AJNR

A Clinical and Imaging Fused Deep Learning Model Matches Expert Clinician Prediction of 90-Day Stroke Outcomes

Yongkai Liu, Preya Shah, Yannan Yu, Jai Horsey, Jiahong Ouyang, Bin Jiang, Guang Yang, Jeremy J. Heit, Margy E. McCullough-Hicks, Stephen M. Hugdal, Max Wintermark, Patrik Michel, David S. Liebeskind, Maarten G. Lansberg, Gregory W. Albers and Greg Zaharchuk

AJNR Am J Neuroradiol published online 8 February 2024
http://www.ajnr.org/content/early/2024/02/08/ajnr.A8140

Published February 8, 2024 as 10.3174/ajnr.A8140

A Clinical and Imaging Fused Deep Learning Model Matches Expert Clinician Prediction of 90-Day Stroke Outcomes

Yongkai Liu, Preya Shah, Yannan Yu, Jai Horsey, Jiahong Ouyang, Bin Jiang, Guang Yang, Jeremy J. Heit, Margy E. McCullough-Hicks, Stephen M. Hugdal, Max Wintermark, Patrik Michel, David S. Liebeskind, Maarten G. Lansberg, Gregory W. Albers, and Greg Zaharchuk

ABSTRACT

BACKGROUND AND PURPOSE: Predicting long-term clinical outcome in acute ischemic stroke is beneficial for prognosis, clinical trial design, resource management, and patient expectations. This study used a deep learning–based predictive model (DLPD) to predict 90-day mRS outcomes and compared its predictions with those made by physicians.

MATERIALS AND METHODS: A previously developed DLPD that incorporated DWI and clinical data from the acute period was used to predict 90-day mRS outcomes in 80 consecutive patients with acute ischemic stroke from a single-center registry. We assessed the predictions of the model alongside those of 5 physicians (2 stroke neurologists and 3 neuroradiologists provided with the same imaging and clinical information). The primary analysis was the agreement between the ordinal mRS predictions of the model or physician and the ground truth using the Gwet Agreement Coefficient. We also evaluated the ability to identify unfavorable outcomes (mRS >2) using the area under the curve, sensitivity, and specificity. Noninferiority analyses were undertaken using limits of 0.1 for the Gwet Agreement Coefficient and 0.05 for the area under the curve analysis. The accuracy of prediction was also assessed using the mean absolute error for prediction, percentage of predictions ±1 categories away from the ground truth (±1 accuracy [ACC]), and percentage of exact predictions (ACC).

RESULTS: To predict the specific mRS score, the DLPD yielded a Gwet Agreement Coefficient score of 0.79 (95% CI, 0.71–0.86), surpassing the physicians' score of 0.76 (95% CI, 0.67–0.84), and was noninferior to the readers ($P < .001$). For identifying unfavorable outcome, the model achieved an area under the curve of 0.81 (95% CI, 0.72–0.89), again noninferior to the readers' area under the curve of 0.79 (95% CI, 0.69–0.87) ($P < .005$). The mean absolute error, ±1ACC, and ACC were 0.89, 81%, and 36% for the DLPD.

CONCLUSIONS: A deep learning method using acute clinical and imaging data for long-term functional outcome prediction in patients with acute ischemic stroke, the DLPD, was noninferior to that of clinical readers.

ABBREVIATIONS: AC = Agreement Coefficient; ACC = accuracy; ±1ACC = mRS accuracy within ±1 score; AIS = acute ischemic stroke; AUC = area under the curve; DL = deep learning; DLPD = deep learning–based predictive model; IQR = interquartile range; MAE = mean absolute error; ROC = receiver operating characteristic

Stroke affects nearly 800,000 people annually in the United States and is a major global cause of disability and mortality.[1] Survivors often face significant functional impairment that impacts their quality of life.[2] Predicting long-term clinical impairment from early-stage information in acute ischemic strokes (AIS) is crucial for enhancing rehabilitation strategies and informing clinical trial designs, resource allocation, and patient expectations.[1-3] However, prediction is complex due to the many factors influencing a patient's eventual disability level and the known weak correlation between initial infarct size and outcome.[4-7]

Although some studies[1,4,8] have attempted to predict long-term functional outcomes, these traditional methodologies have suboptimal performance due to 2 primary factors: their reliance on manually crafted imaging features, which may not be optimal predictors, and the subjective inclusion of clinical measurements,

Received October 27, 2023; accepted after revision December 7.

From the Departments of Radiology (Y.L., P.S., Y.Y., J.O., B.J., J.J.H., S.M.H., G.Z.), Electrical Engineering (J.O.), and Neurology (M.G.L., G.W.A.), Stanford University, Stanford, California; Meharry Medical College (J.H.), Nashville, Tennessee; National Heart and Lung Institute (G.Y.), Imperial College London, London, UK; Department of Neurology (M.E.M.-H.), University of Minnesota Medical School, Minneapolis, Minnesota; Department of Neuroradiology (M.W.), University of Texas MD Anderson Center, Houston, Texas; Neurology Service (P.M), Department of Clinical Neurosciences, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland; and Department of Neurology (D.S.L.), University of California, Los Angeles, Los Angeles, California.

Y. Liu and P. Shah contributed equally to this work.

This work was supported by the US Department of Health and Human Services, National Institutes of Health, National Institute of Neurological Disorders and Stroke, R01-NS066506.

Please address correspondence to Yongkai Liu, PhD, Department of Radiology, 1201 Welch Rd., Stanford, CA 94305-5488; e-mail: yongkliu@stanford.edu; @Focus_on_aca

http://dx.doi.org/10.3174/ajnr.A8140

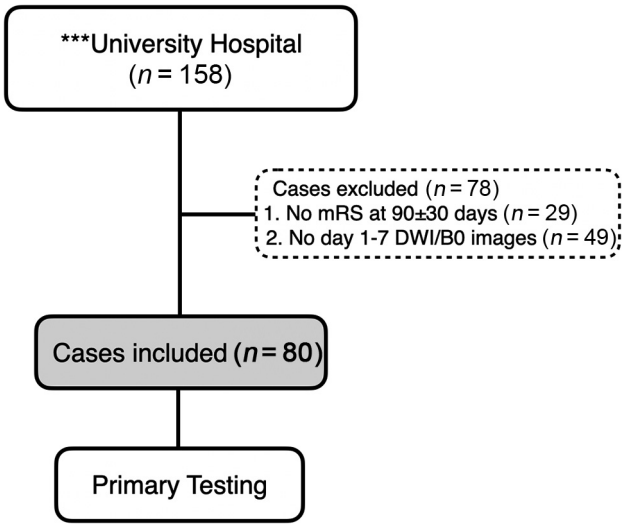

FIG 1. Flow chart for patients in the current study.

some of which may not be available in nonspecialist centers. Selecting and extracting imaging features further compounds these issues, adding yet more subjectivity and often requiring resource-intensive manual postprocessing. In recent years, deep learning (DL), particularly convolutional neural networks, has shown promise in enhancing medical imaging diagnostics and prognostics by adaptively learning from raw images.[9,10] Few existing studies have used DL to discern optimal features from medical imaging for the prediction of long-term disability, especially for the task of predicting the patient's exact score on the 90-day mRS.

The alternative to automated systems might be the predictions of expert physicians, who have significant real-world experience correlating imaging and clinical results with eventual outcomes and may be required to make judgments that can impact individual patients. However, there are few systematic evaluations of prediction of stroke outcome by humans.[11] Benchmarking against clinical readers provides a meaningful context for measuring improvement and may serve as a reference for future research and clinical evaluations. This study aimed to evaluate a previously developed deep learning–based predictive model (DLPD), which uses DWI and clinical variables to predict 90-day clinical outcomes in patients with AIS, using a prospective registry from a comprehensive stroke center. We compared its performance with that of clinical readers, including neurologists and neuroradiologists.

MATERIALS AND METHODS

Patients and MR Imaging Data Sets

This study adhered to the guidelines set forth by the US Health Insurance Portability and Accountability Act of 1996 and received approval from the institutional review board (Stanford University). Under the institutional review board guidelines, we either obtained written informed consent from all subjects or the requirement for consent was waived. The initial study population included 158 patients with AIS from a single registry who were randomly selected from a database of large-vessel occlusion candidates undergoing triage for possible thrombectomy between 2010 and 2019. To assess long-term clinical outcomes, we used the mRS,[12,13] a grading system for measuring disability levels ranging from 0 (no disability) to 6 (death). Inclusion criteria were the acquisition of MR images with DWI acquired between days 1 and 7 after the index event and after all acute therapies were complete and the 90-day mRS evaluation. Routinely collected clinical parameters included the following: age, sex, premorbid mRS, presenting and 24-hour NIHSS, and a history of hypertension or diabetes. Figure 1 provides a detailed flow chart outlining the study subjects.

DLPD

The DLPD model uses DWI and the previously mentioned clinical variables as input to predict 90-day mRS outcome. It was based on a previously developed and validated model that was trained using data from 861 patients across multiple institutions.[14] In brief, it is a fused model that takes DWI and B0 images as input to define deep features relevant to mRS prediction and then fuses these with a separate predictive support-vector machine model using the clinical variables. Part of the nonsensitive code is available at outcome prediction.

Physician mRS Prediction

Outcome prediction using the mRS scale was independently performed by 5 physicians who were given the same information as in the DLPD model. These included 2 neuroradiologists (one being a neurointerventional radiologist) with 13 and 22 years of experience, respectively; 2 stroke neurologists with 8 and 38 years of experience, respectively; and a neuroradiology fellow with 5 years of experience. To compare against the DLPD, we used the consensus mRS prediction using the median score from all the physicians, given the high agreement between them. The consensus mRS score for the physicians was created using a majority score when present (ie, the same score in $\geq$3 readers); otherwise, the median score was used.

Statistical Analysis

The performance in predicting ordinal mRS outcomes was evaluated using several metrics: the Gwet Agreement Coefficient (AC),[15] mean absolute error (MAE), mRS accuracy within $\pm$1 score ($\pm$1ACC), and accuracy (ACC). The Gwet AC, applied with ordinal weighting, quantifies the level of agreement of predictions of both the DLPD and clinicians with the ground truth. The MAE assesses the average absolute difference between the predicted scores and the actual 90-day mRS scores, with a smaller MAE indicating superior performance. $\pm$1ACC evaluates the proportion of predictions that fall within 1 mRS category of the actual score. ACC measures the proportion of predictions that precisely match the actual score. For each of these metrics, the noninferiority of the DLPD compared with the consensus of clinical readers was determined using a predefined margin of 0.1 (MAE) or 10% ($\pm$1ACC, ACC). Additionally, the area under the curve (AUC), sensitivity, and specificity were measured to evaluate the predictive accuracy of the model for unfavorable outcome (mRS $>$2), with a predefined noninferiority margin of 0.05. Analyses for ordinal outcome prediction were performed using

Stata 17.0 (StataCorp),[16] while analyses for unfavorable outcome prediction were performed using Python 3.9.12.

RESULTS

Patient Characteristics

From an initial pool of 158 patients sourced from the stroke registry, a total of 80 patients, median age of 62 years (interquartile range [IQR]: 51–75 years), including 44 men (55%), met the inclusion criteria of the study and were subsequently included in its testing set. Details about the included cohort can be found in Tables 1 and 2.

Performance of Ordinal mRS Prediction

Table 3 compares the DLPD model and the consensus of clinical readers across multiple metrics. The DLPD model had improved values in all evaluated metrics, achieving a Gwet AC of 0.79 (95% CI, 0.71–0.86), an MAE of 0.89 (95% CI, 0.70–1.11), ±1ACC of 81% (95% CI, 73%–90%), and an ACC of 36% (95% CI, 26%–46%). The level of agreement among the 5 clinical readers, as gauged by a strong Gwet AC of 0.83 (95% CI, 0.80–0.86), affirms the consistency in their judgments and justifies the use of a clinical consensus score to compare with the DLPD. The clinical consensus score achieved a Gwet AC of 0.76 (95% CI, 0.67–0.84), MAE of 0.95 (95% CI, 0.75–1.17), ±1ACC of 79% (95% CI, 70%–88%), and ACC of 36% (95% CI, 25%–46%). Noninferiority tests confirmed that the performance of the DLPD model was noninferior to that of the clinicians across all evaluated metrics, except for ACC. The significant *P* values were $P < .001$ for the Gwet AC, $P = .02$ for MAE, and $P < .001$ for ±1ACC, while the *P* value for ACC was not significant ($P = .07$). Figure 2 presents 3 illustrative examples of outcome predictions made by the DLPD and physicians.

Table 1: Summary of the characteristics of patients with AIS included in the Stanford University Hospital cohort ($n = 80$)[a]

Summary	
Characteristics	
Male	44 (55.0)
Age (yr) (median) (IQR)	62 (51–75)
History of hypertension	53 (66.3)
History of diabetes	20 (25.0)
Baseline NIHSS (median) (IQR)	12 (7–17)
24-Hour NIHSS	9 (4–17)
	3.8%[b]
Days after stroke for MR imaging (median) (IQR)	1 (1–3)
90-Day outcome	
Favorable outcome (90-Day mRS≤2)	36 (45.0)
Unfavorable outcome (90-Day mRS >2)	44 (55.0)

[a] Unless otherwise mentioned, data are expressed as number (percentage) of patients.
[b] Percentage of variables missing. If no data are missing, then there will be no percentage reported.

Table 2: MRS score[a]

Scale	Pre morbid mRS	90-Day mRS
0	67 (83.8)	6 (7.5)
1	6 (7.5)	17 (21.3)
2	3 (3.8)	13 (16.3)
3	4 (5.0)	19 (23.8)
4	0 (0.0)	12 (15.0)
5	0 (0.0)	10 (12.5)
6	0 (0.0)	3 (3.8)

[a] Data are expressed as number (percentage) of patients.

Table 3: Performance comparisons for ordinal mRS prediction between the DLPD and the clinical readers[a]

	Gwet AC	MAE	±1ACC (%)	ACC (%)
Model/readers				
Neuroradiologist I	0.70 (0.60–0.80)	1.15 (0.94–1.38)	71 (60–81)	26 (18–36)
Neuroradiologist II	0.69 (0.59–0.79)	1.14 (0.93–1.38)	70 (60–80)	29 (19–39)
Neuroradiology fellow	0.73 (0.65–0.81)	1.04 (0.85–1.24)	74 (6–84)	31 (21–41)
Clinical readers				
Neurologist I	0.75 (0.66–0.84)	1.03 (0.83–1.25)	75 (65–84)	32 (22–44)
Neurologist II	0.77 (0.67–0.86)	0.91 (0.70–1.15)	79 (69–88)	41 (30–51)
Consensus read	0.76 (0.67–0.84)	0.95 (0.75–1.17)	79 (70–88)	36 (25–46)
DLPD	0.79 (0.71–0.86)	0.89 (0.70–1.11)	81 (73–90)	36 (26–46)
P value	$P < .001$	$P = .02$	$P < .001$	$P = .07$

[a] The data in the parentheses represent the 95% confidence interval. The *P* value is for the noninferiority test between the consensus clinical reads and the DLPD with the predefined margin of 0.1 (MAE)/10% (±1ACC, ACC). The Gwet AC for agreement among 5 clinical readers is 0.83 (95% CI, 0.80–0.86), justifying the comparison with a consensus.

Predicting Unfavorable Outcome

Table 4 compares the performance of the DLPD and physicians to predict unfavorable outcome (mRS >2). The DLPD model surpassed the readers by achieving an AUC of 0.81 (95% CI, 0.72–0.89), compared with the physicians' AUC of 0.79 (95% CI, 0.69–0.87). The model was again noninferior to the physicians for this task ($P = .005$). The DLPD model had a higher specificity of 0.81 (95% CI, 0.67–0.92), which was noninferior to the clinical consensus specificity of 0.75 (95% CI, 0.60–0.88) ($P = .03$). The sensitivity of the DLPD model (0.68 [95% CI, 0.54–0.81]) was lower and did not satisfy the noninferiority margin compared with the clinical consensus (0.70 [95% CI, 0.56–0.83]). Figure 3 shows the receiver operating characteristic (ROC) of the DLPD together with the data points representing the individual and consensus physicians. The physicians' operating points are located just beneath the ROC curve of the DLPD, suggesting that for the same level of specificity or sensitivity, the DL model generally achieves slightly better performance.

DISCUSSION

This study demonstrates that a clinical and imaging fused DL model is noninferior to expert physicians in predicting specific mRS outcomes and unfavorable prognoses. Building on our prior work—which established a robust methodology across multiple institutions and demonstrated consistent performance in 2 distinct cohorts—this work not only further enhances the generalizability of the model but also provides critical benchmarks against human expert performance for the task at hand. By evaluating our methodology alongside clinical expert judgments within a unique patient cohort, we investigated the practical implications of our approach in a real-world clinical setting, an element not addressed in our prior work. Slightly better performance was observed in this cohort

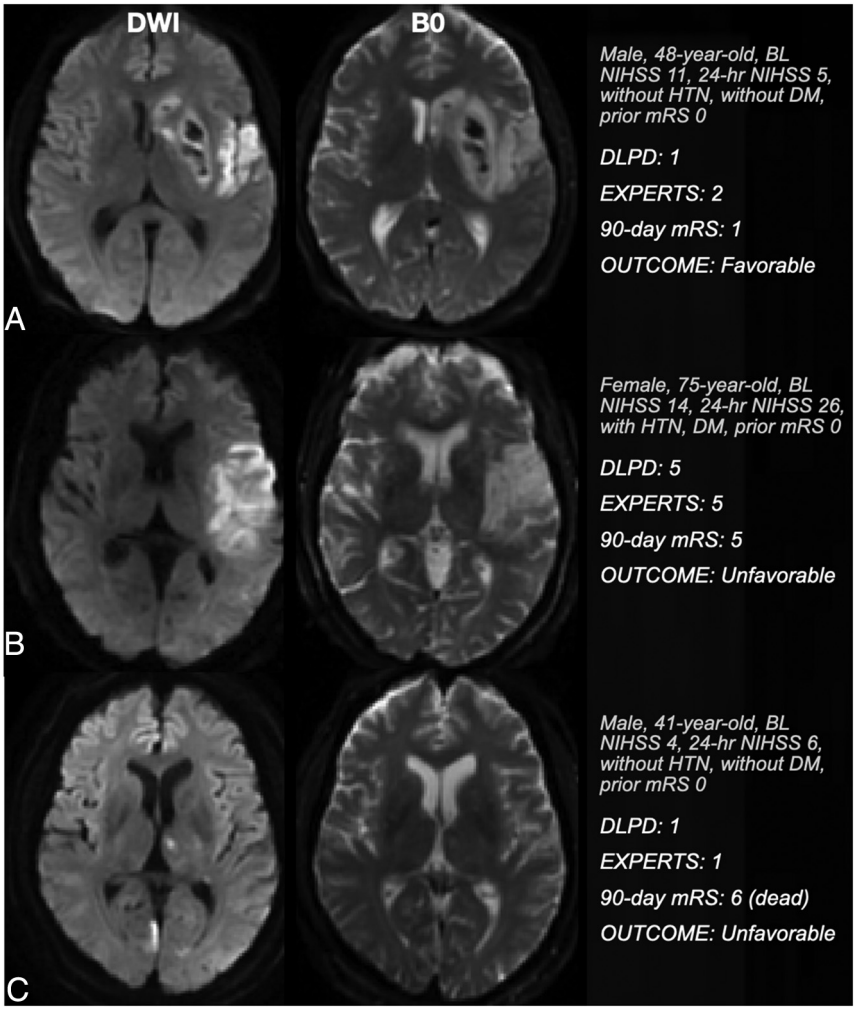

FIG 2. MR images (the first and second columns represent DWI and B0 images, respectively) for 3 patients with diverse clinical histories and 90-day mRS scores. Patient A is a 48-year-old man with a baseline NIHSS of 11, 24-hour NIHSS of 5, and a 90-day mRS of 1. He has no medical history of either diabetes or hypertension. The DL model accurately predicted his score. However, the readers overestimated his score by 1 point. Patient B, a 75-year-old woman, has a medical history that includes diabetes and hypertension and a 90-day mRS of 5. Both the DL model and the readers accurately predicted her 90-day mRS score of 5. Patient C, a 41-year-old man with no history of diabetes or hypertension, has a 90-day mRS score of 6. However, both the DL model and the readers incorrectly predicted his 90-day mRS score as 1. HTN indicates hypertension; DM. diabetes mellitus.

Table 4: Performance comparisons for unfavorable-outcome prediction (mRS >2) between the DLPD and the clinical readers[a]

	AUC	Sensitivity	Specificity
Model/readers			
Neuroradiologist I	0.76 (0.64–0.85)	0.68 (0.53–0.81)	0.69 (0.54–0.83)
Neuroradiologist II	0.81 (0.71–0.89)	0.61 (0.47–0.77)	0.86 (0.73–0.97)
Neuroradiology fellow	0.82 (0.72–0.89)	0.75 (0.62–0.87)	0.69 (0.53–0.83)
Clinical readers			
Neurologist I	0.79 (0.69–0.88)	0.82 (0.7–0.93)	0.53 (0.36–0.68)
Neurologist II	0.77 (0.67–0.86)	0.64 (0.49–0.78)	0.83 (0.7–0.95)
Consensus read	0.79 (0.68–0.87)	0.70 (0.56–0.83)	0.75 (0.60–0.88)
DLPD	0.81 (0.72–0.89)	0.68 (0.54–0.81)	0.81 (0.67–0.92)
P value	*P* = .005	*P* = .25	*P* = .03

[a] The data in the parentheses represent the 95% confidence interval. The *P* value is for the noninferiority test between the consensus clinical read and the DLPD (predefined margin, .05).

compared with the prior work, which may be attributed to several factors. These include the following: 1) natural variations in patient demographics and disease presentations across cohorts; and 2) the single-institution cohort of the current study likely offering more uniform treatment protocols, imaging techniques, and patient management strategies, unlike the multi-institution data sets of our prior work, which presented greater variability. Additionally, this work uniquely applies DL to predict stroke outcomes, achieving an ordinal mRS accuracy rate of 36%, nearly triple the rate of random guessing. Prior studies tackling this task typically had lower accuracy using methods such as linear regression and random forest machine learning with hand-crafted features.[1,8]

The DLPD also provides practical advantages. By eliminating the need for specialized neurologic expertise, it may be useful for facilities that lack immediate access to neurologists or neuroradiologists. Consequently, it broadens the scope of quality care by making sophisticated prognostic information more accessible across diverse health care environments. The DLPD model uses readily accessible imaging and clinical variables and can be easily integrated into the current clinical workflow for predicting 90-day mRS. This model requires minimal preprocessing steps, with the primary requirement being the normalization of DWI and B0 images to a standard template. Unlike approaches that depend on existing radiologic features—potentially introducing added complexity, human effort, and subjectivity—our method automatically uses information from imaging, thereby lending greater objectivity to our outcome-prediction model. For example, a recent multivariate ordinal mRS regression model required inclusion of 19 separate variables, including some derived from imaging, for which interobserver reproducibility has not been reported.[17] In contrast, the current model relies heavily on objective, data-derived imaging features, with only 7 standard clinical measurements.

This study represents the first report of clinical expert performance for 90-day mRS prediction and allows us to

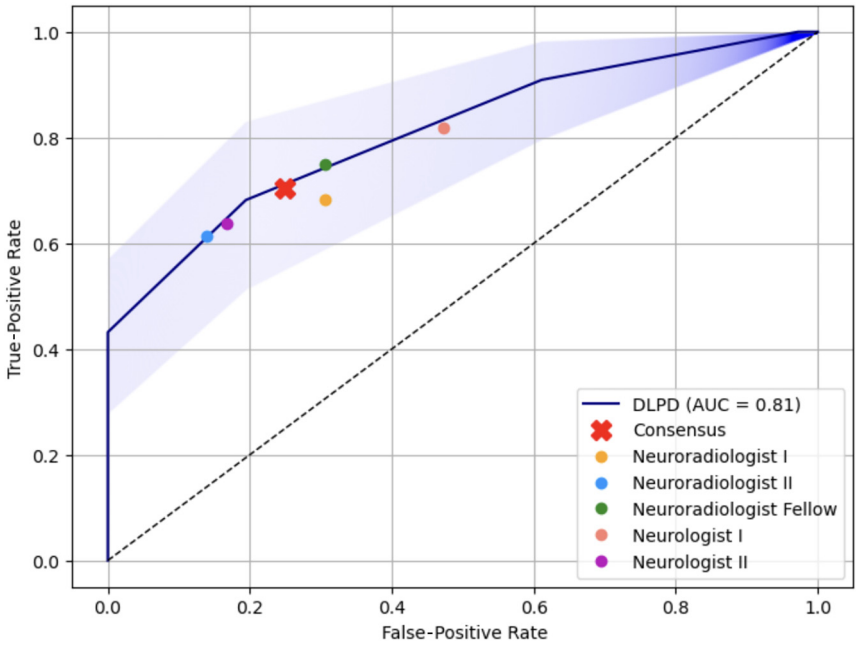

FIG 3. The AUC of the DL-based predictive model for predicting unfavorable outcomes (mRS >2) is shown alongside data points representing the performance of individual clinicians and the consensus of clinicians. The translucent *blue region* denotes the 95% confidence interval for the ROC curve, constructed using bootstrapping.

conduct a comparative analysis with the automated model. Thus, the results can act as a benchmark for future studies and further contextualize the results of the DLPD. One prior study[11] collected outcome-prediction data from treating providers before endovascular therapy and showed relatively poor performance (44% accuracy to predict into mRS bins of 0–2, 3–4, and 5–6). The performance of the readers in the current study was better, probably because they made their assessments after treatment was provided. Also, these authors stressed that the premorbid mRS was an important predictive feature, but it was often unavailable or inaccurately estimated compared with later retrospective assessment. This issue emphasizes the value of using the entire image, which can incorporate both acute and pre-existing lesions to improve prediction.

Potential reasons behind the noninferior performance of the DLPD in outcome prediction are manifold. First, the DLPD may identify and learn from patterns within complex, multimodal data, similar to or better than how physicians apply their medical knowledge and experience.[18,19] It emulates or improves on human readers by evaluating imaging and clinical data in a data-driven manner, considering all available information, from obvious clinical signs to subtle imaging hints. Additionally, these models can discern and understand nonlinear relationships and interactions among numerous variables, mirroring physicians' multidimensional thinking when assessing patient conditions and outcomes. Furthermore, DLPD models may bring additional advantages. Their inherent ability to process and learn from vast amounts of data offers unprecedented scalability. With the growth in available data, the performance of the model can potentially increase, indicating a cycle of continuous advancement that may be challenging to match solely with human expertise.

Our study has the following limitations. First, our patient testing cohort was sourced from a single registry. While this feature mimics how the tool might be used in real practice, performance in other cohorts with different characteristics of severity and age is difficult to assess. However, the model has been previously applied to 2 other clinical cohorts with diverse levels of severity and demonstrated similar performance.[14] Second, the mRS served as our primary outcome measure. While the 90-day mRS is widely used to assess chronic disability severity, its subjective determination of categories and variability in reproducibility among different examiners presents notable challenges. Third, the imaging data used in our study were obtained at least 24 hours after the initial baseline imaging; this timeframe was chosen to minimize the effects of any acute interventions, which were completed at the time of imaging. Future studies could consider using initial, pretreatment imaging combined with different therapies to predict outcomes, potentially informing treatment decision-making. Fourth, including imaging sequences beyond DWI and B0 could yield more detailed insight, though this would require more resources and image postprocessing. Fifth, despite the small number of patients in our study, we want to emphasize several factors: Our evaluation was conducted by 5 clinical readers, ensuring a thorough and nuanced assessment, being particularly noteworthy given the complex nature of the reader study and the demanding schedules of the clinicians.

Reader studies are inherently time-intensive, and coordinating such effort among 5 busy clinicians poses substantial challenges. Nonetheless, our findings demonstrate that the performance of the DL model aligns with that of their clinical evaluations, underscoring its potential for clinical application even within a limited patient cohort. This agreement reinforces our hypothesis that the model can operate at a level comparable with that of humans. Last, it is critical to recognize that outcomes may diverge significantly from predictions due to the multifaceted interplay of medical conditions, social determinants, and systemic health care factors that are not entirely predictable by our algorithms. While our model demonstrates robustness, it is not configured to anticipate every acute medical event or the full range of sociodemographic variables that may substantially affect the clinical course. Therefore, there is a clear need to continuously refine predictive methodologies, possibly incorporating a wider set of variables that capture the complexities of patient trajectories; although again, this suggestion comes with drawbacks related to the complexity of the models and the need to collect information that may be difficult to obtain.

CONCLUSIONS

We demonstrated that a DLPD model that leverages brain MR imaging and routinely obtained clinical information to predict long-term outcomes in patients with AIS generalizes well to another clinical cohort. We have further provided a benchmark of human expert performance on this task and show that the DLPD model is noninferior to predictions made by neuroradiologists and stroke neurologists.

Disclosure forms provided by the authors are available with the full text and PDF of this article at www.ajnr.org.

REFERENCES

1. Xie Y, Jiang B, Gong E, et al. **Use of gradient boosting machine learning to predict patient outcome in acute ischemic stroke on the basis of imaging, demographic, and clinical information.** *AJR Am J Roentgenol* 2019;212:44–51 CrossRef Medline
2. Nichols-Larsen DS, Clark PC, Zeringue A, et al. **Factors influencing stroke survivors' quality of life during subacute recovery**. *Stroke* 2005;36:1480–84 CrossRef Medline
3. Langhorne P, Bernhardt J, Kwakkel G. **Stroke rehabilitation.** *Lancet* 2011;377:1693–702 CrossRef Medline
4. Heo J, Yoon JG, Park H, et al. **Machine learning-based model for prediction of outcomes in acute stroke.** *Stroke* 2019;50:1263–65 CrossRef Medline
5. Brugnara G, Neuberger U, Mahmutoglu MA, et al. **Multimodal predictive modeling of endovascular treatment outcome for acute ischemic stroke using machine-learning.** *Stroke* 2020;51:3541–51 CrossRef Medline
6. Ospel JM, Hill MD, Menon BK, et al; ESCAPE-NA1 Investigators. **Strength of association between infarct volume and clinical outcome depends on the magnitude of infarct size: results from the ESCAPE-NA1 trial.** *AJNR Am J Neuroradiol* 2021;42:1375–79 CrossRef Medline
7. Albers GW, Marks MP, Kemp S, et al; DEFUSE 3 Investigators. **Thrombectomy for stroke at 6 to 16 hours with selection by perfusion imaging.** *N Engl J Med* 2018;378:708–18 CrossRef Medline
8. Zhang MY, Mlynash M, Sainani KL, et al. **Ordinal prediction model of 90-day modified Rankin scale in ischemic stroke.** *Front Neurol* 2021;12:727171 CrossRef Medline
9. Shen D, Wu G, Suk HI. **Deep learning in medical image analysis.** *Annu Rev Biomed Eng* 2017;19:221–48 CrossRef Medline
10. Liu Y, Zheng H, Liang Z, et al. **Textured-based deep learning in prostate cancer classification with 3t multiparametric MRI: comparison with PI-RADS-based classification.** *Diagnostics (Basel)* 2021;11:1785 CrossRef Medline
11. Fargen KM, Kittel C, Curry BP, et al; Satin Research Group. **Mechanical thrombectomy decision making and prognostication: Stroke treatment Assessments prior to Thrombectomy In Neurointervention (SATIN) study.** *J Neurointerv Surg* 2023;15:e381–87 CrossRef Medline
12. 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;38:1091–96 CrossRef Medline
13. Broderick JP, Adeoye O, Elm J. **Evolution of the modified Rankin scale and its use in future stroke trials.** *Stroke* 2017;48:2007–12 CrossRef Medline
14. Liu Y, Yu Y, Ouyang J, et al. **Functional outcome prediction in acute ischemic stroke using a fused imaging and clinical deep learning model.** *Stroke* 2023;54:2316–27 CrossRef Medline
15. Gwet KL. **Computing inter-rater reliability and its variance in the presence of high agreement.** *Br J Math Stat Psychol* 2008;61:29–48 CrossRef Medline
16. StataCorp. *Stata Statistical Software: Release 15*. College Station, Texas: StataCorp LLC. 2017.
17. Chalos V, Venema E, Mulder MJ, et al; HERMES Collaborators; MR CLEAN Registry Investigators. **Development and validation of a post-procedural model to predict outcome after endovascular treatment for ischemic stroke.** *JAMA Neurol* 2023;80:940–48 CrossRef Medline
18. Kamnitsas K, Ledig C, Newcombe VF, et al. **Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation.** *Med Image Anal* 2017;36:61–78 CrossRef Medline
19. Esteva A, Kuprel B, Novoa RA, et al. **Dermatologist-level classification of skin cancer with deep neural networks.** *Nature* 2017;542:115–18 CrossRef Medline