npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01687-y

Hybrid machine learning for real-time prediction of edema trajectory in large middle cerebral artery stroke

Check for updates

Ethan Phillips[1], Odhran O'Donoghue[1], Yumeng Zhang[2], Panos Tsimpos[3], Leigh Ann Mallinger[4], Stefanos Chatzidakis[5,6], Jack Pohlmann[4], Yili Du[7], Ivy Kim[4], Jonathan Song[8], Benjamin Brush[9], Stelios Smirnakis[5,6,10], Charlene J. Ong[4,8,11] ✉ & Agni Orfanoudaki[1,11] ✉

In treating malignant cerebral edema after a large middle cerebral artery stroke, clinicians need quantitative tools for real-time risk assessment. Existing predictive models typically estimate risk at one, early time point, failing to account for dynamic variables. To address this, we developed Hybrid Ensemble Learning Models for Edema Trajectory (HELMET) to predict midline shift severity, an established indicator of malignant edema, over 8-h and 24-h windows. The HELMET models were trained on retrospective data from 623 patients and validated on 63 patients from a different hospital system, achieving mean areas under the receiver operating characteristic curve of 96.6% and 92.5%, respectively. By integrating transformer-based large language models with supervised ensemble learning, HELMET demonstrates the value of combining clinician expertise with multimodal health records in assessing patient risk. Our approach provides a framework for accurate, real-time estimation of dynamic clinical targets using human-curated and algorithm-derived inputs.

Large middle cerebral artery (MCA) infarction is a potentially lethal form of stroke, occurring in between 18% and 31% of ischemic stroke cases involving MCA occlusion[1]. A major driver of poor stroke outcomes is malignant cerebral edema, which can result in a 40–80% risk of neurological deterioration and death[1]. Specifically, space-occupying malignant edema displaces and compresses the surrounding brain tissue, causing further damage referred to as mass effect[1–6].

Early recognition of evolving cerebral edema is imperative as treatments, such as surgical decompression, can reduce the risk of mortality from 80% to 20% in eligible patients[7,8]. Pharmaceutical strategies, such as hyperosmolar therapy, are also widely employed by intensivists in efforts to treat worsening or life-threatening edema[9,10]. However, the course of malignant edema can be unpredictable, varying rapidly over mere hours or more slowly over multiple days. For this reason, stroke patients at risk of malignant edema are often monitored in intensive care unit settings to closely watch for signs of neurological deterioration.

In practice, clinicians use a variety of dynamic information available to them to monitor the risk of edema for individual patients at specific time points, including physical exam assessments, laboratory data, vital signs, and neuroimaging results[9]. Neuroimaging techniques, often in the form of computed tomography (CT) and magnetic resonance imaging, are particularly important for determining the extent of edema as they provide quantitative evidence of edema progression[8,11]. A common clinical marker is the midline shift (MLS) of the septum pellucidum, measured in millimeters of displacement from the cerebral midline[8,11]. MLS is a measurable, quantifiable, and clinically relevant indicator of worsening mass effect, used to standardize communication for edema severity, and thus becomes a key determinant of stroke patient treatment and management decisions[8,12,13]. Prior studies have shown that MLS greater than 5 mm within the first two days is associated with neurological deterioration and early mortality[13]. More recently, MLS as low as 3 mm has been shown to be associated with worse long-term outcomes[12].

Current strategies for monitoring mass effect and other secondary injuries primarily rely on physical examination and confirmatory imaging. However, guidelines recognize that the clinical practice of detecting arousal depression due to mass effect is often inadequate, as it may only become evident after a significant secondary injury has already occurred[1]. This challenge is compounded in patients whose mental status is already

[1]University of Oxford, Oxford, UK. [2]North Carolina State University, Raleigh, NC, USA. [3]Massachusetts Institute of Technology, Cambridge, MA, USA. [4]Boston Medical Center, Department of Neurology, Boston, MA, USA. [5]Brigham & Women's Hospital, Department of Neurology, Boston, MA, USA. [6]Harvard Medical School, Boston, MA, USA. [7]Boston University School of Public Health, Boston, MA, USA. [8]Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA. [9]NYU Langone Hospital, New York, NY, USA. [10]Jamaica Plain Veterans Administration Hospital, Department of Neurology, Boston, MA, USA. [11]These authors contributed equally: Charlene J. Ong, Agni Orfanoudaki. ✉e-mail: cjong@bu.edu; agni.orfanoudaki@sbs.ox.ac.uk

depressed from other factors, including the initial injury, medications, fever, or toxic-metabolic abnormalities. Therefore, in the absence of continuous neuroimaging, early signs of deterioration may go unnoticed. Near-continuous CT imaging is impractical, however, due to constraints on CT availability, risks associated with patient transport, and concerns regarding radiation exposure[14,15]. Imaging in many clinical settings is often limited to once every 24 h, or prompted only by clear signs of clinical deterioration. As a result, current guidelines do not specify the optimal frequency for surveillance CT imaging to monitor edema progression, leaving decisions to clinician discretion and contributing to variability in practice and quality of care[9,13,16].

Data-driven models for edema risk assessment could lead to more personalized and accurate screening policies. However, existing models predicting cerebral edema risk[17–20] rely on structured and curated data collected early in hospitalization to forecast late clinical outcomes, such as death or the need for surgical decompression after medical interventions have taken place (see also Supplementary Table 1). Relevant predictors identified by these models include approximations of higher infarct volume (such as the national institutes of health stroke scale (NIHSS)), baseline laboratory values, and severe early mass effect[17,19–21]. Limited research exists on effectively integrating dynamic changes in these variables into standardized risk assessments, which is critical for providers making real-time decisions across iterative time points. In addition to limited external validation and relatively small sample sizes[17–19], currently available static models often lack utility for personalized decision-making as new information becomes available, and are therefore not frequently consulted by physicians in practice. This is a significant limitation, particularly for patients where treatment decisions, including surgical decompression, are made outside the time frame for which there is high-quality evidence of efficacy (<48 h)[22].

For these reasons, improved methods to estimate cerebral edema trajectory are needed to leverage the rich sources of structured and unstructured data currently available in electronic health records. Readily-available, dynamic, and accurate MLS severity estimation tools could assist medical providers in making more timely treatment decisions, improving operational and clinical outcomes, and delivering more personalized care. Recent efforts to dynamically identify other important hospitalization events, including sepsis and non-neurological clinical deterioration, reinforce the need and potential utility of dynamic forecasting in neurocritical care settings[23,24]. Specifically in the context of cerebral edema, a backward-looking trajectory approach analyzing trends in laboratory and vital sign data identified subtle increases in white blood cell count, temperature, and sodium prior to clinical deterioration events, highlighting potential dynamic biomarkers of worsening cerebral edema[25].

Based on these findings, we curated a retrospective, multi-modal, multi-institutional dataset comprised of both static and time-varying variables from electronic health records, radiographic report texts, and expert-labeled neuroanatomic features derived from radiographic images[21,25,26]. This multi-modal approach builds on prior research showing that integrating data of different types from multiple sources into a single model leads to improved predictive performance over models which used just one data type[27]. Using this dataset, we developed and externally validated the Hybrid Ensemble Learning Models for Edema Trajectory (HELMET) to predict worsening MLS class (0 mm, 0–3 mm, 3–8 mm, and >8 mm) within 8-hour (HELMET-8) and 24-hour (HELMET-24) windows. We employed a combination of machine learning techniques, including large language models for the interpretation of the raw radiographic texts and an ensemble learning algorithm for downstream final predictions from structured data inputs. HELMET provides a paradigm for the development and validation of dynamic prediction scores for complex and volatile targets that are not routinely captured within structured electronic health records.

Our aim is to complement existing non-temporal cerebral edema prediction models and build on their successes by providing more granular data to inform imaging and treatment decisions at critical moments in a generalizable way. By incorporating data from two distinct academic medical centers with differing patient pool demographics, our study demonstrates the generalizability of HELMET in diverse patient populations[28]. Our work can assist in prompting earlier life-saving interventions and more efficient resource use by making edema progression predictions accessible to clinical teams at dynamic time horizons. We believe our findings represent the first step toward developing policies that alert the clinical team to evolving secondary injury and aid in the appropriate use of diagnostic testing.

Results

Study population

The derivation cohort consists of 623 patients with acute MCA ischemic stroke, affecting at least half of the MCA territory, who were retrospectively identified from admissions to Massachusetts General Hospital and Brigham and Women's Hospital—core institutions of the Mass General Brigham healthcare system in Boston, Massachusetts—between January 2006 and July 2021. We also leveraged a prospective external validation cohort of 60 patients with acute MCA ischemic stroke, affecting at least half of the MCA territory, admitted to Boston Medical Center between May 2019 and November 2023, drawn from an existing available dataset originally compiled for pupillometry research. The latter constitutes the largest safety-net hospital in New England with a different racial and socioeconomic patient population makeup than Mass General Brigham. The derivation cohort was used to train the HELMET models, while the Boston Medical Center cohort, selected for its diverse patient population, served as an independent dataset for external validation. The full patient inclusion diagrams for both cohorts are shown in Fig. 1. Information on the exclusion criteria can be found in Section "Patient Identification & Exclusion".

Table 1 summarizes the characteristics of both patient cohorts. For the Mass General Brigham cohort, the average age was 68.0 years, and 48.8% of the patients were female. In the Boston Medical Center cohort, the average age was 67.7 years, and 60.0% of the patients were female. A statistically significant difference in racial composition was found between the two cohorts, with the Boston Medical Center cohort having a higher proportion of non-White patients (p-value < 0.001). The Boston Medical Center cohort also had worse average stroke severity indicators at admission (NIHSS and the Alberta Stroke Programme Early CT Score (ASPECTS)) by a small but statistically significant margin (p-values < 0.005). While no difference was found in the proportion of uninsured patients, the Boston Medical Center cohort had a significantly higher proportion of Medicaid beneficiaries (p-value < 0.001) which reflects the hospital's safety net status.

The number of patients by maximum MLS class reached over their hospitalization is shown for both cohorts in Supplementary Table 5, further showcasing the slight differences in edema severity across the cohorts. Since edema trajectory and overall risk of patient deterioration are highly dependent on initial edema severity, we present summarized characteristics of hospitalization for patients in both cohorts disaggregated by the patient's MLS class after their initial CT scan in Supplementary Table 6. We focused on predictions in the first seven days following patient presentation to the hospital and truncated data beyond this point. For full details of our inclusion criteria and data transformations, please see the Methods section.

To derive dynamic prediction models of edema trajectory, longitudinal patient data were transformed to per-hour *observations* for each patient over the course of their hospitalization (described in Section "Data Curation"). For the Mass General Brigham dataset (derivation cohort), data transformations resulted in 8515 observations (patient-hours) for the 8-hr prediction task and 15,696 observations for the 24-hr prediction task. For the Boston Medical Center dataset (external validation cohort), transformations resulted in 1891 observations for the 8-hr prediction task and 3713 observations for the 24-hr prediction task. Total observations for each outcome class and horizon prediction task are reported in Table 2.

Model performance

The HELMET models were trained using randomized five-fold bootstrapped partitions of the derivation cohort patients, and evaluated on the remaining test set patients from the Massachusetts General

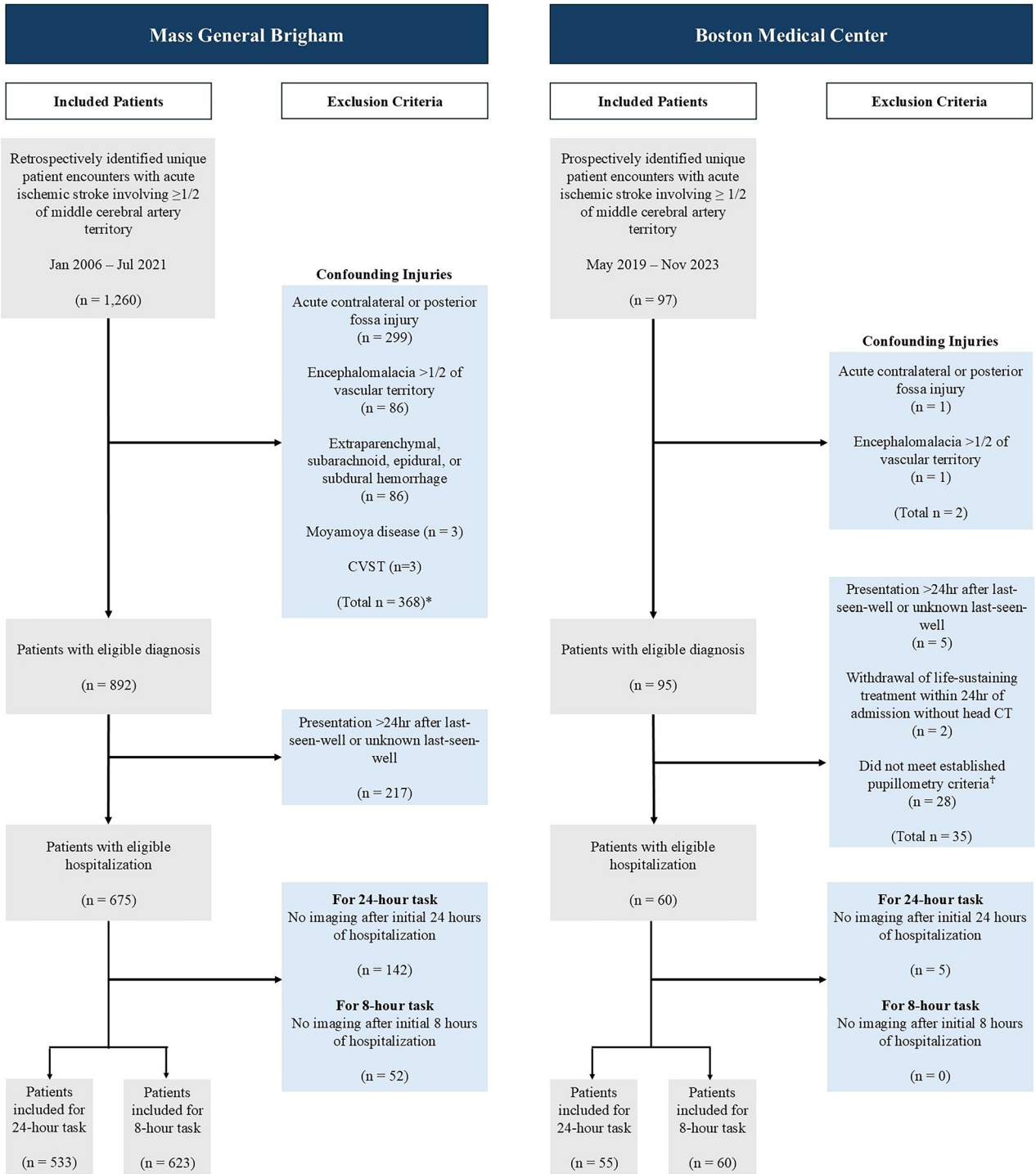

Fig. 1 | Patient inclusion diagrams. After applying exclusion criteria and removing patients with insufficient data, 623 patients were included in the derivation cohort dataset for the 8-hr task, and 60 patients were included in the external validation cohort dataset for the 8-hr task. For the 24-hr task, 533 patients were included in the derivation cohort, and 55 patients were included in the external validation cohort. *Patients may meet multiple exclusion criteria, leading to sums that do not match with total number of excluded patients. †Pupillometry criteria required a minimum of three observations preceding radiographic or clinical evidence of mass effect (e.g. MLS≥5mm), with no data gaps exceeding 24 hrs, and initially reactive, non-sluggish pupils.

Brigham derivation cohort as well as the full external validation cohort from Boston Medical Center. Two separate models were trained to predict MLS severity on 24-hr (HELMET-24) and 8-hr (HELMET-8) horizons, respectively. The HELMET models were retrospectively evaluated in head-to-head comparisons against multinomial regression models trained separately in the derivation and external validation cohorts using the same input features as the pre-existing EDEMA score, originally developed to predict potentially life-threatening malignant edema[17]. By design, the EDEMA baseline models were provided an edge over the HELMET models on the external validation cohort since they were trained separately on both the derivation and external validation datasets, while the HELMET models were only trained on the derivation dataset. Figure 2 provides an illustration of the dataset curation, feature extraction, and model derivation process.

Table 1 | Patient cohort characteristics

Variables	Mass General Brigham	Boston Medical Center	*p*-value	Captured in Medical Record
Number of patients				
8-h task	623	60		
24-h task	533	55		
Patient Demographics				
Age (years), mean (SD)	68.0 (15.3)	67.7 (16.0)	0.858	Yes
Sex - Female	304 (48.8%)	36 (60.0%)	0.128	Yes
Race - White	481 (77.2%)	17 (28.3%)	0.000	Yes
Race - Black	40 (6.4%)	15 (25.0%)	0.000	Yes
Race - Asian	26 (4.2%)	4 (6.7%)	0.000	Yes
Race - Other/Unknown	76 (12.2%)	24 (40.0%)	0.000	Yes
Insurance Status				
Non-Medicaid Insured (Medicare, MA, or Private)	464 (74.5%)	26 (43.3%)	0.005	Yes
Medicaid Beneficiary	136 (21.8%)	31 (51.7%)	0.000	Yes
Uninsured	23 (3.7%)	3 (5.0%)	0.879	Yes
Previous Stroke*	70 (11.2%)	7 (11.7%)	1.000	Yes
History of atrial fibrillation	301 (48.3%)	17 (28.3%)	0.005	Yes
History of hypertension	448 (71.9%)	41 (68.3%)	0.662	Yes
Admission Vitals and Labs				
NIHSS, mean (SD)	17.2 (5.8)	19.6 (6.2)	0.003	Yes
ASPECTS, mean (SD)	4.7 (2.9)	6.1 (2.7)	0.001	Yes
Mean Arterial Pressure (mmHg), mean (SD)	103.3 (17.8)	110.1 (19.1)	0.006	Yes
Systolic Blood Pressure (mmHg), mean (SD)	149.9 (28.7)	158.4 (30.8)	0.031	Yes
Diastolic Blood Pressure (mmHg), mean (SD)	79.7 (15.3)	85.9 (15.7)	0.003	Yes
Heart Rate, mean (SD)	81.8 (19.5)	84.3 (18.8)	0.374	Yes
Body Temperature (°F), mean (SD)	97.7 (1.0)	97.4 (1.1)	0.124	Yes
White Blood Cell Count (1000 cells/μL), mean (SD)	11.7 (8.3)	10.7 (3.8)	0.268	Yes
Blood Glucose* (mmol/L), mean (SD)	149.7 (59.1)	158.0 (67.9)	0.452	Yes
HbA1c* (mmol/mol), mean (SD)	6.2 (1.3)	6.2 (1.4)	0.246	Yes
Osmolality (mOsmol/kg), mean (SD)	298.8 (12.1)	306.8 (16.4)	0.002	Yes
Creatinine (mg/dL), mean (SD)	1.1 (0.8)	1.1 (0.5)	0.575	Yes
Sodium (mEq/L), mean (SD)	137.8 (3.5)	137.6 (3.3)	0.643	Yes
Blood Urea Nitrogen (mg/dL), mean (SD)	20.9 (12.0)	17.7 (8.6)	0.009	Yes
Length of Stay (days), median (IQR)	11 (9)	18 (19)	0.001	Yes
Initial Stroke Characteristics				
Left Hemisphere Stroke	297 (47.7%)	33 (55.0%)	0.342	No
Anterior Cerebral Artery Involved	37 (5.9%)	10 (16.7%)	0.007	No
Vessel Occlusion	298 (47.8%)	29 (48.3%)	0.666	No
0 - None, ICA, or ICA Terminus	457 (73.4%)	29 (48.3)	0.000	No
1 - MCA Horizontal Segment	117 (18.8%)	14 (23.3%)	0.000	No
2 - MCA Insular Segment	42 (6.7%)	12 (20.0%)	0.000	No
3 - MCA Opercular or Cerebral Segments	7 (1.1%)	5 (8.3%)	0.000	No
Ongoing Stroke Characteristics				
CT scans per patient, median (IQR)	4 (3)	5 (3)	0.000	No
Hours Between Scans, mean (SD)	11.3 (9.1)	18.8 (22.7)	0.003	No
First MLS (mm), mean (SD)	3.7 (2.8)	0.4 (0.9)	0.000	No
Maximum MLS (mm), mean (SD)	6.6 (4.3)	6.6 (5.5)	0.560	No
Petechial Hemorrhage	314 (50.4%)	16 (26.7%)	0.000	No
Parenchymal Hemorrhage	60 (9.6%)	16 (26.7%)	0.000	No
Collateral Score	416 (66.8%)	32 (53.3%)	0.051	No
0 - No collaterals	34 (8.2%)	3 (9.4%)	0.285	No
1 - < 50% with > 0% MCA territory	193 (46.4%)	12 (37.5%)	0.285	No

Table 1 (continued) | Patient cohort characteristics

Variables	Mass General Brigham	Boston Medical Center	*p*-value	Captured in Medical Record
2 - > 50% with < 100% of MCA territory	144 (34.6%)	10 (31.3%)	0.285	No
3 - 100% of MCA territory	45 (10.8%)	7 (21.9%)	0.285	No
Cerebral Atrophy	620 (99.5%)	66 (100%)	1.000	No
0 - Normal volume or mild ventricular enlargement	350 (56.5%)	40 (66.7%)	0.164	No
1 - Moderate or severe ventricular enlargement	270 (43.5%)	20 (33.3%)	0.164	No
Treatment				
Medical Thrombolysis*	280 (44.9%)	15 (25.0%)	0.004	Yes
Mechanical Thrombectomy*	128 (20.5%)	48 (80.0%)	0.000	Yes
Hours before Mechanical Thrombectomy, median (IQR)	6.0 (3.0)	5.0 (7.5)	0.297	Yes
Thrombolysis in cerebral infarction (TICI)	127 (20.4%)	47 (78.3%)	0.000	No
TICI 0 - No perfusion	23 (18.0%)	5 (10.6%)	0.022	No
TICI 1 - No distal branch filling	9 (7.0%)	0 (0%)	0.022	No
TICI 2a - <50% filling	28 (21.9%)	5 (10.6%)	0.022	No
TICI 2b - >50% filling	36 (28.1%)	17 (36.2%)	0.022	No
TICI 2c or 3 - 100% filling	31 (24.2%)	20 (42.5%)	0.022	No
Treated with osmotic therapy	221 (35.5%)	34 (56.7%)	0.002	Yes
Hypertonic saline (3%)	76 (12.2%)	27 (45.0%)	0.000	Yes
Hypertonic saline (23.4%)	90 (14.4%)	0 (0.0%)	0.003	Yes
Mannitol	185 (29.7%)	27 (45.0%)	0.021	Yes
Decompressive Hemicraniectomy	73 (11.7%)	13 (21.7%)	0.044	Yes
Clinical Outcomes				
Modified Rankin Score, mean (SD)	5.0 (0.9)	5.0 (0.9)	0.609	Yes
Discharge Disposition	622 (99.9%)	60 (100%)	1.000	Yes
Home	18 (2.9%)	1 (1.7%)	0.000	Yes
Rehabilitation	273 (43.8%)	22 (37.3%)	0.000	Yes
Long Term Care	120 (19.3%)	1 (1.7%)	0.000	Yes
Hospice	38 (6.1%)	8 (13.6%)	0.000	Yes
Death	173 (27.8%)	18 (30.5%)	0.000	Yes
Other	0 (0.0%)	10 (16.7%)	0.000	Yes

Basic patient characteristics across the derivation and external validation datasets. All descriptive statistics were calculated using the 8-h task patient cohorts. Continuous variables are reported as the mean (standard deviation) and used either a two-sample *t*-test or the Mann-Whitney U test for significance testing, depending on normality. Categorical or binary variables are reported as the patient count (proportion) and used the χ^2 test for significance testing.
*Included in EDEMA baseline models.

Table 2 | Observation† counts per prediction class for each task

Data set	Task	Maximum MLS class			
		0mm	0–3mm	3–8mm	> 8mm
Mass General Brigham	8-h	2012	1568	3405	1530
(derivation cohort)	24-h	2638	2713	6867	3478
Boston Medical Center	8-h	523	309	620	439
(validation cohort)	24-h	873	696	1223	921

†Observations are transformed hourly sets of data for a given patient. Each patient may account for up to 8 observations per scan on the 8-h task, and up to 24 observations per scan on the 24-h task.

Each model's performance was assessed using both the original dataset encompassing all observations (overall metrics) as well as a filtered dataset including only observations where the patient's current MLS state differed from their future target class within the predictive window (filtered metrics), to further analyze model performance on observations capturing clinically relevant transitions in MLS state. Figure 3 illustrates in radar plots the overall and filtered performance of each model for each task and cohort. The five radial axes represent the performance metrics of sensitivity, specificity, accuracy, area under the precision-recall curve (AUPRC), and area under the receiver operating characteristic curve (AUROC). Across all measures, higher values are plotted further from the center, indicating superior predictive performance. Since our classification problem required prediction of future MLS among four possible classes, a random guess model would result in an average AUROC of 0.5, average AUPRC of 0.25, accuracy of 0.25, sensitivity of 0.25, and specificity of 0.75. Therefore, performance scores exceeding these thresholds are described as better than random. The exact

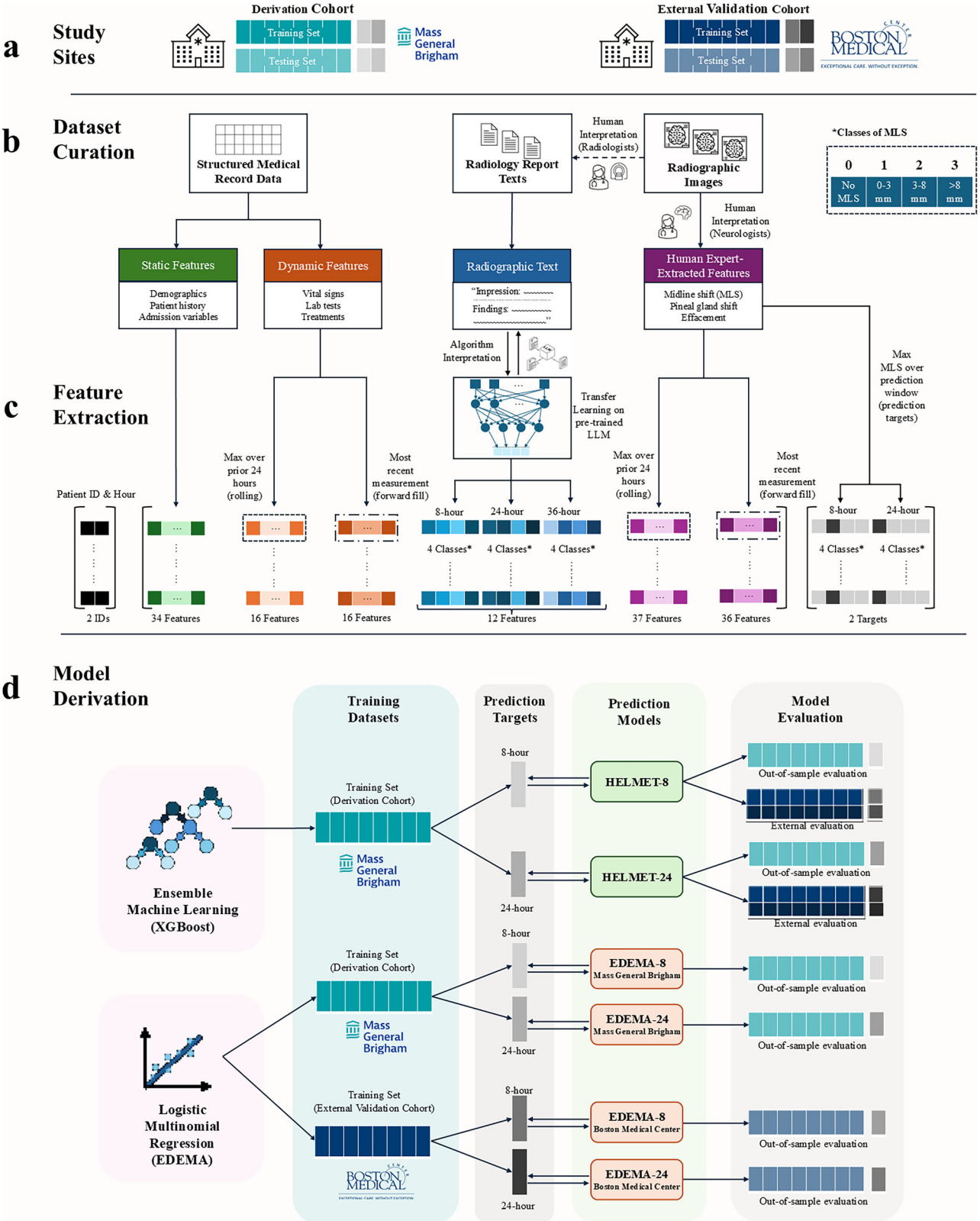

Fig. 2 | Data processing and model development. The figure summarizes the curation and transformation of input data, leading to the training and evaluation of HELMET models and EDEMA baseline models. Across the panels, we show the study site cohorts **a**; multimodal dataset construction using static, time-varying, and text-based variables **b**; transformations applied to the data to generate observations **c**; and subsequent model training **d**.

formulation of composite average metrics is described in Section "Model Evaluation".

In the out-of-sample evaluation of the derivation cohort, the large language model-enhanced HELMET-24 model achieved a mean AUROC score of 96.7%, sensitivity of 91.2%, and specificity of 94.0%. These scores outperform the EDEMA-24 baseline by 18.7 percentage points in AUROC, 53.6 percentage points in sensitivity, and 9.7 percentage points in specificity. On the 8-hr prediction task, HELMET-8 resulted in a mean AUROC of

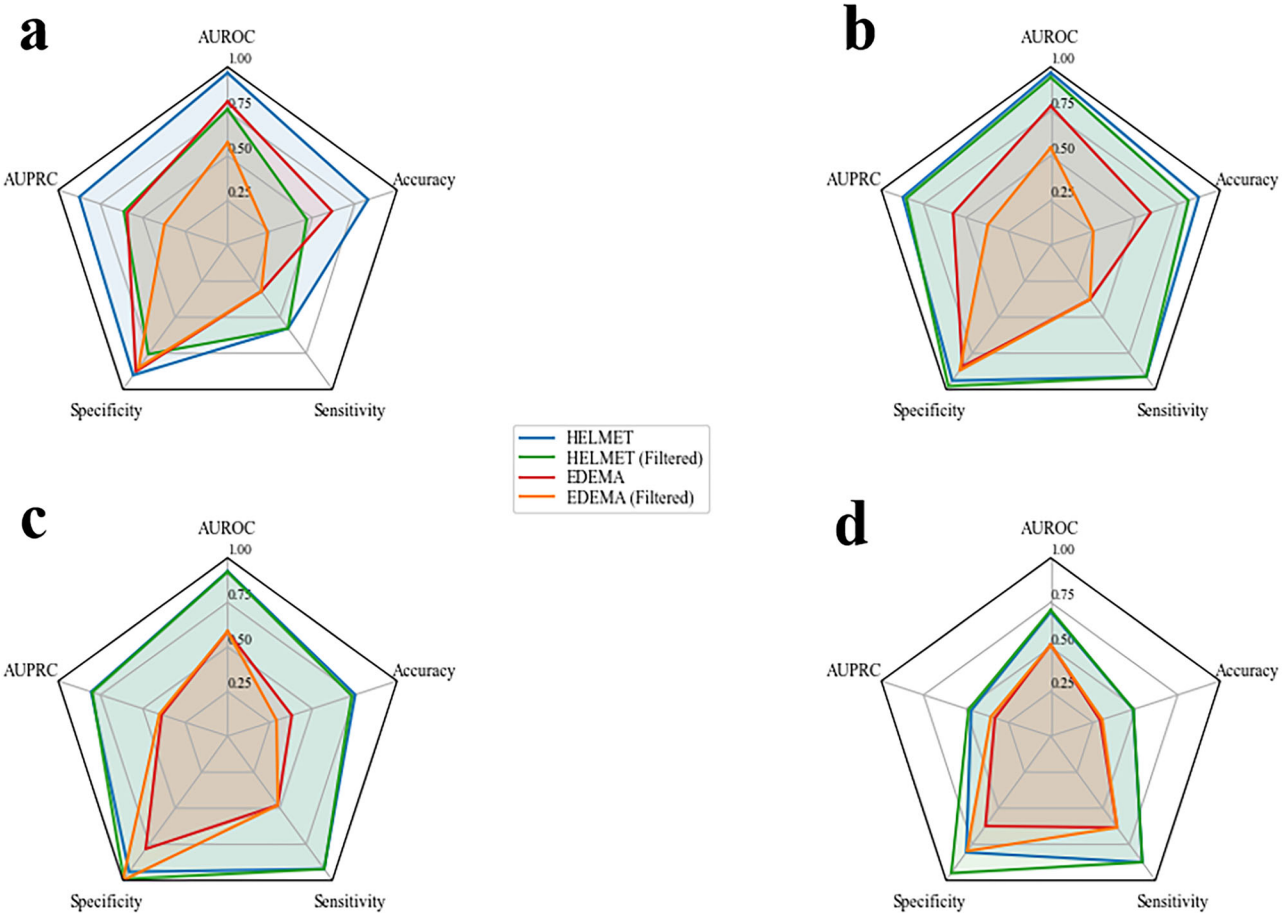

Fig. 3 | Performance comparison. Comparative performance of HELMET models and EDEMA baseline models across all cohorts and prediction tasks. Each radial axis represents a distinct performance metric, with higher performance values further from center. Polyhedrons with greater total area on the plots show higher-performing models, and smaller polyhedrons show lower-performing models. Blue lines represent the overall performance for the HELMET models, green lines represent the filtered performance for the HELMET models, red lines represent the overall performance for the baseline EDEMA models, and orange lines represent the filtered performance for the baseline EDEMA models. Specifically, we summarize model performance across tasks and cohorts in the following panels: **a** derivation cohort for the 8-hr prediction task; **b** derivation cohort for the 24-hr prediction task; **c** external validation cohort for the 8-hr prediction task; **d** external validation cohort for the 24-hr task.

96.6%, sensitivity of 57.7%, and specificity of 90.2%, outperforming the EDEMA-8 model by 16.1, 25.6, and 2.6 percentage points across the three metrics, respectively. In the dataset filtered for changes in MLS severity class, both HELMET-24 and HELMET-8 outperformed baseline EDEMA models. HELMET-24 achieved a mean filtered AUROC of 94.1% compared to 54.7% for the baseline, while HELMET-8 achieved 76.2% compared to baseline performance of 57.5%.

To assess the generalizability of our models, we also evaluated the HELMET models on the entirety of the external validation cohort. We observed that the HELMET models consistently outperform the EDEMA baseline models by a significant margin across both sites and both tasks. HELMET-24 achieved a mean overall AUROC of 69.7% and a mean filtered AUROC of 70.7% on the 24-hr task, outperforming the EDEMA-24 model by 18.4 percentage points on overall AUROC and 19.6 percentage points on filtered AUROC. On the 8-h task, HELMET-8 achieved an overall AUROC of 92.5% and a filtered AUROC of 92.1%, outperforming EDEMA-8 by 34.0 and 33.3 percentage points respectively.

The complete table of performance metrics, including respective 95% confidence intervals, is presented in Supplementary Table 11, and receiver operating characteristic and precision-recall curves are shown in Supplementary Figs. 1 and 2, respectively. AUROC scores for each model and task stratified across various hour-based periods of hospitalization (<24 hrs since last seen well, 24–48 hrs, 48–96 hrs, ≥96 hrs) and stratified by insurance status are also available in Supplementary Tables 12 and 13. As a sensitivity analysis, we also compared the HELMET framework with the EDEMA baseline on a simplified binary classification task using a 5 mm threshold to distinguish severe midline shift. The results of this analysis are summarized in Supplementary Section 7 and Supplementary Table 14.

Model feature interpretation

To determine the relative importance of contributing features to our models, we applied the Shapley Additive Explanation framework on each of the four prediction MLS classes of HELMET-24 and HELMET-8 (see also Section "Feature Importance Analysis"). Figure 4 illustrates the composition of the 20 most important features, ranked by Shapley Additive Explanation values, in the HELMET models for the 24-hr and 8-hr prediction tasks, categorized by the most recent prior MLS class. The Shapley analysis allowed us to gain insight into the interplay of the different data sources comprising the HELMET predictions. Specifically, our results indicate that the large language model predictions on radiological report texts are highly important across both the 8-h and 24-hr prediction task models. Human-extracted features by neurology experts, which may not always be present in the dictated radiology reports, and their associated times make up the second-largest category of high-impact features. We also observe that while dynamic variables (such as laboratory test results and vital signs) contribute significantly to the 8-hr horizon predictions, they are not as important to the 24-hr predictions. Static variables (either demographic variables or measurements from the time of admission) also

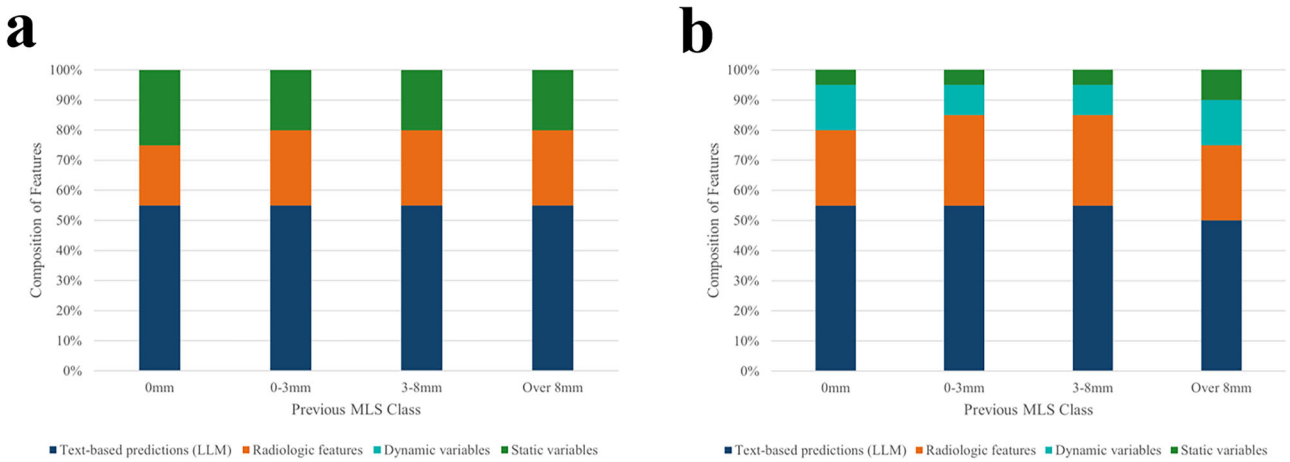

Fig. 4 | HELMET Feature Importance Composition. Comparison of importance of HELMET feature categories across tasks by previous MLS class of hourly patient observations. Dark blue represents proportion of high-importance features from large language model interpretation of radiology report texts, orange shows proportion from human-extracted radiology image features, teal shows proportion of features from dynamic medical records, and green shows proportion from static patient characteristics. first panel **a** shows feature composition for HELMET-24 model, while second panel **b** shows feature composition for HELMET-8 model.

contribute significantly to both models. The detailed Shapley Additive Explanation plots for each task are provided in Supplementary Fig. 3 and the exact definition of variables included in each class of features are described in Supplementary Section 2.

In the 24-hr prediction task, text-based predictions generated by the large language models emerge as key determinants of patient trajectory. Expert-curated radiographic measurements, including the most recent MLS value, the prior MLS value, and the time and value of the first measured MLS, also contribute significantly to HELMET predictions. Our analysis also highlights the role of approximate stroke size based on NIHSS, as well as laboratory markers, such as maximum white blood cell count over the past 24 hours and blood urea nitrogen at admission. Notably, the administration of 23.4% hypertonic saline is the only treatment-related feature ranked among the top predictors. Furthermore, patient age becomes a significant factor in predicting edema progression within higher MLS classes.

In the 8-hr prediction task, the most recent and prior MLS values, along with class probabilities derived from the large language model, are the most significant contributors. Similar to the 24-hr model, the timing and value of the first MLS measurement, as well as the time of the first MLS value exceeding 3 mm, rank among the top features. Laboratory markers, including white blood cell count and blood glucose levels, alongside vital signs such as pulse and temperature, further aid in distinguishing between edema states. Notably, no treatment indicators are among the highest-ranked predictors.

Discussion

Our dynamic time series models leverage nonlinear machine learning algorithms to predict the risk of edema and the trajectory of MLS using a comprehensive multi-modal dataset for two distinct future time windows. To the best of our knowledge, HELMET-8 and HELMET-24 constitute the first dynamic risk models for predicting the trajectory of cerebral edema on an hourly basis, leveraging data from two healthcare systems.

Existing cerebral edema risk scores provide only static predictions of late clinical outcomes, such as death or decompressive hemicraniectomy, primarily using linear techniques[17–20] (with some use of nonlinear models[29]). While useful for initial triage, these models are limited in their utility as clinicians follow individual patients over time and make decisions based on new data. Recent work by our group has highlighted the significance of incorporating post-baseline patient data to enhance the prediction of inpatient outcomes[21]. No studies to our knowledge had utilized granular updating information over the course of hospitalization to estimate the actual state of cerebral edema by objective measurements, such as MLS.

Our analyses in the derivation cohort reveal that HELMET-24 outperformed HELMET-8 by a small margin across all overall metrics. However, the performance difference across tasks was more pronounced when looking at filtered metrics, indicating that the HELMET architecture may be better suited for making predictions over longer time windows. One possible explanation for this observation is that changes in edema trajectory can appear to occur more abruptly when looking at closer time horizons, while such changes in edema trajectory are smoothed over longer prediction windows. Another key contributing factor is likely the frequency of imaging, the primary input feature of our models, which was obtained approximately every 11.3 hrs in the derivation cohort and every 18.8 hrs in the validation cohort. This difference in scanning frequency between the cohorts likely also explains the improved performance of HELMET-24 over HELMET-8 in the external validation cohort. Information on the frequency of laboratory and vital sign data collection is reported in Supplementary Section 2.2 and the missingness analysis in Supplementary Section 3. Subsequent studies could elucidate optimal predictive horizons for these newly developed hybrid models.

By incorporating predictions from fine-tuned large language models into HELMET, we advance the existing literature on multimodal machine learning in medicine[30,31]. Our models were significantly improved by the inclusion of both upstream predictions from large language models fine-tuned on raw radiology reports as well as manually-measured neuroanatomic variables hypothesized as relevant by neurology experts (see Fig. 2). Our hybrid approach highlights the benefit of using clinician-generated raw texts to capture otherwise unmeasured variables and physician beliefs about the patient's trajectory. However, while features derived from the large language model predictions make up a plurality of high-impact features across both tasks (see Fig. 4), the large language model predictions in isolation were inadequate indicators of MLS trajectory (see Supplementary Section 6.1). Our findings underscore that multi-modal hybrid approaches combining both expert-derived features and raw data appear to significantly enhance outcome prediction accuracy in clinical settings and may inform which features should be routinely included in radiology reports.

Our study builds on recent efforts to understand the temporal evolution of physiological markers in patients with large MCA stroke. Ong et al.[25] applied a retrospective, trajectory-based analysis to examine how changes in laboratory and vital sign variables were associated with cerebral edema-related outcomes. The use of multivariable time-dependent Cox regression provided valuable insight into cohort-level risk patterns and the potential prognostic value of dynamic biomarkers. These approaches are complementary, resulting in an in-depth exploration of emerging trends. However, the study by Ong et al.[25] was designed to identify average trends

across populations, rather than to deliver individualized forecasts that update in real time as new data become available. In contrast, the HELMET framework leverages nonlinear machine learning methods to generate patient-specific predictions of edema progression over clinically relevant time horizons. By integrating multimodal inputs, HELMET-8 and HELMET-24 dynamically estimate the severity of midline shift for each patient. This individualized, forward-looking approach enables real-time clinical decision support, complementing earlier trajectory-based work and offering a pathway toward dynamic proactive management of cerebral edema in critical care settings.

The Shapley Additive Explanation analysis (Supplementary Fig. 3) enables us to infer key clinical variables that drive model performance. Dynamic variables capturing previous MLS measurements were the most critical predictors of future MLS status. Additionally, key time-related features, such as the time of the first non-zero MLS and the first MLS exceeding 3 mm, highlight the natural course of edema growth, which typically peaks between 2 and 5 days post-ictus[7,11,13,32–34]. Time is a crucial factor often overlooked in other studies. While clinicians often assess risk based on time subjectively, our model is the first to quantitatively integrate time in a standardized way. In contrast to most risk algorithms that neglect temporal dynamics, our results demonstrate that accurate time quantification is essential for predicting imminent edema progression, aligning with clinical intuition that the same degree of edema is less critical later in the treatment course. In the absence of clear guidelines on imaging timing and its influence on surgical outcomes[9], our model provides valuable guidance for clinicians in predicting MLS worsening and optimizing patient management using real-time data.

Consistent with the literature, our Shapley Additive Explanation analysis highlighted other influential variables, including dynamic white blood cell count, admission blood glucose, and temperature. Elevated glucose at presentation has been previously linked to malignant edema and poor outcomes[17], while recent analyses showed that white blood cell count and temperature increase before radiographic evidence of mass effect[25]. Our results also highlighted that laboratory test results and vital signs have higher predictive power in the 8-hour compared to the 24-hr task, capturing more effectively short-term changes closer in time to the event. Our finding that hypertonic saline administration was also among the top features for HELMET-24 likely reflects the subjective physician risk assessment of the patient and, due to the observational nature of the data, does not provide any causal insights. Intriguingly, we did not observe similar importance of mannitol or other preparations of hypertonic saline, which may reflect its use in clinical practice and should be further studied. Given the heterogeneity in medical treatment patterns[9], the connection between osmotic therapy and clinical intuition should be interpreted with caution.

Our results reveal that human insights and radiographic features extracted from scan images play a complementary role to the large language model predictions, despite using the same foundational data source (CT images). Medical professionals bring nuanced understanding through expert-curated radiographic features, adding a layer of interpretability and context that purely algorithmic approaches may lack, especially in settings with limited sample size. Complementing past studies focused on human-AI interactions[35,36], our work underscores the significant role of human insights in enhancing the predictive power of large language models, particularly in scenarios where critical variables, such as MLS, are not routinely recorded in structured form. The importance of such specific expert-curated radiographic features in our models indicates a possible improvement to radiology reporting whereby clinicians should aim to extract and record additional radiographic features at the time of observation in their report texts.

The synergy between algorithm-derived insights from large language models and human-extracted features highlights the potential of hybrid artificial intelligence systems in clinical settings. These systems leverage the precision and scalability of machine learning while retaining the critical contextual understanding provided by human expertise. This combination is especially valuable in settings such as neurocritical care, where the dynamic nature of conditions like cerebral edema demands a nuanced approach to prediction and intervention. In such highly specific tasks, and in the absence of large databases for research, the successful creation of robust machine learning models may hinge on the consistent and coherent extraction of features. Researchers should aim to work with clinicians to generate more robust datasets of human-extracted variables needed to develop better prediction tools, and future studies in clinical machine learning can benefit from using such datasets by employing hybrid approaches. As deep learning models for radiographic images become more powerful and widespread[37], there may also be additional value in integrating their outputs into hybrid frameworks such as HELMET. Further research should therefore aim to explore combinations of further data modalities and additional interpretive AI-based tools with human-extracted features, with a focus on synergistic integration of automated and expert-driven insights.

Notably, the HELMET models generalize to an external population substantially better than previous baseline models[17,18]. While there is an expected decrease in performance from internal validation to external cohorts due to unmeasured, context-dependent factors, our results demonstrate that model generalizability can be enhanced by leveraging dynamic, time-updated features. Defining an acceptable performance threshold for any predictive model requires not only benchmarking against existing tools, but also understanding how the model compares to clinician gestalt and how it integrates into real-world clinical workflows. Optimizing such models for clinical application will require further prospective studies that assess both their relative accuracy and their impact on decision-making and patient outcomes. When deployed to a new site, we anticipate that HELMET, like other models, will benefit from re-calibration to local practice patterns for maximally effective use. Nevertheless, our external validation results enhance confidence in the applicability of HELMET, as well as the broader use of our dynamic, hybrid approach for developing risk prediction models in clinical practice.

An important implication of our work is that shifting risk prediction from a single baseline assessment to a continuous, longitudinal approach could improve real-time patient triage, optimize imaging resource use, and explore whether early-warning alerts based on these predictions can enhance patient outcomes and care quality. Studying the implementation of these algorithms will be crucial in assessing their clinical utility. Future research could build on existing studies showing the effectiveness of machine learning tools when paired with well-designed clinical interventions. Additionally, sensitive models like these could have a significant impact in smaller or under-resourced settings where neurointensive care may be unavailable or overburdened. Further investigation is also needed to bridge the gap between prediction and action by using machine learning-derived predictions to provide prescriptive recommendations for when scans or clinical interventions should occur.

There are several limitations to our work. The dataset size was constrained by the specific inclusion criteria and the manual labeling of radiographic images. Since imaging was performed at the discretion of treating physicians, the timing of MLS measurements was inconsistent. While MLS is just one indicator of worsening mass effect, and its clinical relevance may vary depending on factors such as age and brain atrophy, it remains a well-established, critical, and measurable radiographic biomarker of cerebral edema.

Similar to other retrospective studies, our dataset contains missing values due to its reliance on hospital electronic health records, which were imputed using widely established techniques. While the impact of missing data in dynamically updated variables was reduced through forward filling and rolling maximums, the high degree of missingness in some static variables (such as body temperature at admission) is a limitation of our data that could have biased our findings. Future studies should seek to collect patient data prospectively to ensure consistent collection of key variables. Such prospective studies should also evaluate how HELMET-8 and HELMET-24 can be integrated into clinical workflows to guide imaging frequency, ICU monitoring, and early interventions. Implementation strategies should focus on real-time clinical decision support, ensuring that predictions are

actionable and seamlessly incorporated into existing care protocols. Additionally, understanding clinician adoption and whether real-time alerts can enhance patient care will be critical for optimizing their deployment.

Related to the limitations due to data missingness, temporal transformations in the data may introduce bias, as patients with more complete data have multiple observations. Our approach using the most recent and maximum values may oversimplify the trajectory of MLS measurements. We also lacked access to key physical examination data, such as neurological deterioration, quantitative pupillometry, and other longitudinal multimodal monitoring methods including electroencephalograms, optic nerve sheath diameter, or direct intracranial pressure monitoring. The integration of these variables along with complementary data modalities, such as the radiographic images, constitute a clear focus for future direction. However, the instruments needed to collect these data are not ubiquitous among all hospital centers or are not routinely used in clinical care (such as intracranial pressure monitoring in ischemic stroke). Therefore, using variables available to most neuroICU centers leads to broader model generalizability and utility.

We also acknowledge that there was a significant increase in reported observations per patient after 2016 due to a change in the electronic medical record system. This created two quasi-distinct data distributions within the training set. We believe that the successful generalization to external validation cohort data suggests that both distributions within the derivation cohort are valuable to the training of our model. Both the derivation cohort data and the external validation cohort data were collected from hospitals in the same metropolitan area (Boston, Massachusetts). While the baseline descriptive statistics presented in Table 1 highlight significant differences in race, ethnicity, stroke severity, and Medicaid enrollment (a proxy for socioeconomic status) across the two patient populations, we acknowledge that geographic similarity or the under-representation of Asian populations may bias the generalizability of our results. Even though our hospital systems differ significantly in racial and socioeconomic composition, future investigations should prioritize external validation among more diverse patient populations to ensure the applicability of such models to a wider range of patient groups and locations.

Our choice of using 8-hr and 24-hr horizon predictions reflects the needs of clinicians to make real-time scanning and treatment decisions over short time windows during a patient's hospitalization. However, these short prediction horizons limit the ability of our models to make longer-term predictions of lethal edema occurance. Further, we lacked reliable functional outcome data from the end of a patient's hospitalization, such as modified Rankin Scale (mRS) values, which limits our models' ability to predict the final patient condition.

Methods

Figure 2 illustrates the study design, including dataset curation, feature extraction, and model development. The summary diagram presents how the multimodal datasets were constructed, transformed, and used for model development and evaluation. Section "Patient Identification & Exclusion" outlines the data sources and patient selection criteria (Panel a). Sections "Data Curation-Large Language Model-Derived Features" describe the structure and processing of the clinical and radiographic datasets, including temporal feature engineering, imputation, and target outcome construction (Panels b and c). Sections "HELMET Model Development, EDEMA Baseline Model Development" detail the model training and evaluation process for the HELMET and EDEMA baseline models (Panel d). Section "Model Evaluation" specifies the model evaluation procedure, while Section "Feature Importance Analysis" presents the interpretability analysis conducted on the HELMET models.

The updated Transparent Reporting of Multivariable Prediction Models for Individual Prognosis or Diagnosis (TRIPOD+AI) and Journal of Medical Internet Research Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research were followed[38,39]. A completed reporting checklist can be found in the Supplementary Information (see Supplementary Fig. 4).

Patient identification & exclusion

The derivation cohort data was obtained from the Massachusetts General Brigham hospital system Research Patient Data Registry and Electronic Data Warehouse, which form the centralized clinical data registries of the organization. To promote the inclusion of diverse and historically marginalized populations, we sourced external validation data from a more racially and socioeconomically diverse patient population. The external validation cohort was obtained from the Boston Medical Center electronic medical records (see panel A of Fig. 2) stored in the hospital's Clinical Data Warehouse. The patient inclusion diagrams for both cohorts are shown in Fig. 1. Ethical approval for the study was granted by the Institutional Review Boards at Massachusetts General Brigham and Boston Medical Center. Informed consent was not required because the only patient data collected was standard of care, no research intervention was implemented, and the study used anonymized, retrospective patient records. The clinical data obtained from the two hospital systems included both structured demographic, vitals, lab, and outcome data and unstructured data in the form of radiology reports and clinical notes. Decision rules regarding ascertainment and cleaning of all data from the electronic health records are described in Pohlman et al.[26].

For the retrospectively identified derivation cohort, we queried the Massachusetts General Brigham data registry for unique patient encounters between January 8th, 2006 and July 5th, 2021 with stroke diagnosis codes and used an established natural language processing model to identify patients with acute MCA stroke involving ≥1/2 of the MCA territory[40,41]. For the prospectively-collected external validation cohort, we leveraged an existing registry of patients with acute MCA ischemic stroke involving at least 1/2 of the MCA territory admitted after May 15th, 2019 and discharged before November 25th, 2023 who also had pupillometry measurements. We excluded patients with confounding injuries, including acute contralateral or posterior fossa injury; encephalomalacia exceeding 1/2 of the vascular territory; extraparenchymal, subarachnoid, epidural, or subdural hemorrhage; moyamoya disease; and cerebral venous sinus thrombosis. Patients were also excluded for having an unknown last-seen-well date or a last-seen-well date more than 24 hrs before presentation. For patients who had hemicraniectomies, data were omitted after a hemicraniectomy had occurred. In the external validation cohort, we also excluded patients for whom life-sustaining treatment was withdrawn within 24 hrs of admission without an interval CT scan and those whose pupillometry data did not meet established criteria. Established pupillometry data criteria included having a minimum of three observations prior to MLS ≥5mm or other evidence of mass effect, no data gaps exceeding 24 hrs, no history of conditions which might effect pupillometry data, and initially reactive and normal pupils. Finally, patients from either cohort were excluded for not having radiographic imaging data after the initial 24 hrs of hospitalization.

Data curation

We used demographic, clinical, and text-based variables at varying time horizons to construct the HELMET models. These variables include time-invariant demographic variables, clinical variables recorded at the time of admission, time-censored dynamic clinical variables changing throughout admission, and radiology reports generated by clinicians at the time of each true scan. The multi-modal structure of the datasets and the operations applied to create the final features are illustrated in panel b of Fig. 2. Further details regarding the curated variables are available in Supplementary Section 2.

Demographic variables included age and sex. Race and ethnicity were not used as inputs to the HELMET models in order to reduce the risk of introducing discriminatory bias into the predictions. Other static admission-related variables included the time of admission, time last seen well, past medical history (including prior stroke, hypertension, and anticoagulant or antiplatelet use), vitals and laboratory blood value readings taken on admission, and the NIHSS score. Static variables from patient medical records accounted for 33 features in the final datasets. Dynamic clinical variables included 16 features, including vital sign data, laboratory

test results, and treatments administered. Features were selected from the patient medical record due to their possible association with malignant edema[1,8,42,43]. Radiographic variables were collected by neurology-specialized team members who labeled the radiographic images at the time of dataset construction. The main variables of interest from the radiographic images were the size of MLS and the size of the pineal gland shift from each scan. We also extracted the time and value of measurements exceeding certain clinically relevant thresholds, such as the first MLS value over 3 mm. A total of 36 human-labeled radiographic variables were included in the dataset (see Supplementary Section 2.3 for further information).

Feature data was aggregated into hourly intervals, and missing values for static variables were imputed using the mean of each feature column, employing the SciKit Learn Simple Imputer[44]. To increase the number of trainable observations per patient, multiple hourly observations were generated, starting from the time of admission and continuing until either discharge or surgical intervention (decompressive hemicraniectomy). Each observation hour was re-indexed based on the number of hours since the patient was last seen well. Static variables were carried forward across all hourly observations, while dynamic variables were assigned to observations corresponding to the specific hour at which they were recorded. A detailed analysis of the cohort's missing data for both datasets is presented in Supplementary Section 3.

Given that each observation was constructed on a single-hour interval, data from previous hours was absent in subsequent observations, and any variables not measured within a particular hour were treated as missing. To enhance model performance, we incorporated historical information into each observation by applying time-based transformations of the dynamic variables. Specifically, time varying features from both electronic medical records and radiographic imaging were carried forward using two methods, resulting in two derived features per original variable: one representing the maximum value over the previous 24 hrs (rolling window), and the other capturing the most recent available measurement (forward-fill). Panel C of Fig. 2 provides an illustration of this process. These transformations yielded 32 features from 16 original medical record variables and 72 features from 36 original radiographic variables. Additionally, a 73rd forward-filled feature was introduced to capture the last known MLS value prior to the most recent measurement, enabling better characterization of the patient's trajectory. The full lists of variables used in training HELMET-8 and HELMET-24 with feature definitions can be found in Supplementary Tables 7 and 8.

Target outcome construction

We chose future MLS as the outcome of interest due to its critical role in determining the severity of ischemic stroke[17,20,45]. We divided the continuous range of MLS values into discrete MLS categories, chosen based on input from collaborating physicians as to the most clinically useful thresholds corresponding roughly to no, mild, moderate, and severe MLS. While MLS ≥5 mm has been previously used to define severe edema[1], more recent research has shown that MLS >3 mm is strongly predictive of poor outcomes[12]. Based on this evidence, 3 mm was set as the first threshold of our edema classes. Preliminary exploration of our dataset revealed that patient MLS typically approached 8 mm before decompressive hemicraniectomy, leading to the choice of 8mm as the upper threshold of the prediction classes. We therefore employed the MLS categories (classes) of no MLS [0 mm], MLS of less than 3 mm (0–3 mm), MLS between 3 mm and 8 mm (3-8mm), and MLS exceeding 8mm (>8mm). To assess the robustness of our framework under a more conventional binary classification target, we also conducted a sensitivity analysis using a 5mm MLS threshold (see Supplementary Section 7).

We defined two prediction targets: the maximum MLS value within the subsequent 8-hr window and the maximum MLS value within the subsequent 24-hr window. These intervals were selected for their clinical relevance, providing a balance between the need for timely diagnostic and therapeutic interventions and the typical cadence of updated clinical data, including laboratory results, imaging studies, and vital signs. The target MLS values were derived as the maximum MLS recorded from radiographic images within the specified prediction window and subsequently mapped to one of four pre-defined MLS classes, thus creating two distinct classification tasks for each observation.

To ensure target validity, we excluded from our analysis observations for which no radiographic scan occurred within the relevant prediction window (e.g., for a 24-hr prediction task, if there was a 40-hr gap between scans, the first 16 hrs post-scan were excluded as valid targets could not be constructed). This approach resulted in the generation of up to eight observations per scan for the 8-hr prediction task and up to 24 observations per scan for the 24-hr task.

Large language model-derived features

Radiology reports, written by radiologists at the time of hospitalization as interpretations of CT scans, offer clinical insights not captured by the quantitative measurements of MLS and pineal gland shift alone. The Clinical-Longformer[46], initially trained on large corpora of clinical text for general medical language modeling, was adapted for multi-class classification using the radiology reports and their corresponding future MLS class labels. This particular pre-trained transformer model was selected based on a review of recent literature[47]. By specifying the intended task of text classification when loading the pre-trained model, a linear classification layer was added at the model head to transform the default output into predictions on the four-class MLS ranges. We fine-tuned three separate text-classification large language models using the derivation dataset to predict the future maximum MLS value within windows of 8 hrs, 24 hrs, and 36 hrs, respectively. Training targets corresponding to the maximum MLS reached over the following 36 hrs were derived using the same process described above for the 8-hr and 24-hr targets, but were only used in training the large language model classifiers. The 36-hr horizon predictions provided more information about long-term MLS trajectory, but were not included for the downstream HELMET models as they were not deemed to be clinically relevant.

These classifiers were trained using raw radiology report texts generated by clinicians at the time of each scan to describe the characteristics and diagnoses associated with a patient's stroke and edema state progression. In order to standardize the reports across hospitals, we cropped the texts to only include the "Findings" and "Impression" sections, which were available in reports from both datasets. Before training, the texts were tokenized using the pre-trained Clinical-Longformer tokenizer[46]. Text data from the derivation set were split by patient into training and test sets, with data from 80% of patients being used for fine-tuning and 20% of patients being reserved for the testing set.

The model's pre-trained weights were updated by minimizing the categorical cross-entropy loss between the predicted class probabilities and the true MLS class labels for each predictive window. The fine-tuning process included six epochs at an initial learning rate of $2*10^{-5}$, using the HuggingFace Transformers package and following previously established methods for the Clinical-Longformer model[47,48]. Fine-tuning was conducted on a Microsoft Azure NC24ads A100 v4 virtual machine using a single Nvidia A100 GPU. The predictive performance of these fine-tuned large language models is reported in Supplementary Section 6.1. By leveraging the pre-trained language model's understanding of clinical terminology and combining it with task-specific data, we enabled the model to effectively learn nuanced patterns in the radiology reports relevant to future MLS prediction.

After transfer learning was complete, the four class probabilities for each of the three large language models (12 total variables) for each radiographic report were then incorporated into the datasets to be used as input features for the downstream ensemble learning models alongside the variables from patient medical records and human-extracted features from radiographic images. We forward-filled any missing hours prior to training of the downstream ensemble learning models.

HELMET model development

To derive HELMET-8 and HELMET-24, multi-class classification models were trained to predict the 8-hr and 24-hr MLS trajectories. By discretizing

continuous MLS values into four predefined classes, the models reframed the regression task into a classification problem, predicting into which of the four MLS ranges a patient's maximum MLS would fall during the specified prediction window.

The HELMET models are trained using the XGBoost algorithm[49], a well-established ensemble learning technique suitable for multi-class classification that leverages tree-based gradient-boosting. We also explored the use of alternative methods, but we opted to use XGBoost as we did not see any significant changes in downstream performance (see Supplementary Section 5.2). We divided the derivation data using five-fold randomized splits. Data were partitioned at the patient level to ensure there was no leakage of observations from the same patient in the training and testing sets of the derivation cohort. Each HELMET model was trained on data from 80% of the patients in the derivation cohort in each of the five splits. Models were then evaluated on data from the remaining 20% of the Massachusetts General Brigham patients for internal validation, and all Boston Medical Center data for external validation. We aggregated performance metrics across all randomized data partitions, allowing us to report confidence intervals around each averaged performance metric.

To prioritize accurate identification of MLS worsening, the algorithm training was based on a modified cross-entropy loss function that gives higher importance to observations with target values different from the immediately preceding MLS class. We refer to these transitioning observations as the "filtered" dataset, with the set of all observations (transitioning and non-transitioning) referred to as the "overall" dataset. The non-transition weight value was tuned as a model hyperparameter. The modified cross-entropy loss function incorporates both filtered and non-filtered observations to account for transitions in the patient's state. Let N represent the total number of observations, and C be the total number of classes. For each observation $i \in \{1, \ldots, N\}$, $y_i \in \{0, 1\}^C$ is the one-hot encoded true label vector, where $y_{i,c} = 1$ if observation i belongs to class c, and $p_{i,c}$ denotes the predicted probability that observation i belongs to class c. Let $\mathcal{F} \subseteq \{1, \ldots, N\}$ represent the set of filtered observations, where a patient's state changes over time, and $\mathcal{NF} = \{1, \ldots, N\} \setminus \mathcal{F}$ denote the non-filtered observations, where no state change occurs. The loss function is given by equation (1), where $w \in (0, 1)$ is a weight applied to reduce the contribution of non-filtered observations. The first term represents the cross-entropy loss for filtered observations, and the second term represents the weighted cross-entropy loss for non-filtered observations. The final values of the weighting term, w, for both models can be found in Supplementary Table 10.

$$L(\mathbf{y}, \mathbf{p}) = -\sum_{i \in \mathcal{F}} \sum_{c=1}^{C} y_{i,c} \log(p_{i,c}) - w \sum_{i \in \mathcal{NF}} \sum_{c=1}^{C} y_{i,c} \log(p_{i,c}), \quad (1)$$

Leveraging the Weights & Biases machine learning training platform[50], we applied Bayesian optimization to fine-tune the hyperparameters of the HELMET models, utilizing filtered AUROC as the optimization objective function[51]. The resulting values of the HELMET models hyperparameters are detailed in Supplementary Section 5.3. We also conducted several sensitivity analyses to test our model structure and development methods, which are explained in further detail in Supplementary Section 5.4. All computational experiments, including model development, validation, and evaluation, were performed using Python 3.11 and the Scikit Learn library[44].

EDEMA baseline model development

To compare HELMET with a baseline, we opted to use linear models akin to the models commonly reported in the existing literature[18,19,45]. Specifically, we developed multinomial regression models that leverage similar independent variables to the EDEMA score[17]. The EDEMA score is a multinomial regression model developed for predicting the adverse event of malignant edema after stroke, leveraging the following variables as input: basal cistern effacement, admission MLS, glucose, previous stroke, and the use of medical thrombolysis or thrombectomy interventions. We approximated the EDEMA score by training multinomial regression models in the derivation and external validation cohorts to predict our target outcomes of interest. By separately training the baseline models to each cohort and prediction task, we derive a different model for each dataset leading to four total baseline models: Massachusetts General Brigham EDEMA-8 and EDEMA-24, as well as Boston Medical Center EDEMA-8 and EDEMA-24 (see Fig. 2). Our baseline models use as input the static features of blood glucose at admission, HbA1C at admission, history of previous stroke, mechanical thrombectomy at any point, and medical thrombolysis using tPA at any point, as well as the dynamic features of most recent blood glucose, presence of basal cistern effacement on the most recent scan, and the MLS measurement from the most recent scan (indicated by asterisks in Table 1). The baseline models were derived using five-fold splits at the patient level for both the derivation cohort and the external validation cohort. For training the EDEMA-8 and EDEMA-24 for Massachusetts General Brigham and the HELMET models, we used the same train-test partitions to maximize comparability.

Model evaluation

The principle target metric of model evaluation during algorithm tuning was filtered area under the receiver operating characteristic curve (AUROC). We selected this criterion given its relatively universal use as a performance indicator[52,53] and its suitability for use in comparison between different model types. To evaluate the predictive performance of the derived models, we report the downstream AUROC, area under the precision-recall curve (AUPRC), accuracy, sensitivity, and specificity of the overall and filtered datasets across both the testing set of the derivation cohort and the external validation dataset. We separately evaluated model performance on the filtered cohort to stress the clinical importance of cases where a patient's MLS class changed. All performance metrics are reported as the mean and the corresponding 95% confidence intervals for the testing sets of the derivation cohort and the external validation cohort.

AUROC and AUPRC are typically defined only for binary classification models. Given that our targets belong to a four-category classification task, we calculated the true positive rate, false positive rate, sensitivity (recall), and precision scores in a one-versus-rest binary classification setup across prediction thresholds ranging between zero and one[44,54]. The resulting metrics were then averaged across the four classes at each given threshold to derive the receiver operating characteristic and precision-recall curves weighted by the number of observations in each class. Accuracy was defined as the proportion of observations where the model correctly assigned the highest predicted probability to the true target class. To adapt sensitivity and specificity for the multi-class setting, we reformulated the task to focus on correctly predicting whether a patient's MLS state would worsen (i.e., increase) within the prediction window. For this binary classification task, true labels were assigned a value of one if the future MLS class exceeded the current MLS class, and zero otherwise. Predicted values were similarly assigned a value of one if the MLS class with the highest predicted probability was greater than the current class, and zero if it was not. Sensitivity and specificity were then computed according to their standard definitions.

Feature importance analysis

We utilized Shapley Additive Explanation analysis to study the relative importance of various features in our final ensemble learning models[55]. The Shapley values measure the marginal contribution of each feature to a prediction in a machine learning model by decomposing the prediction into the sum of effects from each feature, utilizing the principles of cooperative game theory. For each class prediction task, Shapley values were computed to identify the features most influential in predicting across the four MLS classes. Per-class Shapley values were aggregated to generate a ranked list of overall feature importance. While the absolute Shapley values are not directly informative, the relative ranking and composition of top features offer valuable insights into model behavior and potential implications for future clinical practice and edema prediction.

Data availability

The data used in our study come from two academic medical centers in the United States subject to the Health Insurance Portability and Accountability Act. Due to the data use agreement that we have signed with Boston Medical Center and Massachusetts General Brigham, the datasets cannot be made publicly accessible, as they contain protected health information and other sensitive information about the patients. Any user that wishes to gain access to the dataset needs to become HIPAA certified and get approved as an authorized by the collaborating healthcare systems Institutional Review Boards.

Code availability

The code developed for this study will become publicly available via GitHub upon publication of the manuscript by a peer-reviewed journal. Model development and testing were done using Python v3.11 and publicly available packages (including scikit-learn, pandas, numpy, wandb, xgboost, datasets, transformers, evaluate, torch, cupy, and fire).

Received: 12 November 2024; Accepted: 29 April 2025;
Published online: 17 May 2025

References

1. Wijdicks, E. F. et al. Recommendations for the management of cerebral and cerebellar infarction with swelling: a statement for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke* **45**, 1222–1238 (2014).
2. Kimberly, W. T. et al. Association of reperfusion with brain edema in patients with acute ischemic stroke: A secondary analysis of the mr clean trial. *JAMA Neurol.* **75**, 453–461 (2018).
3. Ropper, A. H. Brain edema after stroke: clinical syndrome and intracranial pressure. *Arch. Neurol.* **41**, 26–29 (1984).
4. Ayata, C. & Ropper, A. Ischaemic brain oedema. *J. Clin. Neurosci.* **9**, 113–124 (2002).
5. Huttner, H. & Schwab, S. Malignant middle cerebral artery infarction: clinical characteristics, treatment strategies, and future perspectives. *Lancet Neurol.* **8**, 949–958 (2009).
6. Pullicino, P. et al. Mass effect and death from severe acute stroke. *Neurology* **49**, 1090–1095 (1997).
7. Hacke, W. et al. 'malignant'middle cerebral artery territory infarction: clinical course and prognostic signs. *Arch. Neurol.* **53**, 309–315 (1996).
8. van der Worp, H. B. et al. European Stroke Organisation (ESO) guidelines on the management of space-occupying brain infarction. *Eur. Stroke J.* **6**, XC–CX (2021).
9. Greige, T. et al. Cerebral edema monitoring and management strategies: Results from an international practice survey. *Neuroc. Care* (2024).
10. Hays, A. et al. Osmotherapy: use among neurointensivists. *Neurocr. Care* **14**, 222–8 (2011).
11. Barber, P. A. et al. Computed tomographic parameters predicting fatal outcome in large middle cerebral artery infarction. *Cerebrovasc. Dis.* **16**, 230–235 (2003).
12. McKeown, M. E. et al. Midline shift greater than 3 mm independently predicts outcome after ischemic stroke. *Neurocr. Care* 1–6 (2022).
13. Pullicino, P. M. et al. Mass effect and death from severe acute stroke. *Neurology* **49**, 1090–1095 (1997).
14. Waydhas, C. Intrahospital transport of critically ill patients. *Crit. Care* **3**, R83–R89 (1999).
15. Mettler, F. J., Huda, W., Yoshizumi, T. & Mahesh, M. Effective doses in radiology and diagnostic nuclear medicine: a catalog. *Radiology* **248**, 254–63 (2008).
16. Ong, C., Chatzidakis, S., Ong, J. & Feske, S. Updates in management of large hemispheric infarct. *Semin. Neurol.* **44**, 281–297 (2024).
17. Ong, C. J. et al. Enhanced detection of edema in malignant anterior circulation stroke (edema) score: A risk prediction tool. *Stroke* **48**, 1969–1972 (2017).
18. Cheng, Y. et al. External validation and modification of the edema score for predicting malignant brain edema after acute ischemic stroke. *Neuroc. Care* **32**, 104–112 (2020).
19. Shimoyama, T. et al. The dash score: a simple score to assess risk for development of malignant middle cerebral artery infarction. *J. Neurol. Sci.* **338**, 102–106 (2014).
20. Wu, S. et al. Predicting the emergence of malignant brain oedema in acute ischaemic stroke: a prospective multicentre study with development and validation of predictive modelling. *Eclinicalmedicine* **59** (2023).
21. Stafford, R. et al. Follow-up aspects improves prediction of potentially lethal malignant edema in patients with large middle cerebral artery stroke. *J. NeuroIntervent. Surg.* (2023).
22. Powers, W. et al. Guidelines for the early management of patients with acute ischemic stroke: 2019 update to the 2018 guidelines for the early management of acute ischemic stroke: A guideline for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke* **50**, e344–3418 (2019).
23. Prescott, H. et al. Development and validation of the hospital medicine safety sepsis intitiative mortality model. *Chest* (2024).
24. Gallo, R. et al. Effectiveness of an artificial intelligence-enables intervention for detecting clinical deterioration. *JAMA Intern. Med.* **184**, 557–562 (2024).
25. Ong, C. et al. Association of dynamic trajectories of time-series data and life-threatening mass effect in large middle cerebral artery stroke. *Neuroc.Care* **42**, 77–89 (2025).
26. Pohlman, J. et al. Association of large core middle cerebral artery stroke and hemorrhagic transformaiton with hospitalization outcomes. *Nat. Sci. Rep.* **14** (2024).
27. Kline, A. et al. Multimodal machine learning in precision health: A scoping review. *Nat. Digit. Med.* **5** (2022).
28. Yang, J., Soltan, A. A. & Clifton, D. A. Machine learning generalizability across healthcare settings: insights from multi-site covid-19 screening. *NPJ Digital Med.* **5**, 69 (2022).
29. Hanko, M. et al. Random forest-based prediction of outcome and mortality in patients with traumatic brain injury undergoing primary decompressive craniectomy. *World Neurosurg.* **148**, e450–e458 (2021).
30. Soenksen, L. R. et al. Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ Digital Med.* **5**, 149 (2022).
31. Li, J. et al. Integrated image-based deep learning and language models for primary diabetes care. *Nat. Med.* 1–11 (2024).
32. Wijdicks, E. F. & Diringer, M. N. Middle cerebral artery territory infarction and early brain swelling: progression and effect of age on outcome. *Mayo Clin. Proc.* **73**, 829–836 (1998).
33. Muscari, A. et al. Predicting cerebral edema in ischemic stroke patients. *Neurol. Sci.* **40**, 745–752 (2019).
34. Schwab, S., Aschoff, A., Spranger, M., Albert, F. & Hacke, W. The value of intracranial pressure monitoring in acute hemispheric stroke. *Neurology* **47**, 393–398 (1996).
35. Orfanoudaki, A., Saghafian, S., Song, K., Chakkera, H. A. & Cook, C. Algorithm, human, or the centaur: How to enhance clinical care? *SSRN working paper: papers.ssrn.com/sol3/papers.cfm?abstract_id=4302002* (2022).
36. McLaughlin, B. & Spiess, J. Designing algorithmic recommendations to achieve human-ai complementarity. *arXiv preprint arXiv:2405.01484* (2024).
37. Cui, L. et al. Deep learning in ischemic stroke imaging analysis: A comprehensive review. *BioMed Res. Int.* (2022).
38. Collins, G. S. et al. Tripod+ai statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **385**, https://www.bmj.com/content/385/bmj-2023-078378. https://www.bmj.com/content/385/bmj-2023-078378.full.pdf (2024).

39. Luo, W. et al. Guildines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. *J. Med. Internet Res.* **18**, e323 (2016).
40. Ong, C. J. et al. Machine learning and natural language processing methods to identify ischemic stroke, acuity and location from radiology reports. *PLoS ONE* **15**, e0234908 (2020).
41. Miller, M. I. et al. Natural language processing of radiology reports to detect complications of ischemic stroke. *Neuroc. Care* **37**, 291–302 (2022).
42. Cook, A. M. et al. Guidelines for the acute treatment of cerebral edema in neurocritical care patients. *Neuroc. Care* **32**, 647–666 (2020).
43. Jauch, E. C. et al. Guidelines for the early management of patients with acute ischemic stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke* **44**, 870–947 (2013).
44. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
45. Wu, S. & Anderson, C. S. Precision management of brain oedema after acute ischaemic stroke. *Precis. Clin. Med.* **5**, pbac019 (2022).
46. Li, Y., Wehbe, R. M., Ahmad, F. S., Wang, H. & Luo, Y. A comparative study of pretrained language models for long clinical text. *J. Am. Med. Inform. Assoc.* **30**, 340–347 (2023).
47. Villalobos Carballo, K. et al. Tabtext: A flexible and contextual approach to tabular data representation. *arxiv preprint arXiv:2206.10381* (2022).
48. Wolf, T. et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45 (2020).
49. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794 (2016).
50. Biewald, L. Experiment tracking with weights and biases (2020). https://www.wandb.com/. Software available from wandb.com.
51. Snoek, J., Larochelle, H. & Adams, R. P. Practical Bayesian optimization of machine learning algorithms. *Adv. Neural Inf. Process. Syst.* **25** (2012).
52. Mandrekar, J. N. Receiver operating characteristic curve in diagnostic test assessment. *J. Thorac. Oncol.* **5**, 1315–1316 (2010).
53. Walter, S. Properties of the summary receiver operating characteristic (sroc) curve for diagnostic test data. *Stat. Med.* **21**, 1237–1256 (2002).
54. Su, Q. et al. Faecal microbiome-based machine learning for multi-class disease diagnosis. *Nat. Commun.* **13**, 6818 (2022).
55. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural. Inf. Process. Syst.* **30** (2017).

Acknowledgements

We wish to acknowledge the assistance provided by staff at the Massachusetts General Brigham and Boston Medical Center hospital systems, as well as the technical and computing support received from staff at the University of Oxford Saïd Business School. This work was funded by the NIH/NINDS through CJO's K23NS116033 award. No other authors received external funding for this study.

Author contributions

E.P. co-led model development and evaluation, led the data analysis, and drafted and edited the manuscript. O.O.D. co-led initial model development and evaluation and assisted with the manuscript revision. Y.Z. organized results and figures and revised the manuscript. P.T. proposed ideas and worked on the theoretical underpinnings of this paper. L.A.M. participated in data collection and curation and assisted with manuscript revision. J.P. participated in data collection, dataset creation, and data adjudication. S.C. participated in data collection and adjudication, dataset creation, and manuscript revision. Y.D. participated in data collection, dataset creation, and manuscript revision. I.K. participated in data collection, dataset creation, and data adjudication. J.S. participated in data collection, dataset creation, and data adjudication. B.B. participated in dataset creation, data adjudication, and manuscript revision. S.S. co-led study design and scope; provided clinical insight; and participated in manuscript review. C.J.O. co-led study design, data collection, data curation, and interpretation and provided clinical insight. C.J.O contributed to interpreting the findings and drawing conclusions about the model's effectiveness. C.J.O. actively participated in manuscript drafting and review. A.O. co-led the study design and coordination and guided the data preprocessing and the machine learning model development and training. A.O. contributed to interpreting the findings and drawing conclusions about the model's effectiveness. A.O. actively participated in and led the manuscript writing and revision process.

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/s41746-025-01687-y.

Correspondence and requests for materials should be addressed to Charlene J. Ong or Agni Orfanoudaki.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2025