

Explainable Artificial Intelligence for Prehospital Identification of Stroke Patients Likely to Receive Endovascular Thrombectomy: A Retrospective Real-World Study

Patrick Andreas Eder

eder@ztm.de

Zentrum für Telemedizin Bad Kissingen

Oliver Moser

Zentrum für Telemedizin Bad Kissingen

Asarnusch Rashid

Zentrum für Telemedizin Bad Kissingen

Eckhard Schlemm

University Medical Center Hamburg-Eppendorf

Jasmin Hessler

Zentrum für Telemedizin Bad Kissingen

Holger von Jouanne-Diedrich

TH Aschaffenburg (University of Applied Sciences)

Volker Ziegler

Campus Rhön Klinikum AG

Hassan Soda

Campus Rhön Klinikum AG

Article

Keywords: Telemedicine, Acute Ischemic Stroke, Emergency Medical Services, Machine Learning, Thrombectomy

Posted Date: August 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10095424/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Acute ischaemic stroke due to large-vessel occlusion requires rapid endovascular treatment (EVT). Prehospital stroke scales, though widely used, have not achieved an optimal balance of sensitivity, specificity, and overall discriminative performance, underlining the need for evaluation of machine learning (ML) approaches to improve early identification of EVT candidates. We conducted a retrospective development and internal validation study based on digital real-world data from consecutive EMS-transported patients with suspected stroke ($n = 1,435$, 94 EVT events [6.6%]) from the Stroke Angel initiative in a rural region in Germany (Rhön-Grabfeld), covering the period from January 2015 to June 2021. Predicted probabilities of six ML models were recalibrated using Platt scaling, and clinical relevance was evaluated using decision curve analysis (DCA). SHapley Additive exPlanations (SHAP) values were computed to quantify individual feature contributions to model predictions. Among 287 patients (24 EVT events) in the evaluation cohort, Random Forest achieved the highest AUROC (0.85 [95% CI, 0.78–0.90]), with sensitivity of 0.92 [0.74–0.99], specificity of 0.66 [0.60–0.72]. The 4I-SS (score > 2) yielded an AUROC of 0.82 [0.72–0.89], sensitivity of 0.87 [0.67–0.96], and specificity of 0.69 [0.64–0.74]. Bootstrapped intervals overlapped across all metrics, suggesting no statistically significant superiority.

Introduction

Stroke is a major global health burden and ranks as the second leading cause of death worldwide [1]. Endovascular thrombectomy (EVT) is the standard of care for acute ischemic stroke with large-vessel occlusion (LVO), supported by high-quality randomized evidence and exceptionally low numbers needed to treat [2]. Acute stroke requires immediate identification of patients and their rapid transfer to comprehensive stroke centres for EVT [3]. However, in many stroke systems, stroke assessment does not commence until after hospital arrival, which slows down treatment. Delays could be avoided if emergency medical service (EMS) teams recognized LVO and its severity on scene, prioritize patients instantly, and route them directly to the centre that offers the fastest access to reperfusion therapy [4, 5].

Since only a minority of suspected stroke cases ultimately harbour a proximal LVO, routing all patients directly to a comprehensive stroke (“mothership”) centre may not always be practical or necessary, emphasizing the significance of effective prehospital allocation. Hospitals without EVT capability must expeditiously transfer such patients fast—ideally while thrombolysis continues—to the nearest comprehensive stroke centre with EVT capacity (“drip-and-ship”) [6]. The avoidance of treatment delays necessitates formalized agreements and predefined protocols between the referring hospital, the designated EVT centre, and the EMS [7].

EVT is restricted to a highly selected subset of patients with acute ischemic stroke. In response, severity-based triage strategies commonly use a National Institutes of Health Stroke Scale threshold of ≥ 7 to increase the efficiency of patient selection for advanced imaging and endovascular pathways [2]. Conventional prehospital scales have been designed to detect LVO, yet this approach neglects the

broader determinants of EVT [8, 9]. Prehospital stroke scales, though widely used, have not achieved an optimal balance of sensitivity, specificity, and predictive value. Even prehospital stroke scales computed ex ante demonstrate performance comparable to manually derived scales and do not confer superior LVO-detection [10]. The decision to pursue EVT depends not only on vessel status but also on patient-specific and contextual factors that may not be captured by conventional scales.

Collecting structured prehospital data and reliably linking them to hospital-based outcome measures is challenging due to heterogeneous documentation workflows and limited interoperability across emergency and hospital information systems [11].

Machine-learning (ML) approaches, by integrating multimodal prehospital and hospital data, has the potential to refine prehospital stroke triage, reduce transfers, and increase the likelihood that patients who stand to benefit from EVT reach comprehensive stroke centres faster [12]. These predictions may support prehospital decision-making and better inform paramedics or emergency physicians. Although real-world data (RWD) has been identified as a potential source of ML model development, its utilisation in this field remains scarce [13].

The primary objective is to develop and internally validate ML models that estimate the likelihood of a patient undergoing EVT during the acute hospital admission based on prehospital and early in-hospital RWD. The secondary objective is to assess the calibration of predicted probabilities and to evaluate the impact of classification thresholds on prehospital triage decisions, particularly the balance between over- and undertriage. The models are embedded within an explainable artificial intelligence (XAI) framework to enable interpretation of model outputs and decision logic. This study is designed as a proof of concept for data-driven prehospital EVT triage and aims to explore the potential clinical utility of an explainable decision-support approach.

Results

A total of 3,628 records from Database A (prehospital) and 5,975 records from Database B (hospital) were extracted. Record linkage identified 1,464 matched records (25.4% of Database B). Of these, 29 duplicate index entries were removed, yielding 1,435 unique linked patients comprising the preprocessed data set. Transport mode and admission pathway inclusion criteria were applied prior to linkage to ensure cohort comparability. The final cohort was randomly partitioned into training (n = 918), validation (n = 230), and test sets (n = 287). Among included patients, 94 (6.6%) underwent EVT and 356 (24.8%) received IVT. With 1,148 patients and 70 EVT events, this cohort fell short of the calculated target, increasing the risk of overfitting. To mitigate this, we tuned hyperparameters to constrain model flexibility toward simpler architectures. Baseline characteristics are provided in Table 1.

Table 1 summarizes the performance metrics for all evaluated ML classification approaches, including the 4I-SS reference scale across multiple dimensions of model evaluation.

Figure 2. **ROC and PR curves for identifying EVT candidates within the evaluation cohort.** **a** Panel A shows the receiver-operating-characteristic (ROC) curve with the area under the curve (AUROC) for all machine learning models. The ROC plot displays curves: the 4I-SS reference (yellow) and ML-approaches (shown in distinct colors). The x-axis shows the false-positive rate ($1 - \text{specificity}$, in %), and the y-axis shows sensitivity (true-positive rate, in %). **b** Panel B depicts the corresponding precision–recall curves, with the area under the precision–recall curve (AUPRC) reported for each machine learning model. The x-axis shows recall (sensitivity, in %), and the y-axis shows precision (positive predictive value, in %).

Precision–recall analysis yielded uniformly low AUCPR values across classifiers, consistent with the low EVT prevalence (6.6%). RF and the neural network achieved the highest AUCPR (0.32 and 0.34; 95% CI 0.19–0.51 and 0.17–0.50), though both remained modest with wide uncertainty bounds (Fig. 2). Using a predefined threshold of 0.25, ensemble models demonstrated comparatively stable sensitivity–specificity profiles across CV folds, whereas simpler classifiers showed substantial fold-wise sensitivity variance. Despite this relative robustness, overall discriminative performance remained limited, reflecting the inherent constraints imposed by class imbalance and sparse positive cases.

Figure 3 displays global discrimination and performance under class imbalance, respectively. In the holdout validation cohort, ML classifiers showed moderate-to-high performance for identifying EVT candidates. RF achieved the highest AUROC (0.85, 95% CI 0.78–0.90; NCV 0.80 ± 0.047) and F1 score (0.66, 95% CI 0.60–0.72), with a Brier score of 0.068 (95% CI 0.047–0.096), while other ensemble and neural network models had overlapping CI (AUROC 0.82–0.83, NCV 0.79–0.81; F1 0.63–0.65; Brier 0.068–0.069), indicating no statistically significant differences in overall discrimination. Sensitivity ranged from 0.75 to 0.92, specificity from 0.63 to 0.77 and NPV were consistently high, and the 4I-SS showed comparable performance (AUROC 0.82, 95% CI 0.72–0.89; NCV 0.79 ± 0.033 ; F1 0.69; Brier 0.069), suggesting similar predictive ability across several ML approaches without definitive evidence of superiority. Across NCV folds, the 4I-SS showed the lowest standard deviation (SD) of 0.033, indicating highly consistent performance. Among ML classifiers, RF and XGBoost exhibited relatively low variability, while CART, kNN, Neural Network, and OneR showed higher SDs, suggesting that fold-to-fold stability was generally better for ensemble methods than for other models.

Figure 3 **SHapley Additive exPlanations (SHAP) summary plots illustrating feature contributions to the predictive model.** Features are ranked in order of importance from top to bottom. **a** SHAP summary beeswarm plot ranking features by mean absolute SHAP value. The x-axis represents the SHAP impact on model output, while color denotes normalized feature values. Red values shift predictions toward endovascular thrombectomy, whereas blue values shift toward absence. **b** Top-10 attributes ranked by SHAP values, driving the output. **c** SHAP summary beeswarm plots comparing feature rankings across all six machine learning algorithms, based on SHAP values, to illustrate consistency and variability in predictor influence.

The SHAP analysis across ML-models demonstrates that focal neurological deficits, especially hemiparesis, are the most influential predictors of model output. Cortical signs, such as aphasia,

dysarthria, anomia and brainstem signs, including reduced vigilance, emerged as strong determinants of EVT. Age contributes moderately, with evidence of nonlinear amplification in older patients, while gender exerts negligible influence. RF, XGBoost display sharper threshold effects, whereas the neural network produces smoother gradients and kNN yields more diffuse attributions. The overall concordance across models reinforces the robustness of neurological signs and time criteria as dominant predictors.

Among 287 (100%) test cases, the RF model predicted EVT candidacy with 111 (38,7%) positive and 176 (61,3%) negative classifications (see Fig. 4). Of the 24 (8,4%) patients who ultimately received EVT, 22 (7,7%) were correctly identified (true positives), while 2 (0,7%) were missed (false negatives). The RF model yielded 89 (31,0%) false positives and 174 (60,6%) true negatives. In comparison, the 4I-SS scale predicted EVT in 103 (35,9%) cases and no EVT in 184 (64,1%). It correctly identified 21 (7,3%) true positives and missed 3 (1,0%) EVT cases (false negatives) but produced 82 (28,6%) false positives and 181 (63,1%) true negatives.

DCA showed that all ML models and the 4I-SS provided a positive net benefit at low threshold probabilities, where minimizing undertriage is prioritized (Supplementary Figure S1). At these lower thresholds (up to approximately 12%), kNN and RF tended to yield the highest net benefit, though differences between methods were modest. With increasing threshold probabilities, net benefit declined across all methods, with all ML models converging toward a non-intervention strategy, while the 4I-SS retained a small positive net benefit over a wider range of thresholds. Once the threshold probability exceeded approximately 35% – which clinically translates to accepting fewer than approximately two false-positive mothership transports for one true-positive EVT patient – all strategies lost their clinical utility. Calibration curves revealed that most models overestimated event probabilities, an expected consequence of the resampling applied during preprocessing. Platt scaling effectively corrected this miscalibration, aligning predicted probabilities more closely with observed outcomes (Supplementary Figure S2).

Table 1
Baseline characteristics of patients in development and validation cohorts

n, (%)	Unmatched Datasets	Matched Dataset	Model Development Dataset	Model Evaluation Dataset
	Prehospital n = 3,628 Hospital n = 5,975	1,435 (100%)	1,148 (80%)	287 (20%)
Prehospital Variables				
Age, median [IQR]	78.0 [67.0– 84.0] NA: 25 (0)	79.0 [68.5– 84.0] NA: 0 (0)	79.0 [69.0–84.0] NA:0 (0)	79.0 [68.0– 84.0] NA:0 (0)
Systolic Blood Pressure, median [IQR], min-max	153 [135– 177] 10–262 NA: 165 (5)	160 [140– 180] 84-250 84-250 NA: 55 (4)	160 [140–180] 84-250 NA: 0 (0)	160 [142–180] 95-230 NA: 0 (0)
Diastolic Blood Pressure	85 [74–96] 0-192 NA: 574 (16)	87 [78– 98] 39-142 NA: 245 (17)	86 [77–98] 39-142 NA: 0 (0)	90 [80–99] 44-135 NA: 0 (0)
4I-ISS on scene, n (%)	420 (12)	139 (10)	114 (10)	25 (9)
• 0	919 (25)	450 (31)	369 (32)	81 (28)
• 1	692 (19)	398 (28)	319 (28)	79 (28)
• 2	401 (11)	210 (15)	161 (14)	49 (17)
• 3	214 (6)	126 (9)	99 (9)	27 (9)
• 4	127 (4)	64 (4)	53 (5)	11 (4)
• 5	64 (2)	41 (3)	28 (2)	13 (5)
• 6	13 (0)	7 (0)	5 (0)	2 (1)
• 7	NA: 778 (21)	NA: 299 (21)	NA:0 (0)	NA:0 (0)

n, (%)	Unmatched Datasets	Matched Dataset	Model Development Dataset	Model Evaluation Dataset
	Prehospital n = 3,628	1,435 (100%)	1,148 (80%)	287 (20%)
	Hospital n = 5,975			
Hospital Variables				
Age, median [IQR]	78.0 [66.0–84.0] NA: 84 (1)	79.0 [68.8–84.0] NA: 0 (0)	Dropped clinical age after record linkage	Dropped clinical age after record linkage
Female, n (%)	2,101 (48) NA: 558 (27)	703 (49) NA: 169 (24)	565 (49) NA:0 (0)	138 (48) NA:0 (0)
NIHSS on admission, n (%)	3924 (66)	717 (50)	588 (51)	129 (45)
• 0–4	1543 (26)	550 (38)	429 (37)	121 (42)
• 5–15	508 (8)	68 (12)	131 (11)	37 (13)
• > 15	NA: 0 (0)	NA: 0 (0)	NA:0 (0)	NA:0 (0)
Way of admission, n (%)	1,596 (27)	619 (43)	506 (44)	113 (39)
• Paramedic/ ambulance	1,779 (30)	816 (57)	642 (56)	174 (61)
• Paramedic/ ambulance with Emergency Physician (rendezvous)	698 (12) 1,886 (31)	0 (0) 0 (0)	0 (0) 0 (0)	0 (0) 0 (0)
• Walkin Emergencies	NA: 16 (0)	NA: 0 (0)	NA:0 (0)	NA:0 (0)
• Other transport				
Transport mode, n (%)	2,616 (44)	1205 (84)	957 (83)	248 (86)
• emergency services (primary transport)	1,027 (17) 1,105 (18)	95 (7) 135 (9)	77 (7) 114 (10)	18 (6) 21 (7)
• transfer from another facility (secondary transport)	513 (9)	0 (0)	0 (0)	0 (0)
• referral initiated in primary care	698 (12) NA: 16 (0)	0 (0) NA: 0 (0)	0 (0) NA: 0 (0)	0 (0) NA: 0 (0)
• In-hospital referral				

n, (%)	Unmatched Datasets	Matched Dataset	Model Development Dataset	Model Evaluation Dataset
	Prehospital n = 3,628	1,435 (100%)	1,148 (80%)	287 (20%)
	Hospital n = 5,975			
• Walkin Emergencies				
Years on admission, n (%)	871 (15)	244 (17)	201 (18)	44 (15)
• 2015	883 (15)	246 (17)	192 (17)	54 (19)
• 2016	814 (14)	234 (16)	201 (18)	33 (11)
• 2017	859 (14)	158 (11)	121 (11)	37 (13)
• 2018	1091(18)	233 (16)	182 (16)	51 (18)
• 2019	981(16)	208 (15)	160 (14)	48 (17)
• 2020	474 (8)	112 (8)	91 (8)	20 (7)
• 2021				
Discharge diagnosis (ICD-10), n (%)	3,379 (56)	1011 (70)	803 (70)	208 (72)
	274 (5)	73 (5)	58 (5)	15 (5)
• Stroke (ischemic/unspecified)	51 (0)	1 (0)	1 (0)	0 (0)
• ICB (Intracerebral hemorrhage)	9 (0)	1 (0)	1 (0)	0 (0)
	892 (15)	110 (8)	85 (7)	25 (9)
• SAB (Subarachnoid hemorrhage)	87 (1)	14 (1)	13 (1)	1 (0)
• SDH (Subdural hematoma)	994 (18)	217 (15)	181 (16)	36 (13)
• Stroke mimic	210 (5)	8 (1)	6 (1)	2 (1)
• TGA (Transient global amnesia)	NA:79 (1)	NA:10 (0)	NA:0 (0)	NA:0 (0)
• TIA (Transient ischemic attack)				
• Traumatic brain injury				

n, (%)	Unmatched Datasets	Matched Dataset	Model Development Dataset	Model Evaluation Dataset
	Prehospital n = 3,628	1,435 (100%)	1,148 (80%)	287 (20%)
	Hospital n = 5,975			
IVT, n (%)	723 (12)	356 (25)	289 (25)	67 (23)
ICD-10 Stroke, n (%)	691 (96)	348 (98)	283 (98)	65 (97)
NA	NA: 0 (0)	NA: 0 (0)	NA: 0 (0)	NA: 0 (0)
EVT, n (%)	202 (3)	94 (7)	70 (6)	24 (8)
ICD-10 Stroke, n (%)	200 (100)	94 (100)	70 (100)	24 (100)
NA	NA: 0 (0)	NA: 0 (0)	NA: 0 (0)	NA: 0 (0)

Baseline characteristics of unmatched EMS and clinical data, matched data, as well as those allocated to the training set for development and test set for evaluating the models. Prehospital variables include age, systolic and diastolic blood pressure, and 4I-ISS assessed on scene. Hospital variables include sex, NIHSS on admission, mode of admission and transport, year of admission, discharge diagnosis according to ICD-10 classification, and information on IVT and EVT. Continuous variables are reported as median [IQR]. Categorical variables are presented as counts (percentages). Missing data are indicated as NA: n (%). The table shows that development and evaluation sets demonstrate similar characteristics in relevant variables. These variables serve primarily for descriptive cohort characterization and assessment of dataset balance. The majority were not utilized as input features in the machine learning modeling pipeline. Due to rounding, percentages may not sum to 100%.

Discussion

In this study of consecutive stroke patients at a large, over-regional stroke centre in Germany, we developed and internally validated six ML models to predict the likelihood of EVT using RWD variables available in the prehospital setting. Across all models, discrimination and calibration were comparable to the prehospital LVO scale, and DCA demonstrated similar clinical utility across a range of plausible thresholds. These findings indicate that the predictive signal in routinely collected prehospital and clinical variables is largely captured by existing triage scales, and that complex ML approaches may not confer substantial incremental performance in this context. Overall, the ML models performed comparably to the prehospital stroke scale reference. Although higher mean values were observed for some models, the BCa bootstrap confidence intervals and NCV variability did not support statistically significant differences. The wide confidence intervals reflect the limited number of EVT events in the

evaluation cohort (n = 24), as positive-class-dependent metrics such as sensitivity, PPV, and AUPRC require enough positive cases for stable estimation.

Consequently, clinical implications derived from these results remain constrained by this uncertainty. The low PPV, predominantly driven by the low EVT prevalence inherent to this classification task and consistent with other rare-event prediction studies, highlights the persistent challenge of overtriage in imbalanced RWD, irrespective of modelling strategy. Broad F1 distributions further reflect the difficulty of achieving a stable balance between sensitivity and precision under these conditions. This study reflects clinical practice between 2015 and 2021, a period defined by the early EVT trials (MR CLEAN, ESCAPE, EXTEND-IA, SWIFT PRIME, and REVASCAT) which established thrombectomy for proximal anterior circulation occlusions within 6 hours [14].

Patient selection was intentionally narrow, multimodal imaging was not yet routine, and device technology was less advanced. Although DAWN and DEFUSE-3 expanded the treatment window to 24 hours, widespread adoption of perfusion-based selection and automated imaging software occurred only gradually during the study period [15]. Consequently, the lower EVT rate observed in our cohort accurately represents the treatment landscape of that era rather than a methodological limitation. The subsequent expansion of indications, improvements in workflow, and advances in device technology, further supported by recent trials, have substantially increased EVT eligibility in contemporary practice. The relevance of our findings is therefore not diminished by these developments. Instead, the temporal context provides important interpretive clarity and offer a benchmark for understanding how evolving guidelines and technologies have reshaped stroke care. By capturing outcomes before the widespread implementation of modern imaging and expanded EVT criteria, this study contributes essential historical perspective for evaluating current and future treatment strategies.

In Germany, prehospital patient routing and allocation is largely coordinated through digital bed-capacity and hospital-availability platforms. These systems rely on simplified allocation codes that combine administrative parameters - age, priority code and a short diagnosis or dispatch category - to match patients with hospital reporting available capacity [16]. While moderately efficient for logical coordination, such codes may represent a coarse abstraction of clinical reality and could fail to capture individual disease severity or therapeutic urgency. Consequently, patient assignment may be driven more by capacity management than by clinical need, a limitation particularly critical in acute stroke, where minutes determine outcomes [17]. By contrast, our ML approach leverages individual clinical and physiological features. Embedding such predictive tools within existing EMS dispatch and documentation systems could substantially improve routing precision and reduce treatment delays.

Given the severe consequences of delayed reperfusion, we set model thresholds to minimize under-triage and ensure patient safety, thereby accepting the resulting increase in over-triage as a necessary clinical trade-off. Beyond such binary transport recommendations, the predicted probabilities generated by ML models can further inform clinical decision-making [18]. Across different stroke networks and patient encounters, this enables more nuanced decisions by combining predicted probabilities with other

relevant factors, such as distance to the nearest stroke unit, hospital capacity, or whether transport to a comprehensive stroke centre would place the patient outside the therapy time window.

Mandatory prehospital teleconsultation, repeated checklisting and advanced examinations in borderline cases could be safeguarded against misclassification while preserving efficient resource allocation [19]. Yet it remains limited by these concerns, remote human expertise, delivered via teleconsultation, brings contextual judgement accountability, and the ability to weigh subtle clinical cues, but is constrained by availability and response time. Model attributions are dominated by focal neurological deficits and time-sensitive eligibility variables, with secondary modulation by vital signs and age. This hierarchy mirrors established clinical heuristics for acute stroke, suggesting that algorithmic outputs are aligned with what clinicians already prioritize under time pressure. The interpretability afforded by SHAP, supports transparent integration into prehospital workflows without obscuring clinical judgment. These properties are germane to early adoption, where trust, auditability, and operational clarity are prerequisites for impact [20]. The prominence of these variables suggests that the models are not merely fitting statistical noise but are recapitulating the clinical logic underpinning current guidelines: severe, disabling deficits in the presence of a favourable time window and absence of exclusion criteria.

Prehospital digital documentation is the substrate on which our models operate, and the attribution patterns argue for standardization of neurologic examination fields and precise capture of time of onset. Embedding structured inputs for aphasia/dysarthria, hemiparesis, gaze–head deviation, vigilance level, and pupil reactivity—as mandatory fields with clear definitions—would improve signal quality and reduce missingness. Importantly, missing values in prehospital datasets may not be interpreted as random noise. In many cases, ‘not documented’ may function as an informative proxy, reflecting operational constraints or clinical urgency—for example, when severely affected patients limit the time available for structured data entry, or when uncertainty arises during documentation shortly before hospital arrival.

This interpretation is consistent with the findings of Oh et al. [21], who demonstrated that patterns of missingness in prehospital data can themselves carry predictive information and should be treated as potentially informative rather than ignorable. Time of onset and exclusion flags should be recorded with enforced formats (e.g., verified last-known-well, witnessed onset) to minimize ambiguity. High-fidelity documentation of systolic/diastolic blood pressure and pulse, while secondary, could enable threshold-aware risk adjustment and supports discrimination across diverse acute neurological states. Multiclass models, in contrast to binary classification, would assign probabilities to each differential diagnosis. This probabilistic information could provide EMS personnel with a data-driven foundation for advanced clinical decision making.

One key implication is that systematic interoperability between federated data of the German National Emergency Department Data Registry [22] and the Stroke Angel Initiative could establish a governance-compliant, scalable RWD ecosystem (e.g., HL7/FHIR-based interfaces), enabling harmonized data integration, robust multicentre analytics, and continuous validation and refinement of AI-supported stroke care, e.g., external validation and model updating. While prehospital digital documentation has

been anchored in the physician's statutory duty to document care as required by Section 630f of the German Civil Code (Bürgerliches Gesetzbuch, BGB) or on federal state regulations such as Article 46 Bavarian Emergency Medical Services Act; BayRDG), contemporary quality-management frameworks increasingly emphasize structured and reliable data capture to support the development of AI-driven clinical applications.

Integrating model outputs into standardized stroke prenotification templates would allow EVT centres to anticipate imaging resources, mobilize endovascular teams, and may streamline activation paths. As with Stroke Angel, where structured telemedical input improved process efficiency [23], the addition of interpretable AI-based risk stratification could extend these gains to EVT triage. Although the RF model exhibited a slightly higher NNA (5.0 vs 4.7) compared to the 4I-SS, its improved discriminative performance for identifying EVT candidates suggests that the modest increase in alert burden is acceptable within routine clinical decision making.

The transition from technical "proof-of-concept" to scalable clinical utility represents the primary hurdle in digital emergency medicine. The CAEHR pilot (n = 25) established the feasibility of Stroke Angel, but highlighted significant data attrition, with actionable admission metrics available for only 52% (13/25) of participants despite high consent rates (83.3%) [24]. This finding is indicative of the so-called "implementation gap", a term used to describe the phenomenon whereby organisational friction and manual entry requirements serve to limit data integrity in acute workflows. In contrast, the results of our large-scale evaluation (n = 1,435) demonstrate robust clinical penetration, with a 79% documentation completion rate for the 4I-SS. A critical high-risk cohort has been identified, comprising 28% of patients with a 4ISS > 2, for whom telemedical prenotification is deemed to be of the utmost importance. However, the persistent 21% documentation missingness in our cohort serves as a quantitative "stress-response" metric; it defines the upper limit of manual data entry in high-acuity settings where clinical urgency supersedes structured documentation. Strategic alignment with telemedicine stroke networks can further smooth the feedback loop, using RWD to iteratively retrain and recalibrate the model, thereby closing the cycle of continuous quality improvement. Such initiatives will define the next frontier in prehospital stroke care, where data science and clinical pragmatism converge to save lives [25].

Accurate prehospital identification of LVO remains a major determinant of access to EVT and functional recovery. Conventional stroke scales, while pragmatic, often lack the discriminatory power needed for reliable triage, leading to both under- and over-referral. ML approaches have demonstrated improved diagnostic accuracy of LVOs by leveraging structured clinical and physiological features [26]. However, ML models are inherently limited by their dependence on curated inputs and may struggle in heterogeneous prehospital environments where data are incomplete or noisy. Large language models (LLMs) represent a novel paradigm, capable of synthesizing unstructured data such as paramedic narratives, dispatch transcripts, and multimodal records [27]. The potential to contextualize fragmented information and generate interpretable reasoning is limited, thereby enhancing triage accuracy needs further attention [28].

Unlike conventional ML, LLMs can dynamically adapt to variable data quality and capture subtle clinical cues that structured pipelines may overlook. Nonetheless, concerns regarding reliability, hallucination, and regulatory validation remain significant barriers to clinical deployment. Retrieval augmented generation (RAG), which grounds outputs in curated external knowledge sources, offers a potential pathway to mitigate these risks by improving factual consistency, transparency and auditability [29]. A hybrid strategy that integrates ML's structured-data precision with RAGs contextual reasoning may provide the most robust solution.

The future of EVT triage could not be defined by one approach alone: it can be shaped by the integration of both explainable ML models that can prioritize and flag potential EVT candidates, while experienced neurologists validate and refine these decisions in real time. This synergy combines the strengths of machine precision and human judgement, promising faster, safer, and more trustworthy pathways to EVT. Building on the findings of Schlemm et al. [30], who demonstrated that clinical judgment (CLUE) outperformed established prehospital scales such as RACE and FAST-ED in identifying LVO. Our study extends this discussion by contrasting experiential clinical knowledge with artificial knowledge. CLUE exemplifies the power of accumulated bedside experience, integrating subtle contextual cues that rigid algorithms and scales often miss. In parallel, our ML model translates complex prehospital and hospital-based symptom constellations into reproducible predictions, with SHAP-based feature importance providing transparency into individual triage decisions.

This juxtaposition highlights the limitations of rule-based scales: they are simple and scalable, yet insufficiently precise, whereas clinical expertise is highly accurate but not universally available. Together, these insights suggest that the future of EVT triage will not rest on choosing between human expertise and data-driven approaches, but on strategically combining them to maximize accuracy and equity in acute stroke care.

Strength and Limitations

This study has several limitations. First, as a retrospective, single-centre analysis, this study is inherently vulnerable to selection bias, incomplete ascertainment, and context-specific practice patterns, limiting both causal inference and broader generalisability. However, the study was conducted within a harmonised national network using standardised data structures, which provides a pathway for scalable validation and enhances the potential reproducibility of our findings across additional sites.

Second, due to the design of the study, few demographic data were provided to ensure complete patient anonymisation and further clinical data could not be collected due to the retrospective nature. This limitation will require further consideration to adapt the software in the future.

Third, the results convey that while discriminative ability is similar across models, differences emerge in sensitivity–specificity profiles and precision–recall dynamics, with implications for operational deployment in pre-hospital EVT triage. Models with higher sensitivity increase the likelihood that true

EVT candidates are identified early, enabling direct routing to comprehensive stroke centres and potentially reducing time to reperfusion; however, lower specificity in these models may lead to more false-positive classifications, resulting in unnecessary inter-facility transfers and increased resource utilization. Variations in AUPRC further influence the reliability of positive predictions, affecting the confidence with which prehospital teams can act on model outputs.

This study shows that XAI models can estimate the likelihood of subsequent EVT candidates from integrated prehospital and in-hospital stroke data with promising discrimination. The model does not determine causal eligibility; it quantifies statistical associations between RWD and observed EVT treatment decisions. Feature-level attribution provides transparent, case-specific insight into the drivers of each prediction. XAI outputs via SHAP and DCA enhance interpretability and may support integration into prehospital triage workflows. Prospective and external validation in multicentre settings is warranted to confirm generalizability and broader clinical utility.

Materials and Methods

Study Design

This is a retrospective, single-centre observational study. The Stroke Angel Initiative is a collaborative clinical research project established by emergency medical professionals and researchers to investigate the digital workflow optimization of acute stroke across the Rhön-Grabfeld, Bavaria (Germany) region [23]. The analysis used prospectively collected data from consecutive patients with suspected stroke admitted to the supra-regional stroke centre (approx. annual 1,000 stroke patients) in Bad Neustadt a. d. Saale, Bavaria, Germany.

Reporting follows the TRIPOD + AI guidelines for the development, validation, and updating of AI-based prediction models [31], with the completed checklists provided (Table S4 in the Supplementary Materials). The code written and used in this study is publicly available on GitHub (<https://github.com/stroke-angel-group/prehospital-stroke-triage-R>) and has been archived on Zenodo [DOI]. Instructions for reproducing the analyses are provided in the repository's README file.

This investigation was conducted as a retrospective study using datasets that had been irreversibly anonymized prior to analysis. The retrospective design inherently relies on pre-existing data, without any direct interaction with human participants. The absence of personal identifiers precludes risks to privacy or autonomy, and according to the requirement of the Ethics Committee of the Bavarian Medical Association no formal ethics consultation was required. We confirm that all methods were carried out in accordance with relevant guidelines and regulations. The Ethics Committee of the Bavarian State Chamber of Physicians (Bayerische Landesärztekammer; No. 26042-BO) evaluated and approved all experimental protocols. The committee grants a waiver of informed consent because the study relies on a retrospective design and fully anonymized data. We follow the Declaration of Helsinki and meet all institutional and national ethical requirements. We handled all data in accordance with the General Data

Protection Regulation (GDPR) and institutional data protection standards. The anonymization procedures applied in this study underwent independent evaluation and formal approval by an external institutional data protection officer.

Sample Size Determination and Justification Analysis

We calculated the required sample size using a dual framework addressing both model stability requirements (development cohort) and statistical justification for validation (evaluation cohort). Applying the criteria proposed by Riley et al. [32, 33] indicated a required development cohort of 1,502 patients (101 EVT events). In parallel, a power analysis based on the DeLong method [34] determined a minimum evaluation cohort of 118 patients (8 EVT events). While the evaluation cohort (287 patients, 24 EVT events) satisfied these requirements, the development set was restricted by matching constraints.

Record Linkage and Data Preparation

We probabilistically matched prehospital (Database A) and hospital (Database B) stroke records using age, time of admission and estimated arrival time (see Supplementary File). After matching EMS data with clinical records, the dataset initially contained 252 variables. Among these, 29 variables were contributed by prehospital data, including vital parameters, neurological evaluation results. The remaining 223 variables were derived from the hospital's clinical electronic workflow and included stroke-onset times, ICD-10–based discharge diagnoses, and detailed information on stroke treatments such as intravenous thrombolysis and EVT, together with their respective eligibility criteria. We used only structured data and excluded unstructured and semi-structured sources.

Finally, 14 variables were selected as predictors: age, sex, systolic blood pressure, pulse, blood glucose, body temperature, anticoagulant medication, time of stroke onset (TOO) as interval, exclusion criterion TOO, abnormal pupil reaction (defined as anisocoria or irregular pupil shape), and the four individual items of the prehospital 4-Item Stroke Scale (4I-SS; vigilance, gaze/head deviation, hemiparesis, aphasia/dysarthria). Prehospital neurological assessments were performed by trained paramedics and emergency physicians within the regional EMS. Decisions relied solely on clinical evaluation and radiological findings, which minimizes the risk of label leakage, a common source of bias in medical ML.

A feature selection procedure was used to restrict model development to relevant variables, thereby enhancing clinical interpretability and reducing overfitting. First, we removed all variables that are not available in prehospital settings. We then applied a domain-guided preselection of potential predictors and subsequently performed Recursive Feature Elimination to identify the most relevant variables [35]. We performed outlier detection and re-coded the values of identified outliers as missing. Several clinically relevant variables had incomplete observations. Assuming missingness to be at random after exploratory assessment of missing data patterns, missing values were imputed using a bagged tree model exclusively on the training data and then applied to the test set with the learned parameters, preventing data leakage. To address class imbalance resulting from the 6.6% EVT prevalence, we applied

model-specific resampling techniques, including random undersampling, Synthetic Minority Oversampling Technique (SMOTE) and hybrid resampling while the validation data remained untouched [36]. Predictors underwent algorithm-specific preprocessing, including logistic transformation for the neural network, min–max scaling for k-nearest neighbour (kNN), and no scaling for tree-based models. As all categorical variables were either inherently binary or aggregated to binary form, additional encoding was not required.

Model Development

We developed predictive models using a diverse set of supervised ML algorithms. The selection of algorithms ranged from simple rule-based learners (OneR) to more complex neural network architectures. They were chosen to represent a broad spectrum of learning strategies, including an instance-based method (kNN), tree-based models (classification and regression trees, CART), and ensemble approaches (Random Forest, XGBoost). This multi-algorithm design increased the likelihood of strong predictive performance, enabled comparison across models with different inductive biases, and reduced dependence on any single algorithm when working with complex, non-linear healthcare data [37].

We randomly assigned patients to a training cohort, a validation cohort, and a test cohort according to a priori defined 80/20 split. The validation cohort was used to monitor model performance during training and to select the optimal model, whereas the test cohort was used for the final evaluation. We then performed hyperparameter optimization using EVT-stratified, twice-repeated 5-fold cross-validation in combination with a race-based ANOVA (Analysis of Variance) tuning strategy to maximize AUROC.

Performance Evaluation and Visualization

We assessed model performance on the test cohort using the best-performing models from the 5-fold cross validation. Reported metrics include the AUROC, the area under the precision–recall curve (AUPRC), accuracy, sensitivity (equivalent to recall), specificity, positive predictive value (equivalent to precision), negative predictive value, the F1-score, and the Brier score. The 4I-SS serves as a benchmark for the ML models. It was originally developed to predict acute ischemic stroke, LVO detection and can also support prehospital patient allocation when used with a cutoff greater than 2. The 95% confidence intervals (CI) were computed using the bias-corrected and accelerated (BCa) bootstrap method with 5,000 resamples to address asymmetry observed in the metric distributions. To assess robustness to stochastic variation, the entire nested cross-validation (NCV) pipeline was repeated with five random seeds. Due to computational constraints, confidence intervals were estimated from bootstrapping of pooled out-of-fold predictions rather than across seeds. Seed-to-seed variability was summarized using the standard deviation.

With an EVT prevalence of 6.6% of the final cohort, the baseline AUPRC is 0.066; thus, we assume the observed AUPRC above 0.25 indicates substantial discriminative performance beyond chance. We calculated the number needed to alert (NNA). NNA reflects the burden of false positives expected on the clinical teams responding to the alerts for each correct prediction, which is crucial in balancing the cost of alert fatigue with the benefit of accurate predictions [38]. To quantify the influence of each predictor on the model's output, we applied SHapley Additive exPlanations (SHAP) framed within the broader XAI paradigm [39]. SHAP values summarize how much each feature changes the predicted outcome. Positive values (red) indicate a shift toward predicting EVT, while negative values (blue) indicate a shift toward predicting its absence. Decision curve analysis (DCA) was used to evaluate the tradeoff between undertriage and overtriage across competing triage strategies at clinically relevant probabilities (Supplementary Figure S1).

Calibration of predicted probabilities was assessed using calibration plots (Supplementary Figure S2). As resampling techniques applied during data preprocessing can bias calibration, predicted probabilities were recalibrated using platt scaling with logistic regression fitted on out-of-fold predictions from the training set. This ensured reliable estimation of both Brier scores and net benefit curves. Exploratory subgroup analyses by sex and age group were conducted to assess ensemble-level AUROC disparities (Supplementary Figure S4).

Analytical Tools

All statistical analyses and ML-procedures were carried out in R 4.4.1 (The R Foundation for Statistical Computing, Vienna, Austria). Data linkage, preprocessing, and exploratory analysis were conducted using the Tidyverse package (2.0.0), while Tidymodels provided the framework for ML, including model development, tuning, and evaluation. We used pmsampsize R-package for sample size justification. The ML algorithms were implemented using the following package versions: OneR (2.2), kkn for k-nearest neighbors (1.4.0), rpart for decision trees (4.1.24), xgboost (1.7.9.1), ranger for random forests (0.17.0), and neuralnet for neural networks (1.44.2).

Declarations

Conflict of Interest Statement

All authors declare no financial or non-financial competing interests.

Acknowledges

We thank the sustained and committed contribution of all working group members involved in the diagnosis and treatment of patients in the study.

Data Sharement Agreement

The datasets generated and/or analysed during the current study are not publicly available due to privacy and confidentiality restrictions but are available from the corresponding author on reasonable request.

Funding

This study received no funding.

Author Contribution

PAE, OM, AR, VZ, and HS conceptualized and designed the study and wrote the main manuscript text. HS and VZ collected the data. PAE, OM, and JH curated and cleaned the data. HS, ES, and VZ conducted the clinical evaluation. PAE, OM, HvJD, and JH performed the data and statistical analyses. PAE and OM prepared all figures. PAE, AR, OM, and HvJD developed and evaluated the machine learning models. PAE, HS, AR, ES, and HvJD supervised the project. All authors reviewed the manuscript and approved the final manuscript.

Acknowledgement

We thank the sustained and committed contribution of all working group members involved in the diagnosis and treatment of patients in the study.

Data Availability

The datasets generated and/or analysed during the current study are not publicly available due to privacy and confidentiality restrictions but are available from the corresponding author on reasonable request.

References

1. Feigin, V. L., Owolabi, M. O. & World Stroke Organization–Lancet Neurology Commission Stroke Collaboration Group. Pragmatic solutions to reduce the global burden of stroke: a World Stroke Organization-Lancet Neurology Commission. *Lancet Neurol.* **22** (12), 1160–1206 (2023).
2. Goyal, M. et al. Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomised trials. *Lancet.* 2016;387(10029):1723–1731. (2016).
3. Turc, G. et al. European Stroke Organisation (ESO) - European Society for Minimally Invasive Neurological Therapy (ESMINT) Guidelines on Mechanical Thrombectomy in Acute Ischemic Stroke. *J. Neurointerv Surg.* **15** (8), e8 (2023).

4. Fassbender, K. et al. Prehospital stroke management in the thrombectomy era. *Lancet Neurol.* **19** (7), 601–610 (2020).
5. Milne, M. S. et al. Drip 'n Ship Versus Mothership for Endovascular Treatment: Modeling the Best Transportation Options for Optimal Outcomes. *Stroke* **48** (3), 791–794 (2017).
6. Lima, F. O., Mont'Alverne, F. J. A., Bandeira, D. & Nogueira, R. G. Pre-hospital Assessment of Large Vessel Occlusion Strokes: Implications for Modeling and Planning Stroke Systems of Care. *Front. Neurol.* **10**, 955 (2019).
7. Bösel, J. et al. Access to and application of recanalizing therapies for severe acute ischemic stroke caused by large vessel occlusion. *Neurol Res Pract.* ;5(1):19. 7 Nguyen J, et al. Comparison of Prehospital Scales for Predicting Large Anterior Vessel Occlusion in the Ambulance Setting. *JAMA Neurol.* 2021;78(2):157–164. (2023).
8. Nguyen, T. T. M. et al. Comparison of Prehospital Scales for Predicting Large Anterior Vessel Occlusion in the Ambulance Setting. *JAMA Neurol.* **78** (2), 157–164 (2021).
9. Duvekot, M. H. C. et al. Comparison of eight prehospital stroke scales to detect intracranial large-vessel occlusion in suspected stroke (PRESTO): a prospective observational study. *Lancet Neurol.* **20** (3), 213–221 (2021).
10. Kothari, S. A. et al. Validation of the Ultimate Stroke Scale (USS): A standardized tool for automated large vessel occlusion screening. *Interventional neuroradiology: J. peritherapeutic Neuroradiol. Surg. procedures Relat. neurosciences*, 15910199251362077 (2025).
11. Wendel, J. et al. Improving the interface for information transfer in acute stroke care: a mixed-methods process evaluation of the emergency use case within the CAEHR project. *BMJ Open.* **15**, e105332 (2025).
12. Jalo, H. et al. Early identification and characterisation of stroke to support prehospital decision-making using artificial intelligence: a scoping review protocol. *BMJ Open.* **13** (5), e069660 (2023).
13. Triantafyllidis, A. K. & Tsanas, A. Applications of Machine Learning in Real-Life Digital Health Interventions: Review of the Literature. *J. Med. Internet Res.* **21** (4), e12286 (2019).
14. Jovin, G. T., Chamorro, A., Cobo, E., de Miguel, M. A. & Molina, C. A. Thrombectomy within 8 Hours after Symptom Onset in Ischemic Stroke. *N Engl. J. Med.* **372**, 2296–2306 (2015).
15. Ragoschke-Schumm, A. & Walter, S. D. A. W. N. and DEFUSE-3 trials: is time still important? Die DAWN- und DEFUSE-3-trials: is time still important? *Der Radiologe*, **58**(Suppl 1), 20–23 .
16. Faul, P. et al. Digital referral versus clinical reality: evaluating prehospital trauma allocation with IVENA eHealth®. *Eur. J. Trauma. Emerg. Surg.* **51**, 278 (2025).
17. Saver, J. L. Time is brain-quantified. *Stroke* **37** (1), 263–266 (2006).
18. van Calster, B., Collins, G. S., Vickers, A. J., Wynants, L. & Kerr, K. F. Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: overview and guidance. *Lancet Digit. Health.* **7** (12), 100916 (2025).

19. Mazya, M. V. et al. Implementation of a prehospital stroke triage system using symptom severity and teleconsultation in the stockholm stroke triage study. *JAMA Neurol.* **77** (6), 691–699 (2020).
20. Goisauf, M. et al. Trust, Trustworthiness, and the Future of Medical AI: Outcomes of an Interdisciplinary Expert Workshop *J Med Internet Res* 2025;27:e71236 (2025).
21. Oh Ne, Oh, T. Y. C. et al. Prehospital real-time AI for trauma mortality prediction: a multi-institutional and multi-national validation study. *Nat. Commun.* (2026).
22. Bienzeisler, J. et al. Implementation report on pioneering federated data access for the German National Emergency Department Data Registry. *npj Digit. Med.* **8**, 94 (2025).
23. Eder PA, Laux G, Rashid A, Kniess T, Haeusler KG, Shammass L, ... Stroke Angel Study Group. (2021). Stroke angel: effect of telemedical prenotification on in-hospital delays and systemic thrombolysis in acute stroke patients. *Cerebrovasc Dis*, 50(4),420–428.
24. Hofmann, A. L. et al. Time-critical quality assessment of a digital hub for acute stroke care: Implementing key performance indicators within the CAEHR (CARDiovascular diseases–Enhancing Healthcare through cross-sectoral Routine data integration) project. *Inf. Med. Unlocked.* **61**, 101745 (2026).
25. Pulsiri, N., Vatananan-Thesenvitz, R., Sirisamutr, T. & Wachiradilok, P. Save lives: a review of ambulance technologies in pre-hospital emergency medical services. In 2019 Portland International Conference on Management of Engineering and Technology (PICMET) (pp. 1–10). IEEE.
26. Alobaida, M. et al. Systematic Review and Meta-Analysis of Prehospital Machine Learning Scores as Screening Tools for Early Detection of Large Vessel Occlusion in Patients With Suspected Stroke. *J. Am. Heart Association.* **13** (12), e033298 (2024).
27. Wang, X. et al. Performance of ChatGPT on prehospital acute ischemic stroke and large vessel occlusion (LVO) stroke screening. *Digit. Health.* **10**, 20552076241297127 (2024).
28. Balcarcel, D. R. et al. Feedback loops in intensive care unit prognostic models: an under-recognised threat to clinical validity. *Lancet Digit. Health.* **7** (8), 100880 (2025).
29. Ng, K. K. Y., Matsuba, I. & Zhang, P. C. RAG in health care: a novel framework for improving communication and decision-making by addressing LLM limitations. *Nejm Ai.* **2** (1), Alra2400380 (2025).
30. Schlemm, E. et al. Clinical Judgment vs Triage Scales for Detecting Large Vessel Occlusions in Suspected Acute Stroke. *JAMA Netw. Open.* **6** (9), e2332894 (2023).
31. Collins, G. S. et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *bmj*, 385. (2024).
32. Riley, R. D. et al. Minimum sample size for developing a multivariable prediction model: PART II- binary and time-to-event outcomes. *Stat. Med.* **38** (7), 1276–1296 (2019).
33. Riley, R. D. et al. Importance of sample size on the quality and utility of AI-based prediction models for healthcare. *Lancet Digit. Health.* **7** (6), 100857 (2025).

34. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 837–845 (1988).
35. Xia, S., Yang, Y. & Model-Free, A. Feature Selection Technique of Feature Screening and Random Forest-Based Recursive Feature Elimination. *Int. J. Intell. Syst.* (1), 2400194. (2023).
36. Chawla, N. V., Bowyer, K. W. & Hall, L. O. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 2022, **16**, 321–357 .
37. Alkhamis, M. A. et al. Interpretable machine learning models for predicting in-hospital and 30 days adverse events in acute coronary syndrome patients in Kuwait. *Sci. Rep.* **14**, 1243 (2024).
38. Dewan, M. & Sanchez-Pinto, L. N. Crystal Balls and Magic Eight Balls: The Art of Developing and Implementing Automated Algorithms in Acute Care Pediatrics. *Pediatr. Crit. Care Med.* **20** (12), 1197–1199 (2019).
39. Chen, H. et al. Algorithms to estimate Shapley value feature attributions. *Nat. Mach. Intell.* **5**, 590–601 (2023).

Figures

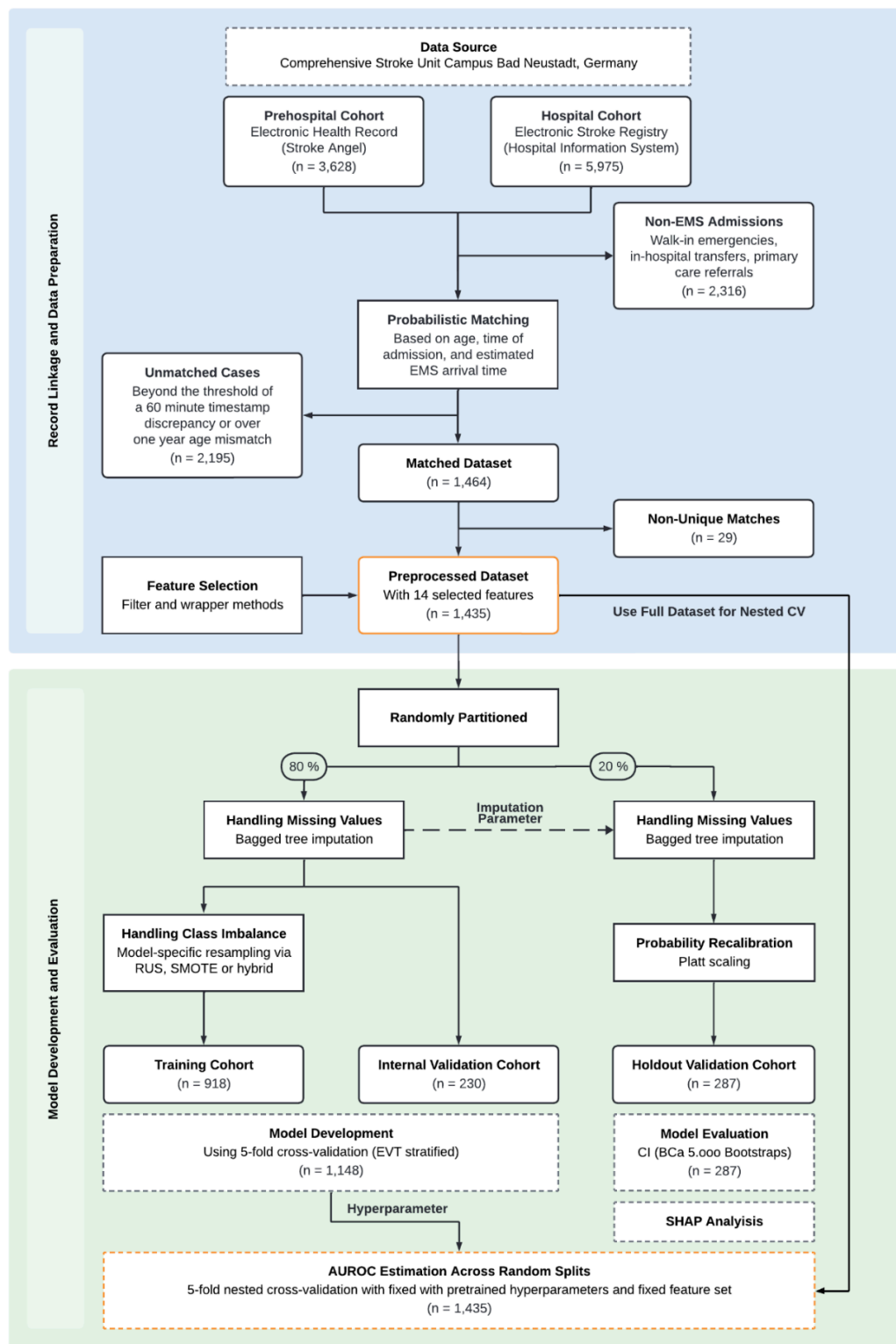


Figure 1

Study flowchart. This figure illustrates the flowchart of this study, including the inclusion and exclusion criteria. AUROC: Area Under the Receiver Operating Characteristic, EVT: Endovascular Therapy, RUS: Random Undersampling, BCa: Bias-Corrected and accelerated, CI: 95% Confidence Intervals, SMOTE: Synthetic Minority Oversampling Technique, SHAP: SHapley Additive exPlanations

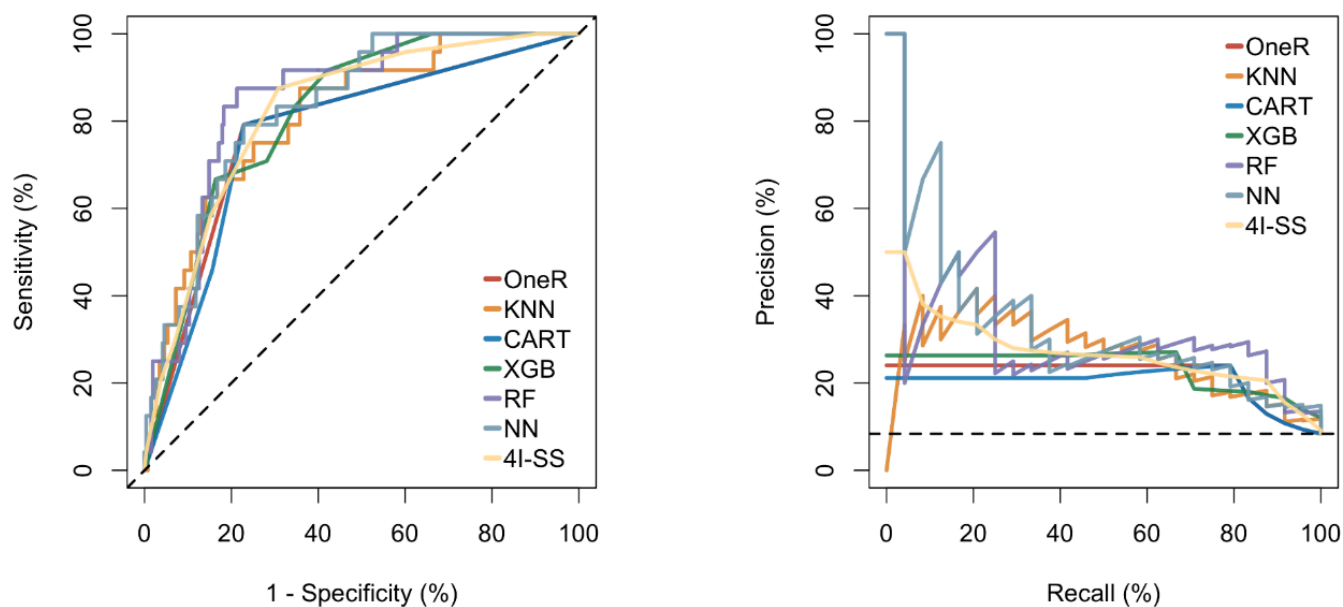


Figure 2

ROC and PR curves for identifying EVT candidates within the evaluation cohort. **a** Panel A shows the receiver-operating-characteristic (ROC) curve with the area under the curve (AUROC) for all machine learning models. The ROC plot displays curves: the 4I-SS reference (yellow) and ML-approaches (shown in distinct colors). The x-axis shows the false-positive rate (1 – specificity, in %), and the y-axis shows sensitivity (true-positive rate, in %). **b** Panel B depicts the corresponding precision–recall curves, with the area under the precision–recall curve (AUPRC) reported for each machine learning model. The x-axis shows recall (sensitivity, in %), and the y-axis shows precision (positive predictive value, in %).

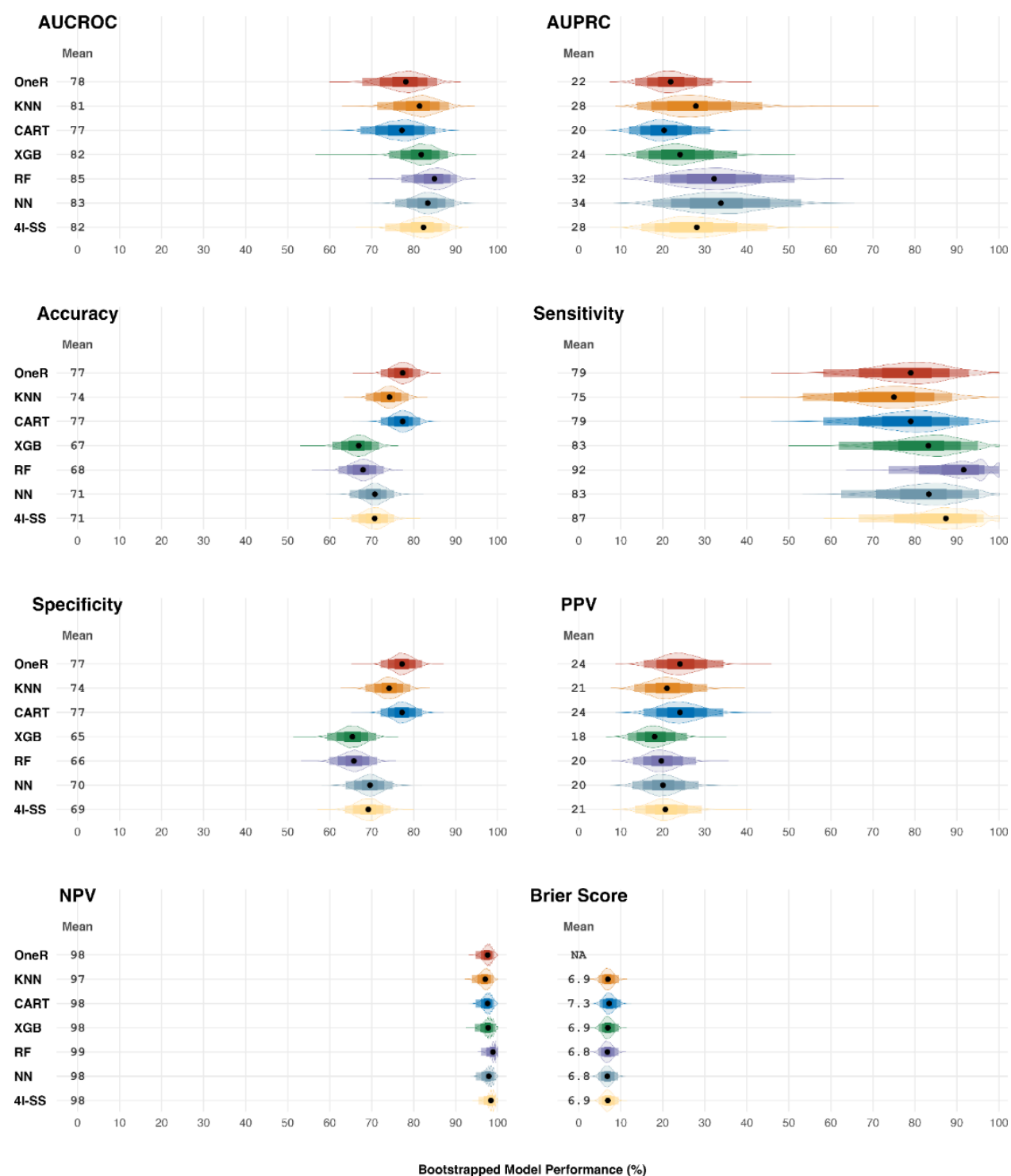


Figure 3

Overview of performance metrics. Predictive out-of-sample performance of machine learning classifiers and the 4I-SS reference to identify endovascular thrombectomy candidates, showing bootstrap distributions ($n = 5,000$ resamples of the test set) for all performance metrics. Black dots denote the bootstrap means, and coloured horizontal bands show the central 50%, 80%, and 95% confidence intervals, with decreasing colour saturation indicating intervals that include a larger proportion of

bootstrap estimates. Given the observed skewness in several metrics, especially AUCPR and sensitivity, bias-corrected and accelerated (BCa) intervals were used for all metrics. Platt-scaled probabilities were used to calculate Brier scores, which are not defined for OneR.

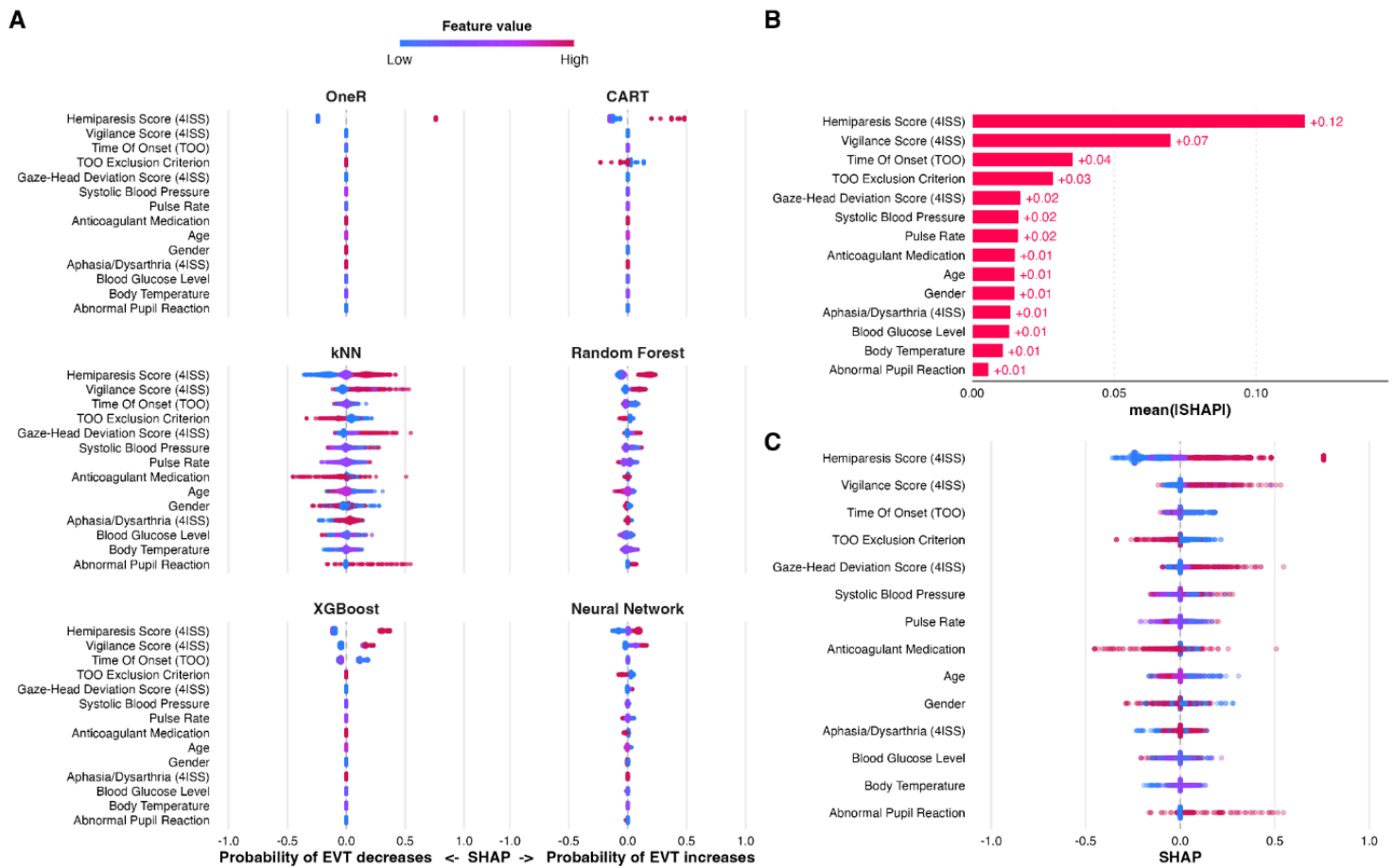


Figure 4

SHapley Additive exPlanations (SHAP) summary plots illustrating feature contributions to the predictive model. Features are ranked in order of importance from top to bottom. **a** SHAP summary beeswarm plot ranking features by mean absolute SHAP value. The x-axis represents the SHAP impact on model output, while color denotes normalized feature values. Red values shift predictions toward endovascular thrombectomy, whereas blue values shift toward absence. **b** Top-10 attributes ranked by SHAP values, driving the output. **c** SHAP summary beeswarm plots comparing feature rankings across all six machine learning algorithms, based on SHAP values, to illustrate consistency and variability in predictor influence.

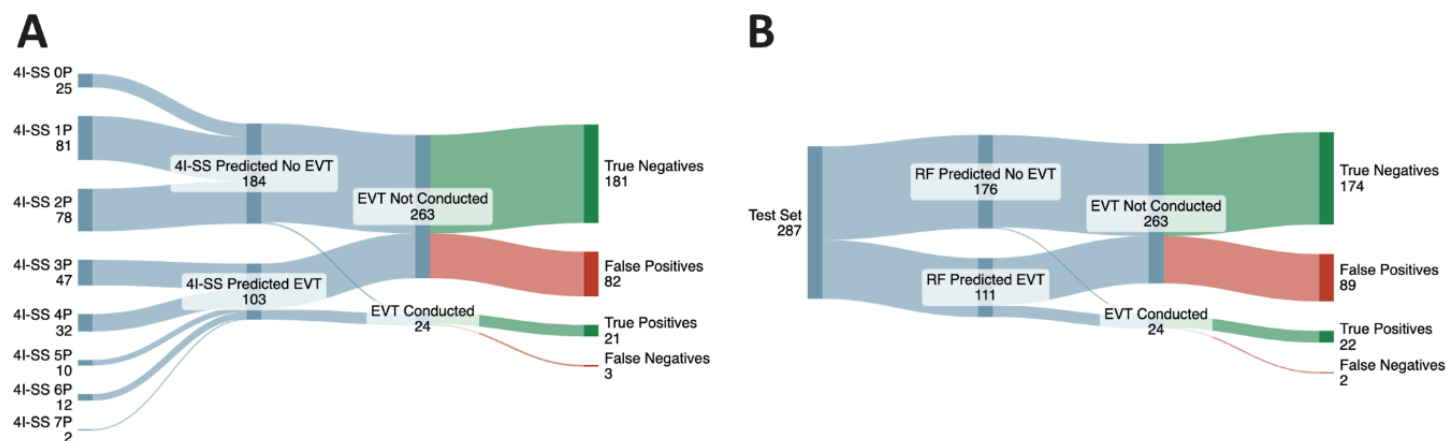


Figure 5

Sankey classification plot. Flow of EVT prediction and outcome classification in **a** the reference (4I-SS) and **b** illustrating the performance of a Random Forest (RF) model in predicting EVT candidacy. The flow shows model predictions (EVT vs. no EVT), actual treatment decisions, and final classification into true positives, false positives, false negatives, and true negatives. Band widths reflect patient counts, enabling visual assessment of triage accuracy and misclassification patterns.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementarymaterialStrokeAngel.docx](#)