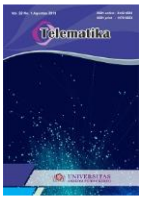

Available online at :
http://ejournal.amikompurwokerto.ac.id/index.php/telematika/

Telematika

Accreditated SINTA “2” Kemenristek/BRIN, No. 85/M/KPT/2020

UNIVERSITAS AMIKOM PURWOKERTO

Optuna Based Hyperparameter Tuning for Improving the Performance Prediction Mortality and Hospital Length of Stay for Stroke Patients

Ades Tikaningsih[1], Puji Lestari [2], Ade Nurhopipah[3], Imam Tahyudin[4], Eko Winarto[5], Nazwan Hassa[6]

[1,3]Department of Informatics, Faculty of Computer Science,
[2,4]Department of Information System, Faculty of Computer Science,
[5,6]Department of Medicine
[1,2,3,4]Universitas Amikom Purwokerto, Banyumas, Indonesia
[5,6]RSUD Banyumas Polyclinic Neural, Banyumas, Indonesia
E-mail: adestikaningsih92@gmail.com[1], puji24318@gmail.com[2], ade_nurhopipah@amikompurwokerto.ac.id[3], imam.tahyudin@amikompurwokerto.ac.id[4], eko_win77@yahoo.co.id[5], nazwanhassa@gmail.com[6]

ARTICLE INFO

History of the article:
Received December 7, 2023
Revised January 2, 2024
Accepted February 16, 2024

Keywords:
Stroke
Machine Learning
Mortality
Length Of Stay
Prediction

Correspondence:
E-mail:
imam.tahyudin@amikompurwokerto.ac.id

ABSTRACT

Cardiovascular disease (CVD) stands as the foremost contributor to worldwide mortality, with strokes as part of significant CVD. Research on potential mortality risks and hospitalizations for stroke patients became crucial as a basis for evaluation to improve the quality and control of stroke patient services. Although machine learning technology has been widely used in health data analysis, understanding the relative performance and characteristics of machine learning (ML) models is still limited. Therefore, the study aims to broaden this understanding by comparing five ML models, namely XGBoost, Random Forest, Decision Trees, CatBoost, and Extra Trees, using stroke patient data from RSUD Banyumas Neural Poliklinik Indonesia. The model performance improvement process is the main focus, involving adjustments using the Optuna tuning library. Through this tuning approach, the key parameters of each ML model are optimally adjusted to improve their performance in predicting mortality risk and the duration of hospitalization for stroke patients. As a result, the XGBoost algorithm proved superior in predicting mortality (accuracy 86%, AUC 0.87) and the duration of hospitalization (accuracy 82%, AUC 0.79). This research has great potential to help hospitals identify high-risk stroke patients and plan more efficient treatment. This approach allows hospitals to use their resources better, improve medical services, and reduce unnecessary treatment costs.

INTRODUCTION

Cardiovascular disease (CVD) is a general term for diseases involving the heart and blood vessels. CVD has become the leading cause of death globally, killing approximately 17.9 million people every year (WHO, 2021). Among the various forms of CVD, stroke is one of the diseases that account for about 12% of all CVD deaths (American Heart Association, 2023). A stroke occurs as a result of a neurological disorder characterized by blockage of blood vessels in the brain. A blood clot formed inside the brain interferes with blood flow, causes blocking of the arteries, and can cause a blood vessel to break (Kuriakose & Xiao, 2020). Stroke is the second cause of death and the third cause of disability worldwide, with 14 million people having a stroke out of approximately 5.5 million people dying (Pacheco-Barrios et al., 2022).

The Global Stroke Fact Sheet released in 2022 revealed that the lifetime risk of stroke has increased by 50% over the past 17 years, and now, one in four people is estimated to have a lifelong stroke. Between 1990 and 2019, there was a 70 percent increase in stroke incidence, a 43 percent increase in stroke mortality, 102 percent higher stroke prevalence, and a 143 percent rise in Disability Adjusted Life Years (DALY). In 2018, Indonesia reported a stroke prevalence of 10.9 per thousand people, approximately 2,120,362 individuals, according to the Ministry of Health. The 2019 Basic Health Research Study (RISKESDAS) indicated a similar prevalence rate but with variations among provinces. Stroke is categorized as one of the top four high-cost catastrophic diseases by the Social Security Maintenance Agency (BPJS). The health service costs for stroke increased from IDR 2.19 trillion in 2017 to IDR 2.57 trillion in 2018, highlighting its financial impact (Venketasubramanian et al., 2022).

Understanding the risk factors that cause strokes opens the door to more effective treatment. An in-depth study highlighted a close relationship between mortality risk and extended hospitalization in stroke patients. The severity of stroke (NIHSS categories 15-19 and ≥20), age over 65, gender, and hyperlipidemia have been identified as significant contributors to mortality (Mohebi et al., 2018). Hospital mortality rates tend to be similar in patients with ischemic and hemorrhagic strokes, with ischaemic stroke accounting for 62.4% of the total cases (Feigin et al., 2021). The importance of these factors is not only limited to influencing the life prognosis but also being a powerful indicator of the duration of treatment needed in hospitals. The close relationship between the risk of death and the length of stay in stroke patients confirms that patients with a higher risk of death tend to require more intensive and prolonged hospital treatment. These findings are strengthened by the correlation between the length of time a patient is treated in the ICU and the likelihood of death, especially if the patient spends more than three days in the intensive care unit (Barsasella et al., 2022). The results of this study align with other findings, showing that a longer duration of hospitalization is associated with higher mortality and readmission rates (Rachoin et al., 2020). Therefore, efforts to reduce the risk of death in stroke patients are essential because they can increase survival and have the potential to reduce the burden of long-term care, increase the efficiency of health resources, and improve the overall quality of service.

With technological advances in the medical field, predicting stroke patients' mortality rate and hospitalization duration can be realized using Machine Learning. Machine learning algorithms have shown great potential in predicting the likelihood of a stroke and leveraging that knowledge to identify high-risk patients (Mridha et al., 2023). Several related studies provide significant insights into predicting mortality and stroke patients' hospitalization length using machine learning. The use of the XGBoost algorithm, which shows superior performance with the lowest Brier score and the highest area value below the ROC curve (AUC) up to 0.833 (Wang et al., 2022), as well as the Naïve Bayes class weighing achieved the highest predictive performance with an AUC of 0.661 (Hung et al., 2020). The Random Forest Model (RF) also showed high performance with an AUC of more than 0.90 in the three groups of deaths evaluated (Fernandez-Lozano et al., 2021). Further research developed a prediction of long-term hospitalization; research in China using the XGBoost algorithm with an AUC value of 0.92 showed the model's ability to predict whether a patient will be hospitalized for more than one day with an accuracy of 85% (R. Chen et al., 2023). The approach using Artificial Neural Network successfully predicted the long-term treatment of stroke patients in hospitals with an AUC of 0.788 (Yang et al., 2023), (Neto et al., 2020).

After observing the potential application of machine learning in predicting the risk of death and duration of stroke patient hospitalization, a deeper understanding of the relative performance and unique characteristics of various machine learning models is still limited and not fully revealed through adequate comparisons. Besides, tuning optimizers to improve the performance of machine learning models in this context have not been fully informed. Model selection can significantly influence prediction results, and choosing the correct algorithm is a critical step in building accurate and relevant models (Jason Brownlee, 2019a). Therefore, this research aims to fill that knowledge gap by applying a comprehensive approach. The research will implement and compare five different machine learning models, namely XGBoost (XGB), Random Forest (RF), Decision Trees (DT), CatBoost, and Extra Trees. CatBoost is known to be the optimal choice for categorical features and large data sets, standing out in speed and accuracy (Jacob Gursky, 2020). Meanwhile, the extra-tree classifier is a more computationally efficient ensemble learning method compared to Random Forest, works well for a wide range of data types, and is relatively easy to use (C. H. Chen et al., 2020), (Saikumar Talari, 2022).

The model performance improvement process is the main focus, involving adaptation using the Optuna tuning library. Optuna is a trending framework explicitly designed to overcome the limitations of optimizing hyperparameters in machine learning models through trial-and-error approaches to optimal performance (Crissman, 2019). Through this tuning approach, the key parameters of each ML model are optimally adjusted for better performance in predicting mortality risk and managing stroke patients. Through this research, it is expected to help hospitals identify high-risk stroke patients against mortality rates so that more appropriate preventive measures can be taken and stroke patient management policies can be better designed. In addition, hospital resources, such as beds and medical personnel, can be managed more efficiently based on more accurate estimates of the duration of hospitalization. These measures are expected to improve the quality of medical care and reduce unnecessary treatment costs.

RESEARCH METHODS

This research adopts the framework of Rui Chen et al.(2023) (R. Chen et al., 2023). The stages of this system are as follows in Figure 1.

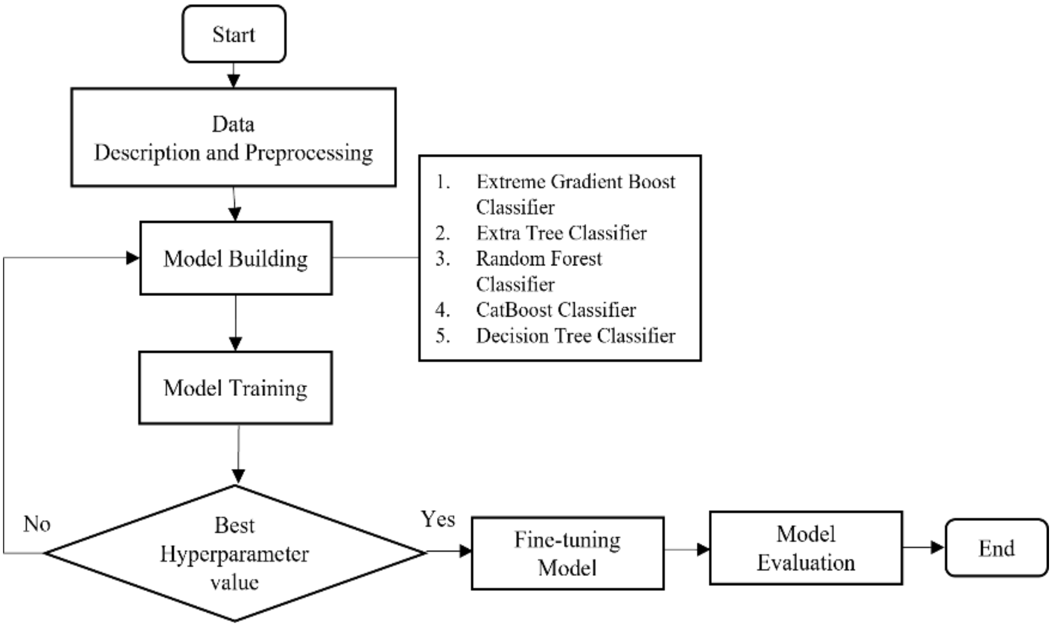

Figure 1. Experimental workflow

1. Data Description

The data used in the study included 106 ischemic stroke patients undergoing treatment at the Regional General Hospital (RSUD) Neural Clinic in Banyumas, Indonesia. The RSUD research institute team collected data in collaboration with doctors and nurses from January 2022 to May 2023. All patients are observed and recorded, including demographic information, medical history, and other clinical records. It is important to note that the data used in this research is entirely original and accessible with permission and cooperation by research ethics. The patient's identity has been kept confidential, and every step has been taken to ensure the security and confidentiality of the data. Table 1 provides an overview of 20 attributes of inpatient data, complete with an in-depth explanation of each attribute.

Table 1. Selected variables of the stroke dataset

Categories	*Variables*	*Description*
Demographic characteristics of the patients	Patient Admission Date	The date when a patient started hospital treatment
	Patient discharge Date	The date when a patient finishes treatment and is allowed to leave the hospital
	Age	Age of Stroke Patient
	Gender	Refers to the category or identity of a person's gender
	Debtors	Source or type of health insurance held by the patient
	Prior Disease History	Information about other previous medical conditions suffered by the patient
	Previous Stroke History	Information on whether the patient has a history of stroke before the observed case
	Patient Conditions	Information on patient health status (health and risk of mortality)
	History Of CVD	Information on cardiovascular disease history, such as previous heart disease or stroke
	Previous Stroke History	Information on whether the patient has a history of stroke before the observed case
Medical Data	Stroke Location	Information about the location of the stroke attack in the patient
	HB(Hemoglobin)	A component of red blood cells that binds to oxygen
	HT(Hematocrit)	The proportion of blood volume filled with red blood cells
	LEU(Leukocytes)	A type of white blood cell
	TR(Thrombocytes)	A type of blood cell involved in blood clotting
	NLR(Neutrophil-Lymphocyte Ratio)	The ratio of neutrophils to lymphocytes in the blood
	CHOL Total	Total cholesterol is the aggregate cholesterol level in the bloodstream.
	HDL(High-Density Lipoprotein)	A type of cholesterol known for its role in reducing excess cholesterol in the blood.
	TG(Triglycerides)	A type of fat in the blood
	LDL(Low-Density Lipoprotein)	A type of cholesterol is considered "bad" because it can clog arteries.

2. Data Preprocessing

Preprocessing data is processed and analyzed to reach a ready-to-use form. The main focus of this research is the development of predictive models related to mortality and Length of Stay (LoS) in stroke patients, which is considered a classification problem. In the context of machine learning, the development of predictive models of classification involves an effort to project the mapping function (f) from the input variable (X) to the discrete output variable(y), often known as the label or category (Jason Brownlee, 2019b). The objective of this mapping function is to predict a class or category corresponding to a particular observation. 1 The input variable in this study includes 19 attributes documented in Table 1. These attributes include demographic information, medical history, and clinical records of patients with ischemic stroke. The most relevant and significant variables will be selected for further analysis through feature selection techniques.

Then, the target variable that is focused in the attempt to predict the death of a stroke patient is the "patient condition" variable. This variable comprehensively contains data on the patient's health status, that is, the patient is healthy and is at risk of mortality. In line with that, the target variable for the long-term prediction of hospitalization is obtained by calculating the number of differences between the patient's entry and exit dates. Patients are then grouped into two categories: standard and non-standard patients. The results of this calculation will go through in-depth analysis to ensure that each patient's long hospitalization duration corresponds to the normal range that has been defined. In particular, patients who have been hospitalized for less than seven days will be categorized as standard patients, while patients who exceed this time interval (more than seven days) or do not meet it will be classified as non-standard patients. The variable "length hospitalization" will be initialized with the result of this categorization.

This basis of thought refers to findings and understanding found in medical literature, which affirms that the period of hospitalization of approximately seven days in patients with ischemic stroke plays a vital role as a protective or protective factor (R. Chen et al., 2023). Patients in intensive care for about seven days have a lower predisposition to possible side effects after leaving the hospital environment. Such side effects may involve several health complications associated with the post-stroke healing process, such as changes in neurological conditions, risk of infectious complications, or even a lower mortality rate. These findings provide the basis for this study in setting such time limits as criteria in categorizing patient hospitalization duration into standard or non-standard.

a. Analysis of Variable Correlation

The correlation analysis, conducted using the pandas corr() function, examines the relationship between "Patient condition" and "Length of hospitalization" with other dataset attributes. Notable correlations include a positive correlation of 0.25 between "Gender" and "Patient condition," strong positive correlations (0.24 and 0.25) for biochemical factors like "LEU" and "NLR," and a negative correlation of -0.05 for "Debtor." "Length of hospitalization" exhibits negative correlations (-0.12 and -0.09) with "Age" and moderate correlations (0.17 and 0.07) with lipid variables like "HDL" and "TG." Cholesterol variables "Chol Total" and "LDL" show robust negative correlations with the patient's condition (0.14 and -0.18). Notably, "Stroke Location" is the most positively correlated attribute with "Length of hospitalization" (0.148774), suggesting a specific stroke location's association with increased duration. Conversely, "NLR" has a negative correlation of -0.092495, indicating lower NLR values may be linked to shorter hospitalization. "Age" shows a minor negative correlation of -0.052316.

b. Missing Value

Handling missing values in data, one of which is replacing them with zero values, is essential in machine learning (Yi et al., 2020). In this study, the missing values were found on several attributes such as "History of Cardiovascular Disease," "Past Medical History," and "Previous Stroke History." Proper handling of missing values is essential to ensure the integrity and relevance of the data before proceeding to the analysis stage. Filling the missing value with a value of 0 is chosen because it has high medical significance in the context of predictive mortality research and extended hospitalization of stroke patients. In some cases, a value of 0 may represent the absence or absence of a particular event. Using a value of 0 in consistency in numerical data formats facilitates the process of statistical analysis and minimizes potential interference that may arise due to missing values (Huey Fern Tay, 2021).

c. Encoding Variable

Category variable encoding is an essential step in data preparation for predictive models, as most machine learning models require numerical data (Garg, 2022). This study used two encoding methods: one-hot Encoding and label encoding. One-hot Encoding involves transforming each category of an attribute into a binary vector. This creates an additional column that represents the membership of a sample in each category. For example, the long-term variables of the hospital, gender, and patient condition are encoded using One-Hot Encoding. Meanwhile, Label Encoding turns the unique values on the column into integers by giving a sequential index starting with 0. For example, the location of a stroke, history of cardiovascular disease, past medical history, and history of previous strokes are encoded using Label Encoding.

3. Machine Learning Classification

a. Extreme Gradient Boosting Classifier (XGB)

Extreme Gradient Boosting (XGB) combines boosting and gradient boosting. XGBoost, with its boosting method, is used to classify errors from the previous model. XGB uses gradient descent, which helps narrow errors occurring when creating or forming a new model. XGB uses advanced regularization (L1 and L2), which enhances model generalization capabilities (Moore & Bell, 2022). XGB has been shown to perform very well when segmenting stroke-infarction regions using clinical and imaging features (Kim, 2016).

b. Categorical Boosting Classifier (CatBoost)

CatBoost is known for its ability to handle large numbers of categorical variables, CatBoost integrates an overfitting detector that stops training when it detects overfitting. This feature improves the model's generalization performance, making it more robust to new health data with different characteristics (Safaei et al., 2022). With the ability to handle categorical features and an intelligent solution to overfitting, CatBoost becomes a strategic choice for classification tasks based on the data health of stroke patients, offering reliability and adaptability in practical use.

c. Random Forest Classifier (RF)

Random Forest (RF) is an ensemble model with a broad scope of use, created to bypass the limitations of ordinary decision tree algorithms. RF techniques involve training many decision-tree learners simultaneously to minimize the bias and variance of models (Ghazwani & Begum, 2023). Thus, RF brings advantages in handling the complexity of stroke patient data and optimizes model performance through a sophisticated ensemble approach. The combination of these features makes RF a powerful and versatile option in the context of this research.

d. Extra Tree Classifier (ET)

ExtraTree is a machine learning ensemble method that instructs multiple decision trees and aggregates the results from the ensemble of trees to obtain an estimate. ET has striking speed and efficiency. ET achieves this by using random splitting and subsampling, enabling it to provide predictions with reduced computing time without sacrificing quality (Sun et al., 2021). Furthermore, ET has effectively handled noisy data and missing values. These advantages make ET suitable for predicting scratches on datasets with missing or incomplete data.

e. Decision Tree Classifier (DT)

Decision Trees (DT), a supervised machine learning algorithm, widely analyzes various variables and separate data into segments or branches, forming a decision tree structure (Ogunleye, 2021). Arranged each branch on a decision tree hierarchically upward, with the top branch representing the final outcome of the model decision. The main advantage of decision trees is their easy-to-understand interpretation, making them a very useful tool for identifying the most important features in stroke risk prediction (Tazin et al., 2021). This clarity of interpretation is beneficial for researchers and health practitioners who need a deep understanding of the factors that contribute to predicted outcomes.

4. Hyperparameter Tuning

In machine learning, hyperparameters are values enhancing model performance. They play a crucial role in algorithm improvement and significantly impact various model tests (Muslim Karo Karo, 2020). This study used the Optuna library, an advanced hyperparameter optimization framework, to optimize mortality prediction and long-term care models for stroke patients. Optuna is a tool experiencing significant development and has three main advantages for model selection or hyperparameter determination. The first main advantage provided by Optuna is the existence of an API that can be defined while the program is running. The second point of advantage is the efficient pruning and sampling mechanism. The third advantage is convenience in setup. The API style that can be specified during program runtime is adopted from deep learning frameworks, allowing users to adjust the hyperparameter search space dynamically. In addition, this ease of setup provides a wide range of settings, from simple experiments to complex distributed computing, under versatile architectures (Akiba et al., 2019), (Chintakindi et al., 2022). By retrying hyperparameter combinations and using truncation to stop training without significant improvement, Optuna is efficient and faster than traditional methods such as GridSearchCV (Lim, 2022).

Figure 2 depicts the critical steps in determining hyperparameters. Figure 2 shows a series of essential steps in determining hyperparameters for a machine learning model using Optuna. The initial stage involves determining hyperparameters for five tree-based machine learning models, namely XGBoost, Random Forest, Decision Trees, CatBoost, and Extra Trees, each with a unique hyperparameter configuration. Afterward, an objective function is defined for Optuna, which acts as a guide in improving the performance of the machine-learning model. This research mainly focuses on improving accuracy, precision, recall, F1 score, and ROC AUC as evaluation metrics. The hyperparameter determination process begins with Optuna study initialization for each model. We construct an objective function for each model that involves initializing the model with the proposed hyperparameters, training it using training data, and evaluating its performance on test data.

We ran Optuna studies with n_trials iterations to find the best combination of hyperparameters for each model. At each iteration, Optuna automatically tries various combinations of hyperparameters and records the evaluation score of the objective function to monitor model performance. After completing the tuning process, we extract each model's best hyperparameter combination from the Optuna study. These models are then reinitialized with the best parameters and retrained on the training data to produce optimized machine models. With this approach, hyperparameter optimization using Optuna efficiently improves the performance of the five machine learning models, has more effective models in processing stroke data, and significantly contributes to improving the quality of mortality risk prediction and management of stroke patients.

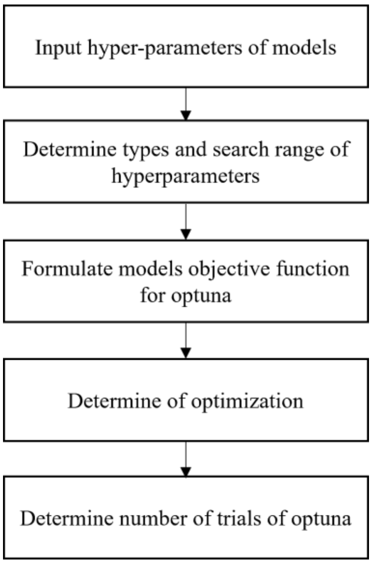

Figure 2. Basic Optuna procedure for hyperparameter tuning

5. Performance Evaluation

To evaluate the predictive model's performance, this study conducted an analysis based on several key metrics to ensure the accuracy and reliability of predicting mortality and length of stay for stroke patients. This evaluation includes precision, recall, F1 score, accuracy, ROC (Receiver Operating Characteristic), and AUC (Area Under the Curve). Accuracy measures the ratio of the number of accurate predictions to the total number of predictions and is calculated based on the formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

Precision measures the relevance of samples the model selects, thereby reducing false positives. The formula expresses Preisis:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

Recall measures the model's ability to identify all positive samples. The formula determines recall:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

The F1 score, which combines precision and recall, provides a balanced evaluation of accuracy and robustness.

$$\text{F1 Score} = \frac{2 \times precision \times recall}{precision + recall} \quad (4)$$

Where TP is True Positive, TN (True Negative), FP (False Positive), and FN (False Negative), additionally, ROC provides visualization of how changing the decision threshold in a classification model affects the True Positive Rate (TPR) and False Positive Rate (FPR). TPR measures the extent to which the model successfully predicts the positive class, while FPR measures the error rate in predicting the negative class. AUC on the ROC curve provides a numerical value to measure the model's overall quality. A bad model has an AUC close to 0, which means it has the worst separability measure. If the AUC is 0.5, the model has no class separation capacity (Teoh, 2018).

RESULTS AND DISCUSSION

Machine learning classification prediction models employ training and testing sets, typically in an 80:20 ratio. The training set (80%) is utilized for model training, while the testing set (20%) assesses the model's performance on unseen data (Joseph et al., 2022). An 80:20 split is adopted in this study, where 80% of the data is used for training while the remaining 20% evaluates the model's accuracy in making predictions.

1. Performance evaluation of the based model

A comparison of evaluation models of mortality prediction and stroke patients' Length of Stay (LoS) was carried out on the base model without hyperparameter optimization tuning. Table 2 presents the evaluation of mortality prediction models; XGB and CatBoost models stand out with the highest degree of precision, 71%, for predicting death. For healthy prediction, the degree of precision of X GBoost and Cat Boost reaches 75% and 67%, respectively. By contrast, DT, RF, and Extra Tree models show lower levels of precision, ranging from 57% to 60%. High precision here highlights the model's accuracy in correctly identifying patients at risk of death based on prediction. In the recall evaluation, the XGB model showed the best performance, with a score of 60% for the prediction of death and 83% for healthy prediction. High recalls indicate the model's ability to recognize most patients at risk of death or being healthy. The F-1 Score, which combines precision and recall, indicates that XGB and CatBoost are the best models for predicting death, scoring 67% and 59%, respectively. Overall, regarding accuracy, the XGBoost model achieved the highest score with 73%, followed by Cat Boost with 68%. The Decision Tree (DT), Random Forest (RF), and Extra Tree models had lower accurations, approximately 59%.

Meanwhile, the evaluation of LoS precision predictions in Table 3 reflects how much the model confirms optimistic predictions. XGB and Extra Tree models show high precision levels, 67% and 62% for "Non-Standard" and 69-71% for "Standard". CatBoost has lower precision, 33%, and 56%, for the categories "Non-Standard" and "Standard". Recall measures how many of the total positive cases were successfully identified by the model. An extra tree had the highest recall, specifically for "Standard" (85%). In comparison, XGBoost (XGB) showed recall of 44% for "Non-Standard" and 85% for"Standard." F1-score, as a combination of precision and recall, shows performance for both the Extra Tree (59-74%) and the XGBoost (XGB) (53-76%). Overall, the XGBoost and the extra tree have an accuracy of about 68%, the CatBoost 50%, while the Decision Tree (DT) and the Random Forest (RF) have different accuracy levels. Overall, the mortality model indicates that the XGB accuracy (73%) is superior to the other models. At the same time, for the long-term hospital prediction, XGB and Extra Tree have the same accuracy, 68%. Nevertheless, it should be noted that the level of precision of these predictions still requires improvements so that both models not only maintain but also improve the prediction performance optimally. By performing careful hyperparameter settings, we can optimize the potential of both models, thus providing excellent reliability in identifying the risk of death and accurate estimates related to the patient's length of stay.

Table 2. Algorithm performance untuned mortality prediction

		Mortality(1)	*Healthy(0)*
Precission	XGB	71	75
	Extra tree	57	60
	CatBoost	71	67
	DT	57	60
	RF	57	60
Recall	XGB	60	83
	Extra tree	40	75
	CatBoost	50	83
	DT	40	75
	RF	40	75

		Mortality(1)	*Healthy(0)*		
F-1score	XGB	67	77		
	Extra tree	47	67		
	CatBoost	59	74		
	DT	47	67		
	RF	47	67		
Accuracy	**XGB**	**Extra tree**	**CatBoost**	**DT**	**RF**
	73%	59%	68%	59%	59%

Table 3. Algorithm performance untuned LoS prediction

		Non-Standard(1)	*Standard(0)*		
Precision	XGB	67	69		
	Extra tree	62	71		
	CatBoost	33	56		
	DT	56	69		
	RF	50	62		
Recall	XGB	44	85		
	Extra tree	56	77		
	CatBoost	22	69		
	DT	56	69		
	RF	33	77		
F-1score	XGB	53	76		
	Extra tree	59	74		
	CatBoost	27	62		
	DT	56	69		
	RF	40	69		
Accuracy	**XGB**	**Extra tree**	**CatBoost**	**DT**	**RF**
	68%	68%	50%	64%	59%

Based on the analysis of the ROC chart, an evaluation of the AUC prediction model for mortality and length of stay revealed significant findings. Figures 3 and 4 show ROC AUC values on mortality predictions and hospitalization length, confirming that the XGBoost model stands out at the highest rate, reaching values of 0.72 and 0.71 in succession compared to other models. The high ROC value of the XGBoost model shows that in mortality forecasts, the model can effectively distinguish between surviving and unsafe patients. Similarly, in the long-term forecast of hospitals, this model can estimate the duration of hospitalization with an advantage in determining patients with standard and non-standard care.

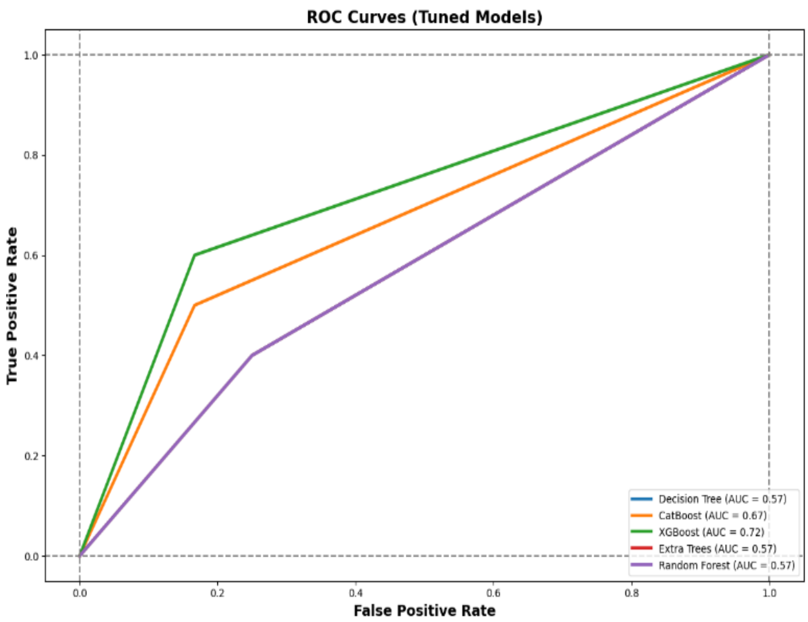

Figure 3. ROC curve for mortality prediction

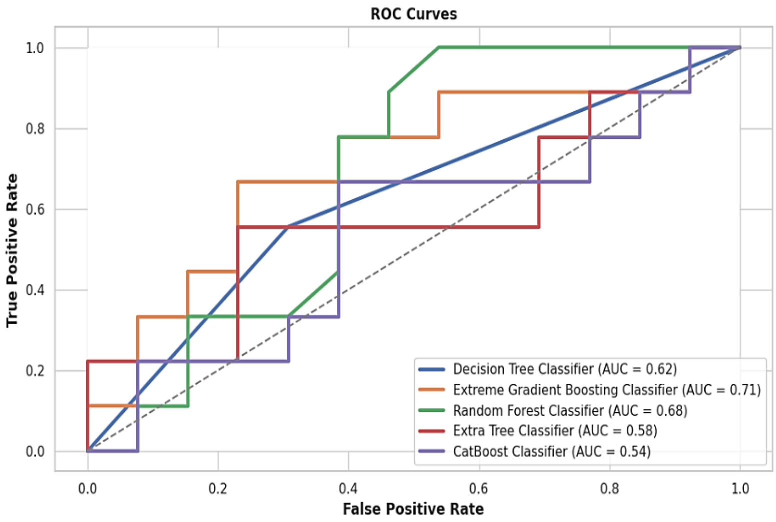

Figure 4. ROC curve for LoS prediction

2. Model Performance Improvement after Tuning

To improve the model's performance based on the model described earlier improved through hyperparameter optimization. Optuna, as the primary setup tool, is used to run automatic searches in the hyperparameter domain. The hyperparameter automatic search process involves repeated iterations, and the final results are documented in Table 4, which records the best hyperparameter values for predicting mortality and hospitalization length. This approach is applied to several models, including Extreme Gradient Boosting (XGB), Extra Tree, CatBoost, Decision Tree Classifier (DT), and Random Forest Classifiers (RF).

Table 4. Best hyperparameter values for mortality and LoS prediction

Algorithm	*Best Hyperparameter Value*
XGB	n_estimator : 267, max_depth : 12, learning_rate : 0.82, subsample : 0.92, colsample_bytree : 0.92, gamma : 0.28, reg_alpha : 0.0
Extra Tree	n_estimator : 392, max_depth : 3, min_samples_split : 0.74, min_samples_leaf : 0.39, max_features : 0.75,min_weight_fraction_leaf : 0.33,criterion : ‘entropy’
CatBoost	max_depth : 10, iterations : 446, learning_rate : 0.04,12_leaf_reg : 6.54 Border_count : 21
Decision Tree	max_depth : 5,min_samples_split : 0.7,min_samples_leaf : 0.6,max_features : 0.38
Random Forest	n_estimator : 91,max_depth : 7,min_samples_split : 0.57,min_samples_leaf : 0.22,max_features : 0.12

By implementing parameter optimization through the tuning process, there was a significant improvement in the evaluation performance of the five models used to predict mortality and the duration of hospitalization of stroke patients, as listed in Tables 5 and 6. Evaluation of the mortality prediction model in Table 5 revealed significant improvements compared to the previous base model. Performance analysis shows that the XGB algorithm stands out as the top choice, with the highest accuracy of 86%. XGB also shows high precision for both the prediction of death (82%) and a healthy condition (91%). In addition, the sensitivity or recall of XGB to predictions of death and health conditions reaches 90% and 83%, respectively. The F-1 score, which measures the balance between precision and recall, also shows good performance for XGB, with values of around 86% and 87%. Extra Tree and CatBoost have a precision of 86% in the mortality category, even though they are susceptible to healthy conditions (73%). In the overall analysis, the XGBoost algorithm can be justified in predicting mortality, considering a good balance between accuracy, precision and recall.

On the other hand, the model's performance in predicting the length of stay for a patient's stroke, as seen in Table 6, also shows improvements on some matrices compared to the base model's results. The algorithm performance results on the stay duration prediction (LoS), as shown in Table 6, provide an in-depth picture. In terms of precision, the XGBoost (XGB) algorithm showed an advantage of 86% for the Non-Standard category and 80% for the Standard category. The achievement of 100% precision by the Random Forest (RF) algorithm for the non-standards category reflects its very high ability to identify cases. Meanwhile, regarding sensitivity or recall, the XGB and Random Forest algorithms lead with high recalls for the Standard categories, reaching 92% and 100%, respectively. The F-1 score, which reflects the balance between precision and recalls, highlights the advantages of XGBoost, with the highest score of 75% for the Non-Standard category and 86% for the Standard category. Regarding accuracy, Extra Tree and Decision Tree also perform well at 68% and 77%, respectively; XGBoost remains the top choice with an 82% accuracy based on comprehensive cross-metric evaluation.

Table 5. Algorithm performance for mortality prediction after tuning

		Mortality(1)	*Healthy(0)*
Precision	XGB	82	91
	Extra tree	86	73
	CatBoost	86	73
	DT	70	75
	RF	83	69
Recall	XGB	90	83
	Extra tree	60	92
	CatBoost	60	92
	DT	70	50
	RF	50	92
F-1score	XGB	86	87
	Extra tree	71	81
	CatBoost	71	81
	DT	70	75
	RF	62	79

Accuracy	**XGB**	**Extra tree**	**CatBoost**	**DT**	**RF**
	86%	77%	77%	73%	73%

Table 6. Algorithm Performance for LoS prediction after tuning

		Non-Standard(1)	*Standard(0)*
Precision	XGB	86	80
	Extra tree	75	67
	CatBoost	50	64
	DT	70	83
	RF	100	62
Recall	XGB	67	92
	Extra tree	33	44
	CatBoost	44	69
	DT	78	77
	RF	11	100
F-1score	XGB	75	86
	Extra tree	46	77
	CatBoost	47	67
	DT	74	80
	RF	20	76

Accuracy	**XGB**	**Extra tree**	**CatBoost**	**DT**	**RF**
	82%	68%	59%	77%	64%

The ROC AUC curve provides a highly informative overview of the model's ability to distinguish between positive and negative categories. In this context, mortality predictions on XGBoost models show a substantial increase from 0.72 on the base model to 0.87. This indicates that the XGB model can more accurately distinguish between mortality and healthy cases, reaching more certainty. Meanwhile, in the long-term hospital prediction, XGB also experienced a significant increase from 0,71 on the basic model to 0,79. This suggests that the XGB model can better distinguish between non-standard and standard categories on duration of stay. This increased ROC AUC indicates that the XGB model, after a hyperparameter setting and optimization process, has successfully improved its ability to make more accurate and reliable decisions.

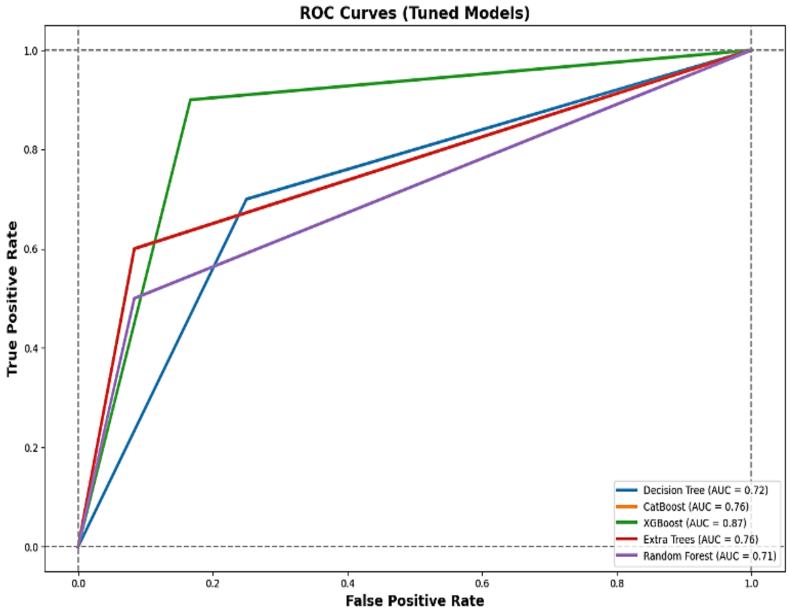

Figure 5. Tuned ROC curve for mortality prediction

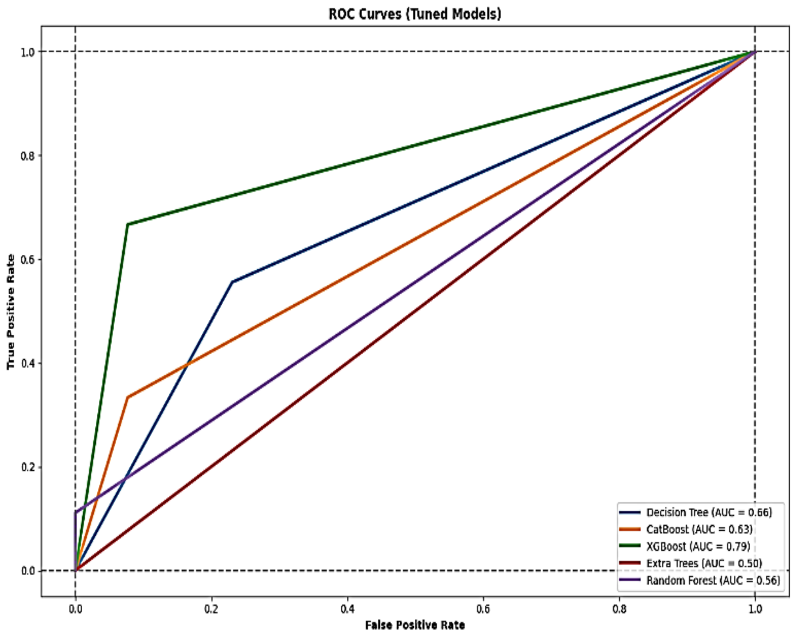

Figure 6. Tuned ROC curve for LoS prediction

3. Feature Importance

In predicting mortality and hospitalization duration for stroke patients, the XGBoost model excels in hyperparameter tuning. In Feature Importance analysis, weights are assigned to features based on their contribution to improved accuracy. This approach deepens our understanding of prediction's most influential risk factors (Oh et al., 2022). The analysis in Table 7 highlights key features influencing stroke patients' mortality and hospitalization duration. In mortality prediction, gender holds the highest influence (score: 0.0966), indicating its significant role. Stroke Location and

Previous Stroke History follow closely with scores of 0.0905 and 0.0867, underscoring their importance. Lipid profiles, represented by HDL, LDL, and TG, exhibit high significance. HDL's positive impact (score: 0.0834), along with LDL (score: 0.0776) and TG (score: 0.0580), provides a comprehensive insight into the lipid profile's relation to mortality. Other features like HB (score: 0.0731), HT (score: 0.0615), neutrophil ratio (score: 0.0610), and LEU (score: 0.0545) also contribute significantly to assessing the risk of mortality. To project the Length of Stay (LOS) for stroke patients, this analysis outlines the contribution of each attribute to predictive performance. This ranking prioritised total cholesterol features with the highest score of 0.1084, followed by HDL and Patient condition with values of 0.1006 and 0.1002, respectively. Biochemical attributes, including LEU, TG, HB, LDL, TR and LRL. Other factors, such as Stroke Location, also influenced the length of stay prediction, with a value of 0.066, respectively. This information aids in understanding and managing the risk of death and projecting the hospitalization duration for stroke patients, facilitating improved intervention and treatment planning.

Table 7. Feature Importance for Mortality and LoS Prediction

Mortality		*Length of Stay (LoS)*	
Feature	*Score*	*Feature*	*Score*
Gender	0.0966	Chol Total	0.1084
Stroke Location	0.0905	HDL(High-Density Lipoprotein)	0.1006
Previous Stroke History	0.0867	Patient Condition	0.1002
HDL(High-Density Lipoprotein)	0.0834	LEU(Leukocytes)	0.0906
LDL(Low-Density Lipoprotein)	0.0776	TG(Triglycerides)	0.0878
HB(Hemoglobin)	0.0731	HB(Hemoglobin)	0.0851
HT(Hematocrit)	0.0615	LDL(Low-Density Lipoprotein)	0.0847
NLR(Neutrophil-Lymphocyte Ratio)	0.0610	TR(Thrombocytes)	0.0742
TG(Triglycerides)	0.0580	NLR(Neutrophil-Lymphocyte Ratio)	0.0717
LEU(Leukocytes)	0.0545	Stroke Location	0.0661

CONCLUSIONS AND RECOMMENDATIONS

Stroke is one of the diseases that has a significant impact on public health. In this context, proper treatment and accurate prediction of mortality and extended hospitalization of stroke patients are critical in efforts to improve the quality of health care. The study aims to predict mortality and hospitalization duration in stroke patients using machine learning methods using the hyperparameter setting approach using the Optuna framework. Of several models evaluated, the Extreme Gradient Boosting model with hyperparameter settings proved to be the best in predicting both. The results showed that XGBoost could predict patient death with an accuracy of 86% and an AUC of 0.87.

Furthermore, the XGBoost model also showed excellent performance in predicting the length of stay after hyperparameter optimization. With an accuracy of 82% and an AUC value of 79, this model has been proven to provide reliable predictions regarding the duration of hospitalization for stroke patients. The findings not only confirm the potential for hyperparameter optimization with Optuna to improve the performance of the XGBoost model but also highlight the advantages of this model in supporting long-term prediction and management of stroke patient hospitals with a high degree of accuracy. This research has some limitations. First, the research was conducted on a relatively small data set of 106 patients. Secondly, the study was conducted in one of the hospitals in Indonesia, which may limit the ability to generalize the findings to other populations. Third, this study does not consider the impact of other factors, such as socioeconomic status and access to health services. Future research should address the limitations of this research by using larger datasets from more diverse populations.

REFERENCES

American Heart Association. (2023). *Heart Disease and Stroke Statistics - 2023*. Professional.Heart.Org. https://professional.heart.org/en/science-news/heart-disease-and-stroke-statistics-2023-update

Barsasella, D., Bah, K., Mishra, P., Uddin, M., Dhar, E., Suryani, D. L., Setiadi, D., Masturoh, I., Sugiarti, I., Jonnagaddala, J., & Syed-Abdul, S. (2022). *A Machine Learning Model to Predict Length of Stay and Mortality among Diabetes and Hypertension Inpatients*. *Medicina (Kaunas, Lithuania)*, *58*(11). https://doi.org/10.3390/medicina58111568

Chauhan, N. S. (2022). *Decision Tree Algorithm, Explained*. Https://Www.Kdnuggets.Com/. https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html

Chen, C. H., Tanaka, K., Kotera, M., & Funatsu, K. (2020). Comparison and improvement of the predictability and interpretability with ensemble learning models in QSPR applications. *Journal of Cheminformatics*, *12*(1), 1–16. https://doi.org/10.1186/s13321-020-0417-9

Chen, R., Zhang, S., Li, J., Guo, D., Zhang, W., Wang, X., Tian, D., Qu, Z., & Wang, X. (2023). A study on predicting the length of hospital stay for Chinese patients with ischemic stroke based on the XGBoost algorithm. *BMC Medical Informatics and Decision Making*, *23*(1), 1–10. https://doi.org/10.1186/s12911-023-02140-4

Crissman, M. on. (2019). *Optuna: An Automatic Hyperparameter Optimization Framework*. Odsc.Com. https://odsc.com/blog/optuna-an-automatic-hyperparameter-optimization-framework/

Feigin, V. L., Stark, B. A., Johnson, C. O., Roth, G. A., Bisignano, C., Abady, G. G., Abbasifard, M., Abbasi-Kangevari, M., Abd-Allah, F., Abedi, V., Abualhasan, A., Abu-Rmeileh, N. M. E., Abushouk, A. I., Adebayo, O. M., Agarwal, G., Agasthi, P., Ahinkorah, B. O., Ahmad, S., Ahmadi, S., … Murray, C. J. L. (2021). Global, regional, and national burden of stroke and its risk factors, 1990-2019: A systematic analysis for the Global Burden of Disease Study 2019. *The Lancet Neurology*, *20*(10), 1–26. https://doi.org/10.1016/S1474-4422(21)00252-0

Fernandez-Lozano, C., Hervella, P., Mato-Abad, V., Rodríguez-Yáñez, M., Suárez-Garaboa, S., López-Dequidt, I., Estany-Gestal, A., Sobrino, T., Campos, F., Castillo, J., Rodríguez-Yáñez, S., & Iglesias-Rey, R. (2021). Random forest-based prediction of stroke outcome. *Scientific Reports*, *11*(1), 1–12. https://doi.org/10.1038/s41598-021-89434-7

Ghazwani, M., & Begum, M. Y. (2023). Computational intelligence modeling of hyoscine drug solubility and solvent density in supercritical processing: gradient boosting, extra trees, and random forest models. *Scientific Reports*, *13*(1), 1–11. https://doi.org/10.1038/s41598-023-37232-8

Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: an interdisciplinary review. *Journal of Big Data*, *7*(1). https://doi.org/10.1186/s40537-020-00369-8

Huey Fern Tay. (2021). *When is it ok to impute missing values with a zero?* Towardsdatascience.Com. https://towardsdatascience.com/when-is-it-ok-to-impute-missing-values-with-a-zero-6d94b3bf1352

Hung, L. C., Sung, S. F., & Hu, Y. H. (2020). A machine learning approach to predicting readmission or mortality in patients hospitalized for stroke or transient ischemic attack. *Applied Sciences (Switzerland)*, *10*(18), 1–13. https://doi.org/10.3390/APP10186337

Jacob Gursky. (2020). *Boosting Showdown: Scikit-Learn vs XGBoost vs LightGBM vs CatBoost in Sentiment Classification*. Towardsdatascience.Com. https://towardsdatascience.com/boosting-showdown-scikit-learn-vs-xgboost-vs-lightgbm-vs-catboost-in-sentiment-classification-f7c7f46fd956

Jason Brownlee. (2019a). *A Gentle Introduction to Model Selection for Machine Learning*. Machinelearningmastery.Com. https://machinelearningmastery.com/a-gentle-introduction-to-model-selection-for-machine-learning/

Jason Brownlee. (2019b). *Difference Between Classification and Regression in Machine Learning*. Machinelearningmastery.com. https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/

Joseph, V. R., Joseph, V. R., & Stewart, H. M. (2022). *Optimal ratio for data splitting*. *February*, 531–538. https://doi.org/10.1002/sam.11583

Kuriakose, D., & Xiao, Z. (2020). Pathophysiology and Treatment of Stroke: Present Status and Future Perspectives. *International Journal of Molecular Sciences*, *21*(20), 1–24.

Lim, Y. (2022). *State-of-the-Art Machine Learning Hyperparameter Optimization with Optuna*. Towardsdatascience.Com. https://towardsdatascience.com/state-of-the-art-machine-learning-hyperparameter-optimization-with-optuna-a315d8564de1

Mohebi, S., Parham, M., Sharifirad, G., & Gharlipour, Z. (2018). *Factors related to 6‑ month mortality after the first‑ ever stroke*. *January*, 1–6. https://doi.org/10.4103/jehp.jehp

Moore, A., & Bell, M. (2022). XGBoost, A Novel Explainable AI Technique, in the Prediction of Myocardial Infarction: A UK Biobank Cohort Study. *Clinical Medicine Insights: Cardiology*, *16*. https://doi.org/10.1177/11795468221133611

Mridha, K., Ghimire, S., Shin, J., Aran, A., Uddin, M. M., & Mridha, M. F. (2023). Automated Stroke

Prediction Using Machine Learning: An Explainable and Exploratory Study With a Web Application for Early Intervention. *IEEE Access*, *11*(June), 52288–52308. https://doi.org/10.1109/ACCESS.2023.3278273

Muslim Karo Karo, I. (2020). Implementasi Metode XGBoost dan Feature Importance untuk Klasifikasi pada Kebakaran Hutan dan Lahan. *Journal of Software Engineering, Information and Communication Technology*, *1*(1), 11–18.

Neto, C., Brito, M., Peixoto, H., Lopes, V., Abelha, A., & Machado, J. (2020). Prediction of Length of Stay for Stroke Patients Using Artificial Neural Networks. *Advances in Intelligent Systems and Computing*, *1159 AISC*(Dm), 212–221. https://doi.org/10.1007/978-3-030-45688-7_22

Ogunleye, B. O. (2021). *Statistical Learning Approaches to Sentiment Analysis in the Nigerian Banking Context A thesis submitted in partial fulfillment of the requirements of Sheffield Hallam University for the degree of Doctor of Philosophy Bayode Oluwatoba Ogunleye October 2021*. *October*.

Oh, T., Kim, D., Lee, S., Won, C., Kim, S., Yang, J., Yu, J., Kim, B., & Lee, J. (2022). Machine learning-based diagnosis and risk factor analysis of cardiocerebrovascular disease based on KNHANES. *Scientific Reports*, 1–11. https://doi.org/10.1038/s41598-022-06333-1

Olson, R. S., La Cava, W., Mustahsan, Z., Varik, A., & Moore, J. H. (2018). Data-driven advice for applying machine learning to bioinformatics problems. *Pacific Symposium on Biocomputing*, *0*(212669), 192–203. https://doi.org/10.1142/9789813235533_0018

Pacheco-Barrios, K., Giannoni-Luza, S., Navarro-Flores, A., Rebello-Sanchez, I., Parente, J., Balbuena, A., de Melo, P. S., Otiniano-Sifuentes, R., Rivera-Torrejón, O., Abanto, C., Alva-Diaz, C., Musolino, P. L., & Fregni, F. (2022). Burden of Stroke and Population-Attributable Fractions of Risk Factors in Latin America and the Caribbean. *Journal of the American Heart Association*, *11*(21). https://doi.org/10.1161/JAHA.122.027044

Rachoin, J.-S., Aplin, K. S., Gandhi, S., Kupersmith, E., & Cerceo, E. (2020). Impact of Length of Stay on Readmission in Hospitalized Patients. *Cureus*, *12*(9). https://doi.org/10.7759/cureus.10669

Safaei, N., Safaei, B., Seyedekrami, S., Talafidaryani, M., Masoud, A., Wang, S., Li, Q., & Moqri, M. (2022). E-CatBoost: An efficient machine learning framework for predicting ICU mortality using the eICU Collaborative Research Database. In *PLoS ONE* (Vol. 17, Issue 5 May). https://doi.org/10.1371/journal.pone.0262895

Saikumar Talari. (2022). *Random Forest vs Decision Tree: Key Differences*. Www.Kdnuggets.Com. https://www.kdnuggets.com/2022/02/random-forest-decision-tree-key-differences.html

Tarwidi, D., Pudjaprasetya, S. R., Adytia, D., & Apri, M. (2023). An optimized XGBoost-based machine learning method for predicting wave run-up on a sloping beach. *MethodsX*, *10*(March), 102119. https://doi.org/10.1016/j.mex.2023.102119

Teoh, D. (2018). Towards stroke prediction using electronic health records. *BMC Medical Informatics and Decision Making*, *18*(1), 1–11. https://doi.org/10.1186/s12911-018-0702-y

Thankachan, K. (2022). *What? When? How?: ExtraTrees Classifier*. Https://Towardsdatascience.Com/. https://towardsdatascience.com/what-when-how-extratrees-classifier-c939f905851c

Venketasubramanian, N., Yudiarto, F. L., & Tugasworo, D. (2022). Stroke Burden and Stroke Services in Indonesia. *Cerebrovascular Diseases Extra*, *12*(1), 53–57. https://doi.org/10.1159/000524161

Wang, W., Rudd, A. G., Wang, Y., Curcin, V., Wolfe, C. D., Peek, N., & Bray, B. (2022). Risk prediction of 30-day mortality after stroke using machine learning: a nationwide registry-based cohort study. *BMC Neurology*, *22*(1), 1–9. https://doi.org/10.1186/s12883-022-02722-1

WHO. (2021). *Cardiovascular diseases (CVDs)*. Www.Who.Int. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Yang, C. C., Bamodu, O. A., Chan, L., Chen, J. H., Hong, C. T., Huang, Y. T., & Chung, C. C. (2023). Risk factor identification and prediction models for prolonged length of stay in hospital after acute ischemic stroke using artificial neural networks. *Frontiers in Neurology*, *14*. https://doi.org/10.3389/fneur.2023.1085178

Yi, J., Lee, J., Kim, K. J., Hwang, S. J., & Yang, E. (2020). Why Not To Use Zero Imputation? Correcting Sparsity Bias in Training Neural Networks. *8th International Conference on Learning Representations, ICLR 2020*, *1*, 1–27.