

Article

Application of the XGBoost Algorithm for Stroke Disease Prediction

Chalista Nesya Prita Wardani ¹, Melanie Putri Salsavina ², Vira Arsy Dwi Pristyanu ^{3*} and Niyalatul Muna⁴.

¹ Health Information Management, Politeknik Negeri Jember, Jember, Indonesia;
g41240939@student.polije.ac.id

² Health Information Management, Politeknik Negeri Jember, Jember, Indonesia;
g41241031@student.polije.ac.id

³ Health Information Management, Politeknik Negeri Jember, Jember, Indonesia;
g41241031@student.polije.ac.id

⁴ Health Information Management, Politeknik Negeri Jember, Jember, Indonesia;
niyalatul@polije.ac.id

* Correspondence: g41241031@student.polije.ac.id

Citation: V. A. D. Pristyanu, C. N. P. Wardani, M. P. Salsavina, and N. Muna, "Application of the XGBoost Algorithm for Stroke Disease Prediction", *nexura*, vol. 1, no. 1, pp. 20–30, Jun. 2026.

Received: 27-05-2026
Accepted: 20-06-2026
Published: 30-06-2026



Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution-ShareAlike 4.0 International License (CC BY SA) license (<http://creativecommons.org/licenses/by-sa/4.0/>).

Abstract: Stroke is one of the non-communicable diseases with a relatively high rate of mortality and disability, making early detection very important to support fast and appropriate patient treatment. This study aims to apply the Extreme Gradient Boosting (XGBoost) algorithm to predict stroke disease based on patient health data. The dataset used was obtained from Kaggle, consisting of 150 patient records, which were divided into 100 training data and 50 testing data. The data processing was carried out using Google Colab, including preprocessing, model training, and performance evaluation stages. The results show that the model achieved an accuracy of 68%, an F1-score of 0.43, and a ROC-AUC of 0.717, indicating that the model has a fairly good classification ability in distinguishing stroke and non-stroke patients. In addition, age, average glucose level, and BMI were the most influential variables in stroke prediction. This study also produced a simple web-based application used to support early stroke detection by allowing input of patient health data and automatically displaying prediction results. Thus, the XGBoost algorithm has potential as a supporting method for early stroke detection using machine learning.

Keywords: machine learning; stroke; XGBoost.

1. Introduction

Stroke is one of the most common non-communicable diseases that causes death and disability worldwide. Stroke is a neurological disorder caused by cerebrovascular disturbances that develop rapidly, last for more than 24 hours, and can lead to death without any other cause besides vascular impairment [1]. The Ministry of Health of the Republic of Indonesia reported that stroke is one of the leading causes of death and disability in Indonesia, with a prevalence reaching 8.3 per 1,000 population in 2024, and it is also classified as a catastrophic disease with the highest healthcare expenditure [2]. Several risk factors that can increase the likelihood of stroke include hypertension, diabetes, heart disease, obesity, smoking habits, lack of physical activity, and unhealthy lifestyles [3].

The high prevalence of stroke indicates that early detection and risk prediction of this disease are very important. Fast and accurate diagnosis can help medical professionals provide earlier treatment, thereby minimizing the risk of complications and death. However, conventional stroke diagnosis still faces various challenges because it requires advanced clinical examinations, the use of medical imaging technology, relatively high costs, as well as the support of medical experts and adequate healthcare facilities [4].

With the advancement of technology, artificial intelligence-based approaches, particularly machine learning, are increasingly being used in the healthcare field. Machine learning is a technology capable of learning patterns from large datasets to generate predictions based on historical data [5]. In the healthcare context, this technology can assist medical professionals in analyzing patient data and predicting disease risk, including stroke, in a faster and more efficient way. One of the most widely used algorithms is Extreme Gradient Boosting (XGBoost). XGBoost is a decision tree-based algorithm designed to achieve optimal predictive performance when processing complex data. It is also equipped with a regularization mechanism that helps reduce the risk of overfitting, resulting in a more stable model that can produce better prediction outcomes [5].

Previous studies have shown that XGBoost performs well in disease prediction. Research on stroke prediction has reported that the XGBoost model can achieve an accuracy of 89.4% on testing data [4]. Another study on heart disease prediction showed that the XGBoost algorithm achieved a classification performance with an accuracy of 73% and an AUC of 0.795 [6]. However, research on stroke prediction still shows variations in model performance influenced by differences in datasets, preprocessing techniques, handling of imbalanced data, and feature selection. In addition, model evaluation in some previous studies has not been conducted comprehensively [7].

Based on these issues, this study aims to apply the XGBoost algorithm to predict stroke risk using patient health data. The dataset used was obtained from the Kaggle platform, consisting of several variables such as gender, age, hypertension, heart disease, work type, residence type, average glucose level, BMI, and smoking status. Data processing was carried out using Google Colab, including preprocessing, model training, and performance evaluation stages to assess the predictive capability of the resulting model.

2. Materials and Methods

This study applies a classification method using the Extreme Gradient Boosting (XGBoost) algorithm to predict the risk of stroke based on patient health data. The research stages are systematically designed to enable the process to be replicated or further developed in future studies. The research procedure includes data collection, data preprocessing, data splitting, model training, model testing, performance evaluation, and result visualization. All data processing and model implementation were carried out using the Google Colab platform as a machine learning development environment. The research workflow is shown in the following figure:

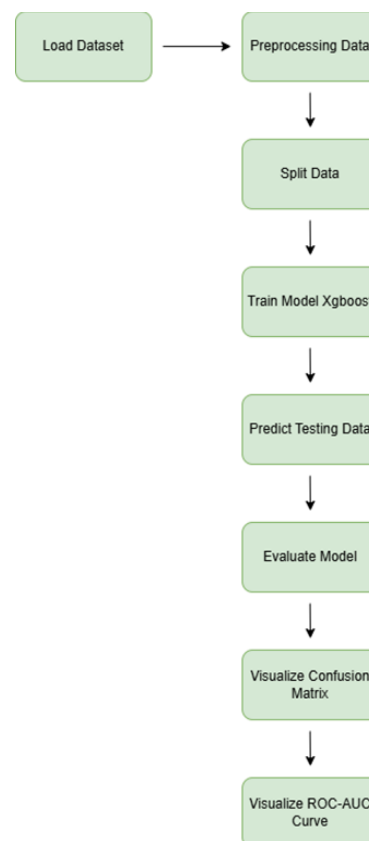


Figure 1. Research Workflow

2.1. Load Dataset

The dataset used in this study was obtained from Kaggle and consists of 150 patient records. The dataset is divided into 100 training data and 50 testing data. The variables used include gender, age, hypertension, heart disease, work type, residence type, average glucose level, BMI, and smoking status. This data is used as the basis for building a stroke prediction model.

Several variables such as age, hypertension, heart disease, blood glucose level, body mass index, and smoking status are known to be associated with stroke risk; therefore, these variables are commonly used in machine learning-based stroke prediction studies [8].

2.2. Preprocessing Data

Data preprocessing is an important stage in machine learning-based research because it improves data quality before model training. Processes such as handling missing values, transforming categorical data using encoding techniques, and data normalization can improve classification model performance. In addition, model evaluation using precision, recall, F1-score, and ROC-AUC is considered effective for measuring the model's ability to distinguish between stroke and non-stroke classes [9].

At this stage, data cleaning is performed to handle missing values, duplicate data, and inconsistent data. Next, categorical data is transformed into numerical data using encoding techniques so that it can be processed by the XGBoost algorithm. In addition, data normalization is applied to improve prediction model performance.

2.3. Split Data

The process of splitting data into training and testing sets is a common stage in machine learning research because it helps evaluate the model's predictive ability more accurately. The use of training and testing data also ensures that the model does not only

learn patterns from specific data but can also generalize to new data with different characteristics [9].

The dataset is divided into 100 training samples and 50 testing samples. The training data is used to build the XGBoost model, while the testing data is used to evaluate its performance. This division helps prevent overfitting so that the evaluation results can more objectively represent the model's performance.

2.4. Train Model XGBoost

In this stage, the model is trained using the XGBoost algorithm with the training data that has been preprocessed. The model learns the relationship between patient health variables and the likelihood of stroke using the XGBoost algorithm, which applies gradient boosting to gradually improve predictive performance [10]. XGBoost is considered capable of producing strong and stable predictive performance while minimizing errors, making the classification model more optimal in distinguishing stroke and non-stroke patients [10].

2.5. Predict on Testing Data

The trained model is then used to predict the testing data to determine stroke and non-stroke classifications. The prediction process is performed by inputting the testing data into the XGBoost model so that it can classify based on learned patterns. This stage evaluates the model's ability to recognize unseen data that was not used during training. The prediction results are used to assess the model's performance and its objectivity in classifying new data [7].

2.6. Evaluate Model

Model evaluation is conducted to determine the performance of the XGBoost algorithm in predicting stroke disease. Performance is measured using the confusion matrix, accuracy, precision, recall, and F1-score to provide a comprehensive analysis of classification results. Additionally, the ROC-AUC curve is used to evaluate the model's ability to distinguish between positive and negative classes.

Accuracy measures overall prediction correctness, precision and recall evaluate the model's ability to identify stroke and non-stroke classes, while F1-score measures the balance between precision and recall. ROC-AUC measures the model's ability to distinguish between classes based on predicted probabilities, where a higher AUC value indicates better classification performance [11].

2.7. Visualize Confusion Matrix

The confusion matrix is visualized to show the number of correct and incorrect predictions for each class, making it easier to analyze classification performance. This visualization helps evaluate the model's ability to distinguish between stroke and non-stroke cases based on prediction results.

2.8. Visualize ROC-AUC Curve

The ROC-AUC curve is visualized to assess the model's ability to distinguish between stroke and non-stroke classes based on prediction results. The ROC (Receiver Operating Characteristic) curve shows the relationship between the true positive rate and false positive rate in a classification model. Meanwhile, the AUC (Area Under Curve) measures the model's overall ability to distinguish between positive and negative classes. An AUC value closer to 1 indicates better classification performance.

3. Results

3.1. Data Processing Results

This study utilizes a stroke disease dataset obtained from Kaggle. The initial dataset contains patient health data with various variables related to stroke risk factors. After the data selection and preprocessing stages, the dataset used in this study consists of 150 patient records, which are then divided into 100 training data and 50 testing data.

The preprocessing stage was carried out to ensure the quality of the dataset before being used in the model training process. This process includes checking missing values, removing inconsistent data, and converting categorical data into numerical data using the Label Encoding method. The variables used in this study consist of:

1. Gender;
2. Age;
3. Hypertention.
4. Hearth Disease
5. Ever Married
6. Work Type
7. Residence Type
8. Average Glucose Level
9. BMI (Body Mass Index)
10. Smoking Status

After all data processing is completed, the dataset is used to build a stroke disease prediction classification model using the XGBoost algorithm. This stage is conducted so that the model can learn the relationship between patient health factors and the likelihood of stroke occurrence. Preprocessing steps such as data cleaning, encoding, and normalization play an important role because data quality significantly affects machine learning model performance. Preprocessing in stroke datasets can improve classification model stability and help the algorithm recognize data patterns more effectively [12].

3.2. Data Splitting Results

The dataset is divided into training and testing data using the train-test split method. The training data is used to train the XGBoost model to recognize dataset patterns, while the testing data is used to evaluate the model's ability to predict new data. The use of the train-test split method in machine learning research is considered effective for model evaluation because the test results are obtained from data that was not used during the training process [12]. In addition, data splitting helps reduce the risk of overfitting so that the model can achieve better predictive performance [12].

Table 1. Data Splitting Results

Data Type	Number of Data	Percentage
Training Data	100	66,7%
Testing Data	50	33,3%
Total	150	100%

Table 1 shows the division of the dataset consisting of 150 data points into two categories, namely training data and testing data. A total of 100 data points (66.7%) are allocated as training data, which are used to train the model so that it can learn the patterns and characteristics of the data. Meanwhile, 50 data points (33.3%) are used as testing data to evaluate the model's performance after the training process. This 2:1 data split ratio is a commonly used approach because it provides a balance between the model learning process and the evaluation of prediction accuracy.

3.3. XGBoost Model Training Results

The classification model in this study was built using the Extreme Gradient Boosting (XGBoost) algorithm. This algorithm was chosen because it has strong capabilities in handling classification problems and can improve prediction performance through a step-by-step gradient boosting method. The use of the XGBoost algorithm in stroke prediction research is considered effective because it can produce good classification performance on patient health data. XGBoost has high predictive ability with a good Area Under Curve (AUC) value in stroke prediction cases [13].

In this study, the XGBoost model was built using several parameters, namely $n_estimators = 100$, $learning_rate = 0.1$, $max_depth = 4$, and $random_state = 42$. These parameters were used to control the model learning process in order to achieve optimal prediction performance. After the training process was completed, the model was then used to make predictions on the testing data.

3.4. Model Evaluation Results

Model evaluation was conducted to determine the performance level of the XGBoost algorithm in predicting stroke disease. The model performance was measured using several evaluation metrics, namely accuracy, precision, recall, F1-score, and ROC-AUC.

Table 2. XGBoost Model Evaluation Results

Evaluation Method	Value
Accuracy	0,68
Precision Stroke	0,55
Precision Stroke	0,35
F1-Score Stroke	0,43
ROC-AUC Probability	0,717

Based on the evaluation results, the model achieved an accuracy of 0.68 or 68%. This result indicates that the model is able to classify stroke and non-stroke data fairly well.

The precision value of 0.55 shows that most of the data predicted as stroke actually belong to the true stroke category. Meanwhile, the recall value of 0.35 indicates that the model still has limitations in detecting all stroke cases in the dataset used.

The F1-score value of 0.43 reflects the balance between precision and recall in the classification model's performance. In addition, the ROC-AUC value of 0.717 indicates that the model has a fairly good ability to distinguish between stroke and non-stroke classes.

3.5. Confusion Matrix Results

The confusion matrix is used to evaluate the performance of the classification model by showing the number of correct and incorrect predictions. The results of the confusion matrix in this study are presented in Table 3. The following is the confusion matrix figure:

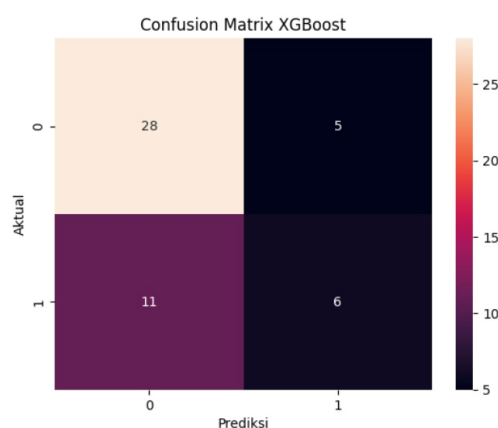


Figure 2. Confusion Matrix XGBoost

Table 3. Confusion Matrix

Actual / Prediction	Non-Stroke	Stroke
Non-Stroke	28	5
Stroke	11	6

Based on the confusion matrix results, it is known that 28 non-stroke data points were correctly predicted as non-stroke, while 6 stroke data points were correctly predicted as stroke. However, the model still produced prediction errors, where 5 non-stroke data points were predicted as stroke and 11 stroke data points were predicted as non-stroke. These results indicate that the model performs better in recognizing the non-stroke class compared to the stroke class. This condition may be influenced by an imbalance in the data distribution between stroke and non-stroke classes in the dataset used.

The results of this study are consistent with previous findings, which show that the XGBoost algorithm is able to produce good classification performance based on confusion matrix analysis and ROC-AUC evaluation. Therefore, this algorithm is considered effective in solving classification problems in the healthcare field [14].

3.6. ROC-AUC Results

The Receiver Operating Characteristic – Area Under Curve (ROC-AUC) test was conducted to evaluate the model's ability to distinguish between the positive class (stroke) and the negative class (non-stroke). In this study, the ROC-AUC calculation was performed based on the predicted probabilities (y_prob) generated by the model.

Table 4. ROC-AUC Results

Method	ROC-AUC Value
y_prob (probabilitas)	0,717

Based on the test results, the ROC-AUC value of 0.717 indicates that the model has a fairly good ability to distinguish between stroke and non-stroke patients. This value suggests that the model's classification performance is better than random classification in identifying both classes.

3.7. Feature Importance Results

The XGBoost model also produces feature importance values that indicate the contribution level of each variable to the stroke prediction results. Based on the test results, it was found that age, average glucose level, and Body Mass Index (BMI) are the

variables with the greatest influence on stroke prediction. The following figure presents the feature importance visualization:

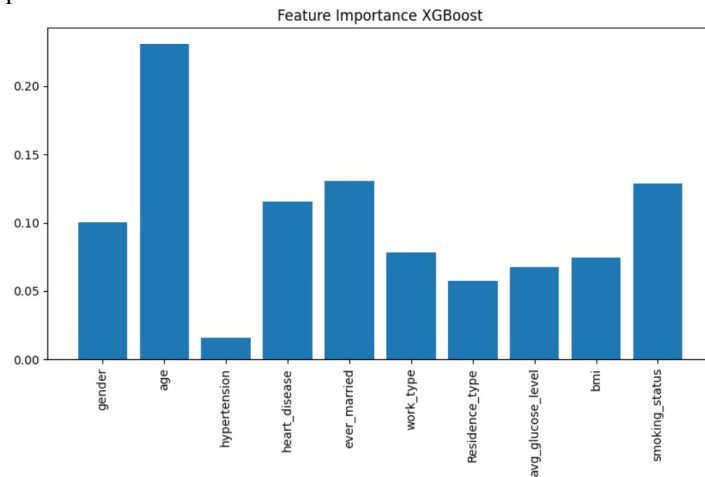


Figure 3. Feature Importance XGBoost

3.8. Stroke Risk Analysis Based on Research Variables

Based on the analysis of the stroke prediction dataset, observations were conducted on several variables, namely gender, age, hypertension, heart disease, marital status, work type, residence type, smoking status, Body Mass Index (BMI), and average glucose level. This analysis aims to identify groups with a higher tendency toward stroke occurrence based on the mean values and percentage distribution of each variable. The results of the analysis are presented in tabular form to facilitate the interpretation of factors associated with stroke risk.

Table 5. Stroke Risk Analysis Results

Variable	Code Description	Stroke Risk Result	Interpretation
Gender	0 = Female, 1 = Male	0 = 0.216; 1 = 0.183	Pada dataset ini, perempuan memiliki persentase kejadian stroke yang sedikit lebih tinggi dibandingkan laki-laki.
Hypertension	0 = No hypertension, 1 = Hypertension	0 = 0.221; 1 = 0.128	Patients without hypertension show a higher stroke percentage in this study dataset.
Heart Disease	0 = No heart disease, 1 = Has heart disease	1 = 0.212; 0 = 0.196	Patients with a history of heart disease tend to have a slightly higher stroke risk..
Ever Married	0 = Not married, 1 = Married	1 = 0.228; 0 = 0.115	Married patients show a higher stroke percentage compared to unmarried patients.
Work Type	0–3 = encoded categories	2 = 0,273; 1 = 0,193; 0 = 0,185; 3 = 0,063	Work category coded as 2 has the highest stroke percentage compared to other categories.

Residence Type	0 = Rural (pedesaan), 1 = Urban (perkotaan)	1 = 0,245; 0 = 0,147	Patients living in urban areas have a higher tendency toward stroke.
Smoking Status	0–3 = encoded categories	0–3 = encoded categories	Smoking status category coded as 0 shows the highest stroke percentage in the dataset.
Age	Patient age value	Stroke = 67.82 years; Non-stroke = 48.13 years	Stroke patients have a higher average age compared to non-stroke patients.
BMI	Body Mass Index	Stroke = 29.36; Non-stroke = 29.33	There is no significant difference in BMI between stroke and non-stroke patients.
Avg Glucose Level	Blood glucose level	Stroke = 121.96; Non-stroke = 127.91	The average glucose levels between stroke and non-stroke patients do not show a large difference.

3.9. Stroke Prediction Application

This early stroke detection application was developed as a simple web-based application as an implementation medium to test the results of data processing and the performance of the model that was previously built using Google Colab with the XGBoost algorithm. The following figure shows the final result of the stroke prediction application:

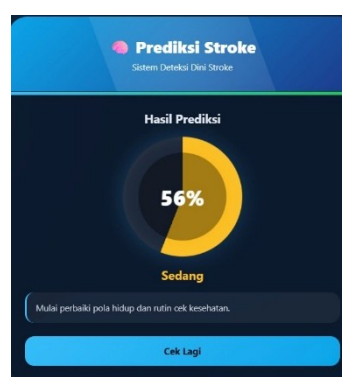


Figure 4. Stroke Prediction Application

4. Discussion

This study applies the Extreme Gradient Boosting (XGBoost) algorithm to predict stroke disease based on patient health data obtained from Kaggle. The dataset used consists of 150 patient records, which are divided into 100 training data and 50 testing data. The variables used include gender, age, hypertension, heart disease, ever married, work type, residence type, average glucose level, BMI, and smoking status.

Before the training process, the data underwent preprocessing stages including handling missing values, cleaning inconsistent data, and converting categorical data into numerical form using Label Encoding. In addition, data normalization was performed to improve model performance. These steps are important because XGBoost can only process numerical data.

The evaluation results show that the XGBoost model achieved an accuracy of 0.68 with a ROC-AUC of 0.717, indicating a fairly good classification performance. A precision value of 0.55, recall of 0.35, and F1-score of 0.43 indicate that the model still has limitations in detecting all stroke cases, although it performs reasonably well in identifying the positive class.

Based on the confusion matrix, the model correctly predicted 28 non-stroke cases and 6 stroke cases. However, 11 stroke cases were misclassified as non-stroke, and 5 non-stroke cases were predicted as stroke. This indicates that the model performs better in recognizing the non-stroke class compared to the stroke class, which may be influenced by data imbalance.

The variable analysis shows that age has the strongest influence on stroke occurrence, with stroke patients having a higher average age compared to non-stroke patients. The variables heart disease, ever married, and residence type also show differences in risk distribution. Meanwhile, BMI and blood glucose level do not show significant differences between the two classes.

5. Conclusions

Based on the data processing and analysis results using the XGBoost algorithm, this machine learning method can be used to assist in predicting stroke disease based on patient health data. The dataset used was obtained from Kaggle, while the data processing was carried out using the Python programming language through Google Colab. The evaluation results show that the model achieved an accuracy of 68% and a ROC-AUC of 0.717, indicating a fairly good classification performance. A precision value of 0.55 indicates that the stroke class predictions are reasonably accurate; however, a recall value of 0.35 shows that the model is still not optimal in detecting all stroke cases. Based on the confusion matrix, the model performs better in recognizing non-stroke data compared to stroke cases, which is likely influenced by data imbalance.

In addition, the variable analysis shows that age, heart disease, marital status, residence type, work type, and smoking status are associated with stroke risk. Age is the most dominant factor, as stroke patients have a higher average age compared to non-stroke patients. The feature importance results also show that age, average glucose level, and BMI contribute the most to the prediction process.

6. Patents

Author Contributions: Conceptualization was performed by C.N.P.W., M.P.S., and V.A.D.P.; methodology was performed by C.N.P.W., M.P.S., and V.A.D.P.; data processing using Google Colab was performed by C.N.P.W., M.P.S., and V.A.D.P.; validation was performed by C.N.P.W., M.P.S., and V.A.D.P.; formal analysis was performed by C.N.P.W., M.P.S., and V.A.D.P.; investigation was performed by C.N.P.W., M.P.S., and V.A.D.P.; writing—original draft preparation was performed by C.N.P.W., M.P.S., and V.A.D.P.; writing—review and editing was performed by C.N.P.W., M.P.S., and V.A.D.P.; supervision was performed by Niyalatul Muna, S.Kom., M.T.; project administration was performed by C.N.P.W., M.P.S., and V.A.D.P. All authors have read and agreed to the published version of the manuscript.

Acknowledgments: The author would like to express gratitude to all parties who have provided support in the preparation of the journal entitled “Application of the XGBoost Algorithm for Stroke Disease Prediction.” Appreciation is also extended to those who assisted in the data collection process, data processing, and the completion of this journal, enabling it to be properly prepared.

References

- [1] A. V. Yozi, A. Stiexs, and F. Anita, "Faktor-faktor yang Berhubungan dengan Kejadian Stroke," *Journal of Qualitative Health Research & Case Studies Reports*, vol. 5, no. 6, pp. 869–881, 2025
- [2] Kementerian Kesehatan Republik Indonesia, "Cegah Stroke dengan Aktivitas Fisik," Kemenkes RI, Oct. 25, 2024. [Online]. Available: Kemenkes RI - Cegah Stroke dengan Aktivitas Fisik. [Accessed: May 18, 2026].
- [3] World Health Organization, "Stroke, Cerebrovascular Accident," WHO, 2024. [Online]. Available: WHO EMRO Stroke Page. [Accessed: May 18, 2026].
- [4] A. D. Rahadjoe, F. I. A. Efendi, A. M. Mahardika, M. R. Firdaus, A. A. D. Maulana, and P. R. I. R. Lapon, "Klasifikasi Penyakit Stroke Menggunakan Algoritma XGBoost Classifier," *Journal of Innovative and Creativity (JOYC)*, vol. 6, no. 1, pp. 407–417, 2026.
- [5] F. Caroline and N. Rachmat, "Comparison of XGBoost and LightGBM Algorithms in Predicting Heart Disease," *Brilliance: Research of Artificial Intelligence*, vol. 5, no. 2, pp. 1232–1239, 2025.
- [6] F. I. Adani and H. Amalia, "Penerapan Metode Algoritma XGBoost untuk Prediksi Risiko Penyakit Jantung," *Jurnal INSAN (Journal of Information Systems Management Innovation)*, vol. 5, no. 2, pp. 117–135, 2025.
- [7] L. R. Sitompul, A. A. Nababan, M. L. Manihuruk, W. A. Ponsen, and Supriyandi, "Comparison of XGBoost, Random Forest, and Logistic Regression Algorithms in Stroke Disease Classification," *Sinkron: Jurnal dan Penelitian Teknik Informatika*, vol. 9, no. 2, pp. 957–967, 2025.
- [8] T. Fang and J. Deng, "An Interpretable Prediction Model for Stroke Based on XGBoost and SHAP," *Journal of Clinical and Nursing Research*, vol. 7, no. 3, pp. 96–106, 2023.
- [9] H. A. Nabila and E. W. Pamungkas, "Perbandingan Algoritma Machine Learning: SVM, Random Forest, dan XGBoost untuk Prediksi Stroke," *Jurnal Cendekia Muda*, vol. 10, no. 2, 2025.
- [9] P. D. Agnesya, A. D. Rahajoe, and A. L. Nurlaili, "Implementation of the XGBoost Algorithm for Acute Respiratory Infection Classification," *Journal of Science and Applied Engineering (JSAE)*, vol. 9, no. 1, pp. 38–46, 2026.
- [10] M. H. Z. Al Faroby, S. Amiroch, B. A. S. Aji, and A. Aritonang, "Classification of IGF1R Ligand Compounds for Identification of Herbal Extracts Using Extreme Gradient Boosting," *Jurnal Informatika*, vol. 15, no. 3, pp. 139–150, 2022.
- [11] S. Mishra and M. Haroon, "Design of FAHP Based Security Framework under Agile Software Development," *International Journal of Intelligent Systems and Applications in Engineering (IJISAE)*, vol. 12, no. 18s, pp. 292–302, 2024.
- [12] R. Oktafiani, A. Hermawan, and D. Avianto, "Pengaruh Komposisi *Split Data* terhadap Performa Klasifikasi Penyakit Kanker Payudara Menggunakan Algoritma *Machine Learning*," *Jurnal Sains dan Informatika*, vol. 9, no. 1, pp. 19–28, 2023.
- [13] C.-C. Chung, E. C.-Y. Su, J.-H. Chen, Y.-T. Chen, and C.-Y. Kuo, "XGBoost-Based Simple Three-Item Model Accurately Predicts Outcomes of Acute Ischemic Stroke," *Diagnostics*, vol. 13, p. 842, 2023.
- [14] A. Cahyana, E. R. Susanto, and Parjito, "Penerapan Algoritma XGBoost untuk Prediksi Diabetes: Analisis Confusion Matrix dan ROC Curve," *Fountain of Informatics Journal*, vol. 10, no. 1, pp. 40–50, 2025.