

Optimizing Stroke Mortality Prediction: A Comprehensive Study on Risk Factors Analysis and Hyperparameter Tuning Techniques

Imam Tahyudin¹, Ades Tikaningsih¹, Puji Lestari¹, Eko Winarto², Nazwan Hassa²

¹ Universitas Amikom Purwokerto, Jl. Letjendpol Soemarto, Purwokerto, Indonesia

² Banyumas Government Hospital, Jl. Rumah Sakit No. 1, Banyumas, Indonesia

Abstract – Stroke is one of the major killer diseases in the world. Understanding the factors that influence the death of stroke patients is vital to improving patient care and outcomes. In this study, we used stroke patient data and machine learning techniques using a variety of algorithms, including Extreme Gradient Boosting, CatBoost, Extra Tree, Decision Tree, and Random Forest, to predict patient death after stroke. After performing hyperparameter settings, the XGBoost model achieved an accuracy of 86% with an AUC of 87. Significant improvements in the accuracy and predictive capability of this model after hyperparameter settings indicate a strong potential for clinical applications. In addition, our findings suggest that factors such as the patient's age, type of stroke, and blood pressure at the time of hospitalization have a significant impact on stroke patients' deaths. By understanding these factors, healthcare providers can improve patient intervention and management to reduce the risk of death after stroke. This research has made an important contribution to the development of a system for predicting the risk of death of stroke patients, which can help doctors and nurses identify high-risk patients and provide appropriate treatment.

Keywords – Predicting death, stroke, machine learning.

1. Introduction

Stroke is the leading cause of death and disability worldwide [1]. In the Global Stroke Fact Sheet released in 2022, it was found that more than 101 million people have experienced stroke, with 6.5 million people dying each year [2]. Between 1990 and 2019, there was a significant increase in stroke incidences of 70%, stroke deaths of 43%, stroke prevalence of 102%, and disability adjusted life years (DALY) increased by 143% [3]. Understanding the type of stroke is also important, with the global prevalence in 2019 reaching 101.5 million, divided into 77.2 million for ischemic stroke, 20.7 million for intracerebral bleeding, and 8.4 million for subarachnoid bleeding [4]. A stroke occurs when the blood supply to the brain is interrupted, so the brain cells cannot obtain the oxygen and nutrients they need. This can be caused by a blockage or rupture of blood vessels in the brain [5].

In Indonesia, stroke was the number one cause of death that killed 328,5 thousand people (21.2% of the total deaths) in 2019 [6]. Between 2013 and 2018, Indonesia witnessed a surge in stroke cases, with the occurrence rising from 8.3 cases per 1,000 individuals to 12.1 cases per 1,000 individuals. The highest incidence of stroke occurred in the East Java region at 12.4% (113.045), in West Java at 11.4% (131.846), and in Central Java at 11.8% (96.794) [7]. Meanwhile, the Central Java Health Service report stated that the prevalence of non-hemorrhagic stroke in Central Java in 2018 was 18,284 cases, which is an increase of 0.05% higher than in 2017. The stroke data in Semarang showed the prevalence of new non-hemorrhagic stroke cases in 2018 at 800 cases [8].

According to clinical reviews, there are various factors that have the potential to influence the prognosis of stroke.

DOI: 10.18421/TEM131-74

<https://doi.org/10.18421/TEM131-74>

Corresponding author: Imam Tahyudin,
Universitas Amikom Purwokerto, Jl. Letjendpol Soemarto,
Purwokerto, Indonesia


Email: imam.tahyudin@amikompurwokerto.ac.id

Received: 27 October 2023.

Revised: 27 January 2024.

Accepted: 05 February 2024.

Published: 27 February 2024.

 © Imam Tahyudin et al; published by UIKTEN. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 License.

The article is published with Open Access at <https://www.temjournal.com/>

In general, these factors can be categorized as basic patient information, complications, stroke subtypes, and treatment plans [9]. The presence and interaction of these diverse factors make the identification and management of stroke patients' prognosis increasingly complex and requires careful therapeutic planning. Taking these variables into account, predicting stroke deaths is a good indicator for evaluating the effectiveness of stroke treatment, and identifying predictors of death in hospitals is important for improving stroke outcomes [10]. However, early detection of stroke is still a challenge, as stroke symptoms can be difficult to recognize, especially in the early stages. Therefore, the use of machine learning technology plays an important role.

Machine learning has shown great potential for predicting the death rate of stroke patients [11]. The complexity of strokes is potentially suitable for the use of machine learning algorithms, which are able to combine a large number of variables and observations into one predictive model [12]. Previous studies have shown that machine learning is able to provide accurate predictions of stroke-related deaths based on a range of clinical information, including demographic data, medical history, and medical examination results.

The primary aim of this study is to investigate the utilization of machine learning methodologies for forecasting stroke-related fatalities at Banyumas Regional General Hospital (RSUD) in Indonesia. Using data from patients who have had a stroke at the Banyumas Medical Center, the study seeks to develop accurate prediction models using various machine learning algorithms. This research is expected to make a significant contribution to more effective prevention and management of stroke risk in this hospital environment. This research provides an important contribution to the development of a system for predicting stroke patient risk of death at RSUD Banyumas, which will later help doctors and nurses identify high-risk patients and provide appropriate treatment.

In addition, the research also broadened our understanding of the risk factors for the death of stroke patients, which could support the development of more effective public health policies in stroke prevention and management efforts at RSUD Banyumas.

2. Related Work

Numerous research endeavors have delved into the utilization of machine learning methods for the assessment and anticipation of mortality risk in individuals afflicted by strokes.

Zhu *et al.* have developed a machine learning model to predict mortality in stroke patients. The data used came from MIMIC-IV with more than 70,000 patients. Several models, including logistic regression, SVM, and random forest, were developed and trained with demographic data, medical history, and laboratory data. As a result, the random forest model obtained the highest accuracy with an AUC of 0.85. This shows the model's ability to accurately classify survivors and deceased patients at 85 percent [11].

Another study by Rahim *et al.* discussed the development of the XGBoost model to predict stroke cases at Dr. Sardjito Central General Hospital in Yogyakarta, Indonesia. The data set used includes information from 200 patients. Research results show that the XGBoost model has a 90% accuracy rate in predicting stroke cases. In addition, the model is also able to identify factors associated with an increased risk of stroke [13].

Tazin *et al.* have also conducted research on the detection and prediction of stroke disease using the robust learning approach in their research. These findings confirm that the decision tree is one of the most effective methods for detecting and predicting stroke disease [14].

The study of Sharma *et al.* presented additional insights regarding the use of extra-tree classifiers in the context of breast cancer prediction. The findings from the study demonstrated that the model extra tree classifiers with 100 estimators achieved an accuracy of 97.35%. Moreover, the model of the extra-tree classifier with 200 estimators reached an accuracy of 96.64%. These results indicated that models with 100 estimators had higher performance in the classification of breast cancers compared to models using 200 estimators [15].

Safaei *et al.* introduced E-CatBoost into the context, a resourceful machine learning framework designed for mortality prediction among ICU patients. The model's efficacy was evaluated through the assessment of its performance using the Area Score under the Receiver Operating Characteristic Curve (AUROC), which ranges from 0.86 to 0.92 for the CatBoost model and from 0.83 to 0.91 for the E-catBoost in the specified disease group. The results showed that both were able to achieve high AUROC scores, with E-KatBoost providing a significant improvement in the accuracy of predicting ICU mortality, especially when measured in the entire patient population [16].

3. Methodology

This research adopts the framework of Rui Chen *et al.* [17] which discusses long-standing forecasts of hospitalization of ischemic stroke patients in China.

The stages of this system are as follows:

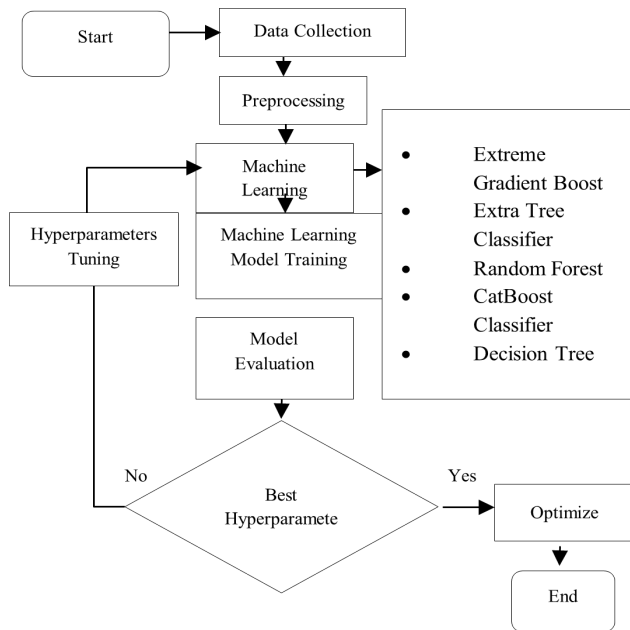


Figure 1. Flowchart

3.1. Data Collection

The study uses data from adult patients receiving treatment at the Neural Clinic of the Regional General Hospital (RSUD) in Banyumas, Indonesia. Data was collected during the period from January 2022 to May 2023. The data collection process involves the observation and recording of medical information by the RSUD research team, including demographic data, medical history, and patient clinical records. In total, there are 106 patient records with 29 attributes available. The attributes of the original stroke patient data can be found in Table 1.

Table 1. Original stroke dataset

Attribute name	Description
Number	Entity number
Enumerator	Observer
Data Collection Date	Data Collection Date
Patient Condition	Patient Health Status
Resp Code	Prescription Code
Entry Date	Date of Patient Admission
Out Date	Date of Patient Discharge
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

Employee	Occupation or profession of the patient
Marital Status	Patient Marital Status
Primary Diagnosis	Diagnosis or primary medical condition suffered by the patient
History Of CVD	Information on cardiovascular disease history, such as previous heart disease or stroke
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 prior to the observed case
Initial Scanning Date	Date of the first examination or scan conducted on the patient
Stroke Location	Information about the location of the stroke attack in the patient
DLO Test Date	Date of Blood Test
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	Platelets, a type of blood cell involved in blood clotting
NLR	Neutrophil-Lymphocyte Ratio, the ratio of neutrophils to lymphocytes in the blood
LIPID Test Date	Date of Blood Lipid Profile Test
CHOL Total	Total Cholesterol, 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 considered "bad" because it can clog arteries.

In this study, the prediction of mortality risk is categorized as a classification problem, to predict patient mortality based on the attributes of the patient condition in the dataset of Table 1, which contains information on patients with mortality status and patients who are healthy or recovering.

3.2. Exploratory Data Analysis (EDA)

EDA is a method employed to scrutinize a dataset with the intention of summarizing its primary characteristics [18].

In the context of this study, EDA is employed to acquire a comprehensive insight into the attributes of the data within a stroke dataset. EDA helps reveal the relationship between variables, providing an initial insight into patterns and trends that may be reflected in the data.

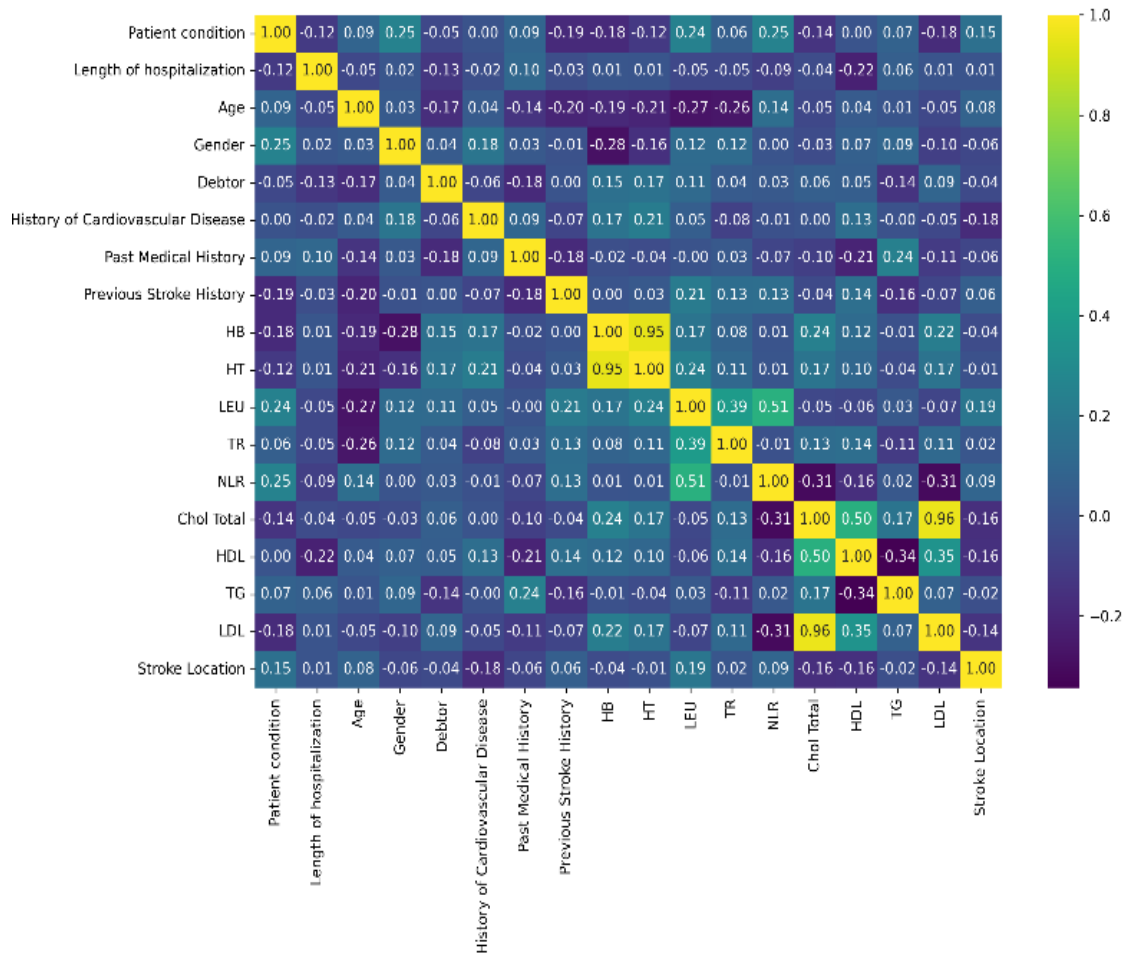


Figure 2. Correlation of stroke dataset

Data correlation visualization is used to evaluate the degree of correlation between variables using a heat map, which describes the relationship of variables in a dataset [19]. This approach is useful for summarizing the correlational matrix in a data set, where each element represents the coefficient of correlation between two variables. In the context of machine learning, heatmaps serve the purpose of revealing highly correlated variables within datasets, both in relation to target variables and among themselves [20].

As seen in Figure 2, correlations between variables in stroke patient data can be identified. For example, hemoglobin (HB) and hematocrit (HT) levels show a very strong positive correlation of around 0.94.

Also, the number of leukocytes (LEUs) and the neutrophil ratio to lymphocytes (NLRs) show a significant positive correlation of about 0.507816. Furthermore, total cholesterol (total cholesterol) and low-density lipoprotein (LDL) also have a high positive correlation, about 0.95. On the other hand, the correlation between the patient's condition and gender (0.25) and age (0.08) is moderate. Age and financial status (debtor) indicate a moderate negative correlation of around -0.16.

However, there are no significant correlations between patient condition and hospitalization length (-0.05) or between age and hospital duration (-0.002). Furthermore, financial status and sex also show low correlations of around 0.035891.

3.3. Feature Selection

Feature selection is a crucial stage in this research that allows us to select the most relevant and informative subset of attributes to analyze predictions of mortality in stroke patients. From the initial data set consisting of 106 patient samples and 29 attributes, the study carried out a careful feature selection based on a number of critical factors. Of the 18 variables, the study divided them into two main categories that facilitated analysis, namely demographic information and medical history.

The first group included attributes related to patient characteristics and their medical history, while the second group contained attributes related to medical test results and health parameters, as listed in Table 2. In this process, we consider the following factors:

- 1) Rational selection of features based on medical and methodological relevance
- 2) Evaluation of data quality to ensure that only attributes with quality data are included in the analysis
- 3) Removal of irrelevant attributes to reduce complexity and noise in the model
- 4) Understanding the potential interaction or correlation between the attributes selected to ensure proper analysis

Table 2. Selected variables of stroke dataset

Categories	Variables
Demographic characteristics of the patients.	<i>Length of hospitalization</i> <i>Age</i> <i>Gender</i> <i>History of Cardiovascular Disease</i> <i>Past Medical History</i> <i>Previous Stroke History</i> <i>Patient Conditions</i>
Medical Data	<i>HB (Hemoglobin)</i> <i>HT(Hematocrit)</i> <i>LEU (Leukocyte)</i> <i>TR (Thrombocyte)</i> <i>NLR(Neutrophil-Lymphocyte Ratio)</i> <i>Cholesterol Total (Cholesterol Total)</i> <i>HDL (High-Density Lipoprotein)</i> <i>TG (Triglyceride)</i> <i>LDL (Low-Density Lipoprotein)</i> <i>Stroke Location</i>

3.4. Data Preprocessing

In the preprocessing phase, several crucial steps were undertaken in this study to prepare the data before employing it in constructing predictive mortality models for stroke patients.

3.4.1. Handling the missing value

Handling the missing data is undeniably one of the foundational challenges in the realm of machine learning. Among many approaches, the simplest and most intuitive way is zero imputation, which treats missing entry values the same as zero [21]. In the context of this study, missing values have been identified on several attributes, including "History of Cardiovascular Disease", "Past Medical History", and "Previous Stroke History". Proper handling of missing values is essential to ensuring the integrity and relevance of the data before proceeding to the analysis stage. The 0-value impotence was chosen because of its high medical relevance in the context of predictive mortality research and hospitalization time for stroke patients.

The patient's medical history stands as a crucial determinant influencing both treatment outcomes and the ability to predict prognosis. A replacement with a value of 0 makes it possible to separate patients who have a history of disease from those who have no history of illness. Identifying a patient who has no history of a particular disease is relevant information for a medical professional who will analyze the results of this study. In addition, replacing 0 with 0 ensures that the data remains in a consistent numerical format.

3.4.2. Encoding variable

The process of dealing with categorical or non-numerical attributes in the datasets is carried out. Categorical variable encoding is an important step in the preparation of data for predictive models, as most machine learning models require numerical data [22]. In this study, the variables length of hospitalization, gender, and patient conditions were encoded using one-hot encoding, while stroke location, history of cardiovascular disease, past medical history, and previous stroke history were coded using coding labels. Encoding categorical variables with the right method is a key to ensuring that the data used in the modeling is an appropriate representation of the information contained in the dataset.

3.4.3. Splitting the Data

Split datasets are intended to divide data sets into training and testing data that will be used for modeling. It is commonly used in 80:20 or 80:10:10 configurations [23]. In this study, the 80:20 split configuration is used, where 80% of the data from the data set is used by machines to determine patterns and 20% will be used as data to make predictions.

3.5. *Machine Learning Algorithm*

In this section, we will provide an overview of the five algorithms that are the focus of this research. These algorithms play an important role in improving model performance, with each adopting a different approach.

3.5.1. *Extreme Gradient Boosting*

Extreme gradient boosting (XGB) is an open-source library that provides efficient and effective implementation of gradient enhancement algorithms [24]. XGB uses advanced regularization (L1 and L2), which enhances model generalization capabilities [25]. The basic concept of boosting is to build more accurate models by combining hundreds of simple tree models with low accuracy, in which each iteration will produce a new tree for the model. The complexity of the tree will affect the outcome [26].

3.5.2. *CatBoost Classifier*

The CatBoost Classifier uses sequential target statistics and sequential improvements that make it good for heterogeneous data categorical values and has strong performance compared to other implementations of gradient improvement decision trees [27]. This feature helps improve the generalization performance of the model and makes it more resilient to new health data [16].

3.5.3. *Random Forest*

The Random Forest (RF) is a popular ensemble model specifically created to overcome the limitations of the traditional decision tree algorithm. The RF technique involves training many decision-tree learners simultaneously to minimize the bias and variance of the model [28]. Large-scale studies in general literature provide evidence that supports some class families like random forest (RF) in terms of classification performance [29].

Random forest algorithms have been known for their high accuracy in predicting diseases in medical research [30]. Random forests offer the advantages of being easily interpretable, swift to train and evaluate, demonstrating proficiency in intricate datasets, and exhibiting resilience against irrelevant features.

3.5.4. *Extra Tree Classifier*

ExtraTrees is an ensemble ML approach that trains numerous decision trees and aggregates the results from the group of decision trees to output a prediction. However, there are few differences between extra trees and random forests. They work in a slightly different way.

Another important difference is in the way they choose where to share information within the decision tree. Extra trees does this in a random way, which means he chooses a random value to divide the feature and create a new branch in the tree. On the other hand, random forest uses a more intelligent algorithm to find and select the best value for dividing the feature [31].

3.5.5. *Decision Tree*

The decision tree algorithm works by dividing the data into several parts recursively and assembling the parts based on the characteristics that most distinguish the class until the termination criteria meet [32]. The described algorithm is a supervised machine learning method known for its capability to partition data into segments or branches, based on various input variables. The branches of the decision tree are arranged upward, with the top branches representing the final result [33].

3.6. *Hyperparameter Tuning*

In machine learning methods, there are a series of parameter values that are thought to improve model performance, known as hyperparameters. Hyperparameters are used to improve algorithm performance, and this has a significant impact on a variety of model tests. The hyperparameter adjustment process can be done manually or by testing a group of hyperparameter on a previously specified parameter [34].

The study uses the Optuna library to optimize the performance of the mortality prediction model for stroke patients. Optuna defines the hyperparameter optimization challenge as the task of minimizing or maximizing a target function, which accepts a set of hyperparameters as input and provides a score as the output (validation). Optuna also provides abbreviation, i.e., automatic initial stop of unpromising trials [33].

Table 3. Tuning parameters of the algorithms

Parameters	Description	Range
<i>n_estimators</i>	The number of trees in the model. This controls how many trees will be used in the ensemble (e.g., in Random Forest or XGBoost).	100 - 300
<i>max_depth</i>	Controls the maximum depth of trees in the model. A larger <i>max_depth</i> value will result in a more complex model.	3 – 12
<i>learning_rate</i>	The learning rate controls how big steps the model will take during the learning process.	0.1 - 1.0
<i>subsample</i>	The proportion of data samples used to train each tree.	0.5 - 1.0
<i>gamma</i>	The minimum threshold for splitting a node.	0.6 - 1.0
<i>colsample_bytree</i>	Proportion of features each tree.	0.0 - 1.0
<i>reg_alpha</i>	Regulation L1 to prevent overfitting	0.0 - 1.0
<i>reg_lambda</i>	Regulation L2 to control complexity	0.0 - 1.0
<i>min_samples_split</i>	Minimum number of samples to divide the node	0.1 – 1.0
<i>min_samples_leaf</i>	Minimum amount of sample in each leaf	0.1 – 0.5
<i>max_features</i>	Maximum characteristics for node separation	0.1 – 0.5
<i>min_weight_fraction_leaf</i>	Total weight ratio for one leaf Criteria for evaluation of node segregation	0.0 – 0.4
<i>criterion</i>	The criterion used to measure the quality of splitting at each node in the decision tree.	“gini”, “entropy”
<i>iterations</i>	The number of iterations or steps taken by the model during the learning process.	100-500
<i>l2_leaf_reg</i>	A specific parameter for setting trees.	1-10
<i>border_count</i>	The number of bins or categories for use on categorical data.	10-255
<i>thread_count</i>	The number of threads or threads to be used by the model during training.	-1
<i>loss_function</i>	The loss function to be optimized during the learning process.	Logloss
<i>random_seed</i>	Setting to ensure that the model's results can be reproduced.	123
<i>verbose</i>	The level of noise or information to be displayed during the training or model evaluation process.	False

3.7. Model Performance Evaluation

To evaluate the performance of the classification used in this study, the data set was divided into two parts: an 80% training data set and a 20% test data set. A confusion matrix is used to compare classification performance. The confusion matrix describes how often models estimate correctly and how often they estimate incorrectly. False positives and false negatives are allocated to poor predictive values, while true positives and negatives are actually placed on the correctly anticipated values [14].

To evaluate performance, we use metrics such as precision, recall, F1 score, and accuracy calculated based on the confusion matrix and its formulas below:

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

$$\text{F1-score} = \frac{2TP}{2TP+FP+FN} \quad (2)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (4)$$

where *TP* = True positive, *TN* = True Negative, *FP* = False Positive, and *FN* = False Negative [35].

In addition, this study illustrates the performance of models using a Receiver Operating Characteristic (ROC) curve that displays a trade-off between sensitivity and specificity at each threshold [36]. The area value below the ROC curve (AUC) represents a higher probability of ranking the randomly selected positive instance than the randomly selected negative instance. A better model than the random one will have an AUC value greater than 0.5, while the perfect model will have an AUC value of 1.0 [37].

4. Results and Discussion

The study applied the 80:20 data division scheme, where 80% of the total sample was used for model training processes, while the remaining 20% was used to test model performance. The initial dataset consisted of 106 data samples. Of these, 84 samples were used as training data (X_train), while 22 samples were used as test data (X_test).

4.1. Model Testing Without Hyperparameter Tuning

The comparison results of the fifth algorithm's performance in predicting mortality in stroke patients are recorded in Table 4. The Extreme Gradient Boosting (XGB) model was able to achieve a significant accuracy of 73% and showed a good level of accuracy, reaching 75% in identifying patients with a healthy status. Moreover, the model also recorded the highest recall rate, reaching 60% in recognizing patients who are truly at high risk of death (mortality). The CatBoost model also showed satisfactory results, with an accuracy rate of 68%.

On the other hand, the decision trees, extra trees, and random forest models showed similar performance, but with a lower degree of precision and a lower F-1 score.

In terms of prediction of mortality, measurements of the ROC AUC are used to assess the extent to which the model can distinguish between survivors and non-survivors with a high degree of accuracy. Figure 3 shows that the XGB and CatBoost models have high ROC AUC values, indicating excellent performance. Therefore, the ROC AUC chart is used to measure and compare the predictive capabilities of both models in terms of mortality prediction.

Table 4. Algorithm performance on untuned mortality prediction

		Mortality (1)	Healthy (0)	Accuracy
Precision	XGB	71	75	XGB 73 %
	Extra tree	57	60	
	CatBoost	71	67	
	DT	57	60	Extra Tree 59 %
	RF	57	60	
Recall	XGB	60	83	CatBoost 68%
	Extra tree	40	75	
	CatBoost	50	83	
	DT	40	75	
	RF	40	75	
F-1score	XGB	67	77	DT 59%
	Extra tree	47	67	
	CatBoost	59	74	RF 59%
	DT	47	67	
	RF	47	67	

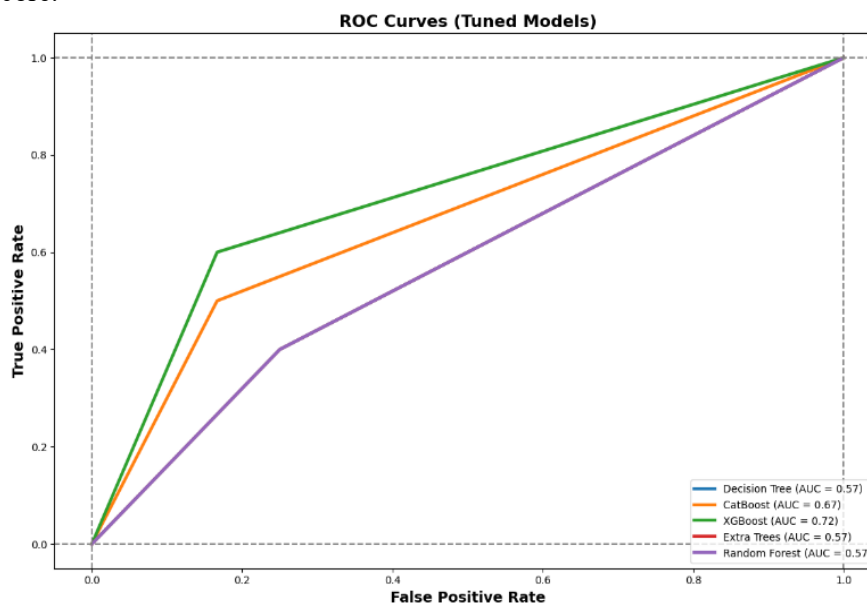


Figure 3. ROC Curve for mortality prediction

4.2. Model Performance after Hyperparameter Setting

Hyperparameter setting is a crucial step in ensuring that the machine learning model achieves optimal performance. In this study, hyperparameter settings were performed on a number of algorithms, including Extreme Gradient Boosting (XGB), Extra Tree, CatBoost, Decision Tree (DT), and Random Forest (RF). The primary mission of this setup is to improve the ability of the model to predict mortality. The best results of the hyperparameter setup process can be found in Table 5. The setup results reflect dedication to optimizing the machine learning model so that predictions of mortality and the duration of stroke patient care are more accurate and efficient.

Table 5. Best hyperparameter values for mortality

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 reg_lambda : 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 l2_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

The performance of the predictive mortality model of stroke patients experienced a significant improvement after hyperparameter settings. Table 7 shows an improvement in the Extreme Gradient Boosting (XGB) model, which achieved high accuracy (86%) and good accuracy (91%) in identifying high-risk patients. The extra tree algorithm managed to recognize low-risk patients with an accuracy of 77%.

The CatBoost model has comparable accuracy, although it has lower performance in identifying patients at high risk (recall 60%). However, the decision tree (DT) and random forest (RF) show lower performance.

In addition, Figure 4 (mortality) shows a significant improvement in model performance after hyperparameter setting in mortality prediction. The XGBoost (XGB) model reached the highest ROC AUC of 0.87, showing excellent ability in predicting mortality. Models such as CatBoost and Extra Tree also experienced significant increases in the ROC AUC after setting. Hyperparameter settings help models give more accurate predictions about the duration of patient care, which is crucial in patient care management.

Table 6. Algorithm performance for mortality after tuning

		Mortality (1)	Healthy (0)	Accuracy
Precision	XGB	82	91	XGB 86 %
	Extra tree	86	73	
	CatBoost	86	73	
	DT	70	75	Extra Tree 77 %
Recall	RF	83	69	CatBoost 77 %
	XGB	90	83	
	Extra tree	60	92	
	CatBoost	60	92	
F-I-score	DT	70	50	DT 73 %
	RF	50	92	
	XGB	86	87	RF 73 %
	Extra tree	71	81	
	CatBoost	71	81	
DT	70	75		
	RF	62	79	

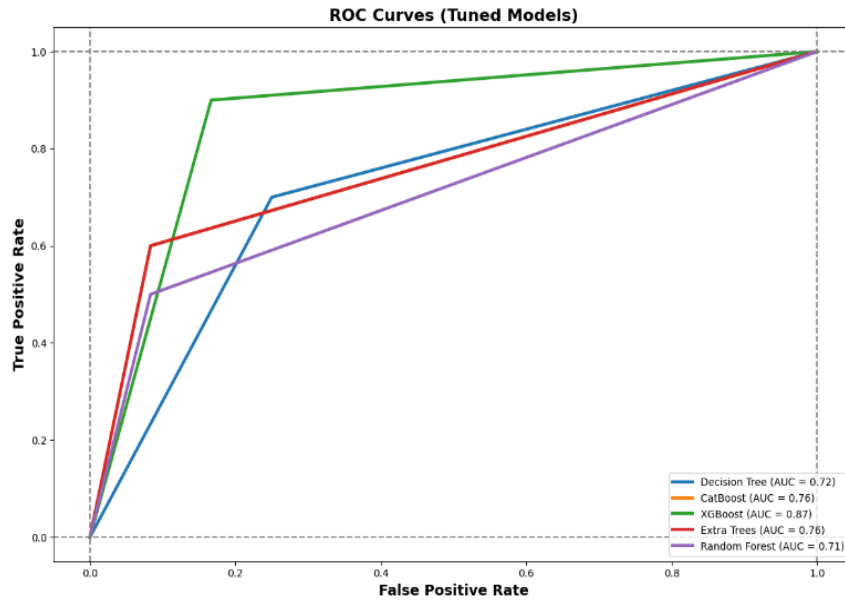


Figure 4. Tuned ROC curve for mortality prediction

4.3. Feature Importance

In this study, the Extreme Gradient Boosting Classifier (XGB) model showed excellent performance, especially after the hyperparameter setting, making it the algorithm with the highest accuracy in predicting mortality and duration of stroke patient care. To identify the features that most influence the risk of mortality and the duration of treatment, we implemented the feature importance approach using the best algorithm, XGBoost. This feature importance analysis allows measurement of the impact of each feature on the accuracy of the model. That is, the features that have a greater influence will have a higher level of importance in the prediction.

The results of the analysis of feature importance in the case of the prediction of mortality can be seen in Figure 5.

The sex of the patient became the most significant attribute, with a difference in the risk of death between men and women. Localization of strokes in the brain, especially in the middle brain and stem, also contributes to a higher risk of death due to the importance of these areas in controlling body functions. Previous stroke history and health factors such as HDL, LDL, hemoglobin, blood pressure, neutrophil-lymphocyte ratio (NLR), triglyceride (TG), and leukocyte count (LEU) also play an important role in predicting mortality. Low HDL levels, high LDL rates, low hemoglobin levels, and high blood pressure are significant risk factors. NLR and the number of leukocytes reflect the inflammatory response in the body, which also affects mortality risk.

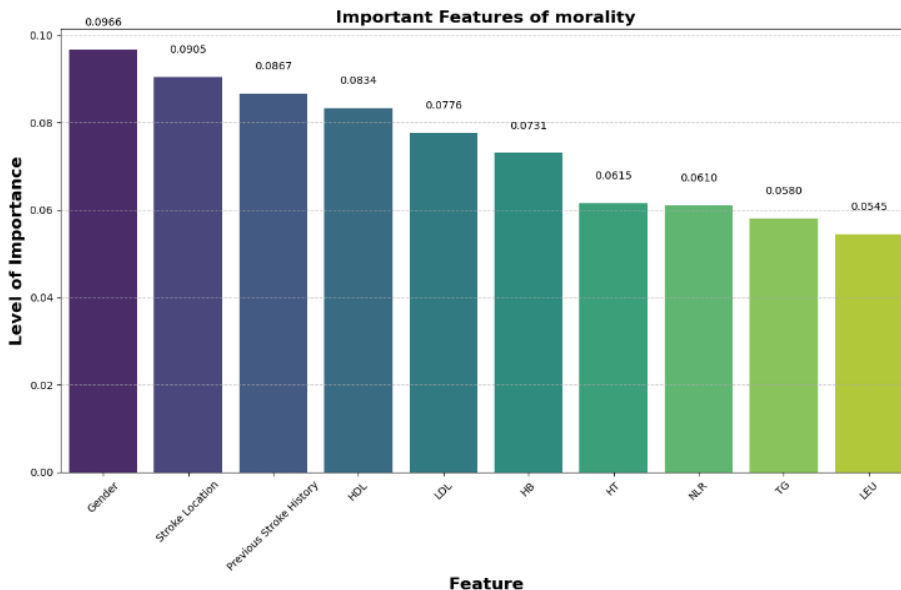


Figure 5. Mortality importance features

4.4. Website Application Forecast Death of Stroke Patient with XGBoost Algorithm

To support the prediction of the mortality rate of stroke patients, a website-based application has been developed using the XGBoost algorithm as the best choice. The purpose of this application is to assist the medical team in their efforts to treat stroke patients so as to avoid death. Here is a brief description of the application used.

The application was developed using the XGBoost algorithm to predict the mortality rate of stroke patients. A web-based interface built with flask as a framework enables the development and provision of an interface so that users can easily fill in patient data, including relevant demographic information and medical history. Next, the application will process the data sent by the user. The data will be channeled to a machine learning model that has been trained using the XGBoost algorithm. The model was trained with previous stroke patient data to understand patterns of death and identify risk factors. After receiving input from the user, the XGBoost model will analyze the attributes entered and predict the death of a stroke patient based on the data.

The results of this prediction will then be sent back to the user via the web interface in the form of an output or report, which will provide information about the likelihood of the patient's death.

Figure 6 shows a view of the stroke patient's data input form. Users will be asked to fill in 17 attributes that include information such as the patient's age, gender, history of cardiovascular disease, previous medical history, and other information. All the inputs required in the input page form are numerical data. Once all of these attributes have been filled in, the data will be taken by the application and processed through a machine learning model using the XGBoost algorithm. The prediction results, as shown in Figure 7, will be shown to users through the web application interface in the form of outputs. This output provides information about the probability of the patient's death.

The application provides two important pieces of information: a patient diagnosed or potentially at high risk of death, represented by the number 1, and a patient who is potentially recovered, represented by the number 0. Thus, the application allows users to predict the death of a stroke patient based on the medical data that has been entered. This could enhance better clinical decision-making as well as provide more effective treatment for stroke patients at Banyumas Hospital.

Figure 6. Input page

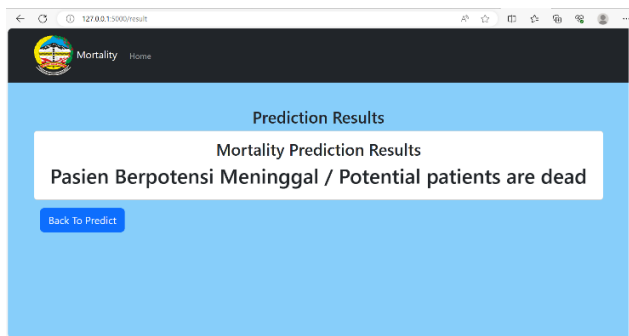


Figure 7. Result page

5. Conclusion

The study used machine learning (ML) algorithms to predict mortality in patients with infrared strokes. Of the five algorithms, the XGBoost model is the best for predicting death in stroke patients. In this study, after setting the hyperparameter, the Extreme Gradient Boosting (XGBoost) model has been able to forecast patient mortality with an accuracy of 86% and an AUC of 87. This achievement describes a significant improvement in the accuracy and predictive ability of the model after hyperparameter settings. In addition, the research has successfully built a website-based application for predicting stroke patients' deaths in Banyumas Hospital.

This research has some limitations. First, the research was conducted on a relatively small data set of 106 patients. Secondly, the study was carried out in one hospital in Indonesia, which may limit the ability to generalize the findings to other populations. Third, it did not take into account the impact of other factors, such as socio-economic status and access to health services, on mortality rates. Future research must overcome the limitations of this research by using larger datasets from more diverse populations. Future research should also consider the impact of other factors, such as socio-economic status, access to health services, and death.

References:

- [1]. Krishnamurthi, R. V., & Feigin, V. L. (2021). Global Burden of Stroke. *Stroke: Pathophysiology, Diagnosis, and Management*, 208–211. Doi: 10.1016/B978-0-323-69424-7.00014-4
- [2]. 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. Doi: 10.1016/S1474-4422(21)00252-0
- [3]. WHO. (2022). *World Stroke Day 2022*. Www.Who.Int. Retrieved from: <https://www.who.int/srilanka/news/detail/29-10-2022-world-stroke-day-2022> [accessed: 20 September 2023].
- [4]. Zhakhina, G., Zhalmagambetov, B., Gusmanov, A., Sakko, Y., Yerdessov, S., Matmusaeva, E., Imanova, A., Crape, B., Sarria-Santamera, A., & Gaipov, A. (2022). Incidence and mortality rates of strokes in Kazakhstan in 2014–2019. *Scientific Reports*, 12(1), 1–12. Doi: 10.1038/s41598-022-20302-8
- [5]. Centers for Disease Control (CDC). (2023). *About Stroke*. Cdc.Gov. Retrieved from: <https://www.cdc.gov/stroke/about.htm#:~:text=A stroke%2C sometimes called a term disability%2C or even death> [accessed: 21 September 2023].
- [6]. Setyopranoto, I., Bayuangga, H. F., Panggabean, A. S., Alifaningdyah, S., Lazuardi, L., Dewi, F. S. T., & Malueka, R. G. (2019). Prevalence of stroke and associated risk factors in sleman district of Yogyakarta Special Region, Indonesia. *Stroke Research and Treatment*, 2019. Doi: 10.1155/2019/2642458
- [7]. Kemenkes RI. (2018). Hasil Riset Kesehatan Dasar Tahun 2018. *Kementrian Kesehatan RI*, 53(9), 1689–1699.
- [8]. Dinas Kesehatan Provinsi Jawa Tengah. (2021). *Dinas Kesehatan Provinsi Jawa Tengah Tahun 2018 - 2023 Dinas Kesehatan Provinsi Jawa Tengah*. Renstra.
- [9]. A. Boehme, C. Esenwa, M. E. (2018). Stroke: Risk Factors and Prevention. *Journal of the Pakistan Medical Association*, 60(3), 412.
- [10]. Ranasinghe, V. S., Pathirage, M., & Gawarammana, I. B. (2023). Predictors of in-hospital mortality in stroke patients. *PLOS Global Public Health*, 3(2), e0001278. Doi: 10.1371/journal.pgph.0001278
- [11]. 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(1). Doi: 10.3389/fneur.2023.1096153
- [12]. 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. Doi: 10.1186/s12883-022-02722-1
- [13]. Rahim, A. M. A., Sunyoto, A., & Arief, M. R. (2022). Stroke Prediction Using Machine Learning Method with Extreme Gradient Boosting Algorithm. *MATRIK: Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, 21(3), 595–606. Doi: 10.30812/matrik.v21i3.1666
- [14]. Tazin, T., Alam, M. N., Dola, N. N., Bari, M. S., Bourouis, S., & Monirujjaman Khan, M. (2021). Stroke Disease Detection and Prediction Using Robust Learning Approaches. *Journal of Healthcare Engineering*, 2021. Doi: 10.1155/2021/7633381

- [15]. Sharma, D., Kumar, R., & Jain, A. (2022). Breast cancer prediction based on neural networks and extra tree classifier using feature ensemble learning. *Measurement: Sensors*, 24, 100560. Doi: 10.1016/j.measen.2022.100560
- [16]. 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*, 17(5). Doi: 10.1371/journal.pone.0262895
- [17]. 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. Doi: 10.1186/s12911-023-02140-4
- [18]. Sarker, I. H. (2021). Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective. *SN Computer Science*, 2(5), 1–22. Doi: 10.1007/s42979-021-00765-8
- [19]. Lovric, A., Granér, M., Bjornson, E., Arif, M., Benfeitas, R., Nyman, K., Ståhlman, M., Pentikäinen, M. O., Lundbom, J., Hakkarainen, A., Sirén, R., Nieminen, M. S., Lundbom, N., Lauerma, K., Taskinen, M. R., Mardinoglu, A., & Boren, J. (2018). Characterization of different fat depots in NAFLD using inflammation-associated proteome, lipidome and metabolome. *Scientific Reports*, 8(1), 1–14. Doi: 10.1038/s41598-018-31865-w
- [20]. Ahmed, M. E. (2023). hyOPTGB: An Efficient OPTUNA Hyperparameter Optimization Framework for Hepatitis C Virus (HCV) Disease Prediction in Egypt. *Research Square*. Doi: 10.21203/rs.3.rs-2768795/v1
- [21]. 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.
- [22]. Garg, S. (2022). *How to Deal with Categorical Data for Machine Learning*. Kdnuggets . Retrieved from: <https://www.kdnuggets.com/2021/05/deal-with-categorical-data-machine-learning.html> [accessed: 25 September 2023].
- [23]. Rácz, A., Bajusz, D., & Héberger, K. (2021). Effect of dataset size and train/test split ratios in qsar/qspr multiclass classification. *Molecules*, 26(4), 1–16. Doi: 10.3390/molecules26041111
- [24]. 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, 102119. Doi: 10.1016/j.mex.2023.102119
- [25]. 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. Doi: 10.1177/11795468221133611
- [26]. Dhingra, C. (2020). *A Visual Guide to Gradient Boosted Trees (XGBoost)*. Medium. Retrieved from <https://towardsdatascience.com/a-visual-guide-to-gradient-boosted-trees-8d9ed578b33> [accessed: 02 September 2023].
- [27]. Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: an interdisciplinary review. *Journal of Big Data*, 7(1). Doi: 10.1186/s40537-020-00369-8
- [28]. Ghazwani, M., & Begum, M. Y. (2023). Computational intelligence modeling of hyoscyne drug solubility and solvent density in supercritical processing: gradient boosting, extra trees, and random forest models. *Scientific Reports*, 13(1), 1–11. Doi: 10.1038/s41598-023-37232-8
- [29]. 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*, 192–203. Doi: 10.1142/9789813235533_0018
- [30]. Trung Pham Dinh a c, Cuong Pham-Quoc a c, Tran Ngoc Thinh a c, Binh Kieu Do Nguyen a b, P. C. K. b. (2023). A flexible and efficient FPGA-based random forest architecture for IoT applications. *Internet of Things*, 22. Doi: 10.1016/j.iot.2023.100813
- [31]. Thankachan, K. (2022). *What? When? How?: ExtraTrees Classifier*. Medium. Retrieved from: <https://towardsdatascience.com/what-when-how-extratrees-classifier-c939f905851c> [accessed: 26 September 2023].
- [32]. Chauhan, N. S. (2022). *Decision Tree Algorithm, Explained*. Kdnuggets. Retrieved from: <https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html> [accessed: 27 September 2023].
- [33]. Ogunleye, B. O. (2021). *Statistical learning approaches to sentiment analysis in the Nigerian banking context*. [Theses, Sheffield Hallam University (United Kingdom)].
- [34]. 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.
- [35]. Imantoko, I., Hermawan, A., & Avianto, D. (2021). Comparative analysis of support vector machine and k-nearest neighbors with a pyramidal histogram of the gradient for sign language detection. *Matrix : Jurnal Manajemen Teknologi Dan Informatika*, 11(2), 107–118. Doi: 10.31940/matrix.v11i2.2433
- [36]. Ivanov, I. G., Kumchev, Y., & Hooper, V. J. (2023). An Optimization Precise Model of Stroke Data to Improve Stroke Prediction. *Algorithms*, 16(9), 417. Doi: 10.3390/a16090417
- [37]. Teoh, D. (2018). Towards stroke prediction using electronic health records. *BMC Medical Informatics and Decision Making*, 18(1), 1–11. Doi: 10.1186/s12911-018-0702-y