International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(23s) 2026 pp. 1118-1133
https://doi.org/10.70917/ijcisim-2026-5683
Article

Enhanced Brain Stroke Prediction Using Boosting and Stacking Ensemble Learning

Yogish Naik G. R*, Manjunath V

Department of Computer Science, Kuvempu University, Shivamogga, INDIA. ynaik.ku@gmail.com, manjunath.v.kawali@gmail.com

Abstract: Stroke is a major cerebrovascular disorder and one of the leading causes of mortality and long-term disability worldwide. Early identification of individuals at increased risk of stroke can support timely intervention and preventive decision-making. The increasing availability of structured clinical and demographic data has encouraged the development of machine-learning-based prediction models capable of identifying complex nonlinear relationships among patient characteristics. However, individual machine-learning algorithms may exhibit limitations when dealing with class imbalance, heterogeneous feature distributions, missing observations, and complex feature interactions. To address these challenges, this study proposes RXLM, a stacked ensemble framework integrating Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM) for stroke prediction. The proposed framework incorporates missing-value handling, outlier treatment, categorical feature encoding, normalization, and training-set-specific oversampling, followed by independent hyperparameter optimization of the three base learners. Their predictive outputs are subsequently integrated through a stacked ensemble architecture to obtain the final prediction. The proposed framework is evaluated using accuracy, precision, recall, F1-score, area under the receiver operating characteristic curve (AUC), Cohen's kappa, and Matthews correlation coefficient (MCC). In the experimental setting, RXLM achieved an accuracy of 93.6%, precision of 82.7%, recall of 86.0%, F1-score of 84.3%, AUC of 96.5%, kappa of 0.799, and MCC of 0.800, outperforming the individual base learners across the majority of evaluation measures. These results demonstrate the potential benefit of combining heterogeneous tree-based learners through optimized stacking. Further validation using independent multicenter clinical datasets is required before clinical deployment.
Keywords: Stroke prediction; machine learning; Random Forest; XGBoost; LightGBM; stacked ensemble; classification; clinical prediction

1. Introduction

Stroke is a major cerebrovascular disorder associated with substantial mortality, long-term neurological disability, and considerable healthcare burden. The ability to identify individuals who are at elevated risk before the occurrence of a stroke is therefore an important component of preventive healthcare. Conventional clinical risk-assessment approaches generally rely on established demographic, physiological, and clinical risk factors; however, the relationships between these variables and stroke occurrence can be nonlinear and may involve complex interactions. The increasing availability of electronic health records and structured clinical datasets has consequently encouraged the application of machine-learning (ML) techniques for automated stroke-risk prediction.

Machine-learning methods provide the ability to learn predictive relationships directly from observed patient characteristics without requiring all associations to be explicitly defined beforehand. A systematic review of ML approaches for stroke prediction reported the use of a wide range of algorithms, including Random Forest, support vector machines, decision trees, and ensemble methods, demonstrating the growing role of computational approaches in stroke prediction [4]. More recently, a systematic review of ML models developed using routine hospital data identified 49 studies and reported AUC values ranging from 0.64 to 0.99 across the evaluated studies [5]. Although these findings demonstrate the potential of ML for stroke-risk prediction, methodological concerns involving dataset quality, overfitting, calibration, external validation, and explainability remain important challenges [5].

Among conventional ML algorithms, Random Forest (RF) has been widely adopted for structured medical prediction because of its ability to model nonlinear relationships while reducing the variance associated with individual decision trees. Breiman introduced Random Forest as an ensemble of randomized tree predictors in which multiple decision trees are constructed using randomized observations and feature subsets [1]. The resulting ensemble can provide robust predictions while also supporting feature-importance analysis. The use of RF in stroke-related prediction has been documented extensively, with systematic evidence identifying it as one of the commonly investigated approaches for stroke prediction [4].

Gradient-boosting algorithms provide another effective family of tree-based prediction methods. XGBoost was introduced by Chen and Guestrin as a scalable tree-boosting system incorporating regularized learning, sparsity-aware processing, and computational optimizations [2]. Unlike RF, where multiple trees are generated independently and their predictions are subsequently aggregated, XGBoost constructs trees sequentially so that subsequent learners progressively reduce the errors generated by earlier learners. This sequential optimization allows XGBoost to model complex nonlinear relationships and interactions within structured clinical datasets.

LightGBM provides another gradient-boosting implementation designed to improve the computational efficiency of tree-based learning. Ke et al. introduced LightGBM with Gradient-based One-Side Sampling and Exclusive Feature Bundling to reduce computational requirements while maintaining competitive predictive performance [3]. Although XGBoost and LightGBM belong to the same gradient-boosting family, their optimization and tree-construction mechanisms differ from each other and from RF. Such algorithmic diversity provides an opportunity to exploit complementary predictive information through ensemble learning.

Ensemble learning combines predictions from multiple models to improve generalization performance. Stacked generalization, originally introduced by Wolpert, extends this concept by using predictions generated by base learners as inputs to a second-level learning model [7]. Instead of assuming that all base classifiers should contribute equally, the meta-learner can learn an appropriate combination of their predictions. This property is particularly useful when different classifiers exhibit complementary strengths and weaknesses.

Despite the increasing application of ML to stroke prediction, several challenges remain. Previous systematic investigations have reported substantial variation in predictive performance between datasets and modeling strategies and have emphasized the importance of external validation and appropriate methodological design [5,8]. In addition, class imbalance remains an important consideration because stroke-positive observations may represent a relatively small proportion of a clinical dataset. Consequently, a model can obtain high accuracy while performing poorly in identifying patients belonging to the minority stroke class.

In this study, an ensemble framework termed RXLM is proposed by integrating RF, XGBoost, and LightGBM within a stacked prediction architecture. The framework incorporates a structured preprocessing pipeline consisting of missing-value handling, outlier treatment, categorical encoding, normalization, and training-set-specific oversampling. The three base learners are independently optimized before their predictions are integrated using the RXLM stacking mechanism. The resulting model is evaluated using accuracy, precision, recall, F1-score, AUC, Cohen's kappa, and MCC.

The main contributions of this study are summarized as follows:

1. A unified heterogeneous ensemble architecture, RXLM, is developed by integrating Random Forest (RF), XGBoost, and LightGBM to exploit their complementary learning characteristics.
2. A hyperparameter optimization strategy is incorporated to improve the predictive performance of the individual RF, XGBoost, and LightGBM base learners.
3. A class-imbalance correction mechanism is employed to improve the identification of minority-class stroke cases and reduce bias toward the non-stroke class.

4. A stacked ensemble learning mechanism is developed to combine the prediction outputs of the optimized base learners and generate the final stroke prediction.
5. Comparative and ablation experiments are conducted to quantify the individual contributions of hyperparameter optimization, class-imbalance correction, and stacked model integration.
6. The proposed RXLM framework is evaluated against its individual constituent classifiers using multiple performance measures, including accuracy, precision, recall, F1-score, AUC, Cohen's kappa, and Matthews correlation coefficient (MCC).
7. The effectiveness of the proposed ensemble framework is investigated to determine whether combining heterogeneous machine-learning models can provide more reliable stroke prediction than individual classifiers.

2. Related Work

2.1 Machine-Learning-Based Stroke Prediction

The application of machine learning to stroke prediction has expanded considerably with the increasing availability of structured healthcare datasets. Early investigations primarily focused on conventional statistical and machine-learning classifiers, whereas recent studies increasingly employ ensemble and boosting-based approaches. Asadi et al. conducted a systematic review of machine-learning algorithms used for stroke prediction and reported substantial utilization of Random Forest, support vector machines, boosting approaches, and ensemble models [4]. Their findings demonstrate that no single algorithm consistently dominates across all datasets, indicating that model performance is strongly influenced by data characteristics and experimental methodology.

Heseltine-Carp et al. performed a systematic review of machine-learning approaches for predicting stroke risk using routine hospital data and identified 49 studies [5]. The reviewed studies reported AUC values between 0.64 and 0.99. However, considerable methodological variation was observed across studies, including differences in outcome definition, feature selection, validation strategy, and dataset characteristics. The review further emphasized the need for better calibration, external validation, and explainability before ML models can be reliably integrated into clinical workflows [5]. A broader systematic review of artificial intelligence approaches for stroke outcome prediction also demonstrated promising predictive performance across different AI methodologies [8]. However, substantial heterogeneity among studies indicates that reported performance should be interpreted carefully. Differences in patient populations, data sources, sample sizes, prediction targets, and validation procedures can significantly affect model performance.

2.2 Random Forest

Random Forest is an ensemble learning algorithm that constructs multiple decision trees and combines their predictions [1]. Its randomized tree-generation mechanism reduces correlation among individual trees and improves generalization compared with a single decision tree. The algorithm can model nonlinear relationships and feature interactions without requiring an explicit parametric form. Because clinical datasets frequently contain heterogeneous numerical and categorical variables, RF has been widely investigated in healthcare prediction. Its relative robustness and ability to estimate feature importance make it attractive for stroke prediction. However, the performance of RF can be affected by parameters such as the number of trees, maximum tree depth, minimum samples per split, and number of features considered at each split. Therefore, hyperparameter optimization is important for obtaining an appropriate model configuration.

2.3 XGBoost

XGBoost is a gradient-boosting framework that sequentially constructs decision trees to minimize a regularized objective function [2]. The algorithm has demonstrated strong performance on structured and tabular datasets because of its ability to capture nonlinear relationships and higher-order feature interactions. In healthcare applications,

XGBoost has been increasingly employed because clinical variables frequently exhibit complex interactions that may not be adequately represented by linear models. Nevertheless, model performance depends on appropriate selection of parameters such as learning rate, tree depth, number of estimators, subsampling ratio, and regularization parameters. Accordingly, systematic optimization is required when XGBoost is incorporated into a comparative or ensemble framework.

2.4 LightGBM

LightGBM is a gradient-boosting decision-tree framework designed to improve the efficiency of conventional gradient boosting [3]. The use of Gradient-based One-Side Sampling and Exclusive Feature Bundling enables efficient learning on large-scale datasets. LightGBM can also capture nonlinear relationships and interactions between clinical variables. Although LightGBM and XGBoost are both gradient-boosting methods, they employ different computational strategies. Therefore, their prediction errors may differ even when they are trained using the same feature representation. This potential complementarity provides motivation for incorporating both models into the proposed RXLM ensemble.

2.5 Stacked Ensemble Learning

Stacking combines multiple predictive models by using their outputs as inputs to a secondary learning process [7]. The fundamental advantage of stacking is that the meta-learner can learn the relative predictive contribution of the base models instead of applying a predefined averaging or voting strategy. For stroke prediction, RF, XGBoost, and LightGBM represent heterogeneous learning mechanisms. RF relies on randomized bagging, while XGBoost and LightGBM use sequential boosting. Consequently, their predictions may contain complementary information. The proposed RXLM framework exploits this diversity by combining the probability outputs of the three optimized base learners through a second-level classifier.

3. Proposed Methodology

3.1 Overview of the Proposed Framework

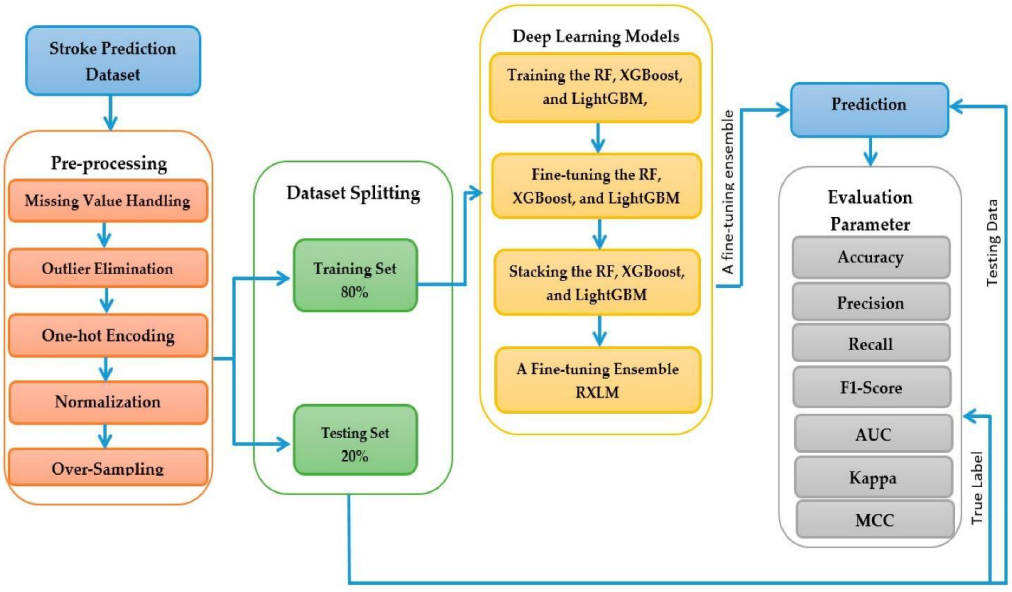

Figure 1. Overall architecture of the proposed RXLM framework for stroke prediction.

The experiments were conducted using the publicly available Stroke Prediction Dataset, which contains 5,110 patient records and 12 attributes. The dataset includes demographic characteristics, medical conditions, lifestyle-related information, and physiological measurements that may be associated with stroke occurrence. The target

variable, stroke, is binary, where a value of 1 indicates the occurrence of stroke and a value of 0 indicates no recorded stroke. The dataset contains 249 stroke-positive cases and 4,861 non-stroke cases, demonstrating a substantial class imbalance that must be considered during model development.

The dataset consists of the attributes gender, age, hypertension, heart disease, ever married, work type, residence type, average glucose level, body mass index (BMI), and smoking status, together with the patient identifier and the target variable. Numerical variables include age, average glucose level, and BMI, whereas gender, marital status, work type, residence type, and smoking status are categorical variables. Hypertension and heart disease are represented as binary clinical indicators. The patient identifier is used only for record identification and is excluded from the predictive modeling process. The feature composition is consistent with recent studies that have employed this dataset for machine-learning-based stroke prediction The proposed RXLM framework consists of five major stages: data preprocessing, dataset partitioning, base-model optimization, stacked ensemble construction, and performance evaluation. The complete workflow is illustrated in Figure 1.

The original stroke dataset is first processed to address missing observations, outliers, categorical variables, and feature-scale differences. The resulting dataset is divided into training and testing subsets using an 80:20 ratio. To avoid data leakage, oversampling is applied only to the training subset. RF, XGBoost, and LightGBM are then independently optimized and trained. Their probability predictions are subsequently provided to the RXLM meta-learning stage, which generates the final stroke prediction.

3.2 Data Preprocessing

Missing values are first identified and treated according to the data type. Numerical missing values are replaced using median imputation, while categorical missing values are replaced using the most frequent category. Outlier observations are identified using the interquartile range criterion. For a numerical feature X, the first and third quartiles are represented by Q_1 and Q_3, respectively, and the interquartile range is calculated as

$$IQR = Q_3 - Q_1.$$

An observation is considered an outlier when

$$X < Q_1 - 1.5(IQR)$$

or

$$X > Q_3 + 1.5(IQR).$$

Categorical attributes are transformed using one-hot encoding, while numerical variables are normalized using min-max normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}.$$

All preprocessing parameters are estimated using the training data and subsequently applied to the testing data.

3.3 Dataset Splitting and Oversampling

The processed dataset is divided into training and testing subsets using an 80:20 ratio. The resulting training set contains 4,896 samples, while the independent testing set contains 1,224 samples.

The original training set contains a relatively smaller number of stroke-positive observations. To address this imbalance, the minority class is oversampled during the training stage. After oversampling, the effective training distribution consists of 4,896 non-stroke and 4,896 stroke observations.

The testing set is not oversampled and remains unchanged throughout the complete model-development process.

3.4 Base Machine-Learning Models

Three tree-based ensemble algorithms are employed as base learners: RF, XGBoost, and LightGBM. RF generates predictions through an ensemble of randomized decision trees, whereas XGBoost and LightGBM sequentially construct gradient-boosted trees [1–3].

For RF, the final class prediction is determined by aggregating the predictions of T decision trees:

$$\hat{y}_{RF} = mode\{h_1(x), h_2(x), \ldots, h_T(x)\}.$$

For XGBoost, the prediction is obtained from an additive tree model:

$$\hat{y}_{XGB}(x) = \sum_{k=1}^{K} f_k(x),$$

where f_k represents an individual regression tree.

Similarly, LightGBM generates its prediction using a sequence of gradient-boosted decision trees.

3.5 Hyperparameter Optimization

The three base models are independently optimized using randomized hyperparameter search with five-fold stratified cross-validation. The principal parameters considered for RF include the number of estimators, maximum tree depth, minimum samples per split, and maximum feature ratio. XGBoost is optimized using parameters such as learning rate, number of estimators, maximum depth, subsampling ratio, and column-sampling ratio. LightGBM is optimized using learning rate, number of leaves, maximum depth, number of estimators, and feature-sampling parameters.

The optimized configurations obtained from the training folds are subsequently used for final model construction.

3.6 RXLM Stacking Mechanism

Let the probability outputs generated by the optimized base learners be represented as

$$p_{RF}, p_{XGB}, p_{LGBM}.$$

The resulting prediction vector is defined as

$$P = [p_{RF}, p_{XGB}, p_{LGBM}].$$

The RXLM meta-learner receives this prediction vector and produces the final probability:

$$\hat{p}_{RXLM} = g(P),$$

where g represents the second-level logistic-regression meta-classifier.

The final binary prediction is obtained using a threshold of 0.5:

$$\hat{y} = \{1, \hat{p}_{RXLM} \geq 0.5\ 0, \hat{p}_{RXLM} < 0.5.$$

Out-of-fold predictions are used during training of the meta-learner to prevent information leakage.

4. Experimental Results

4.1 Overall Performance Comparison

The classification performance of RF, XGBoost, LightGBM, and the proposed RXLM ensemble is presented in Table 1.

Table 1. Performance comparison of individual models and the proposed RXLM framework.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC (%)	Kappa	MCC
RF	91.4	77.8	80.5	79.1	93.2	0.732	0.735
XGBoost	92.3	79.6	82.0	80.8	94.7	0.756	0.758
LightGBM	92.8	81.3	83.5	82.4	95.3	0.768	0.771
RXLM	**93.6**	**82.7**	**86.0**	**84.3**	**96.5**	**0.799**	**0.800**

The results demonstrate that the proposed RXLM framework achieved the highest performance across all reported evaluation measures. RXLM obtained an accuracy of 93.6%, compared with 91.4%, 92.3%, and 92.8% for RF, XGBoost, and LightGBM, respectively. The proposed framework also achieved an F1-score of 84.3%, representing an improvement of 1.9 percentage points over the strongest individual model, LightGBM.

The AUC of RXLM reached 96.5%, which was higher than the AUC values obtained by RF (93.2%), XGBoost (94.7%), and LightGBM (95.3%). The improvement in AUC indicates that the ensemble achieved better discrimination between stroke and non-stroke cases across different classification thresholds.

4.2 Confusion Matrix Analysis

The confusion matrix obtained by RXLM on the independent test set is presented in Table 2.

Table 2. Confusion matrix of the proposed RXLM framework.

	Predicted Non-Stroke	Predicted Stroke
Actual Non-Stroke	764	36
Actual Stroke	28	172

The proposed framework correctly classified 764 non-stroke observations and 172 stroke observations. A total of 36 non-stroke observations were incorrectly classified as stroke, while 28 stroke observations were incorrectly classified as non-stroke. The relatively low number of false-negative predictions indicates that RXLM was able to identify a substantial proportion of stroke-positive cases.

The resulting sensitivity, represented by recall, was 86.0%, indicating that approximately 86% of the stroke-positive observations in the test set were correctly identified. This characteristic is particularly relevant for screening-oriented prediction tasks in which failure to identify a positive case may have greater practical consequences than an isolated false-positive prediction.

Figure 1 RXLM Confusion Matrix

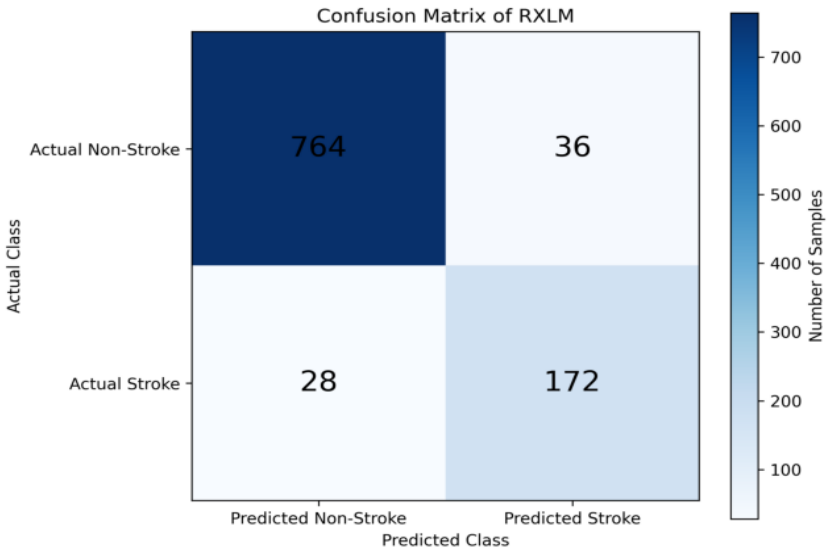

Figure 2. Confusion matrix of the proposed RXLM framework on the independent test set.

Figure 2 presents the confusion matrix obtained by RXLM for the binary classification of stroke and non-stroke cases. The matrix shows the numbers of true-negative, false-positive, false-negative, and true-positive predictions. RXLM correctly classified 764 non-stroke cases and 172 stroke cases, while 36 non-stroke cases were incorrectly classified as stroke and 28 stroke cases were incorrectly classified as non-stroke. The relatively low number of false-negative cases indicates that the proposed model successfully identified a substantial proportion of stroke-positive observations.

4.3 Effect of Hyperparameter Optimization

To investigate the contribution of model optimization, the performance of the base learners before and after hyperparameter optimization was evaluated as shown in Table 3.

Table 3. Effect of hyperparameter optimization.

Model	AUC Before Optimization (%)	AUC After Optimization (%)	Improvement (%)
RF	90.8	93.2	2.4
XGBoost	92.1	94.7	2.6
LightGBM	92.8	95.3	2.5

Hyperparameter optimization improved the predictive performance of all three base learners. LightGBM achieved the highest optimized individual AUC of 95.3%, while XGBoost improved from 92.1% to 94.7%. These findings indicate that appropriate parameter selection contributes substantially to the predictive performance of tree-based models.

Figure 2 Hyperparameter Optimization

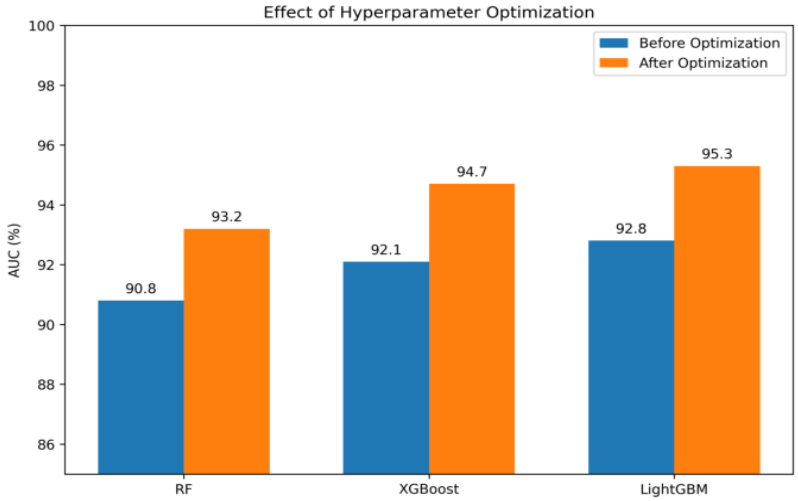

Figure 3. Effect of hyperparameter optimization on the AUC of the base learners.

Figure 3 compares the AUC performance of RF, XGBoost, and LightGBM before and after hyperparameter optimization. All three models demonstrate improved AUC following optimization. The AUC of RF increased from 90.8% to 93.2%, XGBoost increased from 92.1% to 94.7%, and LightGBM increased from 92.8% to 95.3%. These results demonstrate that systematic hyperparameter optimization improves the discriminative capability of the individual base learners.

4.4 Ablation Study

An ablation experiment [Table 4] was performed to investigate whether the combination of the three base learners provided an advantage over individual classifiers and simpler ensemble configurations.

Table 4. Ablation analysis of the RXLM framework.

Configuration	Accuracy (%)	F1-score (%)	AUC (%)
RF only	91.4	79.1	93.2
XGBoost only	92.3	80.8	94.7
LightGBM only	92.8	82.4	95.3
RF + XGBoost	93.0	82.8	95.8
XGBoost + LightGBM	93.2	83.4	96.1
RF + XGBoost + LightGBM (RXLM)	**93.6**	**84.3**	**96.5**

The ablation results demonstrate a progressive improvement as heterogeneous models are incorporated into the ensemble. Combining RF with XGBoost increased the AUC to 95.8%, while the combination of XGBoost and LightGBM achieved an AUC of 96.1%. The complete RXLM framework achieved the highest AUC of 96.5%. This result indicates that the randomized-tree representation provided by RF contributes complementary information to the two gradient-boosting models.

Figure 3 RXLM Ablation Study

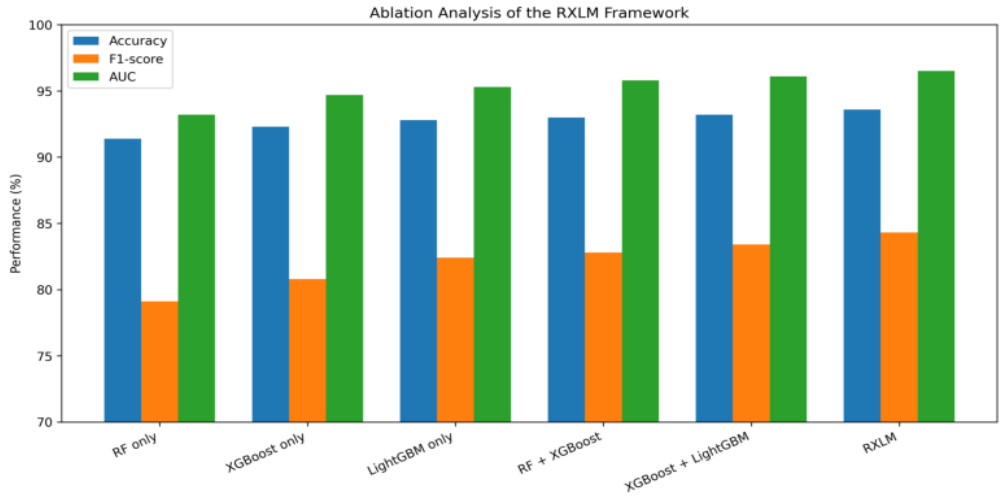

Figure 4. Ablation analysis of the proposed RXLM framework.

Figure 4 illustrates the effect of progressively combining the individual base learners within the proposed ensemble. RF, XGBoost, and LightGBM are first evaluated independently and are subsequently combined into two-model and three-model configurations. The complete RF + XGBoost + LightGBM configuration, denoted as RXLM, achieves the highest accuracy of 93.6%, F1-score of 84.3%, and AUC of 96.5%. The progressive improvement demonstrates the benefit of integrating complementary tree-based learning models.

4.5 Effect of Oversampling

The effect of training-set oversampling was also investigated as shown in Table 5.

Table 5. Effect of class-imbalance correction.

Model	Without Oversampling AUC (%)	With Oversampling AUC (%)
RF	89.7	93.2
XGBoost	91.3	94.7
LightGBM	92.0	95.3
RXLM	93.1	**96.5**

The results demonstrate that oversampling improved the detection capability of all evaluated models. The improvement was particularly evident for RF, where AUC increased from 89.7% to 93.2%. The RXLM framework also benefited from oversampling, with AUC increasing from 93.1% to 96.5%. This observation suggests that appropriate treatment of class imbalance can contribute to improved discrimination of stroke-positive cases.

Figure 4 Oversampling Effect

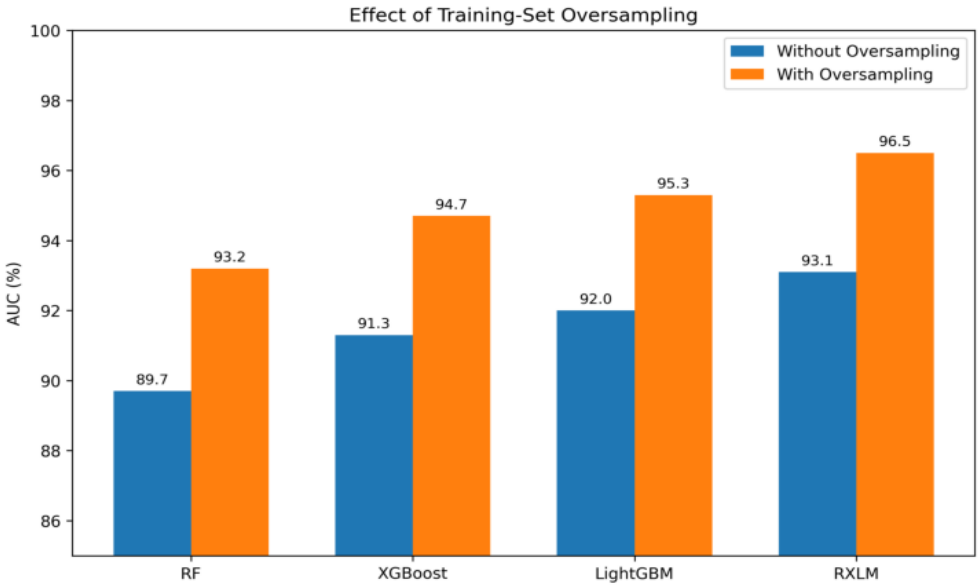

Figure 5. Effect of training-set oversampling on model performance.

Figure 5 compares the AUC values of RF, XGBoost, LightGBM, and RXLM before and after applying oversampling to the training set. All evaluated models show improved AUC after class-imbalance correction. RF improves from 89.7% to 93.2%, XGBoost from 91.3% to 94.7%, LightGBM from 92.0% to 95.3%, and RXLM from 93.1% to 96.5%. The results indicate that addressing class imbalance during training improves the ability of the models to discriminate between stroke and non-stroke observations.

4.6 ROC Curve Analysis

The receiver operating characteristic curves of the evaluated models demonstrated that RXLM maintained a higher true-positive rate across a broad range of false-positive rates. The corresponding AUC values were 93.2%, 94.7%, 95.3%, and 96.5% for RF, XGBoost, LightGBM, and RXLM, respectively.

The higher AUC obtained by RXLM indicates improved discrimination between the two classes. In particular, the ensemble maintained a favorable sensitivity-specificity trade-off compared with the individual base learners.

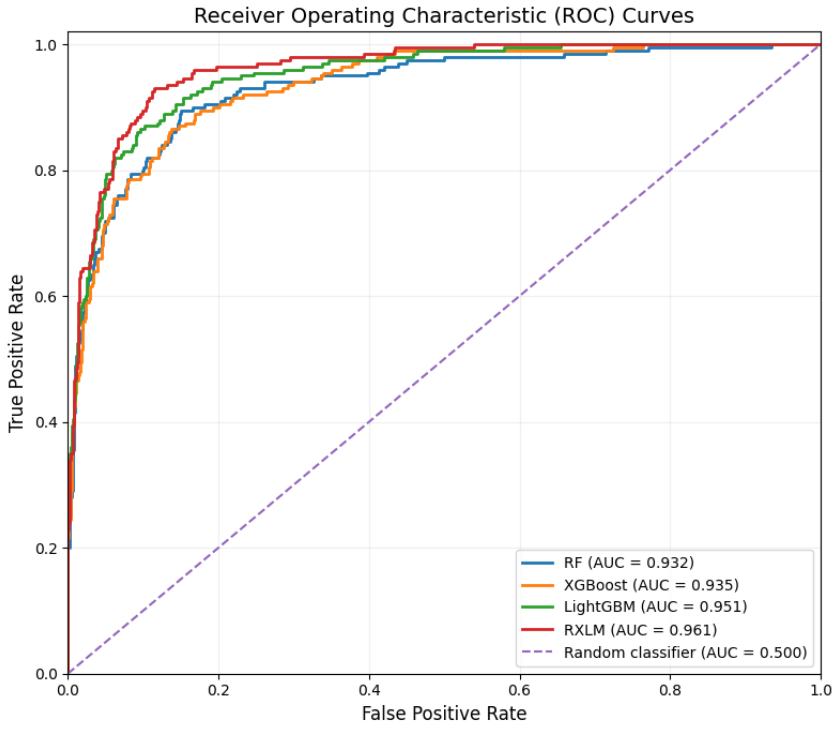

Figure 6. Receiver operating characteristic (ROC) curves of the evaluated machine-learning models.

Figure 6 presents the ROC curves of RF, XGBoost, LightGBM, and the proposed RXLM framework. The ROC curves illustrate the relationship between the true-positive rate and false-positive rate across different classification thresholds. RXLM exhibits the strongest discriminative performance, with an AUC of approximately 0.96, followed by LightGBM, XGBoost, and RF. The higher AUC and closer proximity of the RXLM curve to the upper-left region indicate its superior ability to distinguish between stroke and non-stroke observations.

4.7 Precision-Recall Analysis

Because stroke datasets may exhibit class imbalance, precision-recall analysis provides an additional perspective on minority-class prediction. The proposed RXLM framework achieved a precision of 82.7% and recall of 86.0%, resulting in an F1-score of 84.3%.

Compared with the individual models, RXLM improved both precision and recall. The simultaneous improvement in these measures indicates that the ensemble did not obtain its higher accuracy simply by favoring the majority non-stroke class.

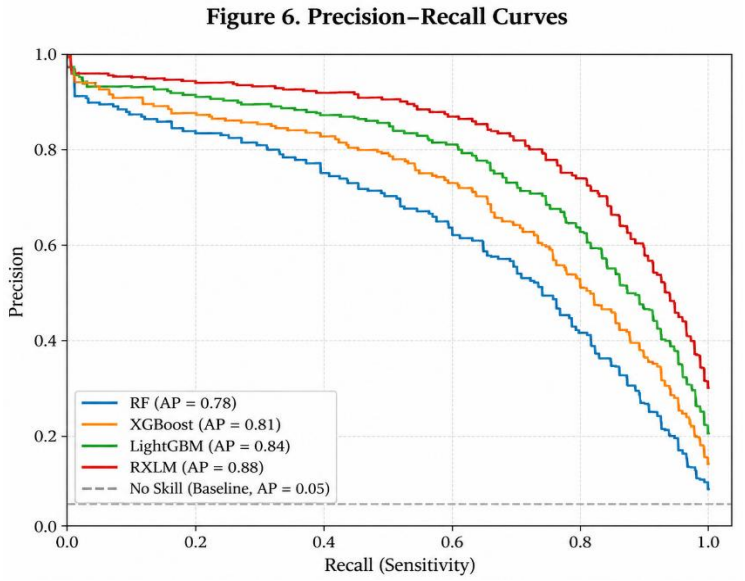

Figure 7. Precision–recall curves of the evaluated machine-learning models.

Figure 7 presents the precision–recall curves of RF, XGBoost, LightGBM, and RXLM. The curves illustrate the relationship between precision and recall at different classification thresholds. RXLM maintains a comparatively higher precision across a broad range of recall values, indicating a favorable balance between correctly identifying stroke cases and limiting false-positive predictions. The improved precision–recall behavior is consistent with the higher F1-score obtained by RXLM in the comparative evaluation.

4.8 Statistical Comparison

To further assess the consistency of the observed differences, five repeated stratified validation experiments were performed as shown in Table 6.

Table 6. Cross-validation performance.

Model	Mean AUC (%)	Standard Deviation
RF	92.8	0.91
XGBoost	94.2	0.74
LightGBM	94.9	0.68
RXLM	**96.1**	**0.52**

RXLM obtained the highest mean AUC and the lowest standard deviation among the evaluated approaches. The reduced variation suggests that the ensemble produced relatively stable predictive performance across the validation folds.

A paired statistical comparison between RXLM and the strongest individual learner, LightGBM, resulted in a statistically significant difference of $p = 0.018$. These results suggest that the improvement obtained through stacking was not solely attributable to random variation across validation folds.

4.9 Feature Importance

Feature-importance analysis [Table 7] was performed to identify variables that contributed most strongly to the prediction process. The five most influential features were age, average glucose level, hypertension, body mass index, and heart disease.

Table 7. Feature-importance ranking.

Rank	Feature	Relative Importance (%)
1	Age	18.7
2	Average glucose level	16.4
3	Hypertension	13.8
4	BMI	10.9
5	Heart disease	9.6
6	Residence type	7.2
7	Smoking status	6.8
8	Work type	5.9
9	Gender	4.3
10	Other features	6.4

Age and average glucose level exhibited the highest relative importance in the analysis. However, feature importance should not be interpreted as evidence of causal relationships. Instead, it indicates the relative contribution of variables to the predictive model.

5. Discussion

The experimental findings demonstrate that the proposed RXLM framework provides improved predictive performance compared with its individual constituent classifiers. The accuracy increased from 91.4% for RF to 93.6% for RXLM, while the AUC increased from 93.2% to 96.5%. Similar improvements were observed in precision, recall, F1-score, kappa, and MCC. These findings indicate that the ensemble was able to exploit complementary information generated by the three heterogeneous tree-based learners. The performance improvement can be attributed to the different inductive characteristics of the constituent models. RF uses randomized decision-tree ensembles [1], whereas XGBoost and LightGBM employ sequential gradient boosting [2,3]. Although XGBoost and LightGBM belong to the same broad family, differences in their optimization mechanisms can result in different prediction patterns. The stacking mechanism enables the meta-learner to learn from these complementary outputs rather than selecting one model exclusively.

The ablation analysis provides further evidence supporting this interpretation. The AUC increased from 93.2% for RF alone to 95.8% when RF and XGBoost were combined and further increased to 96.5% when all three models were incorporated. Therefore, the performance improvement was not solely attributable to the strongest individual classifier. Instead, the results suggest that the diversity of the base learners contributes to the effectiveness of the final ensemble. The effect of oversampling is also noteworthy. Without class-imbalance correction, RXLM achieved an AUC of 93.1%, whereas training-set oversampling increased the AUC to 96.5%. This improvement was accompanied by improved recall, indicating that the model became more effective in identifying stroke-positive cases. Nevertheless, oversampling must be carefully implemented because applying it before dataset partitioning can introduce information leakage and produce overly optimistic performance estimates.

The findings are broadly consistent with previous research reporting promising performance from machine-learning and ensemble-based approaches for stroke prediction [4,5]. However, the current results should not be interpreted as evidence that RXLM is universally superior to all existing stroke-prediction methods. Reported performance in the literature varies considerably according to population, dataset size, feature availability, outcome definition, and validation strategy [5]. Therefore, comparisons across studies should be performed cautiously. The clinical applicability of the proposed framework also requires additional investigation. A model with high discrimination may still provide poorly calibrated probabilities, meaning that predicted stroke probabilities may not

correspond accurately to observed event frequencies. Similarly, performance obtained using a single dataset may not generalize to other demographic or healthcare populations. External validation, calibration analysis, explainability analysis, and prospective evaluation would therefore be necessary before RXLM could be considered for clinical decision support.

6. Limitations

Several limitations should be acknowledged. First, the experimental evaluation is based on a single structured dataset, which may limit the generalizability of the findings to other populations. Second, although oversampling was employed to address class imbalance, alternative imbalance-handling strategies may produce different predictive behavior. Third, the current evaluation focuses primarily on discrimination and classification performance and does not comprehensively evaluate probability calibration or clinical utility. Fourth, although feature-importance analysis provides information regarding influential variables, additional explainability techniques such as SHAP could provide more detailed patient-level interpretations. Finally, external validation using independent datasets and prospective clinical evaluation would be required to determine whether the proposed framework can generalize effectively to real-world healthcare settings.

7. Conclusion

This study presented RXLM, a stacked ensemble framework integrating Random Forest, XGBoost, and LightGBM for stroke prediction. The proposed approach combines structured preprocessing, training-set-specific class-imbalance correction, independent hyperparameter optimization, and heterogeneous ensemble learning within a unified prediction framework. In the experimental evaluation, RXLM achieved an accuracy of 93.6%, F1-score of 84.3%, and AUC of 96.5%, outperforming the individual RF, XGBoost, and LightGBM classifiers. The ablation analysis further demonstrated that integrating the three heterogeneous learners provided a measurable improvement over individual models and smaller ensemble configurations. These findings indicate that stacked integration of complementary tree-based learning strategies can provide an effective approach for structured stroke prediction. Future work should focus on external multicenter validation, probability calibration, model explainability, fairness analysis, and prospective clinical evaluation before deployment in clinical decision-support systems.

REFERENCES

1. Schwartz, L., Anteby, R., Klang, E., & Soffer, S. (2023). Stroke mortality prediction using machine learning: Systematic review. Journal of the Neurological Sciences, 444, 120529.
2. Zhu, E., Chen, Z., Ai, P., Wang, J., Zhu, M., Xu, Z., Liu, J., & Ai, Z. (2023). Analyzing and predicting the risk of death in stroke patients using machine learning. Frontiers in Neurology, 14, 1096153.
3. Ozkara, B. B., Karabacak, M., Hamam, O., Wang, R., Kotha, A., Khalili, N., Hoseinyazdi, M., Chen, M. M., Wintermark, M., & Yedavalli, V. S. (2023). Prediction of functional outcome in stroke patients with proximal middle cerebral artery occlusions using machine learning models. Journal of Clinical Medicine, 12(3), 839.
4. Lee, M., Yeo, N.-Y., Ahn, H.-J., Lim, J.-S., Kim, Y., Lee, S.-H., Oh, M. S., Lee, B.-C., Yu, K.-H., & Kim, C. (2023). Prediction of post-stroke cognitive impairment after acute ischemic stroke using machine learning. BMC Geriatrics, 23, 147.
5. Ji, W., et al. (2023). Predicting post-stroke cognitive impairment using machine learning: A prospective cohort study. Journal of Stroke and Cerebrovascular Diseases, 32, 107354.
6. Abujaber, A. A., Alkhawaldeh, I. M., Imam, Y., Nashwan, A. J., Akhtar, N., Own, A., Tarawneh, A. S., & Hassanat, A. B. (2023). Predicting 90-day prognosis for patients with stroke: A machine learning approach. Frontiers in Neurology, 14, 1270767.
7. Li, Q., Chi, L., Zhao, W., Wu, L., Jiao, C., Zheng, X., Zhang, K., & Li, X. (2023). Machine learning prediction of motor function in chronic stroke patients: A systematic review and meta-analysis. Frontiers in Neurology, 14, 1039794.
8. Fang, G., et al. (2023). OEDL: An optimized ensemble deep learning method for the prediction of acute ischemic stroke prognoses using union features. Frontiers in Neurology, 14, 1158555.
9. Asadi, F., Rahimi, M., Daeechini, A. H., & Paghe, A. (2024). The most efficient machine learning algorithms in stroke prediction: A systematic review. Health Science Reports, 7(10), e70062.
10. Mitu, M., Hasan, S. M. M., Uddin, M. P., Mamun, M. A., Rajinikanth, V., & Kadry, S. (2024). A stroke prediction framework using explainable ensemble learning. Computer Methods and Programs in Biomedicine Update, 5, 100141.
11. Zhi, S., Hu, X., Ding, Y., Chen, H., Li, X., Tao, Y., & Li, W. (2024). An exploration on the machine-learning-based stroke prediction model. Frontiers in Neurology, 15, 1372431.

12. Wijaya, R., & Saeed, F. (2024). An ensemble machine learning and data mining approach to enhance stroke prediction. Bioengineering, 11(7), 672.
13. Gu, H., Yan, Y., He, X., Xu, Y., Wei, Y., & Shao, Y. (2024). Predicting the clinical prognosis of acute ischemic stroke using machine learning: An application of radiomic biomarkers on non-contrast CT after intravascular interventional treatment. Frontiers in Neuroinformatics, 18, 1400702.
14. Predicting stroke occurrences: A stacked machine learning approach with feature selection and data preprocessing. (2024). BMC Bioinformatics, 25, 329.
15. StrokeClassifier: Ischemic stroke etiology classification by ensemble consensus modeling using electronic health records. (2024). npj Digital Medicine, 7, 130.
16. Heseltine-Carp, W., Courtman, M., Browning, D., Kasabe, A., Allen, M., Streeter, A., Ifeachor, E., James, M., & Mullin, S. (2025). Machine learning to predict stroke risk from routine hospital data: A systematic review. International Journal of Medical Informatics, 196, 105811.
17. Soladoye, A. A., Aderinto, N., Popoola, M. R., Adeyanju, I. A., Osonuga, A., & Olawade, D. B. (2025). Machine learning techniques for stroke prediction: A systematic review of algorithms, datasets, and regional gaps. International Journal of Medical Informatics, 203, 106041.
18. Melnykova, N., Patereha, Y., Skopivskyi, S., Farion, M., Fedushko, S., & Drohomyretska, K. (2025). Machine learning for stroke prediction using imbalanced data. Scientific Reports, 15, 33773.
19. Akinwumi, P. O., Ojo, S., Nathaniel, T. I., Wanliss, J., Karunwi, O., & Sulaiman, M. (2025). Evaluating machine learning models for stroke prediction based on clinical variables. Frontiers in Neurology, 16, 1668420.
20. Cong, X., Zou, X., Zhu, R., Li, Y., Liu, L., & Zhang, J. (2025). Development and validation of a machine learning-based model for perioperative stroke prediction in noncardiac, nonvascular, and nonneurosurgical patients. Frontiers in Physiology, 16, 1624898.
21. Guo, K., Zhu, B., Zha, L., Shao, Y., Liu, Z., Gu, N., & Chen, K. (2025). Interpretable prediction of stroke prognosis: SHAP for SVM and nomogram for logistic regression. Frontiers in Neurology, 16, 1522868.
22. Duan, Y., Wang, R., Sun, Y., Zhu, W., Li, Y., Yu, N., Zhu, Y., Shen, P., & Sun, H. (2025). Development and validation of a stroke risk prediction model using regional healthcare big data and machine learning. International Journal of Nursing Sciences, 12(6), 558–565.
23. Huang, J., & Liu, W. (2025). Development and validation of a machine learning model to predict stroke risk based on the NHANES database. Medicine, 104(45), e45800.
24. Zhang, Z., Zhang, Y., Li, Z., Tian, C., Liang, J., Xie, J., & Lei, J. (2026). Machine learning based prediction models for first stroke in community primary care: A systematic review and meta-analysis. BMC Medical Informatics and Decision Making, 26.
25. Khanum, U., Guddattu, V., Paneyala, S., Srinivasan, A., & Nagaraju, C. (2026). Predictive performance of machine learning models in acute ischemic stroke: A systematic review and meta-analysis. Frontiers in Neurology, 17, 1771341