outcome prediction in the LUH dataset, which may be related to the overall reduced severity of the index strokes.

Additional Analyses

Supplemental Table S2 shows the performance of the different models in subgroups of patients who underwent different treatments. Supplemental Table S3 shows the performance of the different models on tertile outcome prediction, showing similar findings to ±1ACC accuracy presented in Table S2. Supplemental Table S4 demonstrates the effect of training on either all available day 1–7 MRI or only the last MRI in this period on the imaging model, demonstrating a small but non-significant improvement in performance using all available day 1–7 MRI studies for training. Supplemental Table S5 demonstrates that the performance of the imaging model improves when the saliency maps demonstrate that

the model identifies the stroke lesion, further demonstrating that the model is identifying important features for outcome prediction.

DISCUSSION

We developed and evaluated a DL-based outcome predictive model, which fused routinely obtained MR images and clinical variables available during the early acute phase of stroke. Overall, the fused models performed the best at 90-day clinical outcome prediction following stroke, especially when combined with 24-hour NIHSS score. This finding was generalizable, seen in the internal multicenter test set as well as in a separate single-site registry from a different country with different severity characteristics. The performance was similar or better than other studies using human-crafted features and included much larger cohorts, both for training and testing. It is worth noting that this study's prediction tasks are twofold: predicting unfavorable outcomes and the exact score on the mRS. While there have been several attempts to predict unfavorable outcomes, studies that aim to predict the exact score on the mRS are still scarce. This study represents one of the first attempts to predict the exact score of the 90-day mRS by utilizing deep learning, a powerful machine learning methodology whose performance scales well with increasing amounts of data.

The fused model, which only used easily accessible imaging and clinical variables, can be seamlessly embedded into the current clinical workflow to achieve 90-day mRS prediction. It entails minimal preprocessing steps, primarily requiring the normalization of DWI and B0 images to a standard template; no subjective human or automated measurements are required. Regarding the clinical variables, we consciously used fewer clinical variables as compared to the previous studies[1,5], limiting ourselves to those that are most routinely available. This should make the model more applicable to sites without support teams to collect such data. Also, the assignment of clinical variables can introduce human variability that could make the model performance suffer during generalization to diverse settings.

When utilizing imaging information for outcome prediction, a previous study (23) demonstrated that infarct volume can serve as an independent predictor of 90-day outcomes, but that it at best explained 41% of the variability in outcomes. Combining location and volume resulted in a significantly better correlation with clinical deficit severity than using volume alone[7]. The use of the ASPECTS score also exhibited the potential for predicting 90-day outcomes[30]. Unlike earlier studies that employed imaging for outcome prediction, our approach uses a deep learning model to extract the most relevant features from the imaging, bypassing the need for traditional human-derived metrics such as volume, location, and ASPECTS. These assessments often suffer from sub-optimal interobserver agreement and only exploit a portion of the available imaging information. In contrast, a deep learning-based, data-driven method can implicitly capture and integrate the essential aspects related to volume, location, or ASPECTS, improving prediction accuracy and consistency. While the model performed better when the saliency maps showed overlap with the infarct, we emphasize that the deep learning model was designed to identify important features for outcome prediction, not stroke segmentation. This allows the model to potentially also include important non-stroke information, such as atrophy, chronic infarcts, and T2-weighted hyperintensities. It is not possible to extract these separate features using deep

learning methodology, but we hypothesize that all contribute potentially to stroke outcomes. Additionally, compared with studies[1,5] using perfusion or angiography imaging, the fused model only took DWI and B0 as inputs, making it more applicable to sites with limited resources.

We did not initially include the 24-hour NIHSS score in the clinical model I, despite evidence from several studies[1,5] that it is a strong predictor of long-term stroke outcomes - a conclusion further validated by Supplementary Table S6. This is because obtaining 24-hour NIHSS may not be a routine clinical practice in some hospitals or healthcare systems[31], and its use may depend on the physician's judgment[31]. Instead, to accommodate different practices, we developed two fused models: one that includes the 24-hour NIHSS score and another that does not. This allows centers to choose the model that best aligns with their routine for obtaining 24-hour NIHSS scores in AIS patients.

Fused Model I, which did not utilize the 24-hour NIHSS, performed similarly to the other models that did include 24-hour NIHSS, including the imaging only model. This underlines two points: first, the imaging model does not require a human to assess NIHSS, thereby reducing variability due to differences in evaluator expertise or training levels. Second, if only initial NIHSS is routinely recorded, the combination of the clinical and imaging information is quite similar to models that required 24-hour NIHSS assessment.

We observed a disparity in the predictive performance of Clinical Model 1 between the LUH and IT cohorts. This inconsistency could potentially be attributed to the comparatively lower severity of patients upon admission in the LUH cohort, which may have resulted in more accurate predictions by the model for this group. The effectiveness of the model could be influenced by the severity of the patient's condition, suggesting that its development or calibration might have primarily focused on less severe cases.

The results of this study indicate that the fused clinical and imaging models outperformed models that included either one or the other approach. This suggests that each model contributes uniquely to clinical outcome prediction. While some clinical information, such as brain age, may be inferred from imaging[32], it can be difficult for the model to accurately learn such information without a large dataset. Imaging, especially DWI, offers a more detailed profile of cerebral tissue viability and perfusion, which cannot be fully captured by clinical variables. It is well known that outcome is only predicted weakly by infarct size[6–9]; the DL methodology allows a way to incorporate location in a purely data-driven way.

Compared to previous stroke outcome prediction methods, our study presents several notable differences and advantages. First, our study achieved the highest AUC (above 0.90 in both internal and external generalization cohorts) for predicting unfavorable stroke outcomes. This performance surpasses all previous studies, though we acknowledge potential differences in test cohort distributions that may affect the fairness of this comparison. Second, by utilizing two separate test cohorts, our study allows for a more robust and generalized assessment of our model across a broader range of patient populations. Third, our study is the first to employ deep learning to extract information from imaging data for outcome prediction, moving beyond the reliance on potentially subjective and

difficult-to-obtain radiological variables. Finally, while previous research mainly focused on predicting unfavorable outcomes by dichotomizing mRS scores, our study not only predicts unfavorable outcomes but also offers predictions of specific mRS scores, providing a more comprehensive approach to stroke outcome prediction. The top-performing fused models achieved an ordinal mRS accuracy rate of 35%, nearly 2.5 times greater than random predictions. This represents one of the initial efforts to predict exact mRS categories using any method, and it is the first to use deep learning. The complexity involved in forecasting a 7-point ordinal scale surpasses that of binary or simpler multi-class prediction tasks, which form the majority of the prior literature on outcome prediction. Thus, although the accuracy might appear modest, it is a promising starting point given the task's challenging nature. However, it is important to clarify that our intent is not to propose immediate clinical application of this model's capability in predicting exact mRS categories.

Supplementary Figure S2 shows the distribution of differences between true and predicted mRS scores for each true mRS score category. We observed the model's performance depends on the true mRS score. For patients with true mRS scores ranging from 0 to 4, the differences are generally closer to 0, indicating accurate predictions and relatively balanced under- and overestimations for these patients. Conversely, for patients with true mRS scores of 5 and 6, the differences are predominantly positive, suggesting that the model tends to underestimate the mRS scores for patients within this higher score range. One potential reason could be the fused model's reliance on acute-phase information for 90-day clinical outcome prediction, while other critical factors after the acute phase, such as rehabilitation techniques, new comorbidities, and emerging diseases, may also impact the long-term disability of stroke patients. This is particularly evident for some of the mRS 6 cases; some had very small strokes and may have passed away from separate causes rather than directly due to their ischemic event (such as the patient shown in Figure 3c). Similar to the findings in a prior study[1], the models do not perform well in these cases. Based on autopsy-verified findings, approximately 60% of AIS patients died from causes other than brain lesions[33]. Therefore, only stroke-related imaging and clinical variables might be suboptimal in predicting death. Nevertheless, the outcome prediction achieved by the model could serve as the baseline prediction based on current acute phase information to offer physicians and patients long-term expectations of stroke-related disabilities and optimize resources for patient rehabilitation planning.

In a separate analysis focusing on excellent clinical outcome (mRS 0–1 versus 2–6), we observed less significant enhancements when using imaging. A potential explanation for the subtler improvements seen with imaging in the Imaging or Fused models within the LUH dataset for either mRS 0–1 or 0–2 may be attributed to the relatively mild severity of the strokes. The disproportionate distribution of mRS scores, heavily skewed towards the 0–2 range, poses a predictive challenge for all models. In contrast, the internal test set exhibited a more broadly distributed range of mRS scores.

Our study has a few limitations; First, we use mRS as the primary outcome measure. Although mRS is currently the most widespread metric to define disability severity, some weaknesses exist, such as subjective determination between categories and reproducibility by different examiners[34]. Since it focuses on activities of daily living, it also combines

many types of disability into the same groups, limiting fine-grained outcome prediction. Second, although we collected training samples from multiple clinical trials, the effect of the training sample size on the prediction performance is still uncertain. Third, the imaging used in the study were obtained after 24 hours since the initial baseline imaging. This was because any acute interventions had concluded by that time. In the future, it would be interesting to investigate using initial, pre-treatment imaging to predict outcome, thereby potentially enabling a role in treatment decision-making. Also, other imaging sequences besides DWI and B0 could be incorporated into the models at the expense of added site requirements and image post-processing. Fourth, our study did not account for variability in treatment strategies employed for the AIS patients, as this was not included in the model. This decision was made because the imaging used was acquired after acute treatment was complete. Additional information about treatment and/or imaging data relevant to treatment (such as perfusion imaging that might signal the presence of continued at-risk tissue) could be included in future models to improve performance, but again, this comes at the expense of greater complexity and demands on the individual sites. Fifth, this study utilizes clinical trials spanning a considerable period, during which the standard of care for acute interventions substantially evolved, especially following the validation of mechanical thrombectomy in 2015 and the treatment window extension in 2018. However, this paper focuses on MRI studies acquired after any acute treatment, and therefore might be less sensitive to different acute treatment modalities.

In conclusion, we developed and evaluated deep learning-based fused models combining brain MR and clinical information for long-term outcome prediction in AIS patients. Overall, the deep learning-based fusion of clinical variables and imaging enables outcome prediction while potentially minimizing user subjectivity and image post-processing requirements. The superiority of the fused model over standalone imaging and clinical models becomes particularly pronounced when the 24-hour NIHSS score is not available, highlighting its potential value in real-world clinical practice, where experienced neurologists may not be available for this assessment. Nonetheless, if resources are available, we recommend obtaining 24-hour NIHSS score to enhance the performance of both the clinical and fused models.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We would like to thank Jarrett Rosenberg and Tie Liang for their invaluable statistical consultation.

SOURCES OF FUNDING

This study was supported by a grant (R01-NS066506) from the National Institutes of Health (NIH).

DISCLOSURES

Max Wintermark serves as a consultant for Subtle Medical, Magnetic Insight, Icometrix, and EMTensor. Patrik Michel receives grant support from the Swiss Heart Foundation, the Swiss National Science Foundation, and the University of Lausanne. David S. Liebeskind consults for Cerenovus, Genentech, Medtronic, Rapid Medical Ltd, and Stryker. Gregory Albers holds equity interests (stocks) in iSchemaView and serves as a consultant for Biogen,

Genentech, and iSchemaView. Greg Zaharchuk consults for Biogen, serves as a Fiduciary officer for ISMRM, receives funding support from GE Healthcare, is a co-founder of Subtle Medical, and holds an equity interest in Subtle Medical.

Nonstandard Abbreviations and Acronyms

±1ACC	accuracy within ±1 score
ACC	accuracy
AIS	acute ischemic stroke
AUC	The area under the receiver operating characteristic curve
CI	confidence interval
DL	Deep Learning
DWI	diffusion-weighted imaging
EVT	endovascular therapy
IV tPA	intravenous tissue plasminogen activator
MAE	Mean absolute error
mRS	modified Rankin Scale
NIHSS	National Institutes of Health Stroke Scale

REFERENCES

1. Xie Y, Jiang B, Gong E, Li Y, Zhu G, Michel P, Wintermark M, Zaharchuk G. Use of gradient boosting machine learning to predict patient outcome in acute ischemic stroke on the basis of imaging, demographic, and clinical information. American Journal of Roentgenology. 2019;212(1):44–51. [PubMed: 30354266]
2. Nichols-Larsen DS, Clark PC, Zeringue A, Greenspan A, Blanton S. Factors influencing stroke survivors' quality of life during subacute recovery. Stroke. 2005;36(7):1480–1484. [PubMed: 15947263]
3. Langhorne P, Bernhardt J, Kwakkel G. Stroke rehabilitation. The Lancet. 2011;377(9778):1693–1702.
4. 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(5):1263–1265. [PubMed: 30890116]
5. Brugnara G, Neuberger U, Mahmutoglu MA, Foltyn M, Herweh C, Nagel S, Schönenberger S, Heiland S, Ulfert C, Ringleb PA et al. Multimodal predictive modeling of endovascular treatment outcome for acute ischemic stroke using machine-learning. Stroke. 2020;51(12):3541–3551. [PubMed: 33040701]
6. Laredo C, Zhao Y, Rudilosso S, Renú A, Pariente JC, Chamorro Á, Urra X. Prognostic significance of infarct size and location: the case of insular stroke. Scientific reports. 2018;8(1):9498. [PubMed: 29934530]
7. Menezes NM, Ay H, Wang Zhu M, Lopez CJ, Singhal AB, Karonen JO, Aronen HJ, Liu Y, Nuutinen J, Koroshetz WJ et al. The real estate factor: quantifying the impact of infarct location on stroke severity. Stroke. 2007;38(1):194–197. [PubMed: 17122428]
8. Ospel JM, Hill MD, Menon BK, Demchuk A, McTaggart R, Nogueira R, Poppe A, Haussen D, Qiu W, Mayank A et al. Strength of association between infarct volume and clinical outcome

depends on the magnitude of infarct size: results from the ESCAPE-NA1 trial. American Journal of Neuroradiology. 2021;42(8):1375–1379. [PubMed: 34167959]

9. Tolhuisen ML, Hoving JW, Koopman MS, Kappelhof M, van Voorst H, Bruggeman AE, Demchuck AM, Dippel DWJ, Emmer BJ, Bracard S et al. Outcome prediction based on automatically extracted infarct core image features in patients with acute ischemic stroke. Diagnostics. 2022;12(8):1786. [PubMed: 35892499]
10. Zhang MY, Mlynash M, Sainani KL, Albers GW, Lansberg MG. Ordinal prediction model of 90-day modified rankin scale in ischemic stroke. Frontiers in Neurology. 2021;12:727171. [PubMed: 34744968]
11. Shen D, Wu G, Suk H-I. Deep learning in medical image analysis. Annual review of biomedical engineering. 2017;19:221–248.
12. Liu Y, Zheng H, Liang Z, Miao Q, Brisbane WG, Marks LS, Raman SS, Reiter RE, Yang G, Sung K. Textured-based deep learning in prostate cancer classification with 3T multiparametric MRI: Comparison with PI-RADS-based classification. Diagnostics. 2021;11(10):1785. [PubMed: 34679484]
13. Thamm T, Guo J, Rosenberg J, Liang T, Marks MP, Christensen S, Do HM, Kemp SM, Adair E, Eyngorn I et al. Contralateral hemispheric cerebral blood flow measured with arterial spin labeling can predict outcome in acute stroke. Stroke. 2019;50(12):3408–3415. [PubMed: 31619150]
14. Lansberg MG, Straka M, Kemp S, Mlynash M, Wechsler LR, Jovin TG, Wilder MJ, Lutsep HL, Czartoski TJ, Bernstein RA et al. MRI profile and response to endovascular reperfusion after stroke (DEFUSE 2): a prospective cohort study. The Lancet Neurology. 2012;11(10):860–867. [PubMed: 22954705]
15. Nogueira RG, Jadhav AP, Haussen DC, Bonafe A, Budzik RF, Bhuva P, Yavagal DR, Ribo M, Cognard C, Hanel RA et al. Thrombectomy 6 to 24 hours after stroke with a mismatch between deficit and infarct. New England Journal of Medicine. 2018;378(1):11–21. [PubMed: 29129157]
16. Lansberg MG, Christensen S, Kemp S, Mlynash M, Mishra N, Federau C, Tsai JP, Kim S, Nogueria RG, Jovin T, Devlin TG, Akhtar N, Yavagal DR, Haussen D, Dehkharghani S et al. Computed tomographic perfusion to Predict Response to Recanalization in ischemic stroke. Annals of neurology. 2017;81(6):849–856. [PubMed: 28486789]
17. Yu Y, Guo D, Lou M, Liebeskind D, Scalzo F. Prediction of hemorrhagic transformation severity in acute stroke from source perfusion MRI. IEEE Transactions on Biomedical Engineering. 2017;65(9):2058–2065. [PubMed: 29989941]
18. Yu Y, Christensen S, Ouyang J, Scalzo F, Liebeskind DS, Lansberg MG, Albers GW, Zaharchuk G. Predicting Hypoperfusion Lesion and Target Mismatch in Stroke from Diffusion-weighted MRI Using Deep Learning. Radiology. 2022;307(1):e220882. [PubMed: 36472536]
19. Quinn TJ, Dawson J, Walters MR, Lees KR. Functional outcome measures in contemporary stroke trials. International Journal of Stroke. 2009;4(3):200–205. [PubMed: 19659822]
20. Banks JL, Marotta CA. Outcomes validity and reliability of the modified Rankin scale: implications for stroke clinical trials: a literature review and synthesis. Stroke. 2007;38(3):1091–1096. [PubMed: 17272767]
21. Drucker H, Burges CJ, Kaufman L, Smola A, Vapnik V. Support vector regression machines. Advances in neural information processing systems. 1996;9.
22. Hara K, Kataoka H, Satoh Y. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition.; 2018:6546–6555.
23. Ganaie MA, Hu M, Malik AK, Tanveer M, Suganthan PN. Ensemble deep learning: A review. Engineering Applications of Artificial Intelligence. 2022;115:105151.
24. Wolpert DH. Stacked generalization. Neural networks. 1992;5(2):241–259.
25. Girshick R Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision.; 2015:1440–1448.
26. Fargen KM, Kittel C, Curry BP, Hile CW, Wolfe SQ, Brown P, Mokin M, Rai AT, Chen M, Starke RM et al. Mechanical thrombectomy decision making and prognostication: Stroke treatment Assessments prior to Thrombectomy In Neurointervention (SATIN) study. Journal of NeuroInterventional Surgery. 2023.

27. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision.; 2017:618–626.

28. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition.; 2016:2921–2929.

29. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological). 1995;57(1):289–300.

30. Esmael A, Elsherief M, Eltoukhy K. Predictive Value of the Alberta Stroke Program Early CT Score (ASPECTS) in the outcome of the acute ischemic stroke and its correlation with stroke subtypes, NIHSS, and cognitive impairment. Stroke research and treatment. 2021;2021.

31. Kasner SE. Clinical interpretation and use of stroke scales. The Lancet Neurology. 2006;5(7):603–612. [PubMed: 16781990]

32. Cole JH, Franke K. Predicting age using neuroimaging: innovative brain ageing biomarkers. Trends in neurosciences. 2017;40(12):681–690. [PubMed: 29074032]

33. VIITANEN M WINBLAD B, ASPLUND K. Autopsy-verified causes of death after stroke. Acta Medica Scandinavica. 1987;222(5):401–408. [PubMed: 3425392]

34. Broderick JP, Adeoye O, Elm J. Evolution of the modified Rankin scale and its use in future stroke trials. Stroke. 2017;48(7):2007–2012. [PubMed: 28626052]

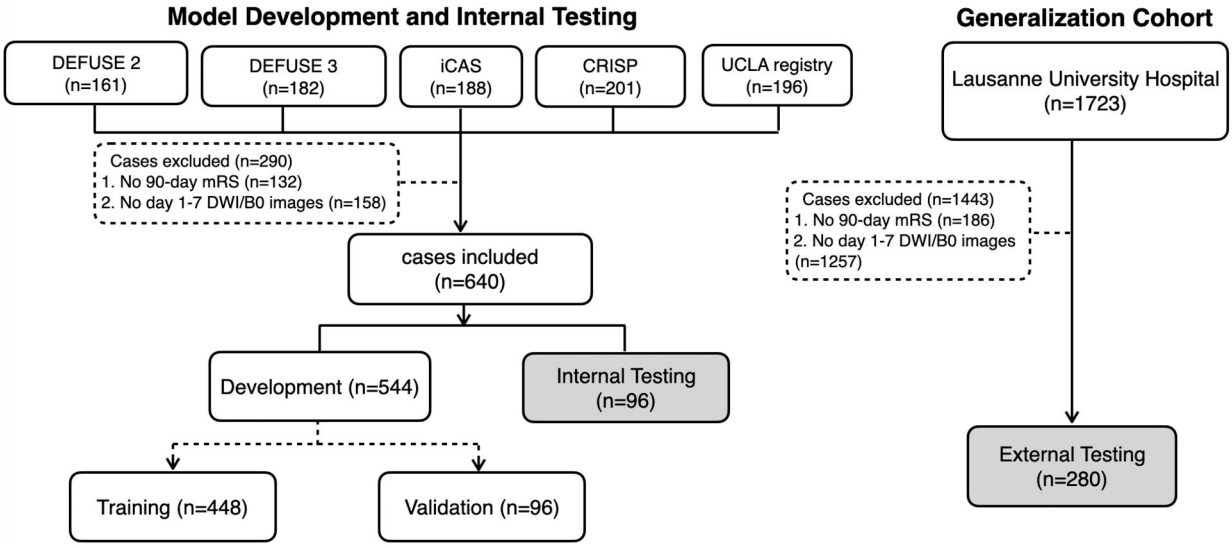


Figure 1:

Training and testing flowcharts for patients in the current study. The gray boxes highlight the test case patients who were completely excluded from all training related to model development.

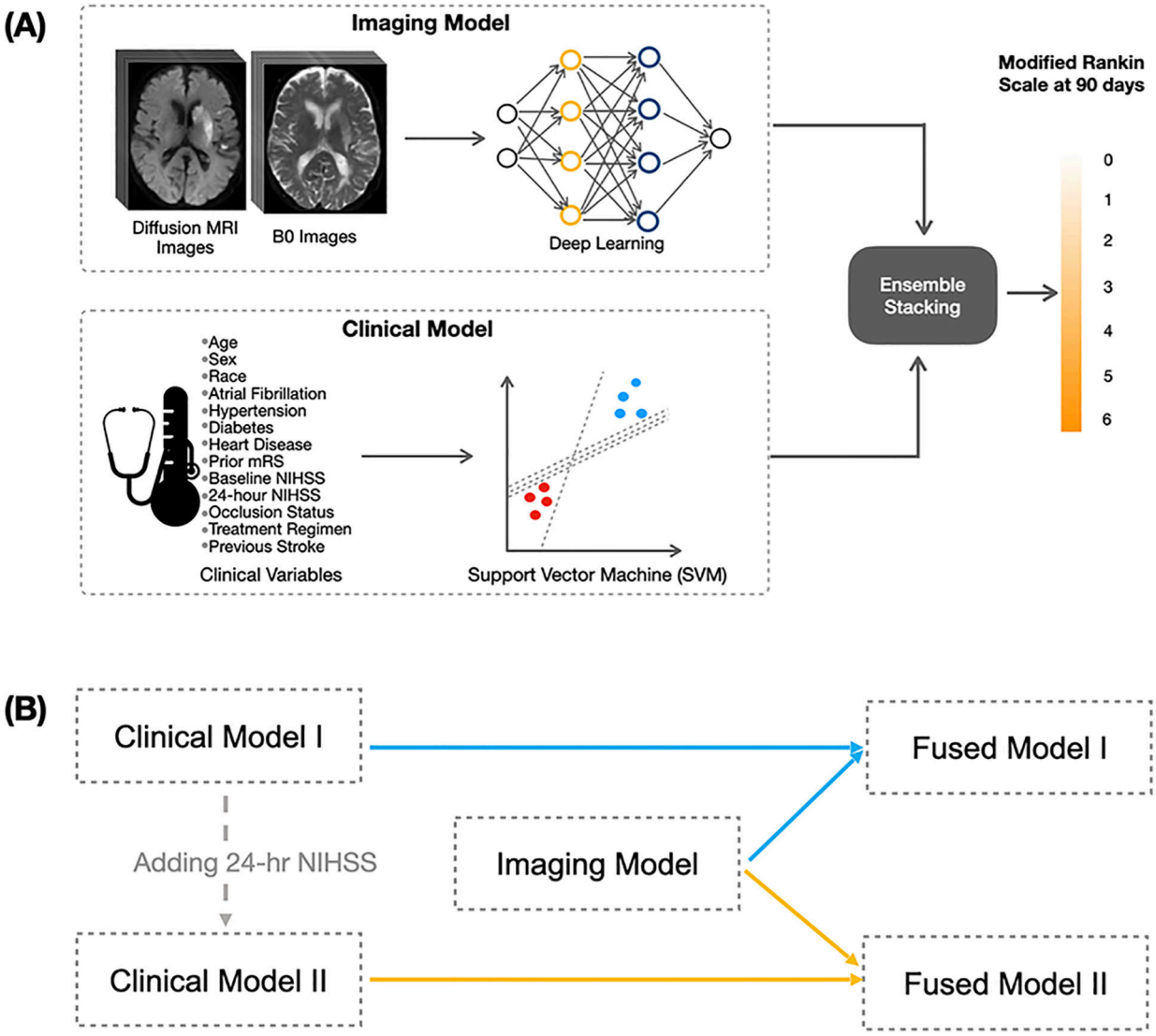


Figure 2:

(A) The overall architecture of the fused model, which comprises a deep learning-based imaging model and a clinical model. The clinical model uses clinical variables as inputs, while the imaging model includes diffusion-weighted and B0 images. **(B)** Detailed compositions of the models employed in the study. Specifically, Clinical Model I incorporates the following clinical information: age, gender, race, baseline NIHSS, prior mRS, medical history (including hypertension, diabetes, atrial fibrillation, heart diseases, and previous stroke), occlusion status, and treatment regimen. Clinical Model II extends Clinical Model I by incorporating NIHSS scores obtained after 24 hours. Fused Models I and II are created by integrating the imaging model with Clinical Models I and II, respectively. All models generate continuous predictions of 90-day mRS outcomes.

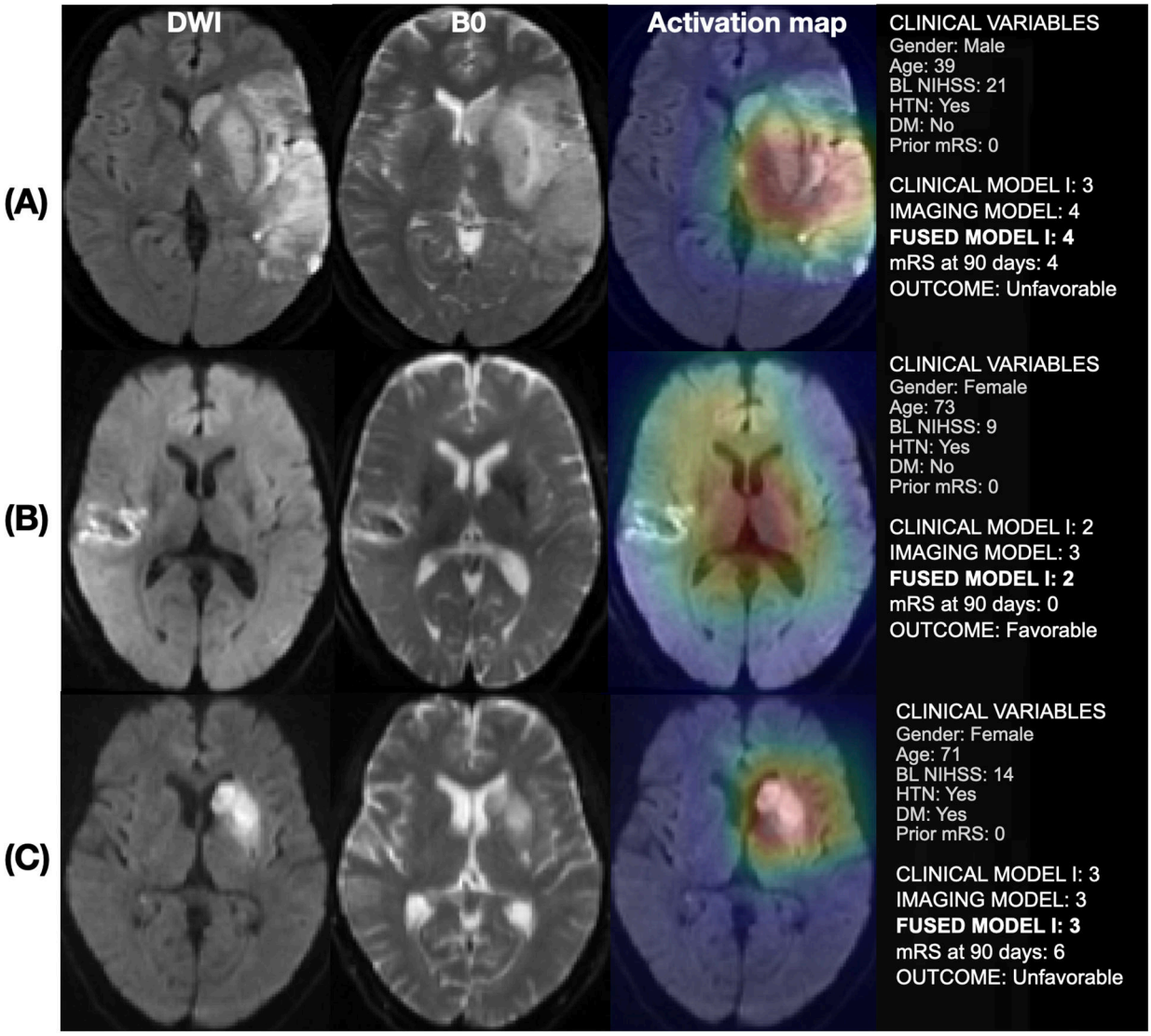


Figure 3:

MR images (first and second columns are diffusion-weighted imaging (DWI) and B0 images) and corresponding saliency activation maps (rightmost column) generated by deep learning-based imaging models for three patients with varying clinical histories and 90-day mRS scores. The activations are color-coded, with red indicating higher attention. All clinical and fused models in this figure are Type I (without 24-hour NIHSS). The predictions of the model are continuous but rounded to the nearest whole number to facilitate comparison with the true mRS. Patient A, a 39-year-old male, has a history of hypertension and a 90-day mRS of 3. The clinical and fused models correctly predict his score, and the imaging model displays high activations around the lesions. Patient B, a 73-year-old female, has no history of diabetes or hypertension and a 90-day mRS of 0. The fused and clinical models predict a score of 2, while the imaging model predicts 3, with low activations around the stroke lesion. Patient C, a 71-year-old female, has histories of diabetes and hypertension and a 90-day mRS of 6. All three models predict a score of 3, and the imaging model shows high activations around the lesions.

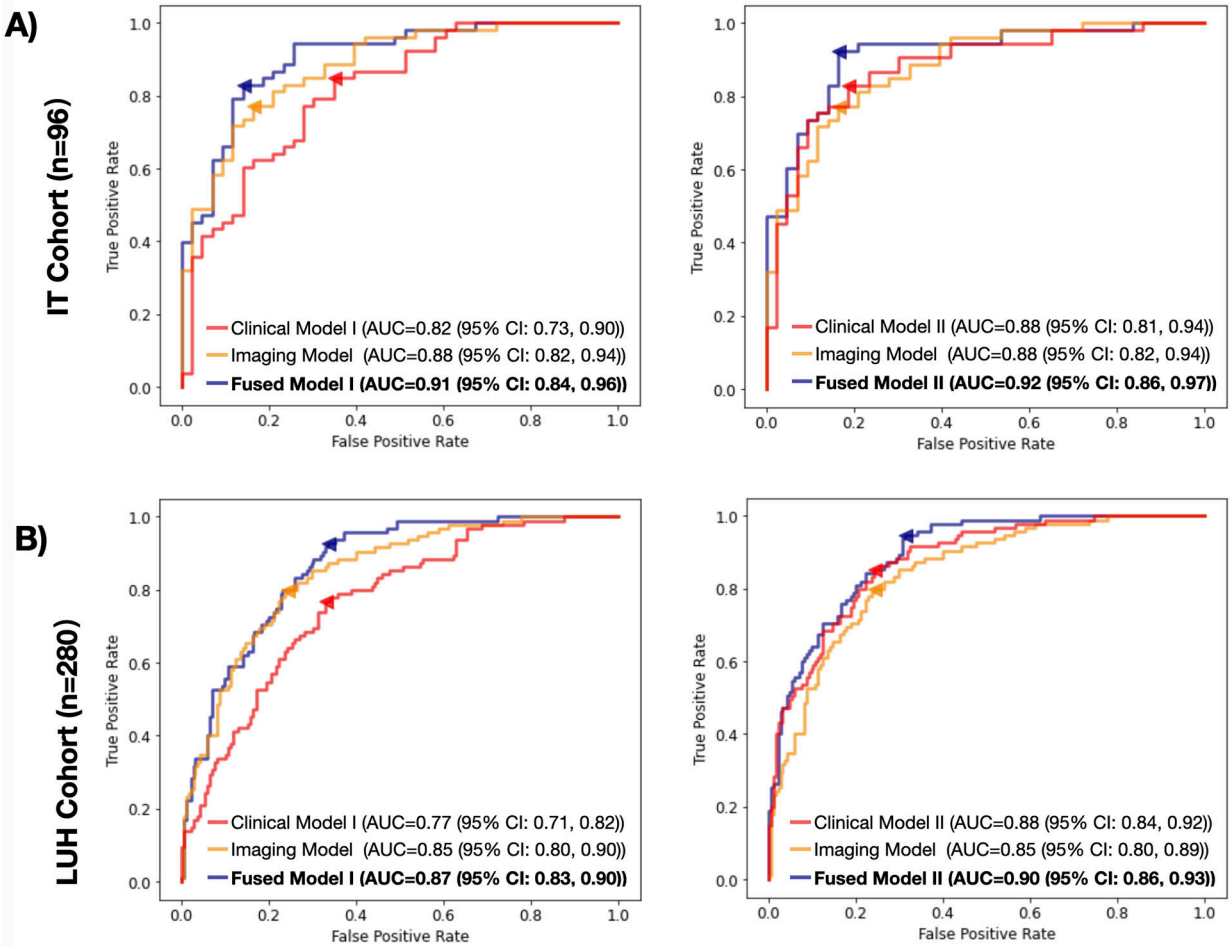


Figure 4:

ROC Comparisons for Clinical Models, Imaging Model, and Fused Models across the two testing cohorts. A) Internal Test (IT) Cohort: Left - Clinical Model I, Imaging Model, and Fused Model I comparisons; Right - Clinical Model II, Imaging Model, and Fused Model II comparisons. B) Lausanne University Hospital (LUH) Cohort: Left - Clinical Model I, Imaging Model, and Fused Model I comparisons; Right - Clinical Model II, Imaging Model, and Fused Model II comparisons. The maximum value of the Youden index was used to determine the optimal cut-off points, shown as solid triangles on the ROC curves.

Table 1:

Summary of the Characteristics of AIS Patients Included in Development, Internal, and External Test Cohorts

	Training &Validation Cohort (n=544)	IT Cohort (n=96)	LUH Cohort (n=280)
Male	271 (50.0) 0.2%[1]	48 (50.0)	159 (56.8)
Age	69 (57, 78)	68 (56, 75)	66 (52, 75)
Hypertension	384 (70.7) 0.2%	73 (76.0)	138 (56.8) 13.2%
Diabetes	134 (24.8) 0.6%	27 (28.1)	47 (19.1) 12.1%
Baseline NIHSS median (IQR)	15.0 (10.0–20.0) 0.2%	15.0 (11.0–21.0)	7.0 (4.0,14.0) 1.8%
24-hour NIHSS	10.0 (4.0–17.0) 15.1%	10.0 (4.0–18.0) 14.6%	5.0 (1.0–11.0) 3.2%
Days After Stroke median (IQR)	1 (1–4)	3 (1–5)	3 (1–5)
Large Vessel Occlusion	497 (91.4) 0.2%[1]	88 (91.7)	192 (68.6)
Treatment Methods			
No Treatment	92 (16.9)	12 (12.5)	167 (59.6)
Only IV tPA	81 (14.9)	14 (14.6)	76 (27.1)
Only EVT	220 (40.4)	34 (35.4)	11 (3.9)
IV tPA & EVT	151 (27.8)	36 (37.5)	26 (9.3)
90-day Outcome			
Favorable Outcome (90-day mRS<=2)	239 (43.9)	43 (44.8)	152 (65.2)
Unfavorable Outcome (90-day mRS>2)	305 (56.1)	53 (55.2)	81 (34.8)
mRS score at 90 days			
0	59 (10.8)	15 (15.6)	45 (16.1)
1	100 (18.4)	17 (17.7)	69 (24.6)
2	80 (14.7)	11 (11.5)	71 (25.4)
3	104 (19.1)	17 (17.7)	45 (16.1)
4	104 (19.1)	18 (18.8)	19 (6.8)
5	34 (6.3)	8 (8.3)	8 (3.0)
6	63 (11.6)	10 (10.4)	23 (8.2)

Unless otherwise mentioned, data are expressed as number (percentage) of patients. Abbreviations: NIHSS, National Institutes of Health Stroke Scale; mRS, modified Rankin scale; LUH, Lausanne University Hospital; IT: Internal Testing. IV tPA: intravenous tissue plasminogen activator; EVT: endovascular therapy.

[1]Percentage of variables missing. If no data is missing, then there will be no percentage reported.

Table 2:

Performance Comparisons for Multinomial Prediction Performance of Clinical models, Imaging model and the Fused model on IT Cohort, LUH cohort.

Models	MAE	±1 ACC	ACC
IT Cohort (n=96)			
Clinical Model I	1.23 (1.03, 1.45)	0.65 (0.54, 0.74)	0.26 (0.18, 0.35)
Clinical Model II (Clinical Model I + 24-hour NIHSS)	1.03 (0.83, 1.23) p^1=.02	0.74 (0.65, 0.82) p^1=.04	0.33 (0.24, 0.44) p^1=.10
Imaging Model	1.04 (0.88, 1.23) p^1=.03; p^2=.89	0.75 (0.67, 0.83) p^1=.06; p^2=.91	0.28 (0.19, 0.36) p^1=.74; p^2=.29
Fused Model I (Clinical Model I + Imaging Model)	0.99 (0.81, 1.19) p^1<.001†; p^2=.26 p^3=.06	0.74 (0.66, 0.82) p^1=.03; p^2=1.00 p^3=.87	0.34 (0.25, 0.44) p^1=.05; p^2=.78 p^3=.07
Fused Model II (Clinical Model II + Imaging Model)	**0.96 (0.77, 1.16)** p^1<.001†; p^2=.002† p^3=.08; p^4=.58	**0.79 (0.71, 0.88)** p^1=.004; p^2=.12 p^3=.35; p^4=.21	**0.35 (0.26, 0.46)** p^1=.04; p^2=.39 p^3=.04; p^4=.73
LUH Cohort (n=280)			
Clinical Model I	1.15 (1.03, 1.27)	0.71 (0.65, 0.76)	0.28 (0.22, 0.33)
Clinical Model II (Clinical Model I + 24-hour NIHSS)	0.91 (0.80, 1.03) p^1<0.001†	**0.84 (0.79,0.88)** p^1<0.001†	0.34 (0.29, 0.40) p^1=.02
Imaging Model	1.03 (0.91, 1.15) p^1=.001†; p^2=.05	0.77 (0.72, 0.82) p^1=.18; p^2=.17	0.35 (0.29, 0.40) p^1=.09; p^2=.78
Fused Model I (Clinical Model I + Imaging Model)	0.99 (0.87, 1.11) p^1<.001†; p^2=.09 p^3=.19	0.79 (0.74, 0.84) p^1<.009†; p^2=.19 p^3=.44	0.35 (0.29, 0.40) p^1=.05; p^2=.85 p^3=.93
Fused Model II (Clinical Model II + Imaging Model)	**0.90 (0.79, 1.01)** p^1<.001†; p^2=.88 p^3=.005†; p^4=.02†	0.83 (0.78, 0.87) p^1<.001†; p^2=.66 p^3=.10; p^4=.13	**0.36 (0.30, 0.41)** p^1=.008; p^2=.48 p^3=.38; p^4=.39

p^1 is for the statistical comparison to the Clinical Model I, p^2 is for the statistical comparison to the Clinical Model II, p^3 is for the statistical comparison to the Imaging Model, p4 is for the statistical comparison to the Fused Model I. Abbreviations: NIHSS, National Institutes of Health Stroke Scale; LUH, Lausanne University Hospital; IT: Internal Testing. MAE: Mean absolute error; ACC: accuracy for a specific mRS; ±1 ACC: mRS accuracy within ±1 score; MAE: Mean Absolute Error.

†P value still held significance ($p < 0.05$) after accounting for multiple comparisons and applying a 5% false discovery rate through the Benjamini-Hochberg procedure.