Social Science and Humanities Journal
Everant Journal
https://sshjournal.com/
Impact Factor: 2024: 6.576
2023: 5.731
ISSN: 2456-2653
Volume 09 Issue 02 January 2025
DOI: https://doi.org/10.18535/sshj.v9i01.1573

Stroke Risk Factor Prediction using Gradient Boost Method

Daniel William[1, a *] | Genrawan Hoendarto[2,b] | Jimmy Tjen[3,c]

[1,2,3]Widya Dharma Pontianak University, Pontianak, West Kalimantan, Indonesia
[a]daniel.william2116@gmail.com, [b]genrawan@widyadharma.ac.id, [c]jimmytjen@widyadharma.ac.id

Received 09-12-2024
Revised 10-12-2024
Accepted 30-12-2024
Published 02-01-2025

Abstract:

Stroke is a major global heart concern, often leading to significant disability or death. Early and accurate prediction of stroke risk can significantly improve patient outcomes. To address this issue, our study employs the Gradient Boosting method to enhance stroke prediction using dataset of 750 records. Key factors analyzed include gender, age, hypertension, heart disease, marital status, work type, residence type, average glucose levels, body mass index, and smoking status; the results identified age as the primary risk factor for stroke, followed by hypertension and smoking history. After preprocessing the data, our model achieves an average accuracy of 77,2% across ten runs, demonstrating strong predictive performance. A decision tree visualization highlights the most critical risk factors associated with stroke. This model aims to assist healthcare professionals in identifying high-risk individuals for early intervention. Additionally, we compare the Gradient Boosting model with other algorithms to determine the most effective predictive approach.

Keywords: Stroke, Gradient Boost, Prediction, Risk Factor.

Introduction:

Stroke continue to be one of the most prevalent causes of mortality and long-term disability worldwide. Each year, approximately 12,2 million new stroke cases are reported globally, with ischemic stroke accounting for the majority [1, 2, 3]. Given the significant public health burden, predicting stroke risk factors accurately is crucial for early intervention and prevention. Key clinical risk factors include age, hypertension, average glucose level, gender, diabetes, hyperlipidemia, even marital status, and lifestyle factors such as smoking and physical inactivity like work types [4, 5, 6, 7, 8]. In recent years, machine learning (ML) techniques have emerged as powerful tools for enhancing the predictive accuracy of stroke risk models. In this study, we aim to apply the Gradient Boosting Method (GBM) to predict stroke risk factors by utilizing a range of clinical and lifestyle parameters, including age, blood pressure, cholesterol levels, and others.

The GBM has gained attention due to its ability to iteratively improve model performance by minimizing prediction errors [9]. Unlike traditional linear models, GBM builds models in sequential manner, where each new model corrects the errors made by the previous one. This enables GBM to capture non-linear relationships among variables and improve accuracy over simpler models such as logistic regression [10]. Additionally, the method's

capability to handle missing data and noise makes it highly suitable for healthcare datasets, where variability in patient information is common. In this study, GBM is applied to a comprehensive dataset containing both clinical and lifestyle data, enhancing its predictive power in stroke risk analysis [11].

Previous studies have demonstrated the potential of machine learning methods in stroke prediction. For instance, recent research has shown that tree-based models, such as Random Forest and Gradient Boosting, outperform traditional statistical methods like Cox regression in predicting cardiovascular events [12]. Similarly, studies have highlighted the superior performance of ensemble methods, particularly GBM, over Support Vector Machines and Neural Networks in predicting stroke and other cardiovascular risk [13]. However, while these models offer high accuracy, they often require careful tuning of hyperparameters and can be computationally expensive when applied to large dataset [14].

One limitation of earlier studies is the lack of interpretability and the complexity of neural networks, which, although powerful, often act as "black-box" models [15]. In contrast, GBM offers a balance between complexity and interpretability, providing insights into the relative importance of different risk factors though feature important scores [16]. This makes GBM particularly advantageous for clinical settings, where both accuracy and interpretability are key. Moreover, GBM's ability to work effectively with smaller datasets without sacrificing performance makes it ideal for stroke risk factors prediction in healthcare systems where large-scale data collective may be a challenge.

This paper is divided into four main sections. The firsts section introduces the research objectives, the problem being addressed, and the proposed solution. The second section provides a detailed explanation of the gradient boost method utilized for predicting stroke risk based on the selected factors. The third section describes the specific algorithm that we developed to fit our study requirements. Finally, the fourth section presents the result and discusses potential avenues for future research.

Method:

In this study, we composed the gradient boost method to produce the stroke risk prediction. The gradient boost algorithm will adapt to previous mistakes from the previous trees make it best to show prediction for specific topics. As the aim for this study is to make the stroke diseases prediction with high accuracy. To provide more of our prediction advantages, we also did other prediction others algorithm method such as decision tree and random forest to compare gradient boost's prediction. The steps that we did in developing our gradient boost algorithm on stroke prediction are:

A. Data Collection:

This study used the dataset of "Stroke Prediction Dataset" that provides various information on patient health records that could be used as this machine learning variable. The dataset is cleaned and preprocessed to optimize the calculation of 10 variables. The dataset provides an important information for calculating and predicting the risk of getting stroke on people as the variables that we selected is the characteristic or the risk factors of stroke disease. The selected variables will be used on training and composing the better machine learning's algorithm model with the variables that also shown in Table 1.

B. Data Preprocessing:

Table 1. Stroke Prediction by Machine Learning's Variables

Name	Description
Gender	Divided by female = 1, male = 2
Age	Age in years
Hypertension	1 = have hypertension, 2 = does not have hypertension
Heart Disease	1 = have heart disease, 2 = does not have heart disease
Ever Married	Ever married or not, 0 = no, 1 = yes
Work Type	Person types of works, 1 = Private, 2 = Self-employed, 3 = Government job, 4 = Never worked. 5 = Children.
Residence Type	Residential area, 0 = rural, 1 = urban
Average Glucose Level	The average glucose level (in mg/dL)
BMI	Body mass index in units of kg/m^2
Smoking Status	1 = smokes, 2 = never smoked, 3 = formerly smoked 4 = unknown.

An important step before using the variables data into the machine learning, preparing the data so it can be used in the prediction algorithm. Before we put our data into our model, the variables that have string values were transformed into number from 1 to 5. After that, we divided the data from the dataset into two section, one section is to train the training model and the other one is to test the model accuracy and efficiency. We also randomize our data from the dataset to optimize our model since the dataset's stroke result is structured from 1 to 0, meaning that it will impact our model result if it not randomized.

C. Data Analysis:

This model used a dataset with 750 data rows to build the model's prediction and training. The model will have a continuous training and test subnet to analyze the prediction according to the built algorithm. The model was constructed with 80% of the data for model's training and the rest 20% to confirm the prediction accuracy. The dataset is also going to be used in the machine learning model to predict the stroke prediction by risk factors with gradient boost method. The prediction target is "stroke" where it described with the number 0 for not stroke and 1 for have stroke in each algorithm that we use to compare each other.

D. Model Training and Validation:

The final goal of the model's training we made is to predict Stroke by its risk factors using the gradient boost algorithm and compare it with the others algorithms we used. The gradient boost algorithm has a better model its ability to give a better performance by each new tree that developed during the model's training. This algorithm performs by adapting each tree that have a mistake

and will create a new tree based on the fixed previous tree. This method will produce the better prediction with each mistake it fixes, using the last tree that did not spot any mistakes as the final prediction.

E. Evaluation:

We also did a thorough evaluation of the model's training and the prediction result. Checking the gradient boost algorithm to the implied dataset to see the result consistency and prediction accuracy. This evaluation in this study includes the methods we were using and the algorithm prediction final score.

F. Performance Evaluation and Sensitivity Analysis:

The gradient boost algorithm is a robust ensemble machine learning technique employed for both regression and classification tasks. This method builds models sequentially, with each new model aiming to correct the errors of its predecessor. The strength of the gradient boosting method lies in its ability to improve prediction by combining several weak models. This approach helps it recognize intricate patterns in the data, making it particularly useful in medical approach [13]. In this study, we utilized Gradient Boosting to develop a predictive model for assessing stroke risk based on various risk factors.

As part of this method, we will provide the prediction sensitivity level along with the confusion matrix calculation. The confusion matrix is a valuable tool for evaluating classification problems, as it measures accuracy through four key values: True Positive (TP), True Negative (TN), False Positives (FP), and False

Table 2. Confusion matrix values

	Prediction No	Prediction Yes
Actual No	80	23
Actual Yes	9	38

Negative (FN). Below are the values for the confusion matrix:

These values are used to calculate the prediction sensitivity level using the following formulas:

1. True Positive Rate (TPR)
   The rate of positive samples correctly predicted as positive, to measure the proportion of actual positives that were correctly identified by the machine Eq. (1):
   $$TPR = \frac{TP}{TP + FN} \quad (1)$$
2. True Negative Rate (TNR)
   The rate of negative samples that correctly predicted as negative, to measure the proportion of actual negatives that were correctly identified by the machine Eq. (2):
   $$TNR = \frac{TN}{TN + FP} \quad (2)$$
3. False Positive Rate (FPR)
   The rate of negative samples that incorrectly predicted as positive, to measure the proportion of actual negatives that were incorrectly predicted as positives by the machine Eq. (3):
   $$FPR = \frac{FP}{TN + FP} \quad (3)$$
4. False Negative Rate (FNR)
   The rate of positive samples that incorrectly predicted as negative, to measure the actual positives that were incorrectly predicted as negatives Eq. (4):
   $$FNR = \frac{FN}{TP + FN} \quad (4)$$

The followings information presents the model's prediction sensitivity rate, derived from the values in the confusion matrix and calculated using the specified formulas:

Table 3. Confusion matrix's performance values

Performance Metrics	Values
True Positive Rate	0,8085 (80,85%)
True Negative Rate	0,7767 (77,67%)
False Positive Rate	0,2233 (22,33%)
False Negative Rate	0,1915 (19,15%)

Result

Table 4. Prediction accuracy results

Attempt	Result
1	77,3333
2	76,6667
3	76,0000
4	77,3333
5	78,6667
6	77,3333
7	76,6667
8	77,3333
9	76,6667
10	78,0000
Accuracy Total Average	77,2000

In this section, we present the result of our machine learning's model employing the gradient boosting algorithm to predict the likelihood of stroke based on various risk factors. Utilizing the sensitivity rates outlined in the previous the methods section, we derived our final predictions from these values. The accuracy of our model in predicting stroke, based on identified risk factors, is the primary focus of this study. To validate our model's performance, we conducted ten iterations to asses prediction accuracy. Below are the results from each run attempts:

Table 5. Comparing each algorithm results

Method	Result
Decision Tree	0,6733 (67,33%)
Random Forest	0,6267 (62,67%)
Gradient Boost (RUSBoost)	0,772 (77,2%)

This table presents the prediction accuracy derived from each run of our model. The results show consistent accuracy, ranging between 76% and 78,67%, indicating that the model performs reliably across iterations. The overall average accuracy across all attempts is 77,2%, underscoring the robustness of our gradient boosting algorithm for machine learning predictions. Additionally, we employed other algorithms for comparison with our gradient boost model, and the results are as follows:

The results of our study indicate a substantial disparity in performance among the algorithms assessed. Notably, the gradient boosting algorithm emerged as the most effective model, achieving an accuracy of 77,2%. This performance not only surpassed the other algorithms tested but also establishes gradient boosting as a robust and reliable method for predicting medical conditions such as stroke.

In contrast, the decision tree and random forest algorithms had lower accuracy, both falling below 70%. The decision tree struggled to capture complex patterns of the data, leading to its reduced performance. The random forest, while generally more stable, also faced challenges like overfitting which affected its accuracy.

We also included the tree graph created during the run with the highest accuracy to illustrate how the gradient boosting algorithm makes decisions and identifies important factors for stroke prediction.

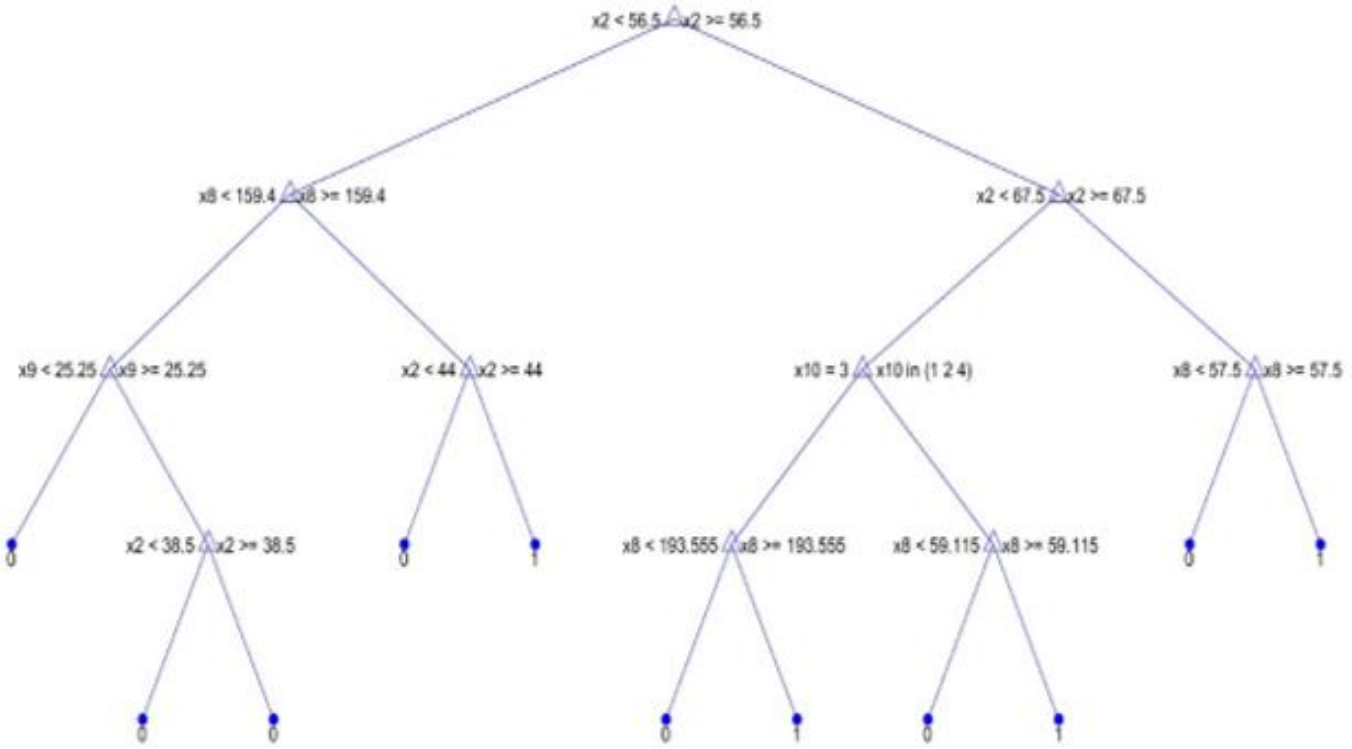

Fig 1. Gradient boost tree graph model tree

In this tree graph, we have simplified the names of each variable to "x" for better clarity. The detailed description of these variables are as follows: x2 represents age, x8 represents average glucose level, x9 represents body mass index, and x10 represents smoking status. The tree model shows that only these four variables are employed to achieve the prediction accuracy.

This model explains that a person can be suspected of experiencing a stroke if they meet the following criteria:

1. Age at or above 44 and below 56,5 years, and average glucose level at or above 159,4 mg/dL.
2. Age above 56,5 and below 67,5 years, smoking status is former smoker, and average glucose level at or above 193,555 mg/dL.
3. Age above 56,5 and below 67,5 years, smoking status includes smokes, never smoked, and unknown, and average glucose level at or above 59,115 mg/dL.
4. Age at or above 67,5 years, and average glucose level is at or above 57,5 mg/dL.

The result of this study indicates that individuals with certain criteria have a high likelihood of experiencing a stroke. These criteria align with the findings from several studies examining the same issue [3], [5], [6], [7]. The findings emphasize the dominant role of age as a primary risk factor for stroke. Age is one of the most influential factors, as individuals tend to experience a decline in overall health with increasing age, particularly in circulation, which becomes narrower and stiffer. This finding significantly highlights the increased likelihood of stroke, particularly among individuals over the age of 50, with the risk doubling with each subsequent decade of life [3, 5, 7].

In addition to age, average glucose levels also play a crucial role in stroke risk, as demonstrated in our model. According to the International Diabetes Federation, abnormal average glucose levels can be the cause of one in three cases worldwide. Interestingly, younger stroke patients are more likely to have diabetes compared to those without the condition [6]. Therefore, both age and average glucose levels are critical risk factors to consider in stroke prevention efforts.

In addition to age and average glucose levels, smoking status is a significant risk factor for stroke.

Smoker faces a higher risk, which increase with the number of cigarettes and duration of smoking. The harmful substances in tobacco contribute to vascular damage and atherosclerosis, elevating stroke susceptibility. Quitting smoking can substantially reduce the risk, underscoring the importance of cessation programs in stroke prevention efforts [17]. Addressing these factors collectively is crucial for effectively reducing stroke incident.

The results from our machine learning model clearly demonstrate its effectiveness in identifying key stroke risk factors from various factors in the dataset. The model consistently highlighted age and average glucose levels as the most influential contributors to stroke risk, followed by smoking history, aligning with established medical knowledge. The model's ability to accurately capture these critical factors reinforces its reliability as a predictive tool, making it an asset for healthcare professionals in stroke prevention and risk assessment.

Summary:

This study aimed to develop a predictive model for estimating stroke risk in individuals using machine learning techniques, specifically the Gradient Boost method. The demonstrated effective variables analysis, achieving an accuracy rate of 77,2% and outperforming other tested algorithms.

Key findings indicated that age is the primary risk factors for stroke, followed by hypertension and smoking history. The model also identified the most influential risk factors among the dataset, highlighting the importance of targeted prevention strategies for high-risk populations.

While the model's accuracy is in the 70% range, it provides a solid foundation for future research and refinement in stroke risk assessment. Future studies should enhance predictive capabilities by incorporating additional variables such as genetic factors, lifestyle choices, and comorbidities. Exploring alternative machine learning algorithms and conducting longitudinal studies could further improve accuracy and provide real-time predictions, ultimately contributing to more effective stroke prevention strategies.

References:

[1] S. K. Feske, "Ischemic Stroke," *Am J Med*, vol. 134, no. 12, pp. 1457–1464, Dec. 2021, doi: 10.1016/J.AMJMED.2021.07.027.

[2] World Health Organization, "World Stroke Day 2022." Accessed: Aug. 05, 2024. [Online]. Available: https://www.who.int/srilanka/news/detail/29-10-2022-world-stroke-day-2022

[3] GBD 2019 Stroke Collaborators, "Global, regional, and national burden of stroke and its risk factors, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019.," *Lancet Neurol*, vol. 20, no. 10, pp. 795–820, Oct. 2021, doi: 10.1016/S1474-4422(21)00252-0.

[4] M. J. O'Donnell *et al.*, "Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): a case-control study.," *Lancet*, vol. 388, no. 10046, pp. 761–75, Aug. 2016, doi: 10.1016/S0140-6736(16)30506-2.

[5] R. M. Carey, A. E. Moran, and P. K. Whelton, "Treatment of Hypertension: A Review," *JAMA*, vol. 328, no. 18, pp. 1849–1861, Nov. 2022, doi: 10.1001/jama.2022.19590.

[6] O. Mosenzon, A. Y. Y. Cheng, A. A. Rabinstein, and S. Sacco, "Diabetes and Stroke: What Are the Connections?," *jos*, vol. 25, no. 1, pp. 26–38, Jan. 2023, doi: 10.5853/jos.2022.02306.

[7] C. A. Simmons, N. Poupore, and T. I. Nathaniel, "Age Stratification and Stroke Severity in the Telestroke Network," *J Clin Med*, vol. 12, no. 4, Feb. 2023, doi: 10.3390/jcm12041519.

[8] D. S. Dhindsa, J. Khambhati, W. M. Schultz, A. S. Tahhan, and A. A. Quyyumi, "Marital status and outcomes in patients with cardiovascular disease,"

Trends Cardiovasc Med, vol. 30, no. 4, pp. 215–220, 2020, https://doi.org/10.1016/j.tcm.2019.05.012.

[9] I. D. Mienye and N. Jere, "A Survey of Decision Trees: Concepts, Algorithms, and Applications," *IEEE Access*, vol. 12, pp. 86716–86727, 2024, doi: 10.1109/ACCESS.2024.3416838.

[10] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Mar. 2016, doi: 10.1145/2939672.2939785.

[11] I. K. Nti, O. Nyarko-Boateng, J. Aning, G. K. Fosu, H. A. Pokuaa, and F. Kyeremeh, "Early Detection of Stroke for Ensuring Health and Well-Being Based on Categorical Gradient Boosting Machine," *Journal of ICT Research and Applications*, vol. 16, no. 3, pp. 313–332, 2022, doi: 10.5614/itbj.ict.res.appl.2022.16.3.8.

[12] J.-J. Beunza *et al.*, "Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease).," *J Biomed Inform*, vol. 97, p. 103257, Sep. 2019, doi: 10.1016/j.jbi.2019.103257.

[13] T. Vu *et al.*, "Machine Learning Approaches for Stroke Risk Prediction: Findings from the Suita Study.," *J Cardiovasc Dev Dis*, vol. 11, no. 7, Jul. 2024, doi: 10.3390/jcdd11070207.

[14] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, 2022, https://doi.org/10.1016/j.inffus.2021.11.011.

[15] C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead.," *Nat Mach Intell*, vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.

[16] S. M. Lundberg *et al.*, "Explainable AI for Trees: From Local Explanations to Global Understanding," May 2019, [Online]. Available: http://arxiv.org/abs/1905.04610

[17] F. Hasnah, Y. Lestari, and A. Abdiana, "The risk of smoking with stroke in Asia : meta-analysis," *Jurnal Profesi Medika : Jurnal Kedokteran dan Kesehatan*, vol. 14, no. 1, Apr. 2020, doi: 10.33533/jpm.v14i1.1597.