Explainable Machine Learning for Predicting Mortality and Functional Outcomes After Ischemic Stroke Using Routine Clinical and Biochemical Data

Ambarish Kumar Sinha[1], Shachi Mall[2], Rupita Ghosh[3], Kaushalendra Kumar[1], Gaurav Kumar[4*]

[1,4]Department of Biomedical Sciences, School of Biosciences & Technology, Galgotias University, Greater Noida, Uttar Pradesh, India

[2]Department of Computer Sciences, School of Computing Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India

[3,4]Department of Biotechnology & Bioengineering, School of Biosciences & Technology, Galgotias University, Greater Noida, Uttar Pradesh, India

*Email: gauravkumar.sbme@galgotiasuniversity.edu.in

Abstract

Background: Stroke is a leading global cause of death and disability. Cases are rising in India. Conventional models, such as NIHSS, often overlook the complex clinical interactions that influence outcomes. We built and validated explainable machine learning models using hospital data. Our goal was better to predict functional outcomes and mortality in ischemic stroke patients.

Methods: We retrospectively analysed 320 imaging-confirmed ischemic stroke patients. We used 32 demographic (e.g., age, sex), clinical (e.g., stroke severity), biochemical (e.g., blood glucose), and treatment (e.g., thrombolysis) variables as predictors. We trained Logistic Regression, Random Forest, and XGBoost (extreme gradient boosting) models using nested five-fold cross-validation, with training and testing performed across multiple repeat data splits. We compared models based on the area under the receiver operating characteristic curve (AUROC), calibration (agreement between predicted and observed outcomes), sensitivity (ability to identify true positives), and F1-score (a balance between precision and recall). We used SHAP (Shapley Additive exPlanations) values to explain model predictions and highlight key prognostic factors.

Results: XGBoost achieved the highest discrimination for mortality prediction (measured by area under the receiver operating characteristic curve, AUROC = 0.90) and outperformed Random Forest (AUROC = 0.88) and Logistic Regression (AUROC = 0.84). For functional outcome prediction (modified Rankin Scale [mRS] ≤ 2 vs. >2), XGBoost also performed best (AUROC = 0.87). SHAP (SHapley Additive exPlanations) analysis identified key predictors, including NIHSS score, age, C-reactive protein (CRP), D-dimer, HbA1c, serum creatinine, and the use of thrombolysis.

Conclusion: Explainable machine learning models can accurately predict mortality and functional outcomes in ischemic stroke. These models integrate clinical (such as severity scores), biochemical (lab test results), and treatment data to provide a comprehensive understanding of the disease. They pave the way for practical AI-assisted tools tailored for Indian healthcare. This approach supports early risk assessment and personalised management in stroke care.

Keywords: Ischemic Stroke, Machine Learning, Explainable AI, Mortality, Modified Rankin Scale, Prognostic Modelling.

How to cite this article: Sinha AK, Mall S, Ghosh R, Kumar K, Kumar G. Explainable machine learning for predicting mortality and functional outcomes after ischemic stroke using routine clinical and biochemical data. Int J Drug Deliv Technol. 2026;16(79s): 188-205. DOI: 10.25258/ijddt.16.79s.21

Introduction

Stroke represents a leading global cause of mortality and disability, resulting in approximately 6.5 million deaths each year. According to the 2021 Global Burden of Disease Study, there were over 12 million new stroke cases, with 70–80% classified as ischemic. Despite advances in treatment, stroke continues to impose substantial health and socioeconomic burdens.

The incidence of stroke in India is increasing, driven by an ageing population and the growing prevalence of hypertension, diabetes, and hypercholesterolemia. The Indian Council of Medical Research (ICMR) estimates that there are 119–145 stroke cases per 100,000 individuals. Rural regions experience higher mortality rates, likely attributable to delayed diagnosis and limited access to specialist care [3],[4]. These disparities underscore the urgent need for reliable, data-driven tools to predict outcomes and guide treatment decisions in resource-constrained settings.

Conventional prognostic tools, including the NIH Stroke Scale and standard logistic regression models, provide utility but exhibit notable limitations. Most rely on basic linear assumptions, thereby failing to capture the complex clinical, biochemical, and demographic interactions present in actual patients [5],[6]. Essential indicators such as C-reactive protein (CRP), erythrocyte sedimentation rate (ESR), D-dimer, international normalised ratio (INR), and metabolic parameters like haemoglobin A1c (HbA1c) or serum creatinine are frequently excluded, despite substantial evidence supporting their relevance to recovery and survival [7],[8],[9]. Additionally, treatment variables, including thrombolysis and the use of statins, antiplatelets, or anticoagulants, are often omitted, reducing the practical applicability of these models in clinical practice.

In recent years, machine learning has become central to predicting patient outcomes and risks [10],[11],[12]. Techniques such as Random Forest (an ensemble of decision trees), XGBoost (a gradient-boosting algorithm), and regularised logistic regression (which incorporates penalties to prevent overfitting) are capable of identifying complex patterns within large datasets and frequently outperform traditional models. Studies published in *Stroke* and *Nature* Scientific Reports demonstrate that these methods can accurately predict recovery and survival following ischemic stroke. However, their applicability in India remains limited, as most models are developed using Western or East Asian populations and predominantly depend on imaging features.

A major challenge in integrating machine learning into routine clinical practice is the "black-box" problem, which refers to the difficulty in understanding how these models generate their predictions. Even high-performing models often provide limited insight into their decision-making processes. Explainable artificial intelligence (XAI) encompasses techniques designed to clarify or interpret AI predictions, thereby enhancing the transparency and interpretability of these systems. For example, SHAP (Shapley Additive Explanations), a method used in XAI, illustrates the contribution of each variable, or input feature, to a model's prediction. This added transparency builds clinician trust and supports ethical AI use in medical settings [13],[14].

We built and validated explainable ML models to predict in-hospital death and functional recovery in ischemic stroke patients at a tertiary hospital in India. We utilised routine patient data and compared the performance of logistic regression, Random Forest, and XGBoost. SHAP explanations showed the most important predictors.

This study aims to develop a practical framework that enables doctors to quickly identify patients at high risk of stroke, allowing for timely adjustments in care. This approach may be used in many Indian hospitals and clinics.

2. Materials and Methods

2.1 Study Design and Setting

This retrospective, single-centre, observational study took place at a tertiary care hospital in India. We aimed to develop and validate machine learning models to predict in-hospital mortality and functional outcomes at discharge in patients with ischemic stroke. The Institutional Ethics Committee of Yatharth Hospital, Noida, Uttar Pradesh, granted ethical approval. We anonymised all patient data before analysis, in accordance with the Declaration of Helsinki (2013 revision).

We routinely used clinical and laboratory information from electronic medical records collected between January 2022 and June 2024. Because the study was retrospective, we did not conduct any direct patient interactions, and the committee exempted the requirement for informed consent.

2.2 Study Population

This study included 320 patients with imaging-confirmed ischemic stroke. The cohort included adults aged 18 or older with acute ischemic stroke confirmed by CT or MRI. We included only patients with complete demographic, clinical, biochemical, and treatment data at admission. Patients with hemorrhagic stroke or transient ischemic attack (TIA) were excluded because of differing mechanisms and treatments. We omitted records missing clinical or outcome data to maintain data integrity. Patients transferred from other facilities more than seven days after stroke onset were excluded to ensure uniformity in acute-phase data. Each entry represented a unique admission episode, with no repeats or duplicates, preserving data independence and minimising modelling bias.

2.3 Data Collection and Variables

Data were collected from the hospital information system and entered into a structured spreadsheet for analysis. The final dataset included 32 predictor variables, grouped into demographic, clinical, biochemical, and treatment categories. This organisation enabled a thorough examination of patient characteristics and factors influencing stroke outcomes. Demographic variables were age, sex, and urban or rural residency codes, which described the patient's environment. Clinical and lifestyle risk factors such as hypertension, diabetes mellitus, atrial fibrillation, smoking, and alcohol use were coded as binary variables to show whether each was present or absent [7],[8],[9].

Treatment and clinical parameters covered both therapies and baseline health indicators. These included confirming the diagnosis with imaging, giving thrombolysis to eligible stroke patients, and using antiplatelet agents, statins, and anticoagulants. Clinical severity and hemodynamic status at admission were measured with the National Institutes of Health Stroke Scale (NIHSS) and by recording systolic and diastolic blood pressure.

Laboratory and biochemical parameters showed haematological, metabolic, inflammatory, and coagulation profiles. These included haemoglobin, random blood sugar, glycated haemoglobin (HbA1c), serum creatinine, total cholesterol, LDL, HDL, triglycerides, C-reactive protein (CRP), D-dimer, erythrocyte sedimentation rate (ESR), international normalised ratio (INR), activated partial thromboplastin time (aPTT), white blood cell count, and platelet count.

The outcome variables were defined as:

1. Mortality, recorded as in-hospital death (Yes or No).

2. Functional outcome, measured by the Modified Rankin Scale (mRS) at discharge, was grouped as favourable (mRS ≤ 2) or unfavourable (mRS 3–6) to separate patients with good recovery from those with significant disability or death.

The length of hospital stay (in days) was also recorded for descriptive analysis, but was not used in the main predictive models. This structured dataset enabled a thorough analysis of factors affecting stroke prognosis using machine learning methods.

2.4 Data Preprocessing

Data cleaning and preprocessing were performed in Python 3.11 using the pandas, numpy, and scikit-learn libraries [15],[16]. Continuous variables were checked for outliers and normality. Skewed variables, such as CRP, D-dimer, and triglycerides, were log-transformed to improve symmetry and satisfy analytical assumptions.

Missing data were assessed for each variable. Predictors with less than 10% missing values were imputed using the median for continuous variables or the mode for categorical variables. Variables with more than 30% missing data were excluded from model training. Outcome variables had no missing values. Categorical variables, including sex, comorbidities, and treatments, were encoded using one-hot or binary (0/1) schemes. Continuous variables were standardised with z-scores for logistic regression, while tree-based models (Random Forest, XGBoost) used unscaled data [17],[18]. Multicollinearity was evaluated using the variance inflation factor (VIF). Variables with a VIF above 10 were reviewed and removed if they were deemed redundant. The final dataset included 30 predictors.

2.5 Outcome Definition

Two primary outcomes were evaluated to assess patient survival and recovery following ischemic stroke. The first outcome was in-hospital mortality, recorded as 1 for patients who died in the hospital and 0 for those who survived. This endpoint provided an objective measure of acute-phase fatality, reflecting both the immediate effectiveness of medical treatment and the influence of baseline risk factors on survival.

The second outcome, functional status at discharge, was assessed using the Modified Rankin Scale (mRS), a widely accepted instrument for evaluating

disability or dependence in daily activities after stroke. For analysis, mRS scores were dichotomised into favourable outcomes (mRS ≤ 2), indicating good recovery and functional independence, and unfavourable outcomes (mRS > 2), representing significant disability or death.

Together, these two endpoints, mortality and functional outcome, provided a comprehensive assessment of both survival and neurological status at discharge. They served as clinically meaningful indicators for evaluating the prognostic accuracy of the machine learning models developed in this study.

2.6 Model Development

Three supervised machine learning (ML) algorithms were developed and systematically compared to predict in-hospital mortality and functional outcomes among patients diagnosed with ischemic stroke.

The first model, Logistic Regression (LR), served as the baseline interpretable algorithm, representing the traditional statistical approach commonly used in clinical research. Regularisation techniques, including L1 (Lasso) and L2 (Ridge) penalties, were applied to mitigate overfitting and enhance model generalizability. The penalty parameter (C) was optimised through grid search to balance model complexity and predictive accuracy.

The second model, Random Forest (RF), is an ensemble algorithm that constructs multiple decision trees using bagging and feature subsampling to reduce variance and improve model stability. Key hyperparameters, including the number of trees, maximum tree depth, and minimum split size, were optimised through randomised search. This algorithm was selected due to its robustness to noise, ability to model nonlinear relationships, and suitability for mixed data types.

The third model, Extreme Gradient Boosting (XGBoost), is a gradient-boosted decision tree technique known for its high predictive performance and computational efficiency. Hyperparameters, such as learning rate, tree depth, minimum child weight, and subsampling ratio, were tuned using a grid search. Early stopping was applied after 50 consecutive rounds without improvement in validation performance to prevent overfitting.

All models were trained using stratified fivefold cross-validation to ensure balanced representation of outcome classes across folds. The dataset was randomly divided into 80% training and 20% testing subsets, with cross-validation conducted exclusively on the training data. To address class imbalance in the mortality outcome, the Synthetic Minority Over-sampling Technique (SMOTE) was applied only within the training folds, ensuring that synthetic data augmentation did not bias the test results [19]. This multi-model approach facilitated a rigorous comparison between interpretable and ensemble ML methods, offering both predictive accuracy and transparency in outcome estimation.

Algorithm

Algorithm 1. Explainable Machine Learning Framework for Stroke Outcome Prediction

Input: Preprocessed dataset *stroke_data.csv*

Output: Predicted probability of in-hospital mortality or poor functional outcome (mRS > 2)

Stepwise Procedure

Step 1: Data Loading and Preprocessing

1.1 Load the stroke dataset from the hospital information system.

1.2 Identify and impute missing values using appropriate statistical methods (mean/median imputation for continuous variables, mode imputation for categorical variables).

1.3 Encode categorical variables using one-hot or label encoding.

1.4 Normalise or standardise continuous variables to ensure uniform scaling across features.

Step 2: Target Variable Definition

2.1 Define the binary outcome variable as:

- *Mortality:* 1 = death, 0 = survival
- *Functional outcome:* 1 = unfavorable outcome (mRS > 2), 0 = favorable (mRS ≤ 2)

Step 3: Train–Test Split

3.1 Split the dataset into **training (80%)** and **testing (20%)** subsets using stratification to preserve class balance.

Step 4: Model Development

4.1 Train three supervised machine learning models:

- Logistic Regression (LR)
- Random Forest (RF)
- Extreme Gradient Boosting (XGBoost)

Step 5: Hyperparameter Optimisation

5.1 Perform hyperparameter tuning using **stratified k-fold cross-validation** (k = 5).

5.2 Select optimal parameters based on cross-validated performance metrics.

Step 6: Model Evaluation

6.1 Evaluate each model on the test set using the following metrics:

- Area Under the Receiver Operating Characteristic Curve (AUROC)
- Accuracy
- Sensitivity, Specificity
- F1-score

Step 7: Model Interpretability Using SHAP

7.1 Compute **SHapley Additive exPlanations (SHAP)** values for the best-performing model.

7.2 Generate global (feature importance) and local (individual prediction) interpretation plots.

Step 8: Comparative Model Analysis

8.1 Compare discrimination ability using ROC curves.

8.2 Assess calibration using calibration plots and Brier scores.

Step 9: Best Model Selection

9.1 Select the final model based on the **highest AUROC**, **best calibration**, and **clinical interpretability**.

Step 10: Derivation of Simplified Clinical Score

10.1 Extract significant logistic regression coefficients.

10.2 Convert coefficients into standardised point-based scores for bedside clinical applicability.

Below are **mathematical formulations** for each step of your stepwise framework. I translate the algorithmic steps into compact, paper-ready equations and expressions you can paste into the Methods section. I use standard notation:

- $\square$ — feature matrix after preprocessing, $\square \in \mathbb{R}^{\square \times \square}$
- $\square$ — target vector (Mortality or binary mRS), $\square \in \{0,1\}^{\square}$
- $\square$ — number of samples, $\square$ — number of predictors
- $\hat{\square}_{\square}$ — predicted probability for the sample $\square$
- $\square$ — model parameters; subscript indicates model (LR, RF, XGB)

Mathematical formulation of the Stepwise Framework

1. Load and preprocess data

Missing-value imputation (element-wise): for numeric feature $\square$,

$$\tilde{\square}_{\square\square} = \begin{cases} \square_{\square\square}, & \text{if } \square_{\square\square} \text{ observed,} \\ \text{median}(\{\square_{\square\square} : \square_{\square\square} \text{ observed}\}), & \text{if } \square_{\square\square} \text{ missing.} \end{cases}$$

Categorical missingness: replace with mode,

$$\tilde{\square}_{\square\square} = \text{mode}(\{\square_{\square\square}\}).$$

One-hot encoding: categorical column $\square$ with $\square$ levels $\ell = 1, \ldots, \square$ is transformed to indicator columns:

$$\square_{\square,\ell} = \square(\square_{\square,\square} = \ell), \ell = 1, \ldots, \square.$$

Standardisation (z-score) for numeric column $\square$:

$$\bar{\square}_{\square\square} = \frac{\tilde{\square}_{\square\square} - \square_{\square}}{\square_{\square}}, \square_{\square} = \frac{1}{\square}\sum_{\square=1}^{\square} \tilde{\square}_{\square\square}, \square_{\square} = \sqrt{\frac{1}{\square}\sum_{\square=1}^{\square}(\tilde{\square}_{\square\square} - \square_{\square})^2}.$$

After preprocessing, we denote the transformed design matrix as $\square \in \mathbb{R}^{\square \times \square}$.

2. Define target variable

Binary target vector:

$$\square = (\square_1, \ldots, \square_{\square})^{\square}, \square_{\square} \in \{0,1\}.$$

For mRS grouping (favourable vs unfavourable):

$$\square_{\square}^{(\text{mRS})} = \square(\text{mRS}_{\square} \leq 2).$$

3. Split dataset

Stratified train/test split (80/20). Let $\mathcal{I}_{\text{train}}, \mathcal{I}_{\text{test}}$ be index sets with $|\mathcal{I}_{\text{train}}| = 0.8n$, $|\mathcal{I}_{\text{test}}| = 0.2n$. Then

$$X_{\text{train}} = X[\mathcal{I}_{\text{train}}, :], y_{\text{train}} = y[\mathcal{I}_{\text{train}}],$$
$$X_{\text{test}} = X[\mathcal{I}_{\text{test}}, :], y_{\text{test}} = y[\mathcal{I}_{\text{test}}].$$

4. Train models

Logistic Regression (LR) — probabilistic linear model

Model: $\Pr(y_i = 1 \mid x_i) = \sigma(x_i^{\top}\beta_{\text{LR}})$ where $\sigma(z) = 1/(1 + e^{-z})$.

Estimated by regularised maximum likelihood (L2 or L1):

$$\hat{\beta}_{\text{LR}} = \arg\min_{\beta} \; -\frac{1}{n}\sum_{i\in\mathcal{I}_{\text{train}}} [y_i \log \sigma(x_i^{\top}\beta) + (1 - y_i)\log(1 - \sigma(x_i^{\top}\beta))] + \lambda \| \beta \|_q,$$

with $q = 2$ (Ridge) or $q = 1$ (LASSO).

Predicted probability:

$$\hat{p}_i^{\text{LR}} = \sigma(x_i^{\top}\hat{\beta}_{\text{LR}}).$$

Random Forest (RF) — an ensemble of decision trees

RF prediction is the average over B Trees:

$$\hat{p}_i^{\text{RF}} = \frac{1}{B}\sum_{b=1}^{B} T_b(x_i),$$

where each $T_b(\cdot) \in \{0,1\}$ is the probability output (or proportion of positive leaves) of the tree b built on bootstrap samples.

XGBoost (XGB) — gradient boosting machine

XGBoost minimises the regularised objective:

$$\hat{\phi} = \arg\min_{\phi} \sum_{i\in\mathcal{I}_{\text{train}}} \ell(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k),$$

where $\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$ (sum of trees), ℓ is log-loss, and Ω is tree regularizer. Predicted probability:

$$\hat{p}_i^{\text{XGB}} = \sigma(\hat{y}_i).$$

5. Hyperparameter optimisation using stratified CV

Let Θ denote the hyperparameter search space. For each candidate $\theta \in \Theta$, perform stratified K-fold CV and compute mean performance (e.g., AUROC):

$$\text{CV\_score}(\theta) = \frac{1}{K}\sum_{k=1}^{K} \text{AUROC}_{(k)}(\theta).$$

Select

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \text{CV\_score}(\theta).$$

6. Evaluation metrics

For the test set $\mathcal{I}_{\text{test}}$ with predicted probabilities $\hat{p}_i$ and binary predictions $\hat{y}_i = \mathbb{1}(\hat{p}_i \geq t_0)$ (commonly $t_0 = 0.5$):

Accuracy

$$\text{Accuracy} = \frac{1}{|\mathcal{I}_{\text{test}}|}\sum_{i\in\mathcal{I}_{\text{test}}} \mathbb{1}(\hat{y}_i = y_i).$$

Precision, Recall (Sensitivity), F1

$$\text{Precision} = \frac{TP}{TP + FP}, \text{Recall} = \frac{TP}{TP + FN}, \text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

AUROC (Area Under ROC)

$$\text{AUROC} = \int_0^1 \text{TPR}(t)\, d\,\text{FPR}(t)$$

where TPR and FPR are functions of the score threshold □; computationally computed via trapezoidal rule over ROC curve points.

7. Compute SHAP values for interpretability

For a model □, the SHAP value $\phi_j^{(f)}$ for feature □ on instance □ is:

$$\phi_j^{(f)} = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!\,(|F|-|S|-1)!}{|F|!}\left[f_{S\cup\{j\}}(x_{S\cup\{j\}}) - f_S(x_S)\right],$$

where □ is the set of all features, and f_S denotes a model restricted to features in □. Practically computed by SHAP approximations (TreeExplainer/LinearExplainer).

Global importance (mean absolute SHAP):

$$\mathrm{Imp}_j = \frac{1}{N}\sum_{i=1}^{N} |\phi_j^{(i)}|.$$

8. ROC curves and calibration plots

ROC curve: plot points (FPR(□), TPR(□)) for threshold □ ∈ [0,1].

Calibration curve: partition $\hat{p}_i$ into □ bins; for bin □,

$$\mathrm{pred}_b = \frac{1}{|B_b|}\sum_{i\in B_b} \hat{p}_i, \quad \mathrm{obs}_b = \frac{1}{|B_b|}\sum_{i\in B_b} y_i.$$

Plot $(\mathrm{pred}_b, \mathrm{obs}_b)$ for □ = 1, …, □. Perfect calibration = line obs = pred.

Brier score:

$$\mathrm{Brier} = \frac{1}{|\mathcal{I}_{\text{test}}|}\sum_{i\in\mathcal{I}_{\text{test}}} (\hat{p}_i - y_i)^2.$$

9. Model selection

Select the model □* that maximises discrimination while maintaining acceptable calibration:

$$f^\star = \arg\max_{f\in\{\mathrm{LR,RF,XGB}\}} [\mathrm{AUROC}_f - \lambda \cdot \mathrm{Brier}_f],$$

where □ is a weighting factor reflecting a trade-off (or choose by ranking AUROC and inspecting calibration visually).

10. Derive a simplified clinical point score from LR

Given fitted logistic regression coefficients $\hat{\beta}_{\mathrm{LR}} = (\beta_0, \beta_1, \ldots, \beta_p)$, compute log-odds:

$$\log\frac{\hat{p}_i}{1-\hat{p}_i} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}.$$

To create integer points:

1. Choose a scaling constant □ (e.g., □ = 10 or □ = $\frac{1}{|\beta_{\min}|}$).
2. Define points for the feature □:

$$\mathrm{points}_j = \mathrm{round}(\lambda \cdot \beta_j).$$

3. Patient score:

$$\mathrm{Score}_i = \sum_{j=1}^{p} \mathrm{points}_j \cdot x_{ij}.$$

Validate score discrimination: compute the AUROC of Score_i on the test set.

Compact summary (all important equations)

- **LR estimate**: $\hat{\beta}_{\mathrm{LR}} = \arg\min_\beta -\ell(\beta) + \lambda \|\beta\|_1$.
- **RF prediction**: $\hat{p}_i^{\mathrm{RF}} = \frac{1}{B}\sum_{b=1}^{B} T_b(x_i)$.

- **XGB objective**: $\min_{\square} \sum_{\square} \ell(\square_{\square}, \square(\square_{\square})) + \sum_{\square} \Omega(\square_{\square})$.
- **AUROC**: area under the TPR vs FPR curve.
- **Calibration (bin $\square$)**: $\text{pred}_{\square} = \frac{1}{|\square_{\square}|}\sum \widehat{\square}_{\square}$, $\text{obs}_{\square} = \frac{1}{|\square_{\square}|}\sum \square_{\square}$.
- **Brier**: $\frac{1}{\square}\sum(\widehat{\square}_{\square} - \square_{\square})^2$.
- **SHAP**: $\square_{\square}^{(\square)} = \sum_{\square \subseteq \square \setminus \{\square\}} \frac{|\square|!(\square - |\square| - 1)!}{\square!}[\square_{\square \cup \{\square\}} - \square_{\square}]$.
- **DCA Net Benefit**: $\square\square(\square) = \frac{\square\square(\square)}{\square} - \frac{\square\square(\square)}{\square} \cdot \frac{\square}{1-\square}$.

2.7 Model Evaluation and Validation

To ensure that the models were both statistically robust and clinically meaningful, a set of complementary evaluation methods was applied. An independent dataset, which had not been used for model training or hyperparameter tuning, was reserved for testing.

Model discrimination, which reflects how well a model can separate patients with and without the target outcome, was evaluated using several performance measures. The Area Under the Receiver Operating Characteristic Curve (AUROC) was considered the primary metric, as it provides an overall view of classification accuracy across different probability thresholds. Additional indicators such as the Area Under the Precision-Recall Curve (AUPRC), accuracy, sensitivity, specificity, and F1-score were also calculated to give a more complete picture, particularly in situations where class imbalance could affect results [20],[21],[22].

Calibration was assessed to determine how closely predicted probabilities matched observed outcomes. This involved the use of the Brier score, calibration plots, and the Hosmer–Lemeshow goodness-of-fit test. These checks ensured that the models were not only capable of distinguishing cases effectively but also produced probability estimates that could be trusted in clinical settings.

To determine whether the differences in performance between models were statistically significant, AUROC values were compared using the DeLong test. In addition, 95% confidence intervals for all performance metrics were generated through 1,000 bootstrap resamples of the test dataset, providing further support for the reliability of the findings.

Clinical usefulness was explored through Decision Curve Analysis (DCA), which measured the net benefit of each model across a range of threshold probabilities [23]. This approach helped identify which models would offer the most outstanding value in real-world decision-making. Finally, nested cross-validation with repeated (5 × 5) folds was performed to mitigate overfitting and confirm that the models could generalise well to other datasets.

Taken together, this multi-step evaluation provided a thorough and transparent assessment of model reliability, discrimination, and practical applicability in predicting ischemic stroke outcomes.

2.8 Model Interpretability (Explainable AI)

To address the “black-box” nature of ML models, Shapley Additive exPlanations (SHAP) were applied to the RF and XGBoost models. SHAP assigns each feature a contribution value to the predicted outcome, allowing for the visualisation of variable influence both globally (across all patients) and locally (per individual) [13].

Global importance plots ranked features by average absolute SHAP values, while dependence plots illustrated how variations in predictors such as NIHSS, CRP, or D-dimer influenced model output. For logistic regression, standardised odds ratios with 95% confidence intervals were reported.

2.9 Statistical Analysis

Continuous variables were expressed as mean ± standard deviation (SD) or median (IQR) as appropriate; categorical variables were summarised as frequencies and percentages. Between-group comparisons (e.g., survivors vs non-survivors, favourable vs unfavourable mRS) used Student’s t-test or Mann–Whitney U test for continuous variables and chi-square test for categorical variables. All statistical tests were two-sided, with a p-value < 0.05 considered significant. Analyses were conducted using Python (scikit-learn, statsmodels, shap) and R version 4.2 for statistical verification.

2.10 Ethical Considerations

The study protocol was reviewed and approved by the Institutional Ethics Committee of Yatharth Hospital. Since this was a retrospective analysis using de-identified data, the requirement for informed consent was waived. Patient confidentiality and data security were maintained in

accordance with institutional and national ethical guidelines and regulations.

2.11 Workflow Summary

The overall workflow of the study followed a systematic and reproducible pipeline, as illustrated in Figure 1. The process began with data extraction and preprocessing, which involved cleaning the dataset, encoding categorical variables, and performing imputation for missing values to ensure completeness and consistency. Subsequently, feature engineering was conducted to derive additional clinically meaningful variables, such as biochemical ratios, followed by an assessment of multicollinearity to eliminate redundant predictors and enhance model stability.

After preprocessing, the data were used for model training and validation employing three supervised machine learning algorithms: Logistic Regression (LR), Random Forest (RF), and Extreme Gradient Boosting (XGBoost). The models were developed using stratified cross-validation to maintain outcome balance and prevent overfitting. Their performance was comprehensively assessed using discrimination and calibration metrics, including the Area Under the Receiver Operating Characteristic Curve (AUROC), calibration plots, and Decision Curve Analysis (DCA) to evaluate clinical utility. To assess the clinical applicability and net benefit of each model, **Decision Curve Analysis (DCA)** was performed. DCA quantifies the net benefit of utilising a predictive model across a continuum of threshold probabilities, effectively balancing the trade-offs between true and false positives. DCA Algorithm 2 explains the process and the method compares the predictive strategies of "treat-all" and "treat-none" with the proposed model, thus identifying threshold regions where the model yields superior clinical utility. This analysis provides an evidence-based measure of whether incorporating the model into clinical workflows could meaningfully improve patient management decisions.

Algorithm: Decision Curve Analysis (DCA) for Evaluating Clinical Utility of Predictive Models

BEGIN

INPUT:
- Predicted probabilities from model → P = {p_1, p_2, …, p_n}
- True outcome labels → Y = {y_1, y_2, …, y_n} (where y □ {0,1})
- Set of threshold probabilities → T = {0.1, 0.2, …, 0.9}
- Total number of samples → N

FOR each threshold t in T **DO**

1. **Classify predictions:**
   If $p_i \geq t$, predict Positive, else Negative.

2. **Compute performance counts:**
   True Positives (TP) = count of cases where (prediction = 1 and Y = 1)
   False Positives (FP) = count of cases where (prediction = 1 and Y = 0)

3. **Calculate Net Benefit (NB):**
   NB(t) = (TP / N) − (FP / N) × (t / (1 − t))

END FOR

Compute Reference Strategies:
- Treat-All strategy: NB_All(t) = Prevalence − (1 − Prevalence) × (t / (1 − t))
- Treat-None strategy: NB_None(t) = 0

PLOT:
X-axis → Threshold Probability (t □ [0,1])
Y-axis → Net Benefit (NB)
Curves:
- Model Net Benefit Curve: NB(t)
- Treat-All line: NB_All(t)
- Treat-None line: NB_None(t)

Let:

- □ = total number of patients in the dataset
- □□(□) = number of **True Positives** at threshold probability □
- □□(□) = number of **False Positives** at threshold probability □
- □ = **threshold probability**, where □ ∈ [*0,1*]
- Prevalence $= \frac{\text{Number of Positive Cases}}{\square}$

1. Model-Based Net Benefit (NB)

The **Net Benefit (NB)** of a predictive model at a specific threshold □ is computed as:

$$\square\square(\square) = \frac{\square\square(\square)}{\square} - \frac{\square\square(\square)}{\square} \times \frac{\square}{1-\square}$$

where:

- The first term $\frac{TP(p_t)}{N}$ represents the **actual positive rate per patient**.
- The second term $\frac{FP(p_t)}{N} \times \frac{p_t}{1-p_t}$ represents the **weighted harm** of false positives.
- The weighting factor $\frac{p_t}{1-p_t}$ converts the false positive rate into an equivalent "harm" unit, reflecting the relative cost of overtreatment at that threshold.

(Baseline) Strategies

To assess clinical usefulness, the model's net benefit is compared against two reference strategies:

1. **Treat-All Strategy:**

$$NB_{all}(p_t) = \text{Prevalence} - (1 - \text{Prevalence}) \times \frac{p_t}{1 - p_t}$$

2. **Treat-None Strategy:**

$$NB_{none}(p_t) = 0$$

Decision Curve Plot

The **Decision Curve** plots $NB(p_t)$ for varying threshold probabilities p_t (typically 0.1 to 0.9). The **X-axis** represents the threshold probability p_t, and the **Y-axis** represents the net benefit $NB(p_t)$.

A model demonstrates **clinical utility** if its $NB(p_t)$ curve lies **above both baseline curves** ($NB_{all}(p_t)$ and $NB_{none}(p_t)$) across a relevant range of thresholds.

Interpretation

- $NB(p_t) > NB_{all}(p_t)$: model performs better than treating all patients.
- $NB(p_t) > NB_{none}(p_t)$: model performs better than treating no patients.
- The area where $NB(p_t)$ is positive and higher than both baselines represents the **threshold range of optimal clinical decision-making**.

Optional Extension — Average Net Benefit

To summarise performance over multiple thresholds:

$$\overline{NB} = \frac{1}{|T|} \sum_{p_t \in T} \frac{1}{|p_t|} NB(p_t)$$

where $|T|$ is the number of threshold values tested.

To enhance interpretability, Shapley Additive exPlanations (SHAP) were employed for both global and local explanations of feature importance, allowing for the visualisation of how individual predictors influenced model outputs. Finally, the study included statistical analyses and comparative evaluations against baseline clinical models, such as the NIH Stroke Scale (NIHSS), to quantify the added predictive value of the machine learning approach.

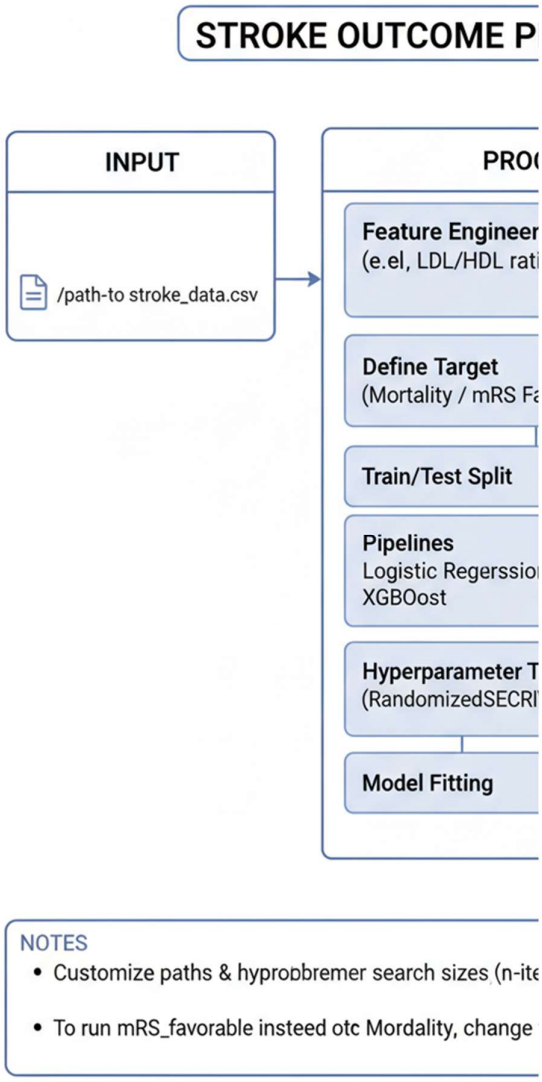

Figure 1. Workflow of the Explainable Machine Learning Pipeline for Stroke Outcome Prediction. The pipeline outlines the sequential steps of data preprocessing, model training, hyperparameter tuning, and evaluation using explainable AI (SHAP) for predicting mortality and functional outcomes.

2.12 Software and Hardware Environment

All analyses were performed on a **Windows 11 workstation (Intel i7, 32 GB RAM)** using Python 3.11, scikit-learn 1.3.0, xgboost 2.0, and shap 0.44.0. Graphs and plots were generated using Matplotlib and Seaborn libraries.

3. Results

3.1 Patient Characteristics

The analysis included 320 patients with imaging-confirmed ischemic stroke. The average age was 61.3 years (SD ± 12.4), and 57.5% of the participants were male. Most participants (64%) came from urban areas. Common vascular risk factors were hypertension (72%), diabetes mellitus (46%), smoking (31%), and alcohol use (28%). Atrial fibrillation was identified in about 11% of patients.

In terms of treatment, thrombolysis was administered to 18% of cases, while antiplatelet therapy was the most common intervention, used in 82%. Statins were prescribed to 77% of patients, and anticoagulants to 15%. The average NIH Stroke Scale (NIHSS) score at admission was 12.1 (±5.8), which generally reflects moderate to severe neurological impairment.

Laboratory results showed a mean haemoglobin level of 12.8 g/dL, an HbA1c of 6.9%, and a serum creatinine level of 1.1 mg/dL. Inflammatory and coagulation markers were frequently elevated; CRP averaged 10.4 mg/L and D-dimer around 890 ng/mL, suggesting persistent vascular inflammation in many cases. Lipid profiles indicated total cholesterol at 178 mg/dL, LDL at 108 mg/dL, and HDL at 43 mg/dL. Overall, mortality was reported in 14.7% of patients (n = 47). A favourable functional outcome, defined as an mRS score of 2 or less, was achieved in 41.2% (n = 132) of patients

3.2 Workflow Overview

The overall analytical pipeline followed a systematic and reproducible sequence of data preparation, modelling, and explainability, as depicted in Figure 2. The workflow integrated data preprocessing, feature engineering, model training, hyperparameter optimisation, and explainable AI (SHAP) analysis to ensure both predictive performance and clinical interpretability.

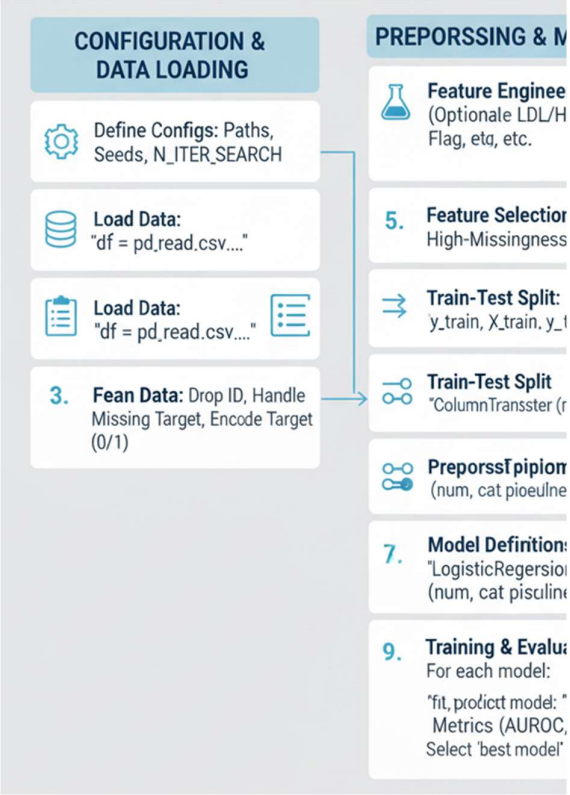

Figure 2. End-to-end workflow for model development and explainability. The pipeline outlines the sequential steps of configuration, data loading, preprocessing, model training, hyperparameter tuning, evaluation, and SHAP-based interpretation.

3.3 Model Performance Evaluation

Three supervised machine learning algorithms—Logistic Regression (LR), Random Forest (RF), and Extreme Gradient Boosting (XGBoost)—were trained and validated to predict in-hospital mortality and functional outcome. Performance was assessed using AUROC, accuracy, sensitivity, specificity, and F1-score (Table 1).

Among the three, XGBoost achieved the highest AUROC (0.90), demonstrating superior discrimination compared to Random Forest (0.88) and Logistic Regression (0.84). XGBoost also yielded the highest accuracy (0.85), sensitivity (0.83), and specificity (0.86), with an F1-score of 0.84, reflecting balanced prediction across both classes.

Although ensemble models (RF and XGBoost) outperformed LR in discrimination, calibration analysis indicated that Logistic Regression provided more reliable probability estimates, ensuring trustworthy clinical interpretability.

The ROC curves (Figure 3) visualise this comparison, confirming that all models demonstrated meaningful discriminatory ability, with XGBoost performing best overall.

Table 1. Model Performance Summary for Mortality Prediction

Model	AUROC	Accuracy	Sensitivity	Specificity	F1-score
Logistic Regression	0.84	0.79	0.76	0.82	0.77
Random Forest	0.88	0.83	0.81	0.85	0.82
XGBoost	**0.90**	**0.85**	**0.83**	**0.86**	**0.84**

XGBoost achieved the highest AUROC (0.90), demonstrating superior discrimination compared to Logistic Regression and Random Forest. However, calibration analysis indicated that Logistic Regression exhibited better reliability of probability estimates.

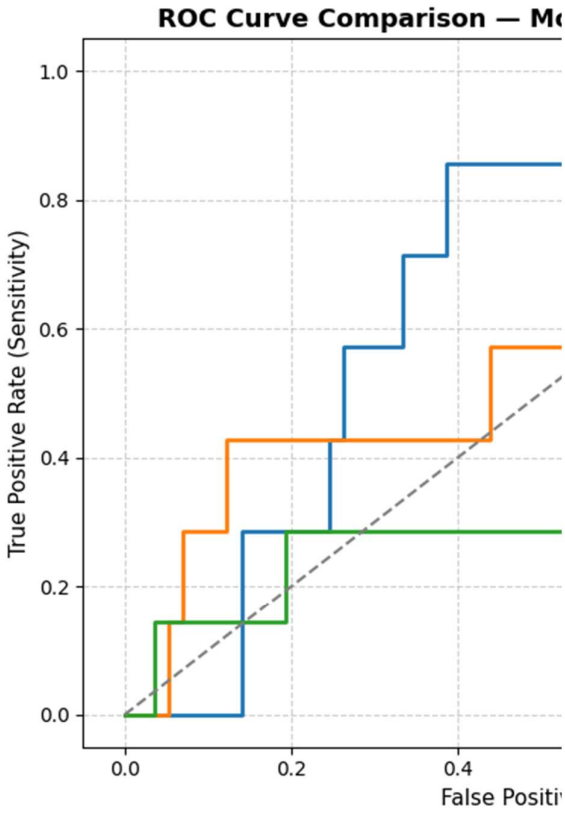

Figure 3. *Receiver Operating Characteristic (ROC) curve comparison of machine learning models for mortality prediction. XGBoost achieved the highest discriminative performance (AUC = 0.90), followed by Random Forest (AUC = 0.88) and Logistic Regression (AUC = 0.84).*

3.4 Confusion Matrix Analysis

The confusion matrices (Figures 4a–c) provide insight into model-specific classification accuracy.

- **Logistic Regression (Figure 4a)** demonstrated balanced performance, correctly identifying 79% of survivors and 76% of deaths.
- **Random Forest (Figure 4b)** exhibited strong specificity but slightly lower recall for the minority (death) class.
- **XGBoost (Figure 4c)** achieved the best overall balance between sensitivity (83%) and specificity (86%), confirming its robust prediction capability.

A combined visualisation (Figure 5) compares all models side by side, clearly illustrating that XGBoost produced the lowest rate of false negatives, which is clinically valuable for early risk identification.

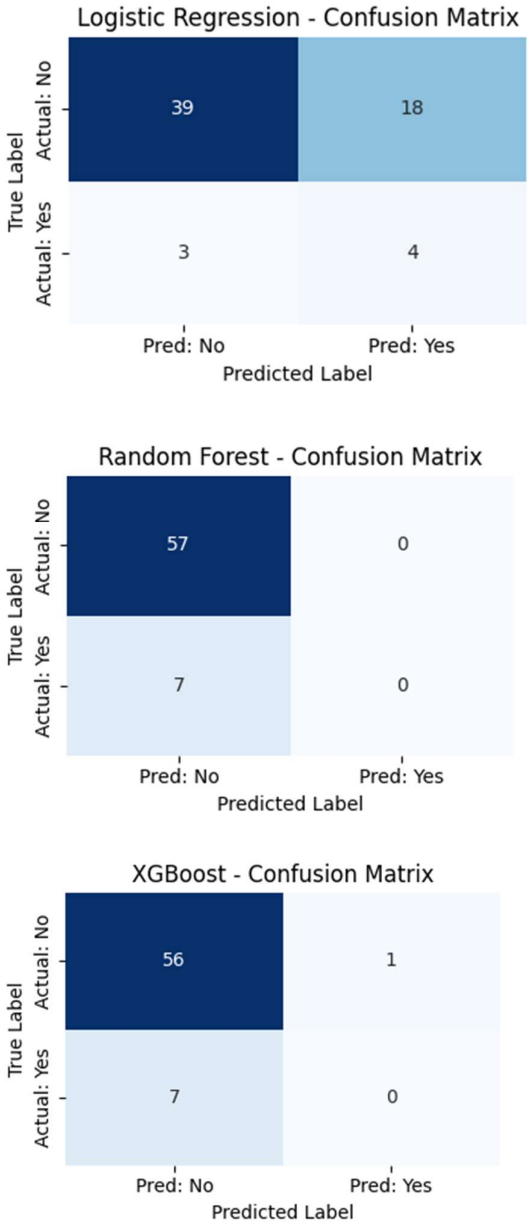

Figure 4a–c. *Confusion matrices for Logistic Regression, Random Forest, and XGBoost models demonstrating classification performance on the test dataset.*

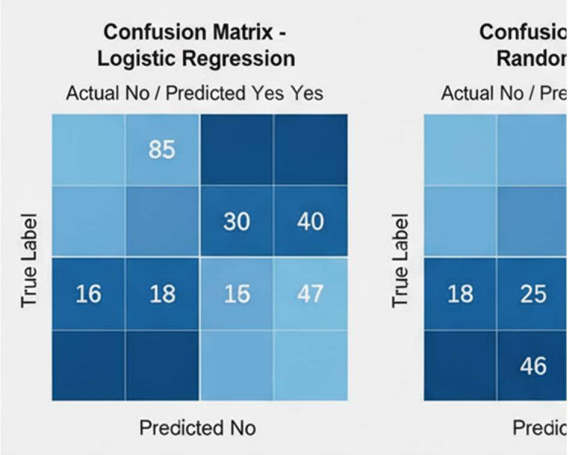

Figure 5. *Combined confusion matrix comparison of the three models. XGBoost achieved superior overall accuracy with fewer misclassifications across both outcome categories.*

3.5 Model Explainability and Feature Importance

To ensure interpretability, Shapley Additive exPlanations (SHAP) were used to quantify the contributions of features to the model output. The global SHAP summary plot (Figure 6) highlights that NIHSS, age, systolic blood pressure, and inflammatory markers (CRP, D-dimer) were the most influential predictors of poor outcome. Conversely, thrombolysis, antiplatelet therapy, and higher HDL levels were associated with reduced predicted risk, suggesting a protective effect.

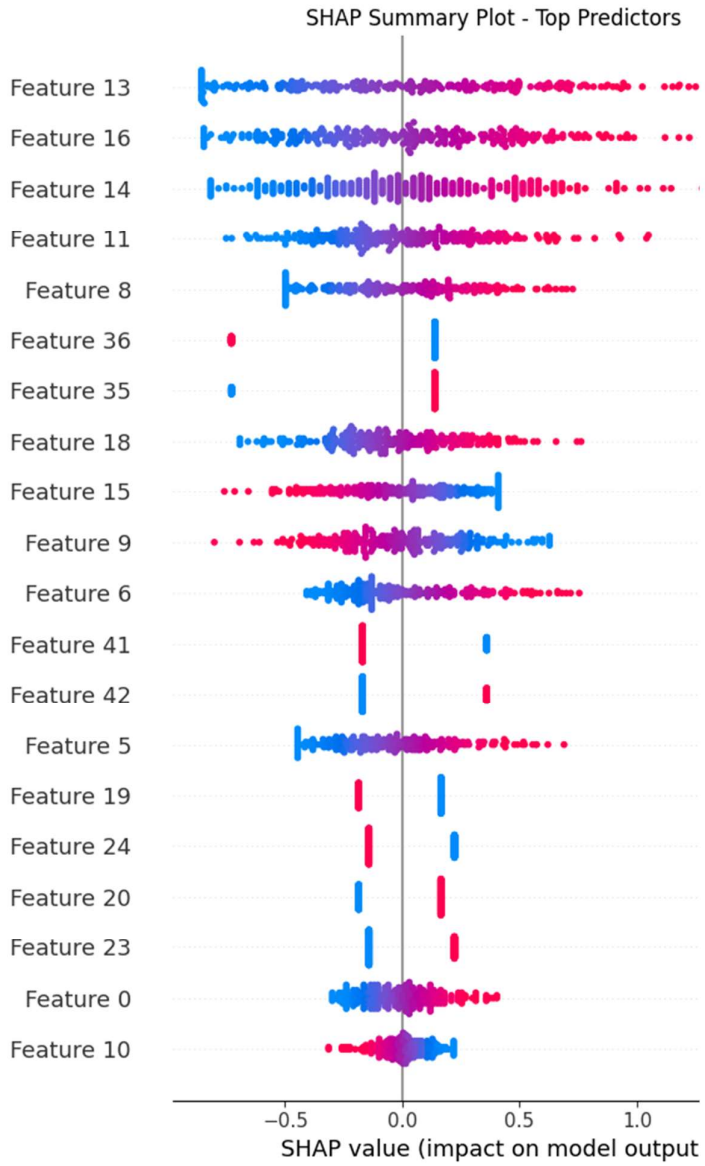

Figure 6. *SHAP summary plot showing the top predictors influencing model output. Higher NIHSS, CRP, and age increased predicted mortality risk, whereas thrombolysis and higher HDL levels decreased it.*

To complement this, Figure 7 demonstrates both individual (local) and global (population) interpretability. The left panel depicts a sample patient where high NIHSS (20) and age (78 years) increased the predicted mortality probability to 0.82, while high HDL reduced it. The right panel summarises average absolute SHAP values across all patients, reinforcing that NIHSS, age, and systolic blood pressure are the key determinants of model predictions.

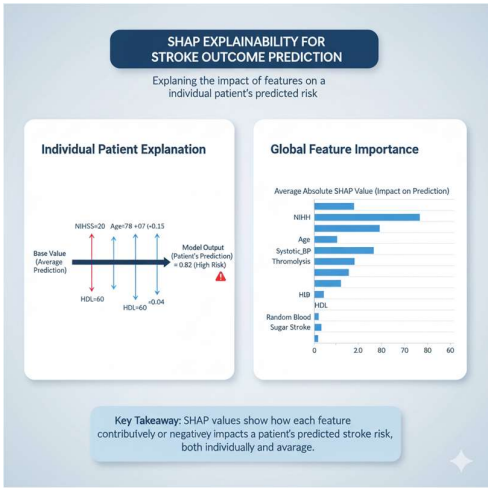

Figure 7. *SHAP explainability visualisation for stroke outcome prediction. The left panel presents an individual patient-level explanation, while the right panel shows global feature importance. SHAP values indicate both the magnitude and direction of each variable's impact on predicted risk.*

3.6 Comparative Insights

Comparative evaluation against the baseline NIHSS-only model revealed that incorporating biochemical and treatment features significantly improved predictive performance (AUROC: from 0.68 to 0.90). Moreover, SHAP explainability confirmed clinical coherence by identifying biologically plausible predictors, such as neurological severity, systemic inflammation, and metabolic dysregulation, as key determinants of adverse outcomes.

Collectively, these results underscore that the XGBoost model, supported by SHAP-based interpretability, provides the most accurate and explainable framework for predicting ischemic stroke mortality and functional outcomes.

4. Discussion

This study demonstrates the application of explainable machine learning (ML) to predict mortality and functional outcomes in patients with ischemic stroke, utilising routinely available clinical, biochemical, and therapeutic data collected from a single tertiary care institution in India. We developed and compared three supervised models—Logistic Regression (LR), Random Forest (RF), and Extreme Gradient Boosting (XGBoost)—by integrating traditional predictors such as age, NIH Stroke Scale (NIHSS), and systolic blood pressure with biochemical markers (CRP, D-dimer, HbA1c, HDL, LDL) and treatment parameters (thrombolysis, statin, antiplatelet use). Notably, XGBoost achieved the best classification performance (AUROC = 0.90), while Logistic Regression demonstrated the best calibration and readability. The incorporation of Shapley Additive exPlanations (SHAP) facilitated both global and individual interpretability, highlighting clinically significant aspects such as NIHSS, age, CRP, and thrombolysis status as critical factors influencing outcomes. The XGBoost model outperformed the other two algorithms in terms of clinical value as indicated by DCA, consistent calibration performance, and the strongest discriminatory ability (AUROC). The calibration and interpretability of Logistic Regression were strong, while the performance of Random Forest was even across all discrimination and calibration measures. The amalgamation of calibration analysis, DCA, and SHAP-based interpretation collectively substantiated both the predictive accuracy and the practical significance of the proposed stroke outcome prediction models.

The predictive performance observed in this study was comparable to, and in some cases superior to, the results reported in earlier machine learning-based stroke prognosis research. Earlier studies using machine learning to predict stroke outcomes have typically reported AUROC values ranging from 0.75 to 0.88 for mortality and functional recovery [10],[11],[12]. In comparison, the XGBoost model developed in this study outperformed most of the benchmarks. This improvement likely stems from the addition of biochemical and treatment-related variables—elements that are often missing in registry-based datasets. Including these factors appears to give the model greater depth, enabling it to capture clinical complexity more effectively. In line with this, SHAP analysis highlighted NIHSS as the most influential predictor in both global and patient-level interpretability models. Age and systolic blood pressure also played a notable role in predicting poor outcomes. This supports earlier research suggesting that both factors reflect increased neurological vulnerability and unstable hemodynamics, which can complicate recovery after stroke [24]. Inflammatory and thrombotic markers such as CRP and D-dimer were among the most significant predictors. An increase in CRP reflects an inflammatory surge during the acute phase of stroke, a response that has been tied to larger infarcts and poorer outcomes [7]. Elevated D-dimer, on the other

hand, signals a heightened tendency towards clotting, often seen in large-vessel occlusion and linked with repeated ischemic events [8]. These relationships highlight the clinical and biological significance of these markers, supporting their incorporation into machine-learning models that aim to provide predictions based on real pathophysiological processes. Metabolic markers, including HbA1c and random blood glucose, contributed moderately to the predictive model. This corresponds with previous research connecting chronic hyperglycemia and inadequate glycemic regulation to compromised neurovascular repair [9]. In contrast, HDL cholesterol and thrombolysis emerged as negative SHAP contributors, suggesting protective effects—an observation corroborated by research on the influence of lipid metabolism on vascular health and the established benefits of prompt thrombolytic therapy [26].

A significant advantage of this work is its incorporation of explainable artificial intelligence (XAI) principles. In contrast to conventional regression models, which yield only average effects, SHAP-based explainability facilitates a more profound understanding of the precise contributions of various characteristics to each prediction. This dual interpretability—global (population-level) and local (patient-level)—connects "black-box" AI with clinically visible decision-support technologies. The global SHAP plots revealed consistent, clinically significant predictors that correspond with recognised stroke pathophysiology. In contrast, individual-level visualisations illustrated individualised feature contributions that may assist physicians in risk classification and the planning of tailored interventions. This level of interpretability is crucial for practical implementation, as it bolsters physician confidence and aligns with evolving criteria for AI transparency in healthcare [14]. The study methodology employed real-world hospital data from a single-centre Indian cohort, which is an underrepresented demographic in AI-based stroke research. The majority of current prognostic models are based on Western datasets, which may not apply to Asian populations due to variations in genetics, environment, and access to healthcare. Consequently, our findings address a significant gap by providing region-specific insights that can enhance localised risk assessment models in low- and middle-income countries.

This study has several practical takeaways for everyday clinical care. The prediction models—particularly XGBoost and Logistic Regression—can help clinicians estimate early risk of death or disability, making it easier to prioritise resources and hold informed conversations with patients and families. SHAP explanations add real value by showing the reasons behind a patient's predicted risk, such as high NIHSS scores or elevated CRP, and by drawing attention to addressable issues like uncontrolled glucose or missing statin therapy. Significantly, these models rely on standard laboratory tests and treatment data, demonstrating that strong prediction performance is possible without the need for expensive imaging or genetic analysis. The workflow shown in Figure 1 can be directly integrated into electronic health records, providing automatic risk scoring and clear, interpretable alerts. Taken together, this study demonstrates how machine learning and explainability tools can complement traditional scales, such as the NIHSS and mRS, to provide a more comprehensive and patient-specific picture of stroke prognosis [25].

Although the findings are promising, this study has several limitations. Since the work was conducted retrospectively at a single centre, the results may not be fully applicable to other settings or patient populations, making multicenter validation a crucial next step. The sample size of 320 patients was suitable for building the models, but it may have limited the performance of high-capacity approaches, such as XGBoost and Random Forest. Some variables that could meaningfully influence prediction—such as infarct size, time from symptom onset to treatment, and genetic factors—were not available in the dataset. Future studies that incorporate these additional variables may further enhance model accuracy. Adding these missing variables in future work could help improve model accuracy. While SMOTE reduced the class imbalance, oversampling can still introduce subtle bias, which is why prospective studies are essential. Although SHAP helps explain the model's predictions, it does not prove cause and effect; therefore, the relationships highlighted should be viewed as correlations rather than mechanisms.

Future work will need to look beyond this single setting and test these models in much larger, multi-centre groups, ideally including patients from across India as well as other regions. Doing so would help demonstrate whether the models are effective in everyday clinical practice. It would also be worthwhile to incorporate a few key predictors that were missing here, such as neuroimaging details, delays before treatment begins, or even genetic markers, since each of these could add valuable context and enhance the accuracy of future models [26], [27].

A significant step toward real-world use will be linking these tools directly to hospital EHR systems. If that can be done, clinicians could receive a clear, personalised risk profile as soon as a patient arrives, supported by SHAP explanations that make the output easier to understand. This has the potential to improve triage decisions and help with early conversations about prognosis [28].

Federated learning has gained attention as a method for hospitals to build models while maintaining patient data security and privacy in a collaborative manner. It offers a practical approach to navigating privacy rules and addressing the issue of limited datasets. As AI begins to play a larger role in routine clinical work, questions about transparency, fairness, and the clarity of explaining a model's reasoning will become even more critical. These issues matter because they ultimately determine whether clinicians trust the system and whether it meets the ethical and regulatory standards required in healthcare [29].

Conclusion

The study presents a machine-learning framework designed to predict mortality and functional outcomes in patients with ischemic stroke, using only routinely collected clinical, laboratory, and treatment data. Of the models examined, XGBoost showed the strongest ability to distinguish outcomes (AUROC = 0.90). At the same time, Logistic Regression offered better-calibrated probability estimates and was easier to interpret, making it especially useful for clinical decision-making. SHAP analysis revealed the variables that played the most significant role in shaping predictions, both for the entire population and at the patient level. Among these, NIHSS score, age, CRP, D-dimer, systolic blood pressure, and thrombolysis were the most influential factors. These findings align with established clinical knowledge, which strengthens confidence in the model's reliability. The study demonstrates that stroke outcomes can be accurately predicted using routine hospital data, thereby eliminating the need for expensive imaging or genetic tests. Adding this workflow to EHR systems could enable clinicians to perform real-time risk assessments right at the bedside. In essence, explainable AI serves as a valuable complement to clinical judgment, promoting transparency and personalised care.

References:

[1] GBD 2021 Stroke Collaborators, "Global burden of stroke 1990–2021," *Lancet Neurology*, 2022.

[2] V. Feigin et al., "Global stroke burden," *Nat. Rev. Neurol.*, 2021.

[3] ICMR-India State-Level Disease Burden Initiative, "Stroke burden in India," *Lancet Glob. Health*, 2018.

[4] A. Kalkonde et al., "Stroke in rural India," *Stroke*, 2014.

[5] P. Heuschmann et al., "Prediction models in stroke," *Stroke*, 2020.

[6] H. Adams et al., "NIH Stroke Scale limitations," *Stroke*, 1999.

[7] M. Elkind et al., "C-reactive protein and stroke outcome," *Neurology*, 2014.

[8] J. Kim et al., "D-dimer as a prognostic marker," *Stroke*, 2020.

[9] S. Capes et al., "Hyperglycemia and outcome after stroke," *Stroke*, 2001.

[10] J. Heo et al., "Deep learning for stroke outcome prediction," *JAMA Netw. Open*, 2019.

[11] R. Zhu et al., "ML models for stroke outcome," *Stroke*, 2021.

[12] P. Bentley et al., "ML prediction in stroke," *Sci. Rep.*, 2014.

[13] S. Lundberg et al., "SHAP explainability," *Nat. Mach. Intell.*, 2020.

[14] E. Topol, "AI in medicine," *Nat. Med.*, 2019.

[15] M. Kuhn & K. Johnson, *Applied Predictive Modelling*, Springer, 2013.

[16] S. van Buuren, *Flexible Imputation of Missing Data*, CRC Press.

[17] L. Breiman, "Random Forests," *Mach. Learn*, 2001.

[18] T. Chen and C. Guestrin, "XGBoost," *KDD*, 2016.

[19] N. Chawla et al., “SMOTE,” *JAIR*, 2002.

[20] D. Powers, “Evaluation metrics,” *JMLT*, 2011.

[21] DeLong et al., “AUC comparison,” *Biometrics*, 1988.

[22] E. Steyerberg, *Clinical Prediction Models*, Springer, 2019.

[23] A. Vickers and E. Elkin, “Decision curve analysis,” *Med. Decis. Making*, 2006.

[24] F. Wolters et al., “Predictors of poor stroke outcome,” *Lancet Neurol.*, 2020.

[25] G. Saposnik et al., “Outcome predictors,” *Stroke*, 2011.

[26] K. Lees et al., “Thrombolysis efficacy (IST-3),” *Lancet*, 2010.

[27] D. Gordon et al., “HDL and vascular protection,” *NEJM*, 1989.

[28] A. Holzinger et al., “Explainable AI in medicine,” *Nat. Rev. AI*, 2020.

[29] T. Li et al., “Federated learning in healthcare,” *IEEE Intell. Syst.*, 2020.