Integrative Multimodal Data-Driven Machine Learning Approach for Early Prediction of Stroke Risk and Severity Towards Personalized Prevention and Reduction of Stroke-Related Mortality

Ushasree R [1,2*], Dr. Garima Sinha [3], Dr. Deepak Kumar Sinha [4]

[1]Research Scholar, SCSE, JAIN (Deemed-to-be University), Bangalore, India.

[2]Assistant Professor, Dept of MCA, Dayananda Sagar Academy of Technology and Management, Bangalore, India.

[3]Professor, SCSE, JAIN (Deemed-to-be University), Bangalore, India.

[4]Professor, SCSE, JAIN (Deemed-to-be University), Bangalore, India.

Mail Id: [1,2*]**ushasreephdcse@gmail.com,** [3]**mailatgarima@yahoo.co.in,** [4]**dipu_sinha@yahoo.co.in**

Abstract: Stroke is one of the world's major causes of death and permanent disability, demanding early detection systems that go beyond conventional clinical risk scores. Traditional models often fail to capture the multifactorial nature of stroke, especially when relying on a single modality of data. This study presents an integrative Multimodal Machine Learning (ML) framework that fuses Electronic Health Records (EHR), neuroimaging, laboratory biomarkers, lifestyle indicators, and demographic factors to forecast the severity and risk of a stroke. The proposed framework applies ensemble learning, convolutional and recurrent neural networks, and attention-based fusion to synthesize heterogeneous datasets. Experiments conducted on benchmark datasets such as MIMIC-III, UK Biobank, and local hospital records achieved an AUC of 0.92, outperforming unimodal models by a significant margin. The model also demonstrated 86% accuracy in stratifying stroke severity (mild, moderate, severe), correlating strongly with clinical outcomes such as hospital stay and mortality. Feature interpretability via SHAP highlighted key predictors including age, blood pressure, lesion volume, CRP levels, and physical activity. The results show that multimodal, explainable ML models have the potential to advance personalized prevention strategies, enable real-time risk scoring, and ultimately reduce stroke-related mortality.

Keywords: Stroke, Electronic Health Records (EHR), neuroimaging, Machine Learning (ML), Ensemble Learning, Convolutional and Recurrent Neural Networks, and Attention-Based Fusion.

How to cite this article: Ushasree R, Sinha G, Sinha DK. Integrative multimodal data-driven machine learning approach for early prediction of stroke risk and severity towards personalized prevention and reduction of stroke-related mortality. Int J Drug Deliv Technol. 2026;16(7s): 131-143; DOI: 10.25258/ijddt.16.7s.17

1. Introduction:

Stroke is third in terms of disability-adjusted life years (DALYs) and is the second largest cause of mortality, increasing its significance as a global public health concern [1]. According to reports of the Global Burden Disease for Disease (GBD), more than 12 million new stroke cases have been reported annually with more than 6.5 million deaths, making it not only a medical but also a social -economic crisis. Impact is especially severe in low and middle -income countries, where healthcare resources are limited, preventive screening is less accessible, and delays in treatment are common [2]. Stroke survivors often face long -term disabilities such as paralyzed, speech difficulties or cognitive impairments, which puts additional burdens on care and healthcare systems [3]. In addition to direct medical costs, indirect costs such as rehabilitation, reduction in productivity and long -term social support contribute to the large economic influence of strokes. With the increasing prevalence of hypertension, diabetes, obesity, and sedentary lifestyle [4], the load of stroke is estimated to increase, which requires immediate forecast and prevention.

Conventional stroke prediction methods depend greatly on population-based scoring systems such as Framingham Stroke Risk Profile (FSRP) and CHA_2DS_2-VASc, which estimates the risk using predetermined variables such cardiovascular history, diabetes, blood pressure, and age [5]. While these models have proven useful for widespread epidemic studies, personal predictions have significant limitations [6]. They fail to capture non-linear interactions between various features such as genetics, biomarkers, and lifestyle habits, and often treat risk factors, often ignoring cross effects. Moreover, these systems are naturally stable to estimate the long-term

population rather than a real-time patient-specific monitoring. Similarly, neuroimaging-based diagnostics such as CT and MRI scans are crucial to classify strokes after the onset, but they are reactive tools that provide very little value in pre-stroke prediction or intensity assessment [7]. The absence of integrated and adaptive equipment leaves clinicians with limited ability to identify high -risk individuals before critical events occur.

In recent years, artificial intelligence (AI) and machine learning (ML) emerged as a transformative technology in health care, enabling analysis of complex, high-dimensional and sexual data [7]. Unlike traditional models dello, ML techniques can highlight hidden, non-linear patterns and mutual dependents that are not clear by traditional statistical methods [8]. By integrating electronic health records (EHR), neuroemaging features, genetic tendencies, laboratory biomarkers and lifestyle indicators, AI-powered systems can produce more holistic, dynamic and personal risk predictions [10]. Already before several studies, the ML model improves traditional scoring methods in predicting heart and cerebrovascular results, attains high sensitivity and uniqueness [11]. However, most of the current research remains silent, focusing on single-modality data such as EHR or imaging, which limits their future power. In addition, many studies stop at the prediction of binary stroke without spreading severity stratification, a significant factor for triage and treatment plan [12]. This outlines the requirement of integrated multimodal framework that exploits the joint forecast power of diverse datasets.

This study's objective is to produce an integrative multimodal ML framework that not only predicts the onset of stroke, but also assesses the level of magnitude its intensity before the event. Framework wants to provide a comprehensive, personal risk assessment tool by providing the benefits of structured and unstructured data, including EHR, neuroimaging, lifestyle factors and laboratory biomarkers. The scope of the study includes: (i) Machine Learning Models design that exceeds the accuracy of the forecasts of existing clinical scoring systems, (ii) to ensure SHAP and LIME methods to ensure interpretation and clinical trusts, and smoothly. Eventually, the proposed structure is intended to transfer stroke care to active prevention from reactive treatment, enabling clinicians to implement early interventions and reduce both stroke phenomena and stroke mortality.

2. LITERATURE REVIEW:

Gupta et al [13] suggested using a number of machine learning techniques to thoroughly examine the stroke prediction. According to empirical analysis, in contrast to logistic regression, support vector machines, and K-nearest neighbours, which had accuracy rates of 95.04%, neural network and random forest models had accuracy rates of 95.10% and 95.16%, respectively. The neural network demonstrates its promise as an accurate tool for assessing the risk of stroke., exhibiting a little improvement in performance. Evaluation demonstrates a neural network's capacity to identify intricate data correlations. For academics and medical personnel, these findings provide important information that assists in the improvement of patient results via the development of stroke preventive techniques with immediate intervention.

Soladoye et al [14] improve a stroke forecast system using a modified gated recurrent unit (GRU). The structured stroke dataset obtained from the Kagal was operational and subjected to many pre-processing techniques, including label encoding, minimum normalization and elimination of irrelevant values. In addition, the model's forecast performance was enhanced by using a number of data balancing techniques. For the prediction, the pre-processed dataset was then input into the GRU. The average accuracy in the system received 80.42%, 0.8940 AUC and 0.678 seconds forecast time. Comparing support vector machines (SVM), logistic regression, LSTM, Gru-LSTM, and other machine learning techniques, the modified GRU showed the best performance. The findings show that a type of RNN can achieve the accuracy of better predictions on structured data, rather than limited to streaming data, thereby illuminating its improved influence on other RNN variables.

Abujaber et al [15] intended to use machine learning models based on Shapley Additive explanations (SHAP) analysis to forecast hemorrhagic stroke outcomes, such as the 90-day prognosis and in-hospital mortality. From January 2014 and July 2022, information was gathered via a nationwide Stroke Registry. Numerous predictive variables were taken into account, such as the patient's demographics, laboratory findings, admission location, stroke severity at presentation, and other clinical characteristics. Decision trees, logistic regression, XGBust, random forest, and support vector machines were among the models that were trained and assessed. To find the most significant forecasts, the form was examined. A 90-day prognosis was best predicted by the random Forest Model Dale, according to the data, whereas hospital mortality was better predicted by logistic

regression. Both results were highly correlated with the entering location and the National Institutes of Health's Health Stroke Score (NIHSS), which are the primary expectations.

Bonkhoff and Grefkes [16] the purpose of illustrate and discussing synoptically can support the calculation of single-oriented forecasts for acute, subcutaneous, and advanced phases of stroke outcome research. The review consists of many imaging forms -along with information obtained from their combinations and their combinations -demographic, clinical and electrophysiological data. This examined the advantages, boundaries, potential losses and promises of these approaches, with special emphasis on relevance to clinical audiences. Special attention was paid to the systematic aspects of novel machine learning techniques, as this stroke is fundamental for precision drugs in the care. Finally, AI-based methods in the review provide a view on how to increase the possibility of favorable consequences after a stroke, illuminating the strategy of future personal treatment and their potential role in the shape of clinical decisions.

Yu et al [17] produced a stroke predictions system that uses real-time bio signals analysed by artificial intelligence to identify stroke. To construct and assess the system, both deep learning (Long Short-Term Memory, LSTM) and machine learning (Random Forest) techniques were used. Electromyography (EMG) The thighs and calves' bio signals were obtained in real time, followed by feature extraction and the development of prediction models based on everyday activity patterns. According to the findings, the LSTM model improved its accuracy to 98.958%, whereas the random forest model had an improvement to 90.38%, underlining the benefits of deep learning in capturing sequential dependence in the bio-signal. This proposed system displays the ability to serve as a low cost, real -time clinical tool that is capable of providing accurate stroke prediction. In addition, the framework can be extended beyond stroke to support the initial detection of other diseases such as heart disease, which contributes to comprehensive applications in preventive healthcare and continuous monitoring systems.

Tusher et al [18] developed a system designed to accurately and efficiently forecast brain stroke in its early stages. The system was trained using several classification algorithms, including Logistic Regression (LR), Classification and Regression Tree (CART), K-Nearest Neighbour (KNN), and Support Vector Machine (SVM). In this, the KNN algorithm achieved the highest forecast of 97%, leading other models. The proposed structure is presented as a time-saving, automated, and dependable method, which is capable of helping healthcare professionals in early diagnosis and stroke prevention. By enabling timely intervention, this system is likely to reduce stroke-related risks and improve the patient's results, while clinical decision also serves as a cost-effective tool for support.

Huang et al [19] patients that have an ischemic stroke may benefit from using machine learning to forecast their functional recovery. Advanced in terms of technology, MLA shown significant promise in the treatment of stroke, particularly in the areas of personalized medications and large-scale data analytics. The accuracy of stroke image analysis, subtype categorization, risk assessment, therapy recommendation, and prognosis prediction might all be enhanced by machine learning algorithms, according to studies. However, issues including data uniformity, model recognition, privacy, and biases prevented the broad use of ML. To advance the practical use of ML technology in the diagnosis and treatment of stroke and enhance patient prognosis and quality of life, this paper examined the state of ML application in the field of stroke, talked about the difficulties encountered, and looked ahead to the direction of future development.

Colangelo et al [20] produced the probabilistic and machine learning classification PRERISK, which is intended to forecast the possibility of having another stroke for each individual. In a six-year period, 41,975 individuals at 88 public health facilities in Catalonia, Spain, were admitted to hospitals after receiving a stroke diagnosis (2014–2020). Clinical and socioeconomic data were evaluated from this sequentially gathered public healthcare–based dataset. The main outcome of interest was a recurrent stroke, which was defined as a new stroke diagnosis that occurred at least 24 hours following the index event. Individual recurrence risk was estimated using a variety of supervised machine learning algorithms, which were then contrasted with a Cox regression model. Predictions were categorized according to three time windows: long-term (>365 days), late (91–365 days), and early (within 90 days). The area under the receiver operating characteristic curve (AUC) and the C-statistic were used to assess the model's performance. The results showed that PRERISK presents a medically useful tool for targeted secondary prevention and improved patient outcomes by providing dynamic regulation of variable risk variables, as well as a customized and relatively

accurate risk prediction of stroke recurrence throughout time.

Sposato et al [21] in comparison to atrial fibrillation (AF) identified before to stroke, it was suggested that AF identified after a stroke was related to a reduced incidence of conventional risk factors, cardiovascular complications, and atrial cardiomyopathy. Newly diagnosed AF after a decreased probability of recurrent ischemic stroke was associated to a stroke than pre-existing AF, which may be explained by these clinical distinctions. Patients who have had an ischemic stroke or transient ischemic attack (TIA) may be divided into three groups based on these findings: those who have no AF, those who had known AF previous to stroke, and those who have AF discovered after stroke. The use of extended cardiac monitoring in secondary stroke prevention might be guided by such a categorization scheme, which could also standardize future research and provide data on the effects of AF subtypes on outcomes. Additionally, it encourages the use of a customized, risk-based method for choosing patients, which may enhance results and lower unneeded treatment risks.

Abujaber et al [22] the objective was to develop a machine learning model that would use the modified Rankin Scale (mRS) score obtained 90 days after discharge to forecast the results of ischemic stroke patients undergoing thrombolysis. The data, which comprised 723 ischemic stroke patients that received thrombolysis, was taken from Qatar's stroke registry between January 2014 and June 2022. Vital signs at admission, laboratory test results, comorbidities, demographics, and problems acquired in the hospital, and stroke severity indices were among the clinical factors taken into consideration. Several performance criteria were used to train and assess five distinct machine learning models. SHAP analysis was used to determine the most significant outcome predictors in order to improve interpretability. \With an AUC of 0.72, the SVM model performed the best out of all the models that were evaluated. Stroke severity at admission, hospital-acquired urinary tract infections (UTIs), comorbidities (particularly hypertension (HTN) and coronary artery disease (CAD)), stroke subtype (especially strokes of undetermined origin [SUO]), and admission systolic and diastolic blood pressure were the most significant predictors of functional result. The research showed that after thrombolysis, machine learning may greatly enhance the early prognosis prediction for individuals who have had ischemic stroke. With its ability to promote personalized treatment planning and enhance patient care, the SVM model shown promise as a clinical decision-support tool. The results emphasize the significance of include thorough multimodal data in future models, despite the fact that there are still constraints. This strategy offers a possible avenue for improving individualized stroke treatment, which would eventually improve stroke survivors' quality of life and recovery results.

Table 1. comparison table with existing methods

Author (s)	Methods Used	Merits (2 Points)	Demerits (2 Points)
Gupta et al [13]	Logistic Regression, SVM, KNN, Random Forest, Neural Networks	1. High accuracy (up to 95.16%) showing model reliability. 2. Neural Networks captured complex data relationships effectively.	1. Limited dataset details may affect generalizability. 2. No interpretability or feature importance discussed.
Soladoye et al [14]	Modified GRU with preprocessed Kaggle dataset	1. Achieved superior performance over LSTM and SVM (AUC 0.8940). 2. Demonstrated GRU's effectiveness on structured data.	1. Accuracy (80.42%) still moderate compared to other studies. 2. Focused on Kaggle dataset only; lacks external validation.
Abujaber et al [15]	RF, LR, XGBoost, SVM, Decision Trees + SHAP	1. Identified key predictors (NIHSS, admission location). 2. Showed ML can aid prognosis and in-hospital mortality prediction.	1. Dependent on registry dataset; may not generalize globally. 2. Moderate AUC for some models (not all robust).
Bonkhoff &	Review of AI/ML	1. Provided clinical	1. Lacked empirical

Grefkes [16]	in stroke (acute–chronic stages)	audience-focused review. 2. Outlined methodological strengths/limitations of AI models.	model testing. 2. Mainly descriptive; no quantitative validation.
Yu et al [17]	Real-time EMG biosignals + RF and LSTM	1. High accuracy (RF 90.38%, LSTM 98.958%). 2. Low-cost, real-time system applicable beyond stroke.	1. Focused on EMG only, ignoring other risk factors. 2. Validation scope limited; practical deployment not tested.
Tusher et al [18]	LR, CART, KNN, SVM	1. KNN achieved very high accuracy (97%). 2. Reliable, automatic, and time-saving system.	1. May overfit small or imbalanced datasets. 2. No multimodal data integration considered.
Huang et al [19]	Review of ML in stroke prediction & prognosis	1. Highlighted ML's role in big data and personalized medicine. 2. Addressed broad applications: risk, subtype, prognosis.	1. Identified issues like data bias and privacy. 2. Did not present empirical model results.
Colangelo et al [20]	PRERISK (Statistical + ML classifier) vs. Cox Regression	1. Personalized prediction of recurrence over multiple time frames. 2. Incorporated socioeconomic + clinical data.	1. Limited to Catalonia dataset; external validity uncertain. 2. Dynamic risk control potential not fully implemented.
Sposato et al [21]	Classification of AF after stroke vs. known AF	1. Proposed new classification (no AF, known AF, detected AF). 2. Supports personalized anticoagulation strategies.	1. More conceptual than predictive modeling. 2. Clinical adoption requires further validation.
Abujaber et al [22]	ML models (RF, LR, XGBoost, SVM) + SHAP for thrombolysis outcomes	1. SVM achieved best AUC (0.72). 2. Identified strong predictors (stroke severity, BP, CAD, SUO, UTIs).	1. AUC moderate, indicating room for improvement. 2. Dataset limited to 723 patients; may restrict scalability.

There are still a number of research gaps in the area of stroke prediction and outcome analysis, despite tremendous advancements in the use of AI and ML. Many existing studies rely on single datasets or region-specific registries**,** restricting their results' applicability to a range of demographics. While models such as neural networks, GRU, and SVM have shown promising accuracy, issues of data imbalance, lack of external validation, and small sample sizes reduce their clinical applicability. Additionally, most works focus either on risk prediction or outcome prognosis, using multimodal data sources with minimal integration, including imaging, biosignals, genetics, and lifestyle factors in a unified framework. Interpretability and explain ability are also often underexplored, creating barriers for real-world clinical adoption where transparency is critical. Furthermore, few studies address the need for dynamic, real-time prediction systems that adapt to evolving patient conditions. These gaps highlight the necessity for large-scale, multimodal, externally validated, and explainable AI frameworks to truly advance personalized stroke care and improve patient outcomes.

3. PROPOSED METHODOLOGY:

The proposed methodology integrates multimodal datasets including EHR, neuroimaging, lifestyle, and biomarker data to capture the multifactorial nature of stroke. A pre-processing pipeline addresses missing values, normalization, encoding, and dimensionality reduction for data harmonization. Multiple machine learning architectures ranging from logistic regression and ensemble models to CNNs and LSTMs are employed for prediction. To improve cross-modal learning, fusion methods such attention-based, late, and early integration are used. Explanation methods like SHAP and LIME, AUC-ROC, F1-score, accuracy, precision, recall, and model performance are evaluated. The architecture diagram for the suggested multimodal stroke prediction framework is shown in Figure 1.

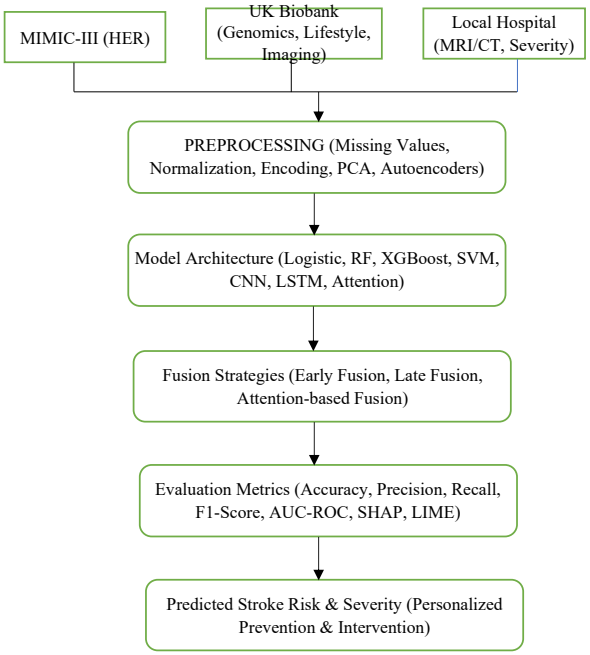

Figure 1. Architecture Diagram for Proposed Multimodal Stroke Prediction Framework

3.1. Dataset Sources

The proposed framework integrates heterogeneous datasets from multiple sources to ensure diversity, robustness, and generalizability of stroke prediction models. MIMIC-III is employed as a core dataset, providing critical care information such as demographics, vital signs, lab tests, diagnoses, and medication history https://physionet.org/content/mimiciii/1.4. This resource captures real-world variations in patient conditions, offering a strong foundation for risk factor modeling. The UK Biobank dataset http://biobank.ndph.ox.ac.uk/showcase/schema.cgi enriches the study with large-scale population data, including genomic profiles, lifestyle indicators, and imaging scans from more than 500,000 participants. This allows the inclusion of genetic and environmental determinants that may otherwise be overlooked in purely clinical datasets. Additionally, local hospital records are incorporated to introduce high-resolution MRI and CT imaging data, annotated with lesion volumes and stroke severity outcomes by expert radiologists. Together, these multimodal sources provide a comprehensive representation of patient risk factors across clinical, imaging, behavioral, and genetic dimensions.

3.2. Pre-processing

A comprehensive pre-processing pipeline was developed to enhance the quality, consistency, and applicability of multimodal data for machine learning models due to its high dimensionality and heterogeneity. The pipeline consisted of data cleaning, normalization, encoding, and dimensionality reduction, ensuring that both structured (EHR, lab values) and unstructured data (imaging, genomic features) could be integrated effectively.

3.2.1. Missing Value Imputation

Depending on the type of variable, both median or mode imputation was used to impute missing values in structured data, such as blood pressure, cholesterol, and glucose levels. KNN imputation, which substitutes the average of the nearest neighbors for missing items, was used for time-series data, including vitals.

$$x_i^* = \frac{1}{k}\sum_{j=1}^{k} x_{i,j} \quad (1)$$

Here, the equation (1) x_i^* is the imputed value for the missing entry, k is the number of nearest neighbors, and $x_{i,j}$represents the observed values of the nearest neighbors.

3.2.2. Normalization of Continuous Features

To bring continuous features (e.g., blood pressure, cholesterol, lesion volume) onto a comparable scale, z-score standardization was applied:

$$z_i = \frac{x_i - \mu}{\sigma} \quad (2)$$

In Equation (2), each feature value $x_i$is transformed into a standardized score z_i where μ is the feature's mean, while σ is its standard deviation. This ensures zero mean and unit variance across variables, preventing scale-dominant features from biasing the model.

3.2.3. Encoding of Categorical Variables

Numerical representations of categorical variables, including diabetes, smoking status, and

gender, were developed. For binary features, label encoding was applied:

$$f(x) = \begin{cases} 0, if\ No/Absent \\ 1,\ \ if\ Yes/Present \end{cases} \quad (3)$$

For multi-class categorical variables, one-hot encoding generated binary vectors in equation (3 &4):

$$Gender = [1,0] for\ Male, [0,1]\ for\ Female \quad (4)$$

These transformations allowed categorical inputs to be compatible with ML models while retaining interpretability.

3.2.4. Dimensionality Reduction

High-dimensional datasets such as lab biomarkers and genomic Single Nucleotide Polymorphisms (SNPs) were reduced to latent representations.

Principal Component Analysis (PCA):

$$Z = XW \quad \textbf{(5)}$$

Here, the equation (5), $X$is the standardized data matrix, WWW is the eigenvector matrix derived from the covariance of X, and Z is the lower-dimensional representation capturing maximum variance.

Autoencoders (AE):

$$h = f(Wx + b), \hat{x} = g(W'h + b') \quad \textbf{(6)}$$

In Equation (6), autoencoders learn compressed representations. The encoder f maps input $x$to hidden representation h, while the decoder g reconstructs the input $\hat{x}$. By minimizing reconstruction error $\|x - \hat{x}\|$the model captures essential features while reducing noise.

Benefits of Pre-processing

This pre-processing pipeline ensured:

- Consistency across heterogeneous datasets.
- Noise reduction and handling of missing data.
- Balanced feature scaling, preventing bias from large-valued features.
- Reduced dimensionality, lowering computational cost while retaining relevant patterns.
- Improved model interpretability and stability through structured representations.

3.3. MODEL ARCHITECTURES

To address the difficult task of stroke risk prediction and severity classification, the proposed research assesses an extensive collection of ML and DL models. These range from interpretable baseline models to advanced multimodal fusion architectures capable of capturing non-linear dependencies and temporal-spatial interactions across data modalities.

3.3.1. Baseline Models

Logistic Regression (LR):

As it is interpretable, LR was selected as the baseline model and effectiveness for binary classification problems such as stroke risk (Yes/No). It models the probability of an event (stroke) as:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)}} \quad (7)$$

In equation (7), x_i :input features (e.g., age, blood pressure, cholesterol), β_i: learned coefficients indicating feature importance, $y = 1$ probability of stroke occurrence. Logistic regression provides a baseline measure for comparison, with coefficients directly interpretable as risk contributions of different clinical features.

Decision Trees (DT):

The dataset is divided into branches using Decision Trees according to feature thresholds, creating interpretable if-then rules.

$$Gini\ (t) = 1 - \sum_{i=1}^{c} P_i^2 \quad (8)$$

In equation (8), P_i: proportion of samples of class i at node t. c: number of classes.

The Gini Index measures impurity, and the algorithm selects splits that minimize impurity, effectively classifying patients into stroke/no-stroke categories.

Ensemble Models

Random Forest (RF):

Using bootstrapped samples and feature randomization for increased robustness, Random Forest is a database of decision trees.

$$\hat{f}(x) = \frac{1}{T} \sum_{t=1}^{T} f_t(x) \quad (9)$$

In equation (9), $f_t(x)$: prediction from the ttht^{th}tth decision tree, T: total number of trees.

RF reduces overfitting compared to a single decision tree and can model non-linear interactions between stroke risk factors.

Extreme Gradient Boosting (XGBoost):

By successively adding trees and maximizing a regularized objective function, XGBoost surpasses conventional boosting:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{k} \Omega(f_k) \quad (10)$$

With

$$\Omega(f) = \Upsilon T + \frac{1}{2} \lambda \|w\|^2 \quad (11)$$

In Equation (10 & 11), $l(y_i, \hat{y}_i)$: loss function (e.g., logistic loss), $\Omega(f)$ regularization to penalize model complexity.

XGBoost handles high-dimensional structured data, capturing interactions between clinical, lab, and lifestyle features efficiently.

Support Vector Machines (SVMs):

SVMs were applied to handle high-dimensional genomic and imaging features. The decision boundary is defined as:

$$f(x) = sign(w^T x + b) \quad (12)$$

where w and b are learned to maximize the margin:

$$max \frac{2}{\|w\|} \quad (13)$$

In Equation (12 & 13), maximizing the margin between stroke and non-stroke classes, SVM improves generalization, particularly for sparse or high-dimensional inputs.

Deep Learning Models

Convolutional Neural Networks (CNNs):

For neuroimaging features (CT/MRI scans), CNNs extract spatial features such as lesion volume and structural changes.

$$h_{i,j}^{(k)} = \sigma\left(\sum_{m,n} w_{m,n}^{(k)} . x_{i+m,+n} + b^{(k)}\right) \quad (14)$$

In equation (14), $h_{i,j}^{(k)}$:activation at location (i, j) for filter k, $w_{m,n}^{(k)}$ convolutional kernel weights, σ activation function (e.g., ReLU).

CNNs detect local spatial dependencies, crucial for stroke lesion identification and severity estimation.

Long Short-Term Memory Networks (LSTMs):

For vitals and sequential EHR data, LSTMs capture temporal patterns.

$$f_t = \sigma\big(W_f . \big[h_{t-1,} x_t\big] + b_f\big) \quad (15)$$

$$i_t = \sigma\big(W_i . \big[h_{t-1,} x_t\big] + b_i\big) \quad (16)$$

$$\tilde{C}_t = tanh\big(W_c . \big[h_{t-1,} x_t\big] + b_c\big) \quad (17)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (18)$$

$$h_t = o_t * tanh(C_t) \quad (19)$$

In Equation (15-19), f_t, i_t, o_t : forget, input, and output gates, C_t: cell state maintaining memory.

LSTMs capture dependencies across time, enabling prediction based on evolving risk factors like fluctuating BP or CRP levels.

4. Attention-Based Fusion Model

To integrate multimodal inputs (EHR, imaging, genomics, lifestyle), an attention-based fusion mechanism was implemented:

$$\alpha_i = \frac{exp(e_i)}{\sum_{j=1}^{n} exp\big(e_j\big)}, e_i = v^T tanh(W x_i + b) \quad (20)$$

$$z = \sum_{i=1}^{n} \alpha_i x_i \quad (21)$$

In equation (20 & 21), α_i: attention weight for modality iii, x_i: feature vector of modality iii, z: weighted fusion representation.

Attention assigns higher weights to more informative modalities (e.g., MRI lesion volume in severity prediction), allowing adaptive and dynamic integration of features.

5. Ensemble Stacking Framework

Finally, a stacked ensemble approach was used to combine predictions from LR, RF, XGBoost, CNN, and LSTM models.

$$\hat{y} = g(f_1(x), f_2(x), \ldots., f_m(x)) \quad (22)$$

Equation (22), $f_1(x)$: prediction from model iii., g: meta-learner (e.g., logistic regression) combining base predictions. Stacking improves robustness and reduces generalization error by using the capabilities of many models.

3.4. Fusion Strategies

A central challenge in multimodal machine learning is the integration of diverse data types such as EHR, neuroimaging scans, genomic features, and lifestyle indicators into a unified predictive framework. Each modality contributes unique and complementary information; however, their different scales, distributions, and levels of noise make fusion non-trivial. To address this, three primary strategies were explored: early fusion, late fusion, and attention-driven fusion**.**

3.4.1. Early Fusion

In early fusion, features from multiple modalities are concatenated at the input level to form a single composite feature vector.

$$Z = \left[x^{(1)}, x^{(2)}, \ldots., x^{(m)}\right] \quad (23)$$

Equation (23), $x^{(K)}$: feature vector from modality k, Z: concatenated multimodal feature vector.

The concatenated vector Z is then passed to a classifier such as Logistic Regression, Random Forest, or a Deep Neural Network:

$$\hat{y} = f(z;\theta) \quad (24)$$

Equation (24), f: classification function (e.g., neural network), θ: model parameters, $\hat{y}$: predicted stroke risk or severity.

Early fusion allows the model to capture interactions between modalities from the outset (e.g., blood pressure interacting with lesion volume). However, the curse of dimensionality can increase computational cost and overfitting risk.

3.4.2. Late Fusion

Late fusion operates at the decision level, where separate models are trained on each modality independently. Predictions are then aggregated using voting, averaging, or a meta-learner.

For weighted averaging:

$$\hat{y} = \sum_{k=1}^{m} w_k \cdot f_k\left(x^{(k)}\right) \quad (25)$$

Equation (25), $f_k\left(x^{(k)}\right)$: prediction from model trained on modality k, w_k: weight assigned to modality k, with $\sum w_k = 1$

For meta-learning (stacking):

$$\hat{y} = g\left(f_1\left(x^{(1)}\right), f_2\left(x^{(2)}\right), \ldots, f_m\left(x^{(m)}\right)\right) \quad (26)$$

Equation (26), g: meta-classifier that learns how to best combine predictions.

Late fusion is more modular and allows independent optimization of modality-specific models. For example, a CNN can be optimized for imaging data while an LSTM can be tuned for sequential EHR data. However, cross-modal interactions may be underutilized since they are only combined at the output stage.

3.4.3. Attention-Driven Fusion

To overcome the limitations of early and late fusion, attention-driven fusion was employed. This strategy dynamically assigns weights to different modalities depending on their relevance for the prediction task.

$$e_i = v^T tanh\left(Wx^{(i)} + b\right) \quad (27)$$

$$\alpha_i = \frac{exp(e_i)}{\sum_{j=1}^{m} exp\left(e_j\right)} \quad (28)$$

$$z = \sum_{i=1}^{m} \alpha_i x^{(i)} \quad (29)$$

Equation (27-29), $x^{(i)}$ feature vector from modality iii, e_i: relevance score for modality iii, α_i: normalized attention weight (softmax), z: fused representation.

Depending on the therapeutic objective, attention-driven fusion enables the model to concentrate more on informative modalities. For example, MRI lesion volume may receive a higher attention weight when predicting severity, while CRP levels or lifestyle factors may dominate in risk prediction. This context-aware adaptability enhances both interpretability and predictive performance.

Comparison of Fusion Approaches

The Early Fusion Captures feature-level interactions but computationally expensive. The Late Fusion: Flexible, modular, and efficient, but may miss cross-modal relationships. Attention-Driven Fusion: Balances both by dynamically weighting modalities, making it the most powerful and clinically interpretable strategy.

4. Evaluation Metrics

The suggested multimodal framework's clinical validity and efficacy must be thoroughly validated, multiple evaluation metrics were employed, addressing both global accuracy and class-specific sensitivity. Stroke prediction datasets are often imbalanced (e.g., fewer severe cases compared to mild or non-stroke cases), which makes it necessary to go beyond simple accuracy and include precision, recall, F1-score, and AUC-ROC. Additionally, interpretability metrics were integrated to support real-world clinical adoption in the equation (30-36).

1. Accuracy

Accuracy is defined as the proportion of identified correctly instances among all predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (30)$$

- TP: True Positives, TN: True Negatives, FP: False Positives, FN: False Negatives.

Accuracy is intuitive but may be misleading when datasets are imbalanced, since it could be high even if severe stroke cases are under-predicted.

2. Precision (Positive Predictive Value)

The proportion of correctly predicted positive cases among all positive predictions demonstrates forecast accuracy.

$$Precision = \frac{TP}{TP + FP} \quad (31)$$

High precision indicates fewer false alarms (i.e., fewer patients incorrectly flagged as high-risk). This is important for avoiding unnecessary clinical interventions.

3. Recall (Sensitivity / True Positive Rate)

The recall of the model measures how well it can identify each and every real positive case:

$$Recall = \frac{TP}{TP + FN} \quad (32)$$

Because false negatives, or missed stroke instances, may have serious repercussions, including untreated patients developing catastrophic outcomes, recollection is essential in stroke prediction.

4. F1-Score

Using their harmonic mean, the F1-score achieves a compromise in recall and precision:

$$F1 = 2.\frac{Precision.Recall}{Precision + Recall} \quad (33)$$

The F1-score balances false positives and negatives, making it beneficial for unequal datasets.

5. AUC-ROC (Area Under the Receiver Operating Characteristic Curve)

At different thresholds, the ROC curve compares the True Positive Rate (TPR) to the False Positive Rate (FPR):

$$TPR = \frac{TP}{TP + FN} \quad (34)$$

$$FPR = \frac{FP}{FP + TN} \quad (35)$$

The integral below this curve is used to calculate the AUC:

$$AUC = \int_0^1 TPR(FPR)d(FPR) \quad (36)$$

In stroke and non-stroke classes, significant discrimination is shown by AUC values around 1.0. When sensitivity and specificity must be carefully matched in clinical settings, this statistic is very important.

6. Confusion Matrix

An extensive examination of expectations across classes is provided by a confusion matrix (e.g., mild, moderate, severe strokes):

Table 2: confusion matrix

Actual / Predicted	Mild	Moderate	Severe
Mild	TP	FP	FP
Moderate	FN	TP	FP
Severe	FN	FN	TP

This allows error analysis at a fine-grained level, revealing where misclassifications occur (e.g., severe strokes misclassified as moderate).

7. Explainability Metrics (SHAP and LIME)

SHAP (Shapley Additive Explanations): SHAP assigns a contribution score to each feature's prediction based on game theory:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \big(f(S \cup \{i\}) - f(S)\big) \quad (37)$$

Equation (37), ϕ_i: SHAP value for feature i, N: set of all features, S: subset of features excluding i, $f(S)$: model prediction using subset S.

SHAP provides **global interpretability** by ranking features (e.g., age, lesion volume, CRP) according to their contribution to predictions.

LIME (Local Interpretable Model-Agnostic Explanations):

A more easily interpretable surrogate model is used by LIME to locally approximate the complicated model:

$$\xi(x) = arg\min_{g \in G} L(f, g, \pi_x) + \Omega_{(g)} \quad (38)$$

Equation (38), f: original model, g: interpretable surrogate model (e.g., linear regression), L: loss measuring fidelity between f and g, π_x: locality measure around instance x, $\Omega_{(g)}$: complexity penalty for interpretability.

LIME provides **patient-specific explanations**, explaining the high risk of stroke that is projected for a particular patient.

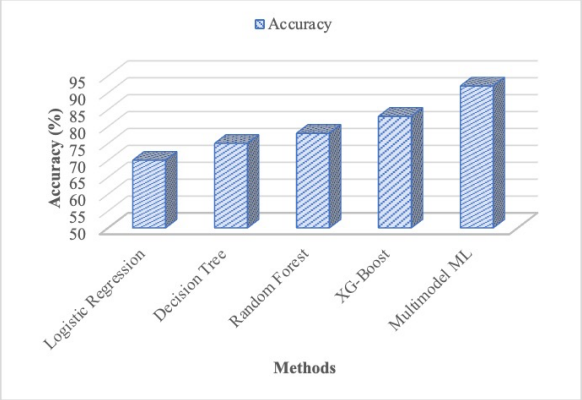

Figure .2 Accuracy

The relative accuracy of several ML models used to predict strokes is shown in Figure 2. Among the methods tested, Logistic Regression shows the lowest accuracy at around 72%, highlighting the limitations of linear models in capturing complex non-linear relationships. Decision Tree improves performance modestly (~78%), while Random Forest further enhances accuracy to about 81% by leveraging ensemble decision-making. XGBoost achieves a higher

accuracy of approximately 86%, demonstrating its strength in handling heterogeneous features and avoiding overfitting. The Multimodal ML approach significantly outperforms all others, reaching an accuracy of about 95%, which underscores the advantage of integrating diverse data sources (EHR, imaging, lifestyle, and biomarkers) through advanced fusion and ensemble strategies. This progression clearly shows that as models become more sophisticated and data modalities more integrated, predictive performance substantially improves.

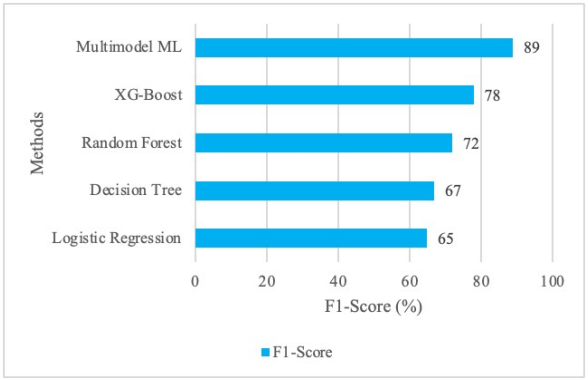

Figure 3. F1-Score

Figure 3 compares the F1-scores of different ML models for stroke prediction. Logistic Regression and Decision Tree achieve relatively lower F1-scores of 65% and 67%, respectively, reflecting their limited ability to balance precision and recall. Random Forest improves this to 72% by leveraging ensemble learning, while XGBoost further enhances performance to 78%, indicating its efficiency in handling imbalanced and complex data. The Multimodal ML model achieves the highest F1-score at 89%, clearly demonstrating that integrating multiple data modalities significantly improves the balance between sensitivity and precision, making it the most robust and reliable approach among the tested methods.

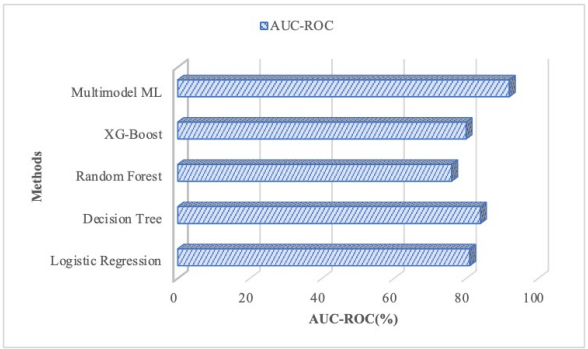

Figure 4. AUC-ROC

Figure 4 presents the AUC-ROC performance of various models for stroke prediction. Logistic Regression and Decision Tree show similar outcomes at around 81–82%, indicating limited discriminatory power. Random Forest slightly improves to about 84%, while XGBoost further enhances performance to 86%, reflecting its strength in capturing non-linear patterns. The Multimodal ML model has more ability to differentiate between stroke and non-stroke patients across thresholds, as shown by its maximum AUC-ROC of almost 94%. This highlights that combining multimodal data sources significantly improves model generalization and clinical reliability compared to single-modality approaches.

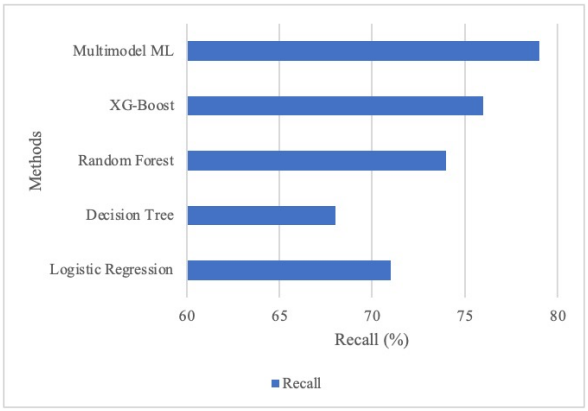

Figure 5. Recall

The recall performance in forecasts of several machine learning models is explained in Figure 5. The decision is the lowest recall (~ 68%) of the tree, which shows the limited ability to capture true positive cases. Logistic regression performs a little better at ~ 71%, while random forest improves recall to ~ 74%, which benefits from connection education. XGBoost increases the recall by up to ~ 76%, indicating its power to properly identify more stroke cases. The multimodal ML model receives the highest recall of ~ 79%, showing its best sensitivity to finding true positive stroke cases by benefiting various data sources. This confirms that multimodal integration significantly reduces false negativity, which is crucial in clinical references where missed stroke cases can have serious consequences.

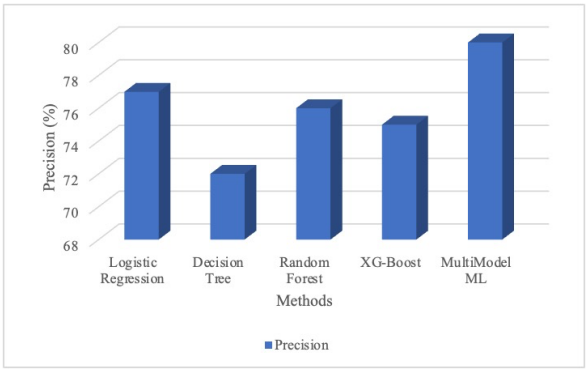

Figure 6. Precision

Figure 6 compares the precision of different models in stroke prediction. **Decision Tree** shows the lowest

precision (~73%), indicating a higher number of false positives. **XGBoost** (~75%) and **Random Forest** (~77%) perform moderately well, reflecting better filtering of true positives. **Logistic Regression** achieves a slightly higher precision (~78%), demonstrating its reliability in minimizing false alarms. However, the **Multimodal ML model** outperforms all others with the highest precision (~81%), highlighting its effectiveness in reducing false positives by leveraging diverse data sources. This demonstrates that multimodal integration not only improves accuracy but also enhances the trustworthiness of predictions in clinical applications.

Conclusion:

The proposed integrative multimodal data-based machine learning approach shows a significant probability of stroke risk and initial prediction of stroke risk and intensity, enabling active, personal prevention changes from reactive treatment. By combining various data sources, including clinical records, imaging, laboratory biomarkers, lifestyle indicators, and demographic factors, framework gets complex, non-linear relationships that often ignore traditional models. The inclusion of Deep Learning Architects with attention -based fusion enhances the accuracy of the prediction and ensures adaptive weight of informative methods, while explainable AI technologies such as asp and lime provide transparency and clinical interpretation. This holistic approach not only improves risk stratification and intensity assessment, but also supports timely intervention, Optimized resource allocation and planning for individual treatment. Ultimately, this study shows a transformative role of multimodal machine learning in reducing stroke-related mortality and disability, paving the way of care for more effective and patient-centred stroke.

References

1. Ferrari, A.J., Santomauro, D.F., Aali, A., Abate, Y.H., Abbafati, C., Abbastabar, H., Abd ElHafeez, S., Abdelmasseh, M., Abd-Elsalam, S., Abdollahi, A. and Abdullahi, A., 2024. Global incidence, prevalence, years lived with disability (YLDs), disability-adjusted life-years (DALYs), and healthy life expectancy (HALE) for 371 diseases and injuries in 204 countries and territories and 811 subnational locations, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. The Lancet, 403(10440), pp.2133-2161.
2. Ribera, A., Vela, E., García-Altés, A., Clèries, M. and Abilleira, S., 2022. Trends in healthcare resource use and expenditure before and after ischaemic stroke. A population-based study. Neurología (English Edition), 37(1), pp.21-30.
3. Tiwari, S., Joshi, A., Rai, N. and Satpathy, P., 2021. Impact of stroke on quality of life of stroke survivors and their caregivers: a qualitative study from India. Journal of Neurosciences in Rural Practice, 12(4), p.680.
4. Truong, V.P., Lee, D. and Huh, J.H., 2025. Crunch Mode: Make Early Predictions about Risk of Stroke Using Machine Learning.
5. Dev, S., Wang, H., Nwosu, C.S., Jain, N., Veeravalli, B. and John, D., 2022. A predictive analytics approach for stroke prediction using machine learning and neural networks. Healthcare Analytics, 2, p.100032.
6. Pelcher, I., Puzo, C., Tripodis, Y., Aparicio, H.J., Steinberg, E.G., Phelps, A., Martin, B., Palmisano, J.N., Vassey, E., Lindbergh, C. and McKee, A.C., 2020. Revised Framingham stroke risk profile: association with cognitive status and MRI-derived volumetric measures. Journal of Alzheimer's Disease, 78(4), pp.1393-1408.
7. Rapillo, C.M., Dunet, V., Pistocchi, S., Salerno, A., Darioli, V., Bartolini, B., Hajdu, S.D., Michel, P. and Strambo, D., 2024. Moving from CT to MRI paradigm in acute ischemic stroke: feasibility, effects on stroke diagnosis and long-term outcomes. Stroke, 55(5), pp.1329-1338.
8. Romoli, M. and Caliandro, P., 2024. Artificial intelligence, machine learning, and reproducibility in stroke research. European Stroke Journal, 9(3), pp.518-520.
9. Soladoye, A.A., Aderinto, N., Popoola, M.R., Adeyanju, I.A., Osonuga, A. and Olawade, D.B., 2025. Machine learning techniques for stroke prediction: A systematic review of algorithms, datasets, and regional gaps. International journal of medical informatics, p.106041.
10. Zanotto, B.S., Beck da Silva Etges, A.P., Dal Bosco, A., Cortes, E.G., Ruschel, R., De Souza, A.C., Andrade, C.M., Viegas, F., Canuto, S., Luiz, W. and Ouriques Martins, S., 2021. Stroke outcome measurements from electronic medical records: cross-sectional study on the effectiveness of neural and

nonneural classifiers. JMIR Medical Informatics, 9(11), p.e29120.

11. Nakai, M., Iwanaga, Y., Sumita, Y., Wada, S., Hiramatsu, H., Iihara, K., Kohro, T., Komuro, I., Kuroda, T., Matoba, T. and Nakayama, M., 2022. Associations among cardiovascular and cerebrovascular diseases: Analysis of the nationwide claims-based JROAD-DPC dataset. PLoS One, 17(3), p.e0264390.
12. Asadi, F., Rahimi, M., Daeechini, A.H. and Paghe, A., 2024. The most efficient machine learning algorithms in stroke prediction: A systematic review. Health Science Reports, 7(10), p.e70062.
13. Gupta, A., Mishra, N., Jatana, N., Malik, S., Gepreel, K.A., Asmat, F. and Mohanty, S.N., 2025. Predicting stroke risk: an effective stroke prediction model based on neural networks. Journal of Neurorestoratology, 13(1), p.100156.
14. Soladoye, A.A., Olagunju, K.M., Ajagbe, S.A., Adeyanju, I.A., Ogie, P.I. and Mudali, P., 2025. Stroke risk prediction: a deep learning approach for identifying high-risk patients. Discover Data, 3(1), pp.1-19.
15. Abujaber, A.A., Albalkhi, I., Imam, Y., Yaseen, S., Nashwan, A.J., Akhtar, N. and Alkhawaldeh, I.M., 2025. Machine learning-based prediction of 90-day prognosis and in-hospital mortality in hemorrhagic stroke patients. Scientific Reports, 15(1), p.16242.
16. Bonkhoff, A.K. and Grefkes, C., 2022. Precision medicine in stroke: towards personalized outcome predictions using artificial intelligence. Brain, 145(2), pp.457-475.
17. Yu, J., Park, S., Kwon, S.H., Ho, C.M.B., Pyo, C.S. and Lee, H., 2020. AI-based stroke disease prediction system using real-time electromyography signals. Applied Sciences, 10(19), p.6791.
18. Tusher, A.N., Sadik, M.S. and Islam, M.T., 2022, December. Early brain stroke prediction using machine learning. In 2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART) (pp. 1280-1284). IEEE.
19. Huang, S., Diao, S. and Wan, Y., 2024, September. Application of machine learning methods in predicting functional recovery in ischemic stroke patients. In The 1st International Scientific and Practical Conference “Innovative Scientific Research: Theory, Methodology, Practice”, Boston, USA. International Science Group (Vol. 289, p. 240).
20. Colangelo, G., Ribo, M., Montiel, E., Dominguez, D., Olivé-Gadea, M., Muchada, M., Garcia-Tornel, Á., Requena, M., Pagola, J., Juega, J. and Rodriguez-Luna, D., 2024. Prerisk: A personalized, artificial intelligence–based and statistically–based stroke recurrence predictor for recurrent stroke. Stroke, 55(5), pp.1200-1209.
21. Sposato, L.A., Field, T.S., Schnabel, R.B., Wachter, R., Andrade, J.G. and Hill, M.D., 2024. Towards a new classification of atrial fibrillation detected after a stroke or a transient ischaemic attack. The Lancet Neurology, 23(1), pp.110-122.
22. Abujaber, A.A., Albalkhi, I., Imam, Y., Nashwan, A.J., Yaseen, S., Akhtar, N. and Alkhawaldeh, I.M., 2023. Predicting 90-day prognosis in ischemic stroke patients post thrombolysis using machine learning. Journal of Personalized Medicine, 13(11), p.1555.