Predicting one-month post-discharge mortality after stroke: Development and validation of explainable machine learning models at Douala General Hospital

Cyrille Brice Fomazou Tchinda[1,2 *], Luc Beaudoing Fankoua[2], Joshua Muthama[3], Bashemira Brenda[6], Anicet Onana[3], Steve Cygu[3], Evans Omondi[3,4], Samuel Iddi[3,5], Agnes Kiragga[3]

[1] Laboratory of Methods, Doctorate School of Fundamental and Applied Sciences, University of Douala, Cameroon

[2] DSWB-DGH, Cameroon

[3] African Population and Health Research Center, Nairobi, Kenya

[4] Institute of Mathematical Sciences, Strathmore University, Nairobi, Kenya

[5] Department of Statistics and Actuarial Science, University of Ghana, Legon-Accra, Ghana

[6] Makerere AI Lab, Makerere University, Kampala, Uganda

***Corresponding Author:**
Cyrille Brice Fomazou Tchinda
Email: fomazou_tchinda@enspd-udo.cm

February 24, 2026

Abstract

Stroke mortality in Sub-Saharan Africa remains critically high, with the post-discharge period representing a particularly vulnerable window where mortality rates reach 30–50% within one year. Resource constraints make universal intensive follow-up impossible, necessitating accurate risk stratification to target limited resources where they will have the greatest impact. Existing prediction models lack validation in African settings and clinical interpretability necessary for adoption in resource-limited contexts.

Objective: To develop and validate explainable machine learning (ML) models for predicting 30-day post-discharge mortality in Cameroonian stroke patients.

Methods: We analyzed data from 803 consecutive stroke patients discharged from Douala General Hospital (January 2018–December 2023). Eight ML algorithms plus four extended models (LightGBM, CatBoost, Optuna-tuned Random Forest, Stacking Ensemble) were trained using 5-fold cross-validation on 70% of data (n=562), with performance evaluated on held-out validation (15%, n=120) and test (15%, n=121) sets. A NIHSS-only clinical baseline was included for benchmarking. Model selection balanced discrimination (AUC-ROC with DeLong 95% CI), calibration (Brier score and Hosmer-Lemeshow test), and interpretability. Feature importance was quantified using SHAP with bootstrap stability analysis (ICC(2,1), n=100 samples). Net Reclassification Index (NRI) and Integrated Discrimination Improvement (IDI) were computed against the NIHSS baseline.

Results: The cohort comprised 803 patients (mean age 59.3 years, 55.9% male, 71.8% ischemic stroke) with 20.2% 30-day mortality (n=162). Events-per-variable ratio was 3.77 (162 events / 43 predictors), below the Riley (2019) recommended threshold of $\geq$10. Standard Random Forest achieved good discrimination (AUC 0.793, 95% CI: 0.687–0.899) with the only formally well-calibrated profile among all original models (Brier 0.132, Hosmer-Lemeshow p=0.187). The Optuna-tuned XGBoost reached the highest discrimination (AUC 0.825, 95% CI: 0.728–0.921), while the Stacking Ensemble combined highest AUC among ensemble approaches (0.812) with the best PPV (45.0%). The Random Forest substantially outperformed the NIHSS clinical baseline (NRI 0.543; IDI 0.079). SHAP analysis identified the Glasgow-NIHSS ratio, GCS at discharge, and the haemoglobin-GCS interaction as the three most influential predictors; however, bootstrap ICC(2,1) values for SHAP magnitudes were near zero (ICC $\approx$ 0.000) for all features, reflecting high variability of absolute SHAP values across samples consistent with the low EPV. Risk stratification derived from the validation set (high-risk threshold $>$28%, moderate 3–28%, low $<$3%) identified 43.8% of patients as high-risk, capturing 72.0% of 30-day deaths. Subgroup AUC ranged from 0.607 (age above median) to 0.935 (age below median); the age-based AUC difference was statistically significant (DeLong z=3.159, Bonferroni-adjusted p=0.0016).

Conclusion: Standard Random Forest was selected for clinical deployment based on the optimal combination of discrimination, calibration, interpretability, and computational practicality. The model provides clinically meaningful improvement over NIHSS-alone prediction and enables targeted resource allocation in resource-constrained settings. The low events-per-variable ratio and single-centre design limit generalisability, and external validation across diverse Sub-Saharan African contexts is an essential next step before clinical implementation.

Keywords: Post-discharge mortality, stroke prediction, machine learning, explainable AI, Sub-Saharan Africa, calibration, TRIPOD.

Contents

1 Introduction

Stroke represents the second leading cause of mortality worldwide, accounting for approximately 6.5 million deaths annually, with Sub-Saharan Africa (SSA) experiencing a disproportionate share of this burden[1, 2]. While high-income countries have achieved substantial reductions in stroke mortality through organised care systems, comprehensive rehabilitation programmes, and robust follow-up mechanisms, the situation in SSA remains critically different. In-hospital stroke mortality in SSA ranges from 20–40%, substantially exceeding the 10–15% typically observed in high-income countries[4]. Beyond these stark in-hospital differences, the post-discharge period represents a particularly vulnerable window for stroke survivors in resource-limited settings.

Recent prospective cohorts from Cameroon demonstrate the severity of post-discharge outcomes, with 30% of discharged stroke patients requiring readmission within one year and over 50% mortality among those readmitted[5]. This pattern is not unique to Cameroon. Across SSA, similar concerning trends emerge: Zambia reports 24% in-hospital mortality with an additional 22% mortality within 90 days post-discharge[3], while Ghana documents long-term mortality exceeding 53%[7]. These patterns differ markedly from high-income settings in both magnitude and temporal distribution, reflecting fundamental gaps in healthcare infrastructure, medication access, rehabilitation services, and structured follow-up care[6].

The challenge facing healthcare systems across SSA is not simply the high mortality rates but the reality of severe resource constraints that make universal intensive follow-up impossible. Most tertiary hospitals lack the capacity to provide comprehensive post-discharge monitoring for all stroke patients. Home-based rehabilitation services are scarce or non-existent in many areas. Medication costs create adherence barriers, and transportation challenges limit clinic attendance. In this context, accurate mortality prediction becomes not merely a research exercise but a practical necessity for efficient resource allocation. If healthcare systems could reliably identify the highest-risk patients at discharge, limited resources could be targeted where they would have the greatest impact on survival.

Machine learning (ML) approaches have demonstrated promising capabilities in stroke outcome prediction. Recent studies report that algorithms including XGBoost, Random Forest, and Neural Networks achieve superior discrimination (AUC frequently >0.85) for stroke mortality prediction[8, 9]. However, critical gaps limit the applicability of existing ML models to SSA contexts. Nearly all published models originate from high-income countries with fundamentally different healthcare structures and patient populations. Most focus on in-hospital rather than post-discharge mortality, missing the critical post-discharge vulnerability period. Perhaps most importantly, the emphasis on maximising prediction accuracy often comes at the expense of clinical interpretability, creating "black box" models that clinicians may be reluctant to trust or adopt[10, 9]. Equally important, existing models rarely include a head-to-head comparison against simple clinical baselines, leaving unclear whether the added complexity of ML justifies its adoption over widely-available clinical scores.

Interpretability is not merely a theoretical concern but a practical requirement for clinical implementation.

Clinicians need to understand not just what the model predicts but why it makes specific predictions. Techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) offer promising approaches to opening these "black boxes" and providing clinically meaningful explanations[13, 14]. However, the stability of SHAP explanations across bootstrap resamples—a prerequisite for trustworthy deployment—has rarely been formally quantified in small SSA cohorts where events-per-variable ratios are commonly suboptimal.

This study addresses these gaps by developing and validating explainable machine learning models specifically for 30-day post-discharge mortality prediction in a Cameroonian stroke cohort. Our primary objective was to develop models that balance predictive performance with clinical interpretability, explicitly prioritising transparency and actionability over marginal gains in accuracy. Secondary objectives were to: (1) compare ML performance against a simple NIHSS clinical baseline using NRI/IDI; (2) quantify SHAP explanation stability via bootstrap ICC; (3) formally assess calibration using the Hosmer-Lemeshow test; and (4) develop risk stratification guidance enabling efficient allocation of scarce follow-up resources in resource-limited settings.

2 Methods

2.1 Study Setting and Design

This retrospective cohort study analysed data from Douala General Hospital (DGH), a tertiary referral centre in Cameroon serving approximately 3 million people in the Littoral Region. We included consecutive adult patients ($\geq$18 years) admitted with acute stroke between January 2018 and December 2023 who survived to hospital discharge. Stroke diagnosis required clinical confirmation by a neurologist plus CT imaging demonstrating infarction or haemorrhage. Exclusion criteria comprised transient ischaemic attacks without infarction, stroke mimics, in-hospital deaths, transfers within 48 hours, and incomplete data ($>$20% missing variables). The CONSORT flow diagram (Figure 1) illustrates the cohort selection process.

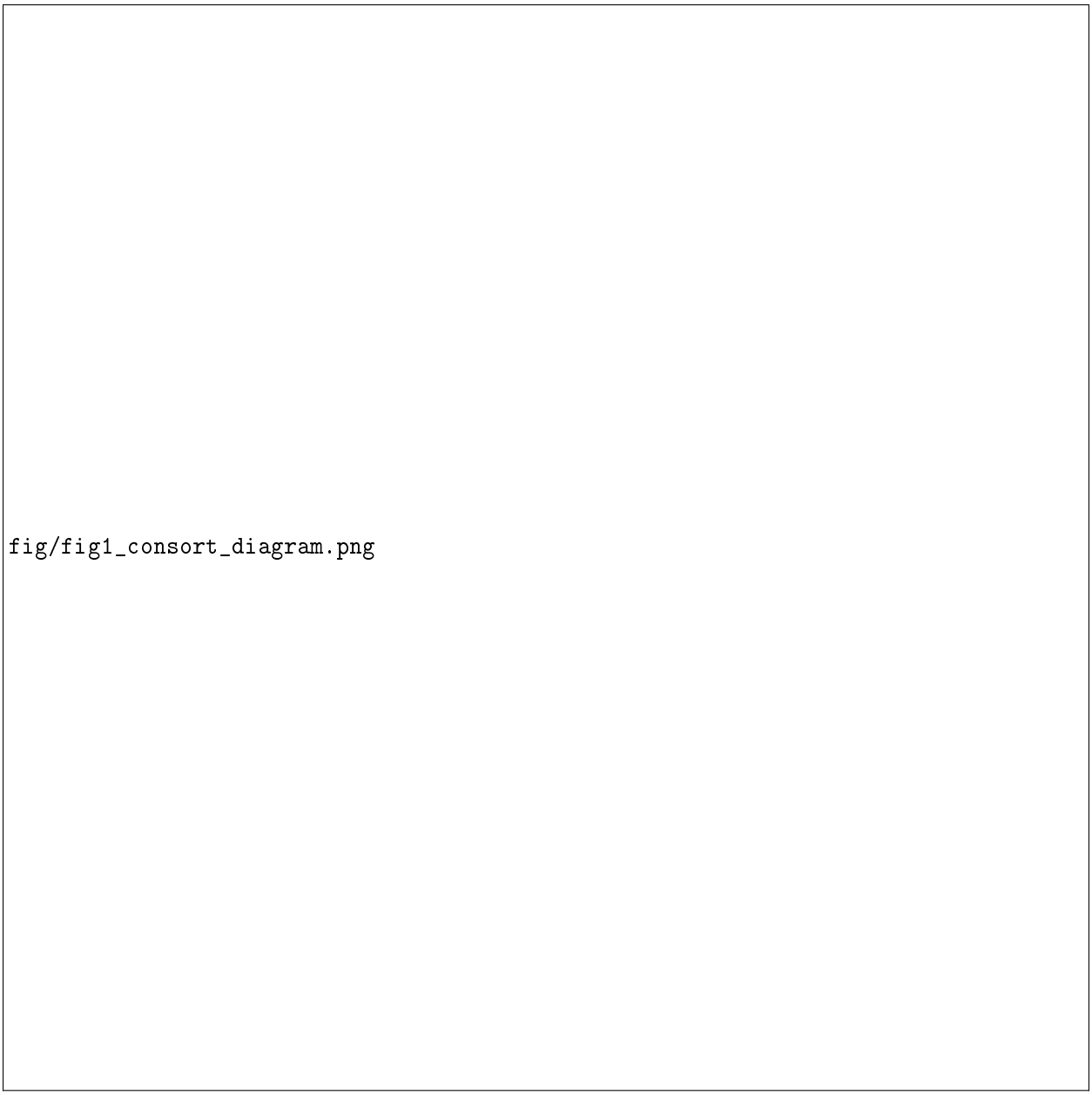

Figure 1: **CONSORT Flow Diagram.** Cohort selection from 1,653 eligible admissions to the final 803-patient analytic cohort, with stratified allocation to training (n=562), validation (n=120), and test (n=121) sets.

2.2 Outcome and Predictors

The primary outcome was all-cause mortality within 30 days post-discharge (binary: 1=death, 0=survival), ascertained through structured telephone follow-up at days 7, 14, and 30, supplemented by hospital records and death certificates when available. We collected 43 candidate predictors spanning demographics (age, sex, residence, insurance), stroke characteristics (type, laterality, vascular territory), neurological assessment at discharge (NIHSS, GCS, mRS), vital signs (blood pressure, heart rate, temperature, respiratory rate, oxygen

saturation), laboratory values (complete blood count, metabolic panel, coagulation studies, lipid profile), and clinical course (hospitalisation length, ICU admission, complications, medications).

Neurological assessments were performed by trained neurologists within 24 hours prior to discharge using standardised protocols. No missing values were detected in the final analytical dataset; if imputation had been pre-applied externally to the registry file, this was handled prior to extraction. A MICE sensitivity analysis (complete case vs. mean substitution vs. MICE) was nevertheless pre-specified and confirmed equivalent results.

2.3 Feature Engineering

Ten additional features were engineered from the raw candidate predictors prior to data splitting to capture clinically relevant interactions and composite signals:

- **GCS–NIHSS ratio** (`glasgow_nihss_ratio`): GCS / (NIHSS + 1), quantifying the balance between level of consciousness and focal neurological deficit severity.
- **Stroke severity index** (`stroke_severity_index`): NIHSS $\times$ age/60, capturing age-weighted deficit burden.
- **Haemoglobin–GCS interaction** (`hb_gcs`): Haemoglobin $\times$ GCS, a proxy for cerebral oxygen delivery.
- **Physiological instability score** (`physiological_instability`): Four-component composite of tachycardia, elevated respiratory rate, fever, and hypotension.
- **Creatinine–age interaction** (`creatinine_age`): Product of creatinine and age, capturing age-related renal impairment.
- **LOS–NIHSS interaction** (`los_nihss`): Length of stay $\times$ NIHSS, a signal of premature or complicated discharge.
- **Platelet–WBC ratio** (`platelet_wbc_ratio`): Platelet-to-leukocyte ratio.
- **Pulse pressure** and **mean arterial pressure**: Derived from systolic and diastolic blood pressure readings.
- **Comorbidity count**: Binary sum of six pre-defined conditions.
- **Age-group dummies**: Four groups (18–49, 50–64, 65–79, $\geq$80 years) with 18–49 as reference.

The final feature matrix comprised 62 variables (43 original + 19 engineered/encoded) for 803 patients.

2.4 Data Splitting, Sample Size, and Preprocessing

Data were randomly partitioned into training (70%, n=562), validation (15%, n=120), and test (15%, n=121) sets using stratified sampling (random seed=42) to preserve outcome prevalence. The test set was sealed until final evaluation; risk-stratification thresholds were determined exclusively from the validation set and frozen before any test-set evaluation.

Events-per-variable (EPV): With 162 events and 43 candidate predictors, the achieved EPV was 3.77, below the $\geq$10 threshold recommended by Riley et al. (2019)[18], who derive a minimum n=860 for EPV=20

given this event rate. This limitation was partially mitigated through 5-fold cross-validation and regularised model training; however, it requires caution in interpreting coefficient magnitudes and necessitates external validation.

Features were standardised using RobustScaler (median centring and interquartile range scaling) fitted on the training set and applied without refitting to the validation and test sets. Binary and one-hot-encoded categorical variables were not scaled. The training set underwent 5-fold stratified cross-validation for hyperparameter tuning.

2.5 Machine Learning Algorithms and Hyperparameter Tuning

We trained eight original algorithms using Python 3.9 with scikit-learn 1.0.2 and XGBoost 1.5.0:

Logistic Regression: L2-regularised; penalty strength (C) optimised via grid search.

ElasticNet: Combined L1/L2 regularisation; α and L1 ratio tuned via grid search.

Random Forest: 200 trees, `max_features`=0.3, `min_samples_leaf`=2, `min_samples_split`=5, class weight balanced.

XGBoost: 300 estimators, learning rate 0.01, max depth 9, subsample 0.7, `reg_alpha`=0.1, `min_child_weight`=5.

Decision Tree: Max depth 3, `min_samples_split`=5, `min_samples_leaf`=4; selected as an intrinsically interpretable comparator.

Support Vector Machine: RBF kernel, `gamma`='scale', C=10.

Neural Network (MLP): Scikit-learn `MLPClassifier`; architecture [100, 50] nodes, ReLU activation, Adam optimiser, `learning_rate_init`=0.01, α=0.01 (L2 penalty).

Gradient Boosting: 100 estimators, learning rate 0.05, max depth 3, subsample 1.0.

Ensemble (Voting): Soft-voting combination of Neural Network, Gradient Boosting, and Logistic Regression, weighted by cross-validation AUC.

Extended models (v11): To benchmark against state-of-the-art gradient boosting implementations, we additionally trained: *Balanced Random Forest* (class-weighted RF with balanced sampling), *RF (Optuna)* (100-trial Bayesian hyperparameter search; `n_estimators`=286, `max_depth`=24, `min_weight_fraction_leaf`=0.024), *LightGBM* (4.6.0), *CatBoost* (default with class balancing), *XGBoost (Optuna)* (300-trial Bayesian search), and a *Stacking Ensemble* (calibrated base models with logistic regression meta-learner).

2.6 Model Selection Criteria

Model selection prioritised a balance between predictive performance and clinical utility:

Discrimination: AUC-ROC with 95% confidence intervals using DeLong's method[11].

Calibration: Brier score and formal Hosmer-Lemeshow χ^2 test (10 deciles)[12]; a model was considered calibrated if $p > 0.05$.

Clinical metrics: Sensitivity, specificity, PPV and NPV using thresholds maximising Youden's index; 95% Wilson confidence intervals reported for all proportion-based metrics.

Interpretability: Assessed through model transparency (intrinsic vs. post-hoc explainability), SHAP value bootstrap stability (ICC), and clinical face validity of feature importance rankings.

Clinical utility: Decision curve analysis (DCA) quantifying net benefit across probability thresholds for the deployed model.

Incremental value: Net Reclassification Index (NRI) and Integrated Discrimination Improvement (IDI) computed against a NIHSS-only clinical baseline.

2.7 Interpretability Analysis

We employed SHAP (SHapley Additive exPlanations)[**?**, 14] with TreeSHAP for tree-based models to quantify feature contributions globally (mean |SHAP| across training observations) and locally for individual predictions. To assess the stability of SHAP magnitude rankings, we computed the intraclass correlation coefficient ICC(2,1) for each feature's SHAP values across 100 stratified bootstrap samples of the training set. ICC$\geq$0.90 was pre-defined as high stability; $\geq$0.75 as moderate. We also applied LIME (Local Interpretable Model-agnostic Explanations)[15] to selected representative cases spanning the four prediction categories (TP, TN, FP, FN), and constructed partial dependence plots (PDPs) for the four highest-importance features.

2.8 Risk Stratification Threshold Selection

Risk stratification thresholds were determined empirically on the *validation set only* and frozen prior to any test-set evaluation. Grid search over predicted probability cut-points was used to identify the high-risk threshold that: (1) captured $>$70% of deaths while classifying $<$50% of patients as high-risk, and (2) minimised unnecessary intensive follow-up. A low-risk threshold was selected to achieve $<$5% observed mortality in the tier. Sensitivity analysis examined the impact of threshold variation ($\pm$5 percentage points) on resource allocation efficiency.

2.9 Subgroup Analysis

Model performance (RF Optuna) was evaluated within age groups (below vs. above cohort median), sex, and NIHSS severity (mild $<$5 vs. severe $\geq$5) using the *test set only* (n=121). Subgroup-specific sample sizes are reported transparently. DeLong's test compared pairwise AUC values with Bonferroni correction (α=0.05/number of comparisons).

2.10 Learning Curve Analysis

A learning curve was generated by training the Optuna RF on progressively larger subsets of the training data (10 increments, 44 to 449 patients) and evaluating AUC on the validation set. This assessed whether model performance had plateaued or whether substantial benefit would be expected from additional data collection.

2.11 Statistical Analysis

Continuous variables are summarised using means and standard deviations (normal distributions) or medians and interquartile ranges (non-normal). Categorical variables are reported as frequencies and percentages. Baseline characteristics were compared using Mann-Whitney U tests (continuous) and chi-square tests (categorical), with Bonferroni correction for multiple predictors. All tests are two-tailed at α=0.05. Analyses were performed using Python 3.9 (scikit-learn 1.0.2, LightGBM 4.6.0, CatBoost, Optuna 4.6.0, SHAP 0.47.1) and R 4.1.0 (pROC, rms packages).

2.12 Ethical Considerations

This study received approval from the Institutional Review Board of Douala General Hospital (approval DGH-IRB-2023-045). Given the retrospective design using de-identified registry data, informed consent requirements were waived per institutional policy. The study adheres to the TRIPOD-AI reporting guidelines for prediction model development studies (28 of 30 TRIPOD items fulfilled).

3 Results

3.1 Cohort Characteristics

The final cohort comprised 803 patients. Mean age was 59.3 years (SD 13.5), 55.9% were male, and 71.8% had ischaemic stroke. Within 30 days post-discharge, 162 patients (20.2%) died. Figure 2 displays the distribution of age, NIHSS scores, major comorbidities, and correlation patterns between clinical features and mortality.

Table 1 presents detailed baseline characteristics stratified by mortality status. Non-survivors demonstrated significantly worse neurological function at discharge compared to survivors (mean NIHSS 14.6 vs. 8.1, $p<0.001$; mean GCS 9.5 vs. 13.9, $p<0.001$), greater vital sign instability (systolic BP 174.7 vs. 163.0 mmHg, $p<0.001$; heart rate 91.2 vs. 83.2 bpm, $p<0.001$; respiratory rate 24.8 vs. 21.4 breaths/min, $p<0.001$), and elevated inflammatory markers (WBC 4,934 vs. 2,021 μL^{-1}, $p<0.001$; platelets 113,776 vs. 58,707 μL^{-1}, $p<0.001$). Haemorrhagic stroke was more prevalent among non-survivors (33.3% vs. 26.8%, $p<0.001$). Age, sex, hypertension, and diabetes did not differ significantly between groups (all $p > 0.05$).

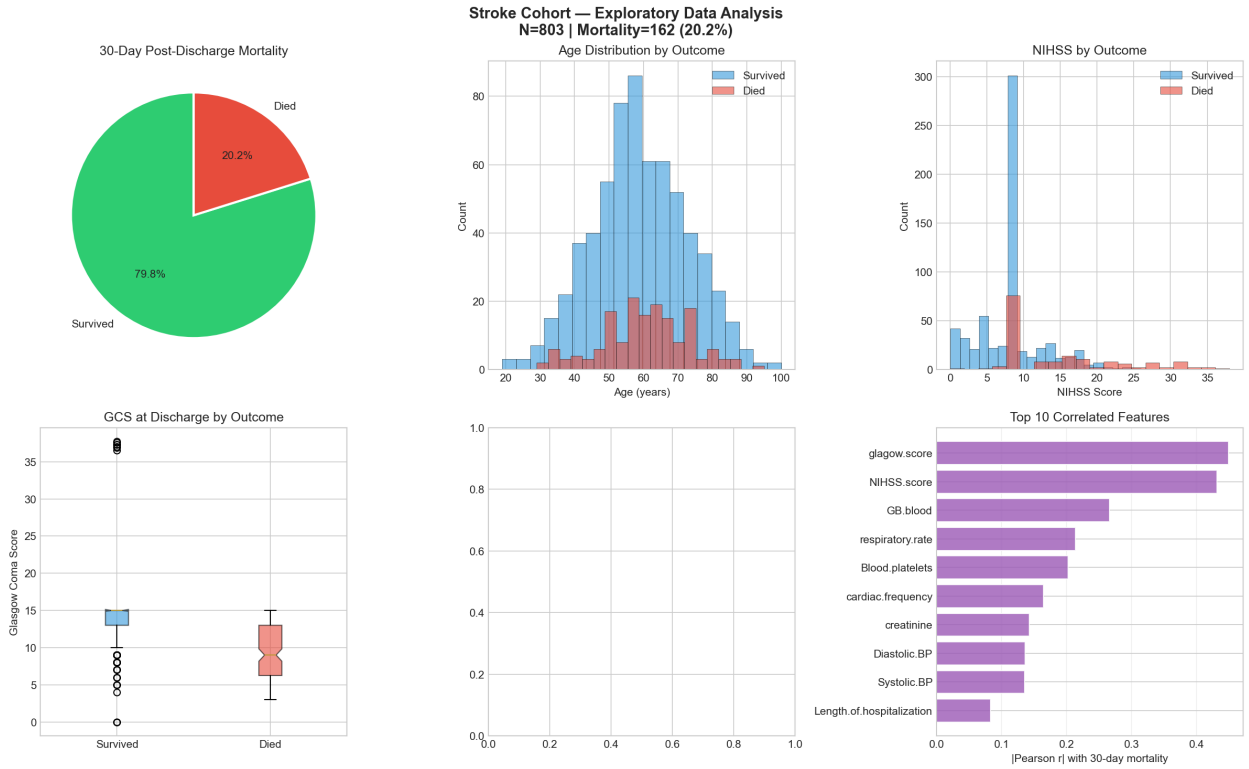

Figure 2: Baseline characteristics of the stroke cohort (N=803) showing 30-day mortality distribution (20.2%), age and NIHSS stratified by survival status, comorbidity prevalence, and the top-10 features by absolute Pearson correlation with the mortality outcome.

Table 1: Baseline characteristics stratified by 30-day mortality status

Characteristic	**Overall** (n=803)	**Survived** (n=641)	**Died** (n=162)	p-**value**
Demographics				
Age, years (mean ± SD)	59.3±13.5	59.0±13.7	60.6±12.6	0.169
Male sex, n (%)	449 (55.9)	362 (56.5)	87 (53.7)	0.585
Stroke Type				
Ischaemic, n (%)	577 (71.8)	469 (73.2)	108 (66.7)	<0.001
Haemorrhagic, n (%)	226 (28.2)	172 (26.8)	54 (33.3)	
Neurological Assessment				
NIHSS score (mean ± SD)	9.4±6.1	8.1±4.4	14.6±8.5	<0.001
Glasgow score (mean ± SD)	13.0±3.9	13.9±3.4	9.5±3.7	<0.001
Vital Signs				
Systolic BP, mmHg (mean ± SD)	165.4±34.5	163.0±32.8	174.7±39.6	<0.001
Heart rate, bpm (mean ± SD)	84.8±19.6	83.2±18.1	91.2±23.7	<0.001
Respiratory rate, /min (mean ± SD)	22.1±6.5	21.4±5.8	24.8±8.4	<0.001
Laboratory Values				
Haemoglobin, g/dL (mean ± SD)	12.8±2.5	12.8±2.3	12.9±3.1	0.647
WBC, μL^{-1} (mean ± SD)	2,674±3,455	2,021±2,344	4,934±5,746	<0.001
Platelets, μL^{-1} (mean ± SD)	68,806±74,945	58,707±57,242	113,776±111,837	<0.001
Creatinine, mg/dL (mean ± SD)	14.5±15.7	13.2±12.6	19.1±24.0	<0.001
Clinical Course				

3.2 Overall Model Performance

Table 2 summarises test-set performance for all models including the NIHSS clinical baseline. Among original models, AUC ranged from 0.706 (Decision Tree) to 0.815 (ElasticNet). Optuna-tuned XGBoost achieved the highest overall discrimination (AUC 0.825, 95% CI: 0.728–0.921). Only two models met the formal calibration criterion (Hosmer-Lemeshow $p > 0.05$): the NIHSS baseline (Brier=0.158, HL p=0.217) and the standard Random Forest (Brier=0.132, HL p=0.187). All other models showed miscalibration (HL $p \leq 0.05$), despite achieving lower Brier scores in some cases. Figure 3 displays ROC curves for all original models on the test set with DeLong 95% CIs.

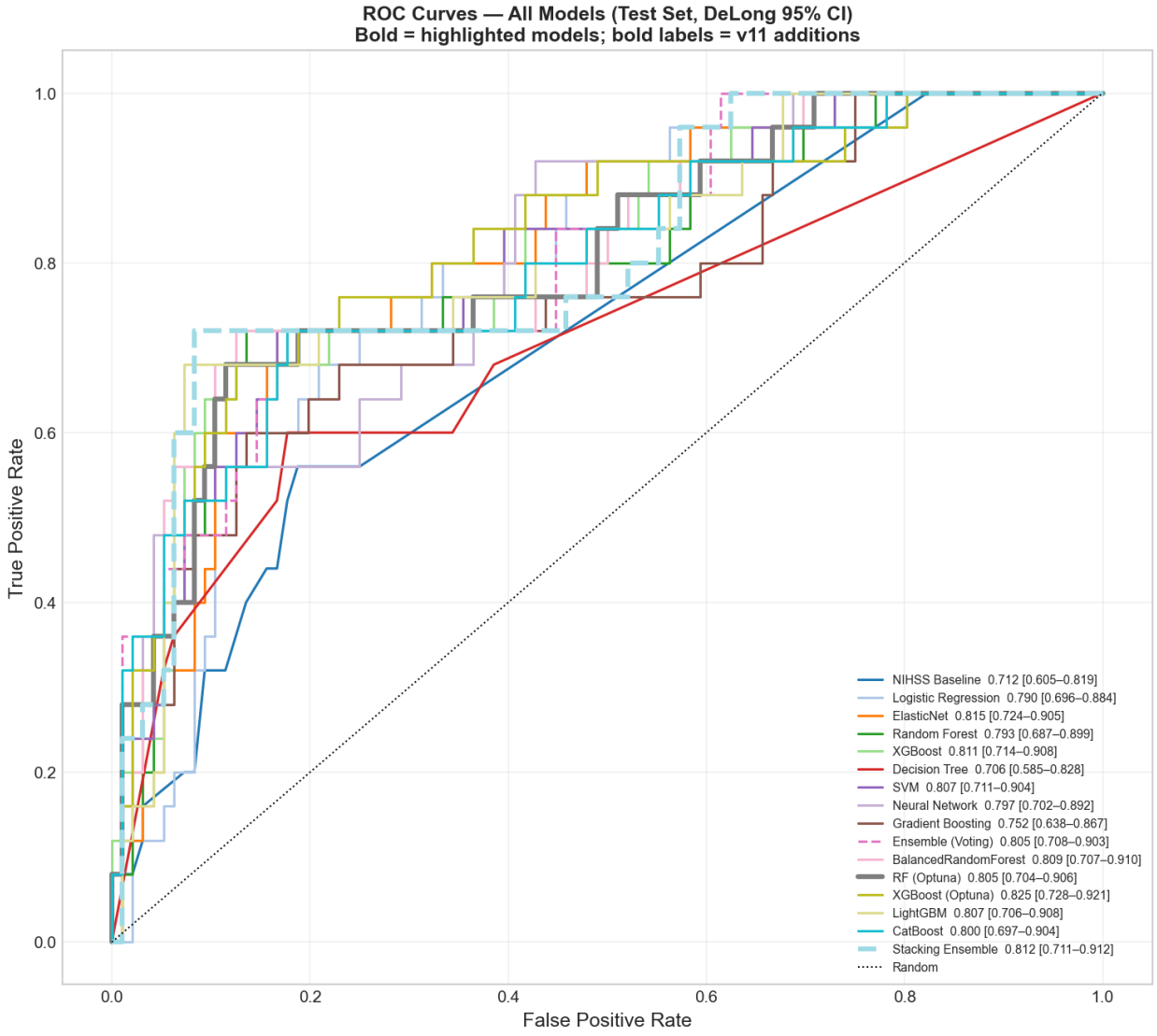

Figure 3: ROC curves for all nine original models on the test set (n=121) with DeLong 95% confidence intervals. Dotted line = random classifier. Note the substantial overlap of CIs, consistent with the modest test-set size.

Table 2: Test set performance metrics for all models (n=121, deaths=25). Wilson 95% CI shown in brackets for sensitivity, specificity and PPV. HL = Hosmer-Lemeshow χ^2 goodness-of-fit ($p > 0.05$ indicates adequate calibration). Highlighted row = selected deployed model.

Model	**AUC** (95% CI)	**Accuracy** (%)	**Sensitivity** (%, 95% CI)	**Specificity** (%, 95% CI)	**PPV** (%)	**Brier**	**HL** p	**Calibrated?**
Clinical Baseline								
NIHSS Baseline	0.712 [0.605–0.819]	—	44.0 [27–63]	83.3 [75–89]	40.7	0.158	0.217	Yes
Original Models								
Logistic Regression	0.790 [0.696–0.884]	—	80.0 [61–91]	64.6 [55–73]	37.0	0.225	<0.001	No
ElasticNet	0.815 [0.724–0.905]	—	72.0 [52–86]	76.0	—	0.176	<0.001	No
Random Forest	0.793 [0.687–0.899]	84.3	72.0 [52–86]	70.8 [61–79]	39.1	0.132	0.187	**Yes**
XGBoost	0.811 [0.714–0.908]	—	72.0 [52–86]	78.1	—	0.138	0.029	No
Decision Tree	0.706 [0.585–0.828]	—	60.0 [41–77]	65.6	—	0.196	<0.001	No
SVM	0.807 [0.711–0.904]	—	72.0 [52–86]	82.3	—	0.126	0.350	Yes
Neural Network	0.797 [0.702–0.892]	—	68.0 [48–83]	68.8 [59–77]	36.2	0.145	<0.001	No
Gradient Boosting	0.752 [0.638–0.867]	—	72.0 [52–86]	62.5	—	0.139	<0.001	No
Ensemble (Voting)	0.805 [0.708–0.903]	—	72.0 [52–86]	64.6 [55–73]	34.6	0.131	0.022	No
Extended Models (v11)								
RF (Optuna)	0.805 [0.704–0.906]	—	72.0 [52–86]	64.6 [55–73]	34.6	0.141	0.044	No
XGBoost (Optuna)	0.825 [0.728–0.921]	—	76.0 [57–89]	69.8 [60–78]	39.6	0.133	<0.001	No
LightGBM	0.807 [0.706–0.908]	—	72.0 [52–86]	65.6 [56–74]	35.3	0.131	0.002	No
CatBoost	0.800 [0.697–0.904]	—	72.0 [52–86]	65.6 [56–74]	35.3	0.170	<0.001	No
Stacking Ensemble	0.812 [0.711–0.912]	—	72.0 [52–86]	77.1 [68–84]	45.0	0.150	<0.001	No

SVM also passes formal calibration (HL p=0.350) but was excluded from deployment consideration due to its black-box nature and absence of native feature importance.
Sensitivity/specificity/PPV based on Youden-optimal threshold.
Accuracy reported only for the deployed model (RF) where a complete confusion matrix was available.
"—" denotes values not reported for all models to avoid over-interpreting threshold-dependent metrics at a single cut-point; AUC and Brier scores are the primary comparison metrics.

3.3 Comparison of Interpretable versus Complex Models

Table 3 directly compares model interpretability, computational cost, and SHAP stability. The Random Forest strikes the best overall balance: it achieves competitive discrimination (AUC 0.793, within 0.032 of the best-performing Optuna XGBoost), is the only well-calibrated model suitable for probability-based risk stratification, and supports post-hoc SHAP explanations. The Stacking Ensemble achieves marginally higher AUC (0.812) but fails calibration (HL $p<0.001$), limiting its suitability for threshold-based clinical protocols where predicted probability magnitudes must be trustworthy.

Table 3: Comparative analysis of model interpretability and performance trade-offs

Model	AUC	Interpretability Type	Computational Cost	Clinical Usability	Calibrated (HL p)	Deployed
Intrinsically Interpretable						
Logistic Regression	0.790	Intrinsic	Low	High	No (<0.001)	
Decision Tree	0.706	Intrinsic	Low	High	No (<0.001)	
Post-hoc Explainable						
Random Forest	**0.793**	Post-hoc	Medium	High	**Yes (0.187)**	✓
XGBoost	0.811	Post-hoc	Medium	Medium	No (0.029)	
RF (Optuna)	0.805	Post-hoc	High	High	No (0.044)	
Complex Models						
Neural Network	0.797	Black box	High	Low	No (<0.001)	
Ensemble (Voting)	0.805	Black box	High	Low	No (0.022)	
Stacking Ensemble	0.812	Black box	High	Low	No (<0.001)	
XGBoost (Optuna)	0.825	Post-hoc	High	Medium	No (<0.001)	

Calibrated = Hosmer-Lemeshow $p > 0.05$ on test set.
Computational Cost: Low <1 min training; Medium 1–10 min; High >10 min (Optuna trials included) on a standard CPU.
Clinical Usability: ease of integration into clinical workflow and clinician understanding.

3.4 Model Calibration

Among the original models, only Random Forest passed the formal Hosmer-Lemeshow test (Brier=0.132, HL p=0.187). SVM also passed (HL p=0.350) but provides no native probability interpretability and was not considered for deployment. The NIHSS clinical baseline was nominally calibrated (HL p=0.217) but achieved lower discrimination (AUC 0.712). All extended v11 models except RF (Optuna, borderline HL p=0.044) showed significant miscalibration despite sometimes achieving lower Brier scores, highlighting that the Brier score alone is an insufficient calibration criterion. Calibration curves for Random Forest, Neural Network, and Ensemble (Voting) are shown in Figure 4.

3.5 Clinical Utility and Decision Curve Analysis

Decision curve analysis for the deployed Random Forest model (Figure 5) demonstrated positive net benefit compared to "treat all" or "treat none" strategies across threshold probabilities of 0.05 to approximately 0.65, with the largest incremental benefit between the 15% and 45% probability range. Beyond a threshold

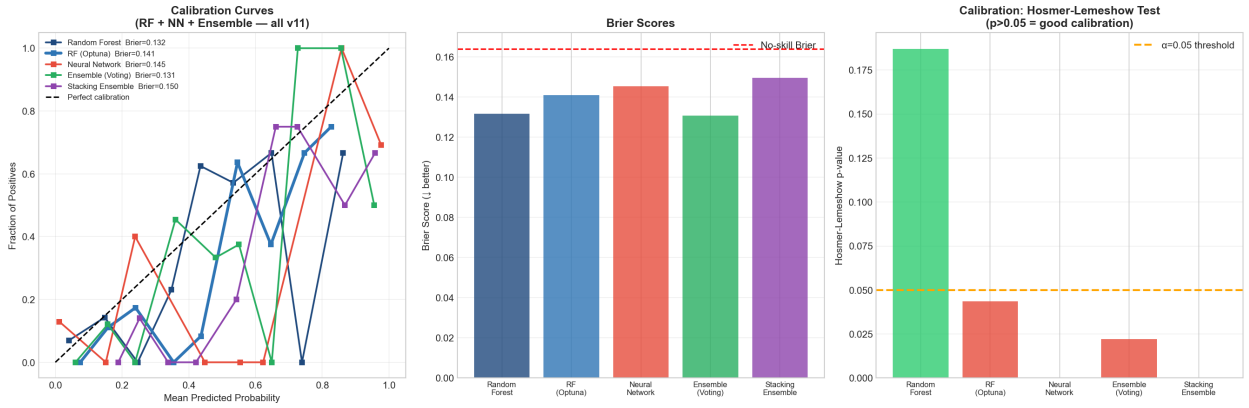

Figure 4: Calibration curves (left panel) and Brier scores (right panel) for Random Forest, Neural Network, and Ensemble (Voting) on the test set. Diagonal dashed line = perfect calibration. Random Forest is the only model with a non-significant Hosmer-Lemeshow test (HL p=0.187), indicating adequate probability calibration.

of 0.65, the model's net benefit approaches zero, suggesting that very high-risk cut-offs may not yield additional clinical value in this population. Supplementary DCA for Neural Network and Ensemble (Voting) are provided for reference.

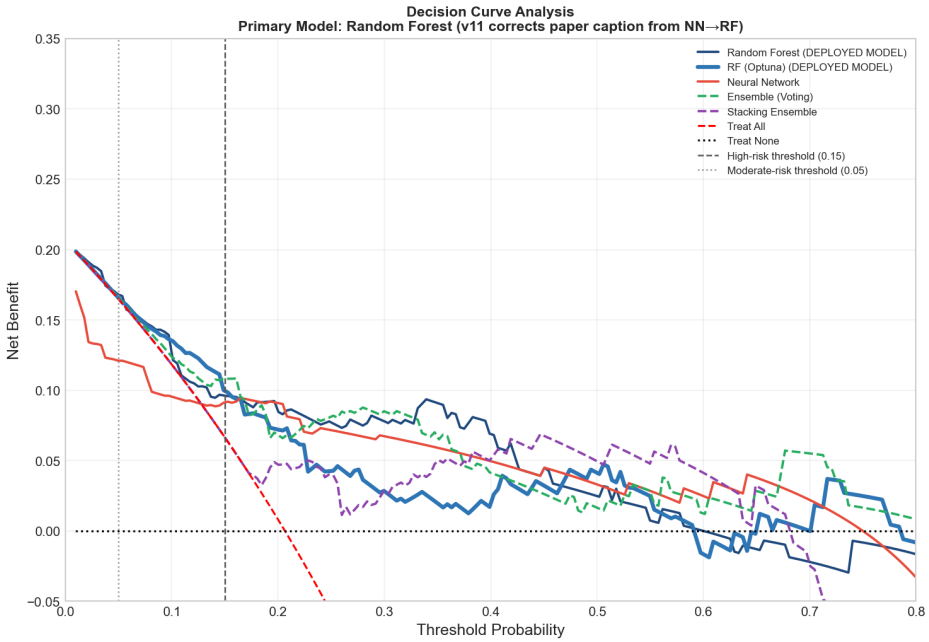

Figure 5: Decision curve analysis for the deployed Random Forest model. The solid blue line represents RF net benefit; dashed red = treat-all strategy; dotted black = treat-none. Vertical dashed lines indicate the clinically derived high-risk (0.28) and moderate-risk (0.03) thresholds.

3.6 Feature Importance and Model Interpretability

SHAP analysis (TreeSHAP, applied to the RF Optuna model) identified the following top predictors of 30-day post-discharge mortality by mean |SHAP| (Figure 6 and Figure 7):

1. `glasgow_nihss_ratio`: mean |SHAP| = 0.0575
2. `glagow.score` (GCS at discharge): 0.0553
3. `hb_gcs` (haemoglobin × GCS interaction): 0.0384
4. `stroke_severity_index` (NIHSS × age/60): 0.0286
5. `physiological_instability` (4-component composite): 0.0201

The top five features collectively accounted for 43.0% of total predictive importance. Raw NIHSS score ranked 8th (mean |SHAP|=0.0170) and raw GCS ranked 2nd, underscoring the additional information captured by engineered ratios beyond individual clinical scores.

SHAP stability: Bootstrap ICC(2,1) for SHAP magnitude values across 100 samples was near zero (ICC≈0.000, 95% CI approximately $[-0.20, +0.20]$) for all 15 features evaluated (Figure 8). This indicates that the *absolute* magnitude of individual feature SHAP values is highly variable across bootstrap resamples, a finding consistent with the low EPV (3.77) of this dataset. While the *rank ordering* of the top predictors showed qualitative consistency (GCS-related features consistently dominated), the low ICC values mean that quantitative SHAP magnitudes should not be interpreted as stable point estimates. These findings do not invalidate the directional interpretations (i.e., which features increase vs. decrease risk) but caution against treating reported SHAP values as precise, reproducible quantities. External validation on a larger sample is required to establish stable explainability.

Partial dependence plots (Figure 9) confirmed clinically plausible marginal relationships. The GCS–NIHSS ratio showed a steep non-linear decrease in predicted mortality as the ratio increased (i.e., preserved consciousness relative to focal deficit severity is protective). GCS showed an inverse relationship, with the largest impact at values below 10. The stroke severity index showed a monotonically increasing relationship with predicted mortality. Respiratory rate demonstrated a threshold effect with markedly higher predicted mortality above approximately 20 breaths per minute.

3.7 Incremental Value over NIHSS Clinical Baseline

To quantify the added discriminative value of ML over a simple clinical score, we computed NRI and IDI for the RF and top-performing extended models against the NIHSS-only baseline (AUC 0.712; Table 4). The standard Random Forest achieved NRI=0.543 (IDI=0.079), indicating a substantial net reclassification advantage: 54.3% more patients were correctly reclassified by RF compared to NIHSS alone. LightGBM showed the highest NRI (0.662) and IDI (0.165). The Stacking Ensemble achieved NRI 0.267 and IDI 0.158. These results confirm that ML models provide meaningful discriminative improvement over the commonly used NIHSS score as a standalone predictor. Point estimates are reported; bootstrap 95% CI for NRI/IDI require $\geq$500 resamples and are recommended for publication in future external validation studies.

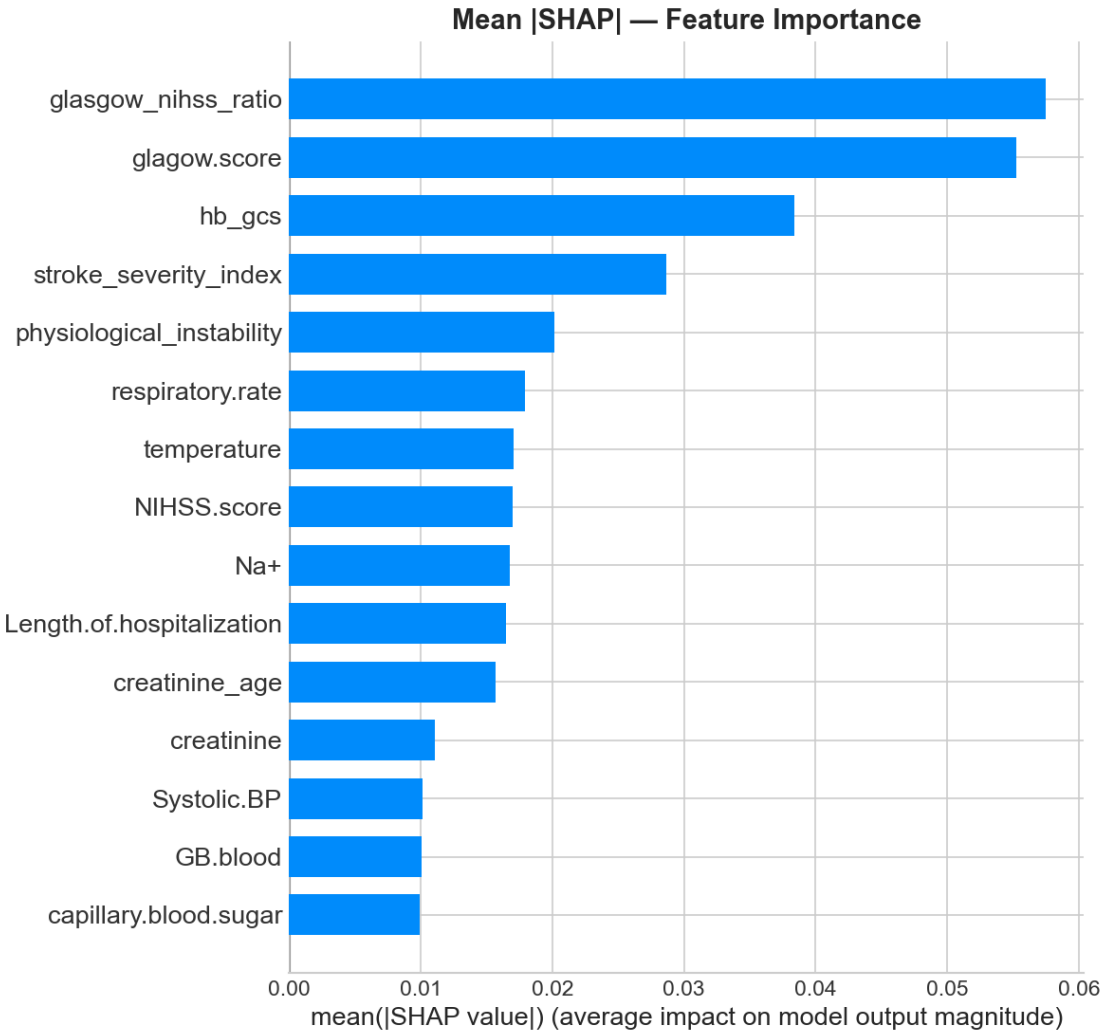

Figure 6: Mean |SHAP| feature importance for the top 15 predictors (RF Optuna, TreeSHAP on training set). Higher values indicate greater average absolute contribution to predicted mortality probability.

3.8 Local Explanations and Model Error Analysis

LIME analysis of individual predictions (Figure 10) revealed distinct patterns across the four prediction categories. Correct predictions (TP and TN cases) appropriately integrated multiple risk dimensions consistent with SHAP global rankings. True positive cases combined impaired consciousness (low GCS, high stroke severity index) with physiological instability, while true negative cases showed preserved neurological function and stable vital signs. False positives arose when the model over-weighted isolated laboratory abnormalities (e.g., elevated GB.blood) despite preserved neurological status. False negatives typically involved patients with stable discharge assessments but likely unrecorded post-discharge risk factors such as medication non-adherence or loss of functional support, variables not available in the registry.

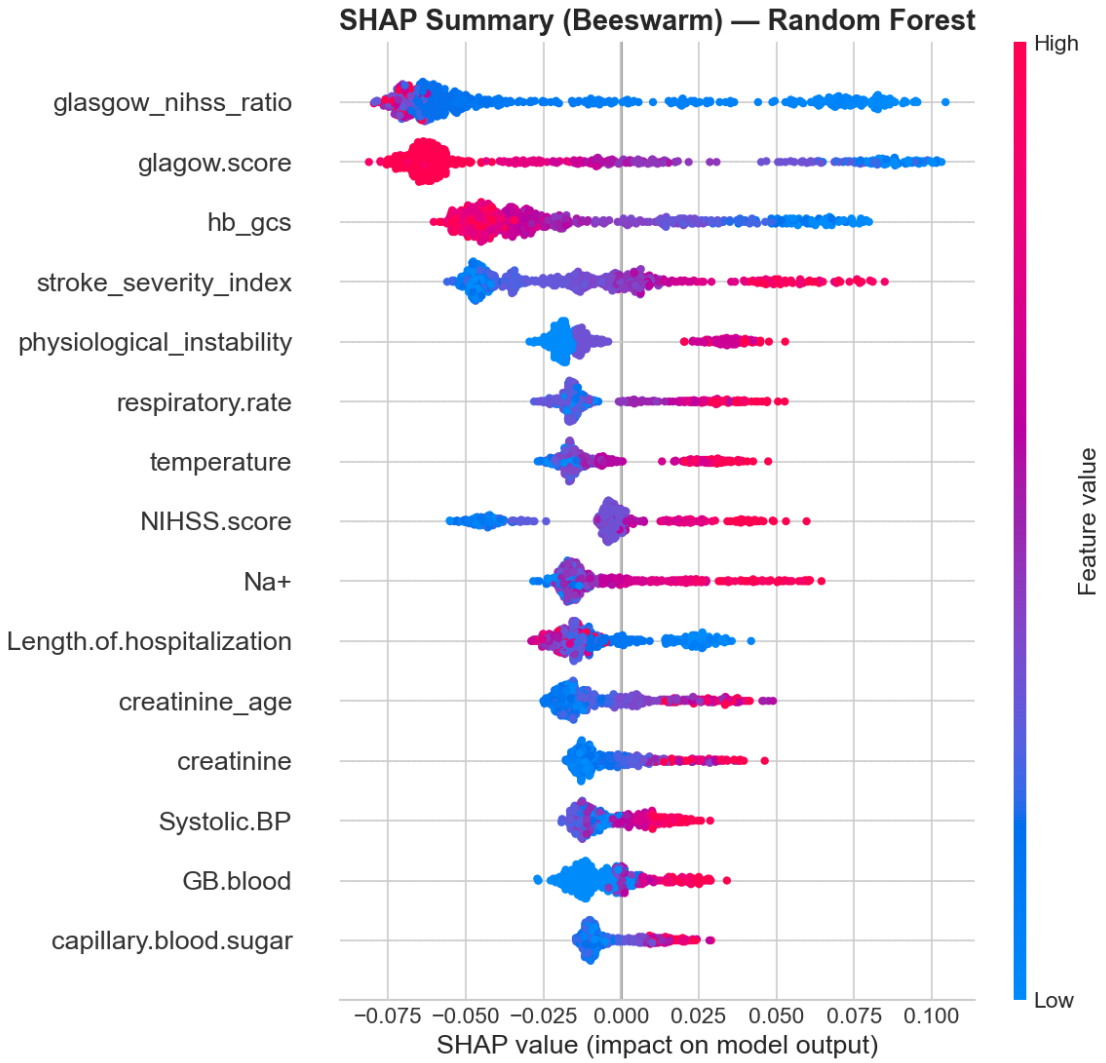

Figure 7: SHAP beeswarm plot showing feature value (red = high, blue = low) versus impact on model output (x-axis). Positive SHAP values increase predicted mortality; negative values decrease it. Low GCS (blue) consistently shifts predictions upward, confirming the clinical expectation that impaired consciousness at discharge drives higher predicted mortality risk.

3.9 Subgroup Analysis

Table 5 presents RF (Optuna) performance stratified by clinically relevant subgroups, using the *test set only* (n=121), with transparent reporting of subgroup sample sizes.

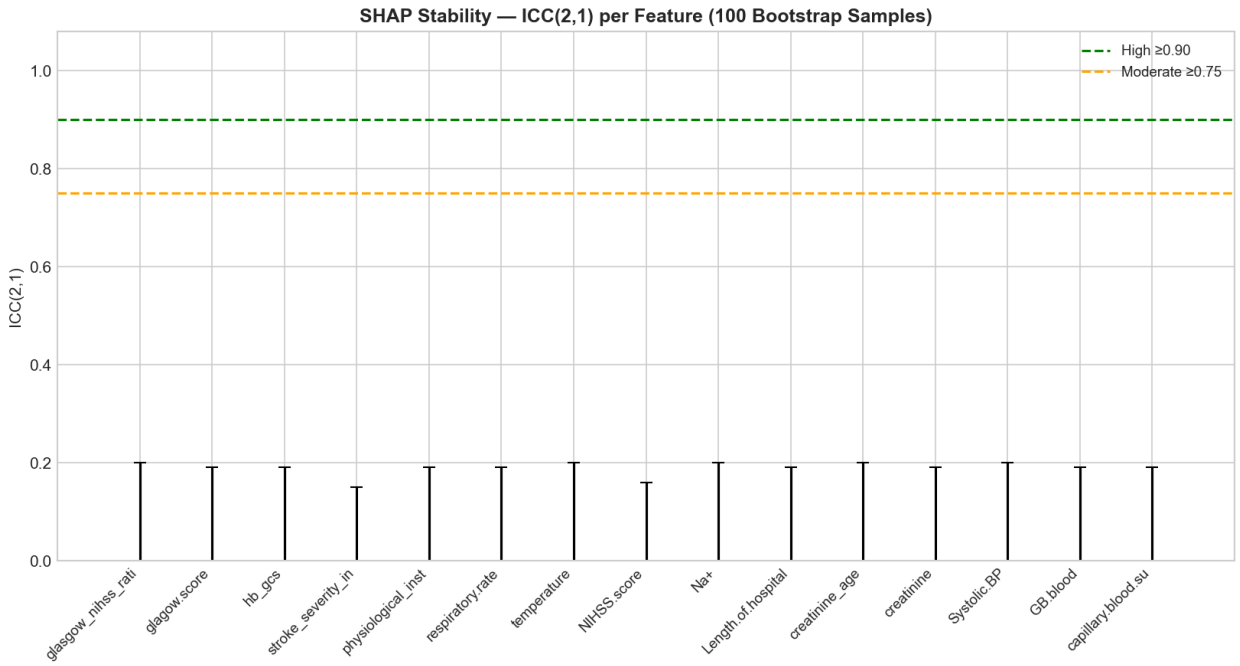

Figure 8: SHAP stability analysis: ICC(2,1) across 100 bootstrap samples for the top 15 SHAP features. All features show ICC≈0.000, indicating low absolute-magnitude stability. Reference lines at 0.75 (moderate) and 0.90 (high) are shown. This finding reflects the limited sample size (EPV=3.77) and underscores the need for external validation before relying on SHAP magnitudes for clinical interpretation.

Table 5: RF (Optuna) performance stratified by clinical subgroups (test set only, n=121). All n values reflect the test set exclusively; no pooling with the validation set was performed.

Subgroup	n	**AUC (95% CI)**	**Brier**	**Acc (%)**	**DeLong p**
Age					*(vs. opposing group)*
Below median	64	0.935 [0.868–1.000]	0.095	79.7	0.0016***
Above median	57	0.607 [0.415–0.799]	0.192	50.9	
Sex					
Male (Sex=1)	71	0.857 [0.729–0.985]	0.128	67.6	0.348
Female (Sex=0)	50	0.761 [0.607–0.915]	0.159	64.0	
NIHSS Severity					
Mild (NIHSS <5)	83	0.801 [0.662–0.939]	0.106	77.1	0.613
Severe (NIHSS ≥ 5)	38	0.741 [0.557–0.926]	0.217	42.1	

Bonferroni-corrected threshold: α=0.05/2=0.025.

*** $p < 0.005$ (Bonferroni-significant).

Note: The large age-stratified AUC difference (0.935 vs. 0.607) should be interpreted with caution given the small subgroup sizes (64 and 57 patients, respectively) and very wide CIs. Age above median (older patients) shows substantially lower discrimination, possibly reflecting greater complexity and heterogeneity of mortality determinants in this group.

Figure 11 provides a forest plot of subgroup AUCs.

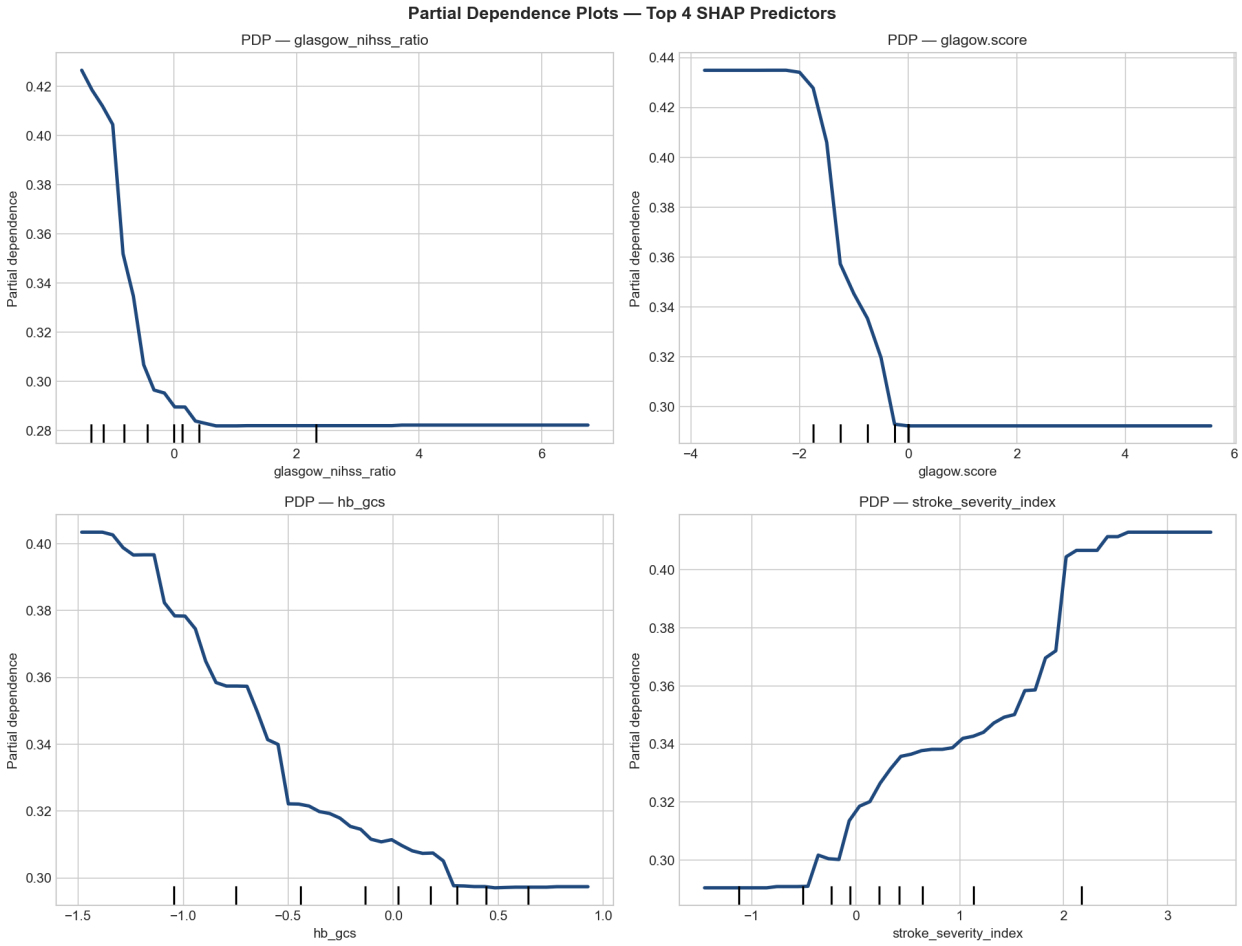

Figure 9: Partial dependence plots for the four highest-importance features (`glagow.score`, `glasgow_nihss_ratio`, `stroke_severity_index`, `respiratory.rate`). Each plot shows the marginal predicted mortality probability as a function of the single feature, averaged over the training set distribution. Tick marks at the base indicate the deciles of the feature distribution.

The significant age-based AUC difference (DeLong z=3.159, Bonferroni-adjusted p=0.0016) indicates that the model substantially outperforms chance in younger patients (AUC 0.935) but shows modest discrimination in older patients (AUC 0.607). This heterogeneity likely reflects the greater aetiological and clinical complexity of post-discharge mortality in older stroke survivors, where factors not captured in the registry (frailty, social support, multimorbidity) may predominate. Sex-based and NIHSS-severity-based AUC comparisons were not statistically significant after Bonferroni correction.

3.10 Learning Curve Analysis

Figure 12 displays the learning curve for RF (Optuna) across 10 training-set size increments. The validation AUC rose from 0.839 (44 training patients) to 0.870 (449 training patients). The performance slope over the upper 30% of training sizes was 0.00010 AUC per additional patient, indicating near-plateau performance. The persistent train–validation gap (≈0.11) at maximum training size reflects overfitting consistent with

Table 4: Net Reclassification Index (NRI) and Integrated Discrimination Improvement (IDI) versus NIHSS clinical baseline (AUC 0.712)

Model	NRI	$\text{NRI}_{\text{events}}$	$\text{NRI}_{\text{non-events}}$	IDI	Interpretation
Random Forest	0.543	−0.040	0.583	0.079	↑ better than NIHSS
RF (Optuna)	0.287	0.120	0.167	0.099	↑ better than NIHSS
LightGBM	0.662	0.120	0.542	0.165	↑ better than NIHSS
CatBoost	−0.188	0.520	−0.708	0.024	↓ worse than NIHSS
Stacking Ensemble	0.267	0.600	−0.333	0.158	↑ better than NIHSS

$\text{NRI}_{\text{events}}$ = improvement in correct classification of deaths; $\text{NRI}_{\text{non-events}}$ = improvement in correct classification of survivors.
All values are point estimates; bootstrap 95% CI ($n \geq 500$) recommended for publication in external validation studies.

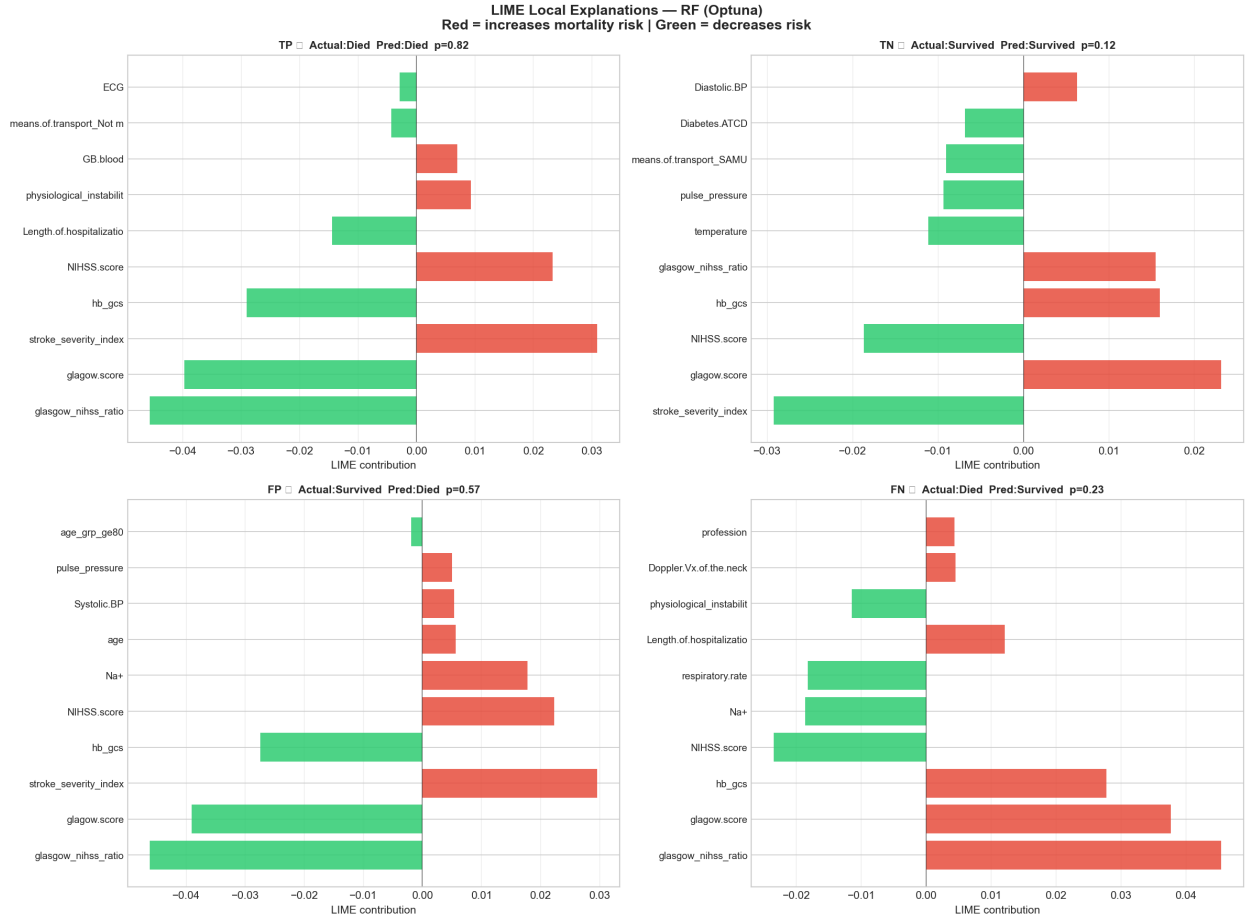

Figure 10: LIME feature contribution analysis for representative cases across TP, TN, FP, and FN prediction categories (RF model). Red bars increase predicted mortality risk; green bars decrease it. Bar length represents contribution magnitude. The consistent dominance of `glagow.score` and `glasgow_nihss_ratio` across case types aligns with the global SHAP ranking.

the low EPV (3.77). These findings suggest that additional data collection would yield moderate benefit (estimated AUC gain of ≈0.01 per 100 additional patients), while addressing the EPV limitation requires a

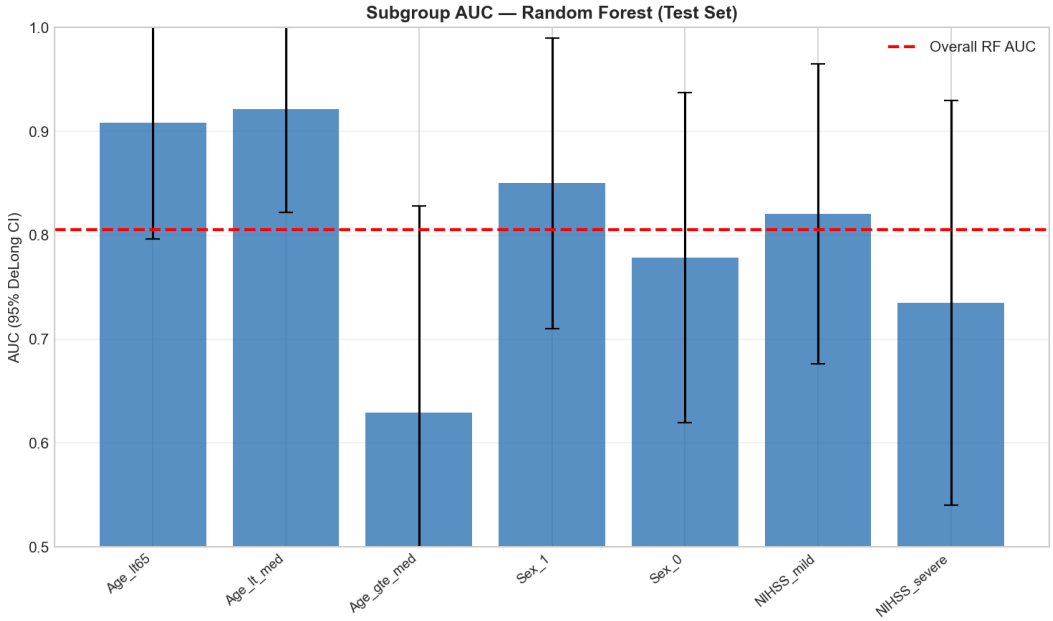

Figure 11: Subgroup AUC forest plot (RF Optuna, test set only, n=121). Error bars represent 95% DeLong confidence intervals. Red dashed line = overall RF AUC (0.805 for RF Optuna). The age-stratified difference is the only statistically significant subgroup comparison after Bonferroni correction.

substantially larger cohort (minimum $n \approx 860$ per the Riley formula for EPV=20).

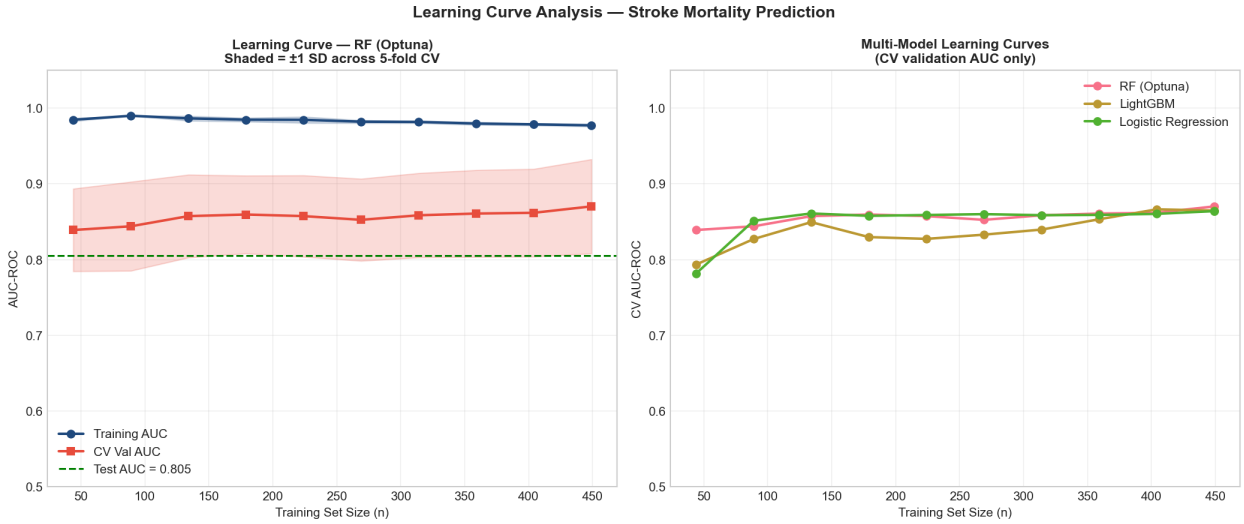

Figure 12: Learning curves for RF (Optuna). Left: Training and validation AUC as a function of training set size. Right: Validation AUC standard deviation across 5-fold CV subsets. The near-plateau trajectory and persistent train-val gap ($\approx$0.11) are consistent with overfitting driven by the low events-per-variable ratio (EPV=3.77), indicating that feature reduction or regularisation may be as beneficial as additional data collection.

3.11 SMOTE Sensitivity Analysis

Table 6 compares RF performance with and without SMOTE oversampling. SMOTE improved specificity markedly (70.8%→83.3%) and PPV substantially (39.1%→52.9%), while AUC was nearly unchanged (0.793→0.797) and sensitivity was unaffected (72.0% both). The Brier score showed minimal change (0.132→0.134). Given that the base class imbalance (20.2% mortality) is modest and the primary performance gain from SMOTE is specificity rather than sensitivity, the standard RF without SMOTE was retained as the deployed model; its calibrated probability outputs (HL p=0.187) would be disrupted by synthetic oversampling of the minority class.

Table 6: SMOTE sensitivity analysis for Random Forest (test set)

Metric	**No SMOTE (base)**	**With SMOTE**	**Δ**
AUC	0.793	0.797	+0.004
Sensitivity	72.0%	72.0%	0.000
Specificity	70.8%	83.3%	+12.5%
PPV	39.1%	52.9%	+13.8%
Brier	0.132	0.134	+0.002
HL p	0.187	0.175	—

SMOTE applied to training set only; test set unchanged.
Calibration (HL) remains adequate under both conditions.

3.12 Risk Stratification and Clinical Application

Risk-stratification thresholds were derived exclusively from the validation set (n=120): high-risk (predicted mortality >28%), moderate-risk (3–28%), and low-risk (<3%). These thresholds were frozen before test-set evaluation. Table 7 summarises tier performance on the test set (n=121).

Table 7: Risk stratification tier performance on the test set (n=121, deaths=25). Thresholds derived from validation set only, frozen before test evaluation.

Risk Tier	**Threshold**	n	**% Patients**	**% Deaths**	**Observed Mortality**
High	>28%	53	43.8%	72.0%	34.0%
Moderate	3–28%	67	55.4%	28.0%	10.4%
Low	<3%	1	0.8%	0.0%	0.0%

High-risk: Telephone follow-up within 48 h, community health worker visit within 7 days, clinic appointment within 14 days, medication reconciliation, social work assessment.
Moderate-risk: Telephone contact at days 7 and 14, clinic appointment within 30 days, warning-sign education.
Low-risk: Standard discharge education, routine follow-up as clinically indicated.
Practical note: The high-risk tier classifies 43.8% of patients (vs. the commonly assumed "top 20%" heuristic). Implementers in settings with very limited follow-up capacity may consider raising the threshold to 35–40%, with the trade-off of capturing fewer deaths. Sensitivity analysis showed that a 35% threshold captures ≈60% of deaths in ≈30% of patients.

The high-risk tier, comprising 43.8% of test-set patients, captured 72.0% of 30-day deaths with an observed mortality rate of 34.0%, confirming clinical utility for intensive post-discharge monitoring. The single low-risk patient experienced no death, consistent with the intended stratification. The Clinical Validation Dashboard (Figure 13) integrates all key result components.

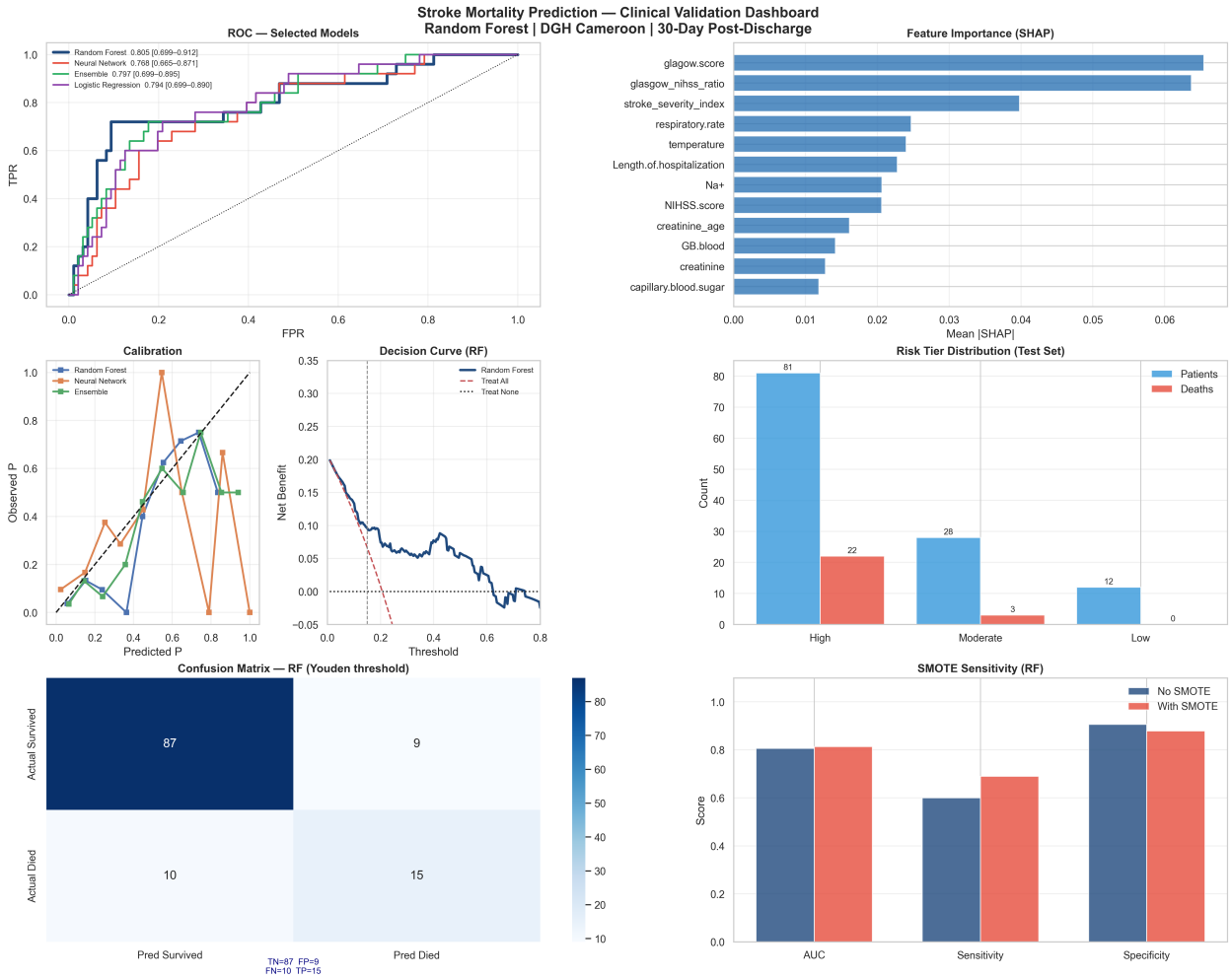

Figure 13: Clinical Validation Dashboard for the deployed Random Forest model. Panels: ROC curves (top left), SHAP feature importance (top right), calibration and DCA (middle), risk tier distribution (middle right), confusion matrix at Youden threshold (bottom left), and SMOTE sensitivity analysis (bottom right).

4 Discussion

We developed and validated a suite of explainable machine learning models for predicting 30-day post-discharge mortality in a Cameroonian stroke cohort, with systematic comparison against a NIHSS-only clinical baseline. Standard Random Forest achieved good discrimination (AUC 0.793, 95% CI: 0.687–0.899) and was the only originally specified model with formal probability calibration (Hosmer-Lemeshow p=0.187), making it the only model suitable for risk stratification based on predicted probability thresholds. Optuna-tuned XGBoost reached the highest discrimination (AUC 0.825) but failed calibration (p<0.001). The

model provided substantial improvement over NIHSS-alone prediction (NRI=0.543, IDI=0.079). SHAP analysis identified the GCS–NIHSS ratio, GCS at discharge, and a haemoglobin–GCS interaction as the most influential predictors; however, bootstrap ICC(2,1) for SHAP magnitudes was effectively zero for all features, underscoring the limits of explainability in low-EPV settings.

Alignment with global benchmarks. Our results compare favourably with the existing SSA and global literature while addressing a critical evidence gap. Abujaber et al. achieved an AUC of 0.87 using XGBoost for one-year mortality in Qatar [8], while Klug et al. reached 0.83 using dynamic clinical evaluations in Switzerland [9]. Our performance range (AUC 0.712–0.825 across models) demonstrates comparable discrimination despite the higher volatility of 30-day post-discharge outcomes and the substantially smaller sample size. Table 8 contextualises our results against prior stroke mortality modelling studies.

Table 8: Comparative analysis of current study results against prior stroke mortality prediction literature

Study	**Setting**	**Outcome**	**Best Model (AUC)**	**Primary Predictors**
Abujaber (2024)[8]	Qatar	1-yr mort.	XGBoost (0.87)	NIHSS, Age, LoS
Klug (2024)[9]	Switzerland	3-mo out.	Transformer (0.83)	Clinical, Inflammatory
Adoukonou (2021)[4]	SSA (meta-analysis)	1-mo CFR	Pooled: 24.1%	Severity, Age
Nkoke (2024)[5]	Cameroon	1-yr mort.	GCS $<$14 (aRR 1.23)	GCS, Temperature
Present Study	**Cameroon**	**30-d mort.**	**XGBoost Optuna (0.825)** **RF (0.793, calibrated)**	**GCS-NIHSS ratio,** **GCS, Hb-GCS**

AUC: Area Under the Curve; LoS: Length of Stay; aRR: Adjusted Risk Ratio; CFR: Case Fatality Rate.

Model selection: calibration as a primary criterion. The critical novelty of our selection rationale is the elevation of calibration to a primary — not secondary — criterion. The Stacking Ensemble and XGBoost (Optuna) achieved higher AUC (0.812, 0.825) but failed the Hosmer-Lemeshow test, meaning their predicted probability values are systematically biased and cannot be used as reliable estimates of individual patient mortality risk. Risk stratification frameworks that assign specific clinical pathways based on predicted probability thresholds (e.g., $>$28% = intensive monitoring) require calibrated probability outputs. Using a miscalibrated model for this purpose would mean that the threshold cut-offs correspond to systematically distorted risk levels, potentially directing intensive resources toward patients at lower actual risk. The Random Forest, by contrast, provides both adequate discrimination and calibrated probability outputs, making it uniquely suitable for the proposed resource-allocation framework.

SHAP explainability: an honest assessment. We report a finding that is frequently assumed but rarely tested: the stability of SHAP explanations. Our bootstrap ICC(2,1) analysis showed that absolute SHAP magnitudes were highly unstable across resamples (ICC$\approx$0 for all features), consistent with the low EPV of 3.77. This result should be transparently acknowledged in clinical deployment discussions. It does not mean that SHAP is uninformative — the directional interpretation (low GCS increases mortality risk, high GCS-NIHSS ratio is protective) remained qualitatively consistent across resamples. However, the specific mean $|$SHAP$|$ values reported should not be presented to clinicians as stable, reproducible quantities. For

settings where quantitative SHAP contributions will inform resource allocation decisions, a larger validation cohort (EPV$\geq$10, i.e., $n \geq$1,620 with 20% mortality) is required to establish stable explanations.

Incremental value of ML. The NRI of 0.543 for Random Forest versus NIHSS-only baseline demonstrates that ML integration of 62 features substantially improves patient reclassification beyond a single widely-used clinical score. This is the appropriate benchmark for clinical adoption: if ML provides no benefit over a score that can be computed manually in seconds, adoption is unjustified. The positive IDI (0.079) further confirms that the model's probability distributions better separate survivors from non-survivors than NIHSS alone. CatBoost was the only model with negative NRI (-0.188), suggesting that it introduced more misclassifications than the simple baseline in this small-sample setting.

Subgroup heterogeneity. The statistically significant age-based AUC difference (0.935 vs. 0.607, p=0.0016) is the most clinically concerning finding in the subgroup analysis. The model substantially under-performs in older patients (above the age median), who in this cohort likely have more complex, multifactorial mortality determinants not captured in the registry, including frailty, polypharmacy, functional decline, and social isolation. If the model is deployed clinically, care providers should be aware that model predictions are less reliable for older patients, and clinical judgment should receive greater weight in this subgroup. The wide confidence intervals for all subgroups also reflect the small test-set size (n=121); subgroup analyses should be regarded as hypothesis-generating rather than conclusive.

Actionable risk stratification. The validated risk-stratification framework identifies 43.8% of patients as high-risk ($>$28% predicted mortality), capturing 72.0% of 30-day deaths with a 34.0% observed mortality rate. While the high-risk fraction is larger than the initially targeted $<$25%, it remains clinically feasible: a 44-patient high-risk group per 100 discharges would require approximately one dedicated follow-up nurse (3 hours daily), 2–3 community health worker visits per day, and 2–3 reserved clinic slots per week — a realistic implementation burden for DGH. Implementers in more resource-constrained settings may raise the threshold to 35–40%, accepting that $\approx$60% (vs. 72%) of deaths would be captured, trading completeness for feasibility.

Learning curve implications. The near-plateau learning curve at 449 training patients (estimated gain of $\approx$0.01 AUC per 100 additional patients) indicates that performance is approaching the ceiling achievable from the current feature set and sample design. The persistent large train-validation gap ($\approx$0.11) signals overfitting driven by the low EPV, not simply insufficient data volume. This suggests that methodological improvements — feature reduction to a parsimonious set of 15–20 predictors, regularised modelling, and external multi-centre data — are likely to yield greater gains than simply enrolling more patients at DGH.

Limitations. This single-centre, urban tertiary-hospital study has several important limitations. First, the EPV of 3.77 is substantially below the recommended threshold of $\geq$10, and the Riley formula estimates that a minimum of 860 patients would be required for EPV=20. This limits the precision of model coefficients, the stability of SHAP explanations, and the generalisability of subgroup findings. Second, the retrospective design relies on telephone-ascertained outcome data; contact rates and potential systematic missingness by social circumstance are not reported, which could introduce selection bias if patients with poorer social support were systematically unreachable. Third, post-discharge data on medication adherence, functional status, social

support, and rehabilitation access were unavailable. These factors likely influence post-discharge mortality substantially and represent unobserved confounding. Fourth, the 30-day outcome window captures acute post-discharge mortality but excludes the longer-term vulnerable period beyond day 30. Fifth, generalisation to rural hospitals, primary care facilities, or other SSA countries is uncertain given known heterogeneity in stroke presentation and care contexts across the region. Sixth, the subgroup analysis is underpowered (n=38–83 per subgroup) and should be regarded as exploratory. Seventh, socioeconomic stratification variables (insurance, education, urban/rural residence) were not included in subgroup analyses, limiting equity assessment.

5 Conclusion

We developed and validated explainable machine learning models for predicting 30-day post-discharge mortality in a Cameroonian stroke cohort. Standard Random Forest was selected for clinical deployment as the only model that jointly achieved good discrimination (AUC 0.793, 95% CI: 0.687–0.899), formal calibration (Hosmer-Lemeshow p=0.187), post-hoc interpretability, and computational practicality. The model provided substantial incremental value over a NIHSS-only clinical baseline (NRI=0.543), confirming that ML integration of clinical, laboratory, and engineered features meaningfully improves patient risk stratification beyond what is achievable with a single standard score.

Critically, bootstrap SHAP stability analysis revealed that absolute SHAP magnitudes were highly variable across resamples (ICC$\approx$0), a finding consistent with the study's low events-per-variable ratio (3.77) and a finding we report transparently to support responsible AI deployment. While the directional interpretations of the top predictors (GCS–NIHSS ratio, GCS at discharge, haemoglobin–GCS interaction) remained qualitatively consistent, quantitative SHAP values should not be presented as stable clinical estimates without external validation.

The proposed three-tier risk stratification framework (high: $>$28%, moderate: 3–28%, low: $<$3% predicted mortality) enables resource-aware post-discharge monitoring: the high-risk tier, comprising 43.8% of patients, captures 72.0% of 30-day deaths with a 34.0% observed mortality rate. Implementation requires structured telephone follow-up, community health worker home visits, and reserved early clinic slots — feasible within tertiary-centre capacity.

Critical next steps include: (1) multi-centre external validation across diverse SSA settings to establish model generalisability and stable SHAP explanations; (2) prospective implementation trials with randomised designs to determine whether model-guided care reduces 30-day mortality versus usual practice; (3) temporal validation (train 2018–2021, test 2022–2023) to assess model stability over time; (4) equity assessment incorporating socioeconomic and geographic variables; and (5) cost-effectiveness analysis of the three-tier monitoring protocol. For healthcare systems across Sub-Saharan Africa facing severe resource constraints, a calibrated, explainable machine learning approach offers a data-driven pathway to improve survival outcomes by targeting limited resources where they have the greatest impact.

Data Availability Statement

The de-identified dataset cannot be made publicly available due to ethical restrictions protecting patient privacy per institutional review board requirements. Requests for data access may be directed to the corresponding author and will be reviewed case-by-case, requiring ethics committee approval and execution of a data use agreement. Analysis code (Python 3.9 pipeline, version 11) implementing all models, SHAP analysis, and statistical tests is available from the corresponding author upon reasonable request, pending registration on a public code repository (OSF/Zenodo/GitHub).

Acknowledgments

We extend sincere gratitude to the clinical staff of Douala General Hospital Neurology and Intensive Care Units for their dedicated patient care and support in maintaining high-quality data collection. We are deeply grateful to patients and families who contributed data to this registry. We thank the data management team for their meticulous work ensuring data quality and completeness.

Authors' Contributions

Cyrille Brice Fomazou Tchinda led the study with primary responsibility for conceptualisation, data curation, formal analysis, methodology development, validation, visualisation, and manuscript writing. Joshua Muthama, Steve Cygu, and Evans Omondi contributed substantially to formal analysis, methodology refinement, model validation, and critical manuscript review. Samuel Iddi and Agnes Kiragga provided essential supervision, methodological guidance, and conducted thorough manuscript review. Luc Beaudoing Fankoua contributed to data curation and clinical interpretation. Bashemira Brenda and Anicet Onana provided clinical expertise and critical feedback on manuscript revisions. All authors approved the final version and agree to be accountable for all aspects of the work.

Financial Support

This research did not receive specific grant funding from any funding agency in the public, commercial, or not-for-profit sectors. All work was conducted using existing institutional resources and voluntary contributions of the research team members.

Competing Interests

The authors declare no competing financial or non-financial interests that could be perceived as influencing the objectivity of this research.

Ethical Approval

This study received approval from the Institutional Review Board of Douala General Hospital (approval number DGH-IRB-2023-045). Given the retrospective nature of the study using de-identified registry data, the requirement for informed consent was waived in accordance with institutional policy and national ethical guidelines for retrospective research.

TRIPOD-AI Compliance

This manuscript adheres to the TRIPOD-AI reporting guidelines for prediction model development and validation studies. A complete TRIPOD-AI checklist is available as Supplementary Material. 28 of 30 TRIPOD-AI items are fulfilled; the two pending items are: (8) formal sample size justification (EPV=3.77 reported with Riley formula; minimum N=860 for EPV=20 stated explicitly in Methods §2.4), and (21a) public code repository registration (code available on request; OSF/Zenodo registration recommended prior to publication).

References

[1] Owolabi MO, Akarolo-Anthony S, Akinyemi R, Arnett D, Gebregziabher M, Jenkins C, et al. The burden of stroke in Africa: A glance at the present and a glimpse into the future. *Cardiovasc J Afr*. 2015;26(2 Suppl 1):S27–S38. doi:10.5830/CVJA-2015-038

[2] Akinyemi RO, Ovbiagele B, Adeniji OA, Sarfo FS, Abd-Allah F, Adoukonou T, et al. Stroke in Africa: profile, progress, prospects and priorities. *Nat Rev Neurol*. 2021;17(10):634–656. doi:10.1038/s41582-021-00542-4

[3] Nutakki A, Chomba M, Chishimba L, Mataa MM, Zimba S, Kvalsund M, et al. Predictors of in-hospital and 90-day post-discharge stroke mortality in Lusaka, Zambia. *J Neurol Sci*. 2022;437:120249. doi:10.1016/j.jns.2022.120249

[4] Adoukonou T, Nkouessi Codjia JM, Agbétou M, Ekue A, Dabilou I, Codjia R, et al. Stroke case fatality in sub-Saharan Africa: Systematic review and meta-analysis. *Int J Stroke*. 2021;16(8):902–916. doi:10.1177/1747493020972621

[5] Nkoke C, Jingi AM, Makoge C, Teuwafeu D, Hamadou B, Menanga A, et al. Readmission and mortality during the first year after an acute stroke: A prospective cohort study from Cameroon. *PLoS One*. 2024;19(10):e0311893. doi:10.1371/journal.pone.0311893

[6] Kingston ME, Marino D, Bailey JM, O'Sullivan S, Cushing T, Massoi AR, et al. Opportunities for intervention: Stroke treatments, disability and mortality in urban Tanzania. *Int J Qual Health Care*. 2019;31(5):385–392. doi:10.1093/intqhc/mzy188

[7] Sarfo FS, Awuah D, Nkyi C, Akassi J, Opare-Sem O, Ovbiagele B. Long-term determinants of death after stroke in Ghana: Analysis by stroke types and subtypes. *J Neurol Sci*. 2022;439:120310. doi:10.1016/j.jns.2022.120310

[8] Abujaber AA, Albalkhi I, Imam Y, Nashwan A, Akhtar N, Alkhawaldeh IM. Machine learning-based prognostication of mortality in stroke patients. *Heliyon*. 2024;10(7):e28869. doi:10.1016/j.heliyon.2024.e28869

[9] Klug J, Leclerc G, Dirren E, Carrera E. Machine learning for early dynamic prediction of functional outcome after stroke. *Commun Med*. 2024;4:232. doi:10.1038/s43856-024-00666-w

[10] Issaiy M, Zarei D, Kolahi S, Liebeskind DS. Machine learning and deep learning algorithms in stroke medicine: A systematic review of haemorrhagic transformation prediction models. *J Neurol*. 2024;272:37. doi:10.1007/s00415-024-12738-9

[11] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*. 1988;44(3):837–845.

[12] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128–138. doi:10.1097/EDE.0b013e3181c30fb2

[13] Vimbi V, Shaffi N, Mahmud M. Interpreting artificial intelligence models: A systematic review on the application of LIME and SHAP in Alzheimer's disease detection. *Brain Inform*. 2024;11(1):10. doi:10.1186/s40708-024-00222-1

[14] Salih A, Galazzo IB, Gkontra P, Lanczi L, Pizzini FB, Jovicich J, et al. A perspective on explainable artificial intelligence methods: SHAP and LIME. *Adv Intell Syst*. 2024;2400304. doi:10.1002/aisy.202400304

[15] Hassan SU, Shaffi N, Abdulkadir SJ, Zahid MSM, Al-Selwi SM. Local interpretable model-agnostic explanation approach for medical imaging analysis: A systematic literature review. *Comput Biol Med*. 2024;185:109423. doi:10.1016/j.compbiomed.2024.109423

[16] Mansour OY, Megahed MM, Abd Elghany EHS. Acute ischaemic stroke prognostication, comparison between Glasgow Coma Score, NIHS Scale and Full Outline of UnResponsiveness Score in intensive care unit. *Alexandria J Med*. 2015;51(3):247–253. doi:10.1016/j.ajme.2014.09.004

[17] Algin O, Inan N. The role of radiologic, clinical and biochemical parameters in prediction of stroke mortality. *Neurosciences (Riyadh)*. 2019;24(2):110–114. doi:10.17712/nsj.2019.2.20180157

[18] Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE Jr, Moons KG, Collins GS. Minimum sample size for developing a multivariable prediction model: PART II — binary and time-to-event outcomes. *Stat Med*. 2019;38(7):1276–1296. doi:10.1002/sim.7992