Earthline Journal of Mathematical Sciences
E-ISSN: 2581-8147
Volume 14, Number 1, 2024, Pages 133-150
https://doi.org/10.34198/ejms.14124.133150

Predicting Functional Outcome After Ischemic Stroke Using Logistic Regression and Machine Learning Models

Grace Chinwe Okoye[1,*] and Prof Edith Uzoma Umeh[2]

[1] Department of Statistics, Faculty of Physical Sciences, Nnamdi Azikiwe University, Awka, Nigeria; e-mail: chinwe.okoye@unizik.edu.ng

[2] Department of Statistics, Faculty of Physical Sciences, Nnamdi Azikiwe University, Awka, Nigeria; e-mail: eu.umeh@unizik.edu.ng

Abstract

This research employed binary logistic regression and machine learning techniques; Decision Tree, Random Forest, and Support Vector Machine (SVM), to predict functional outcomes following ischemic stroke. The main goal was to determine the most suitable model for the dataset through a comprehensive performance evaluation. Four models were examined for predicting post-ischemic stroke functional outcomes: Decision Tree, Random Forest, Logistic Regression, and SVM. The evaluation involved metrics such as Accuracy, Precision, F1-Score, and Recall. The Logistic Regression model achieved the highest accuracy at 90%, accurately predicting outcomes in 90% of cases. However, it had lower precision (50%), indicating an increased rate of false positive predictions. On the other hand, the SVM model displayed the highest precision (71.3%), implying fewer false positive predictions. It also attained the highest F1-Score (77.5%), indicating a strong balance between precision and Recall compared to the other models. Notably, the Logistic Regression model achieved perfect Recall (100%), correctly identifying all positive outcomes, while the Random Forest model showed significant recall performance (93.2%). Conversely, the Decision Tree model

Received: September 7, 2023; Revised & Accepted: November 14, 2023; Published: November 17, 2023
2020 Mathematics Subject Classification: 62-XX.
Keywords and phrases: Ischemic stroke, logistic regression, machine learning, support vector machine, model evaluation.

*Corresponding author

exhibited moderate accuracy (66.11%) but lower precision (66%), F1-Score (6.15%), and recall (3.2%), suggesting challenges with false positives and false negatives. Choosing the best model depends on analysis priorities. For accurate identification of positive outcomes, the Logistic Regression model's perfect recall is advantageous. For balanced performance, the SVM model's high F1-Score makes it a compelling option.

1 Introduction

The digital revolution has ushered in an era of information transformation, driven by advanced data analysis algorithms. Machine learning, a subset of this field, leverages mathematical and statistical models to make predictions and gain insights from data. In healthcare research, predictive modeling techniques, including logistic regression, decision trees, random forests, and support vector machines, have gained prominence for their ability to forecast outcomes such as stroke recovery. This study explores the application of machine learning to predict functional outcome after ischemic stroke, with a focus on enhancing predictive model accuracy. Ischemic stroke is a significant global health concern, and early intervention is critical for improving patient outcomes. By leveraging patient characteristics, risk factors, and imaging data, machine learning algorithms can aid in prognosis and treatment decisions. The aim of this research is to investigate the efficacy of various machine learning algorithms, including logistic regression, support vector machines, decision trees, and random forests, in predicting functional outcome after ischemic stroke. Accurate predictions in this context hold the potential to significantly impact patient care and outcomes.

2 Materials and Methods

2.1 Source of Data

The data in this study comprises real-life data obtained from Nnamdi Azikiwe University Teaching Hospital Nnewi, Anambra State, Nigeria, covering patients

diagnosed with stroke from 1st January 2014 to 31st July 2019. The dataset encompasses information from a total of 601 patients and includes various personal characteristics such as patient's age, gender, smoking, heart rate, chest pain, cholesterol, blood pressure, blood sugar, and stroke.

2.2 Methods

This research applies a combination of statistical and machine learning techniques to predict functional outcomes following ischemic stroke.

2.2.1 Sampling Techniques

The data analysis involves the application of sampling techniques to enhance model robustness. The techniques employed include boosting, bootstrapping, and bagging.

2.2.2 Model Evaluation Metrics

To assess the predictive performance of the models, several evaluation metrics are utilized, including: Precision, Recall (Sensitivity), F1-Score, and Confusion Matrix which provides information on true positives, true negatives, false positives, and false negatives.

2.2.3 Logistic Regression

Logistic regression is employed to model the relationship between independent variables and the binary dependent variable, which represents functional outcome after ischemic stroke. Logistic regression in this research is aimed at developing a model that can accurately classify or predict the likelihood of specific outcomes after ischemic stroke. The interpretability of logistic regression results can provide

valuable insights for clinical decision-making, enhance prognostic capabilities, and contribute to the field of stroke management and treatment strategies.

Logistic regression is a method for fitting a regression curve, $y = f(x)$, when y is a categorical variable. The typical use of this model is predicting y given a set of predictors x. The predictors can be continuous, categorical, or both. The sigmoid function maps the predicted values to probabilities. The probability is calculated as:

$$\hat{y}_j = \frac{e^{x \cdot b_j}}{1 + e^{x \cdot b_j}}.$$

In logistic regression, the logistic function is defined as:

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}.$$

To fit the model, the method of maximum likelihoods can be used. The logistic function will always produce a sensible prediction:

$$\frac{P(X)}{1 - P(X)} = e^{\beta_0 + \beta_1 x},$$

where $\frac{P(X)}{1-P(X)}$ is called the odds and can take any value between 0 and ∞. By taking the log of both sides:

$$\log\left(\frac{P(X)}{1 - P(X)}\right) = \beta_0 + \beta_1 x.$$

The left-hand side is called the log odds or logit, and the logistic regression model has a logit that is linear in X.

2.2.4 Estimating the Logistic Regression Coefficients

β_0 and β_1 in the logistic regression model are unknown. We seek estimates for β_0 and β_1 such that the predicted probability $\hat{p}(x_i)$ of stroke for each individual using the logistic regression formula corresponds as closely as possible to individuals' observed stroke status. In other words, we try to find $\hat{\beta}_0$ and $\hat{\beta}_1$ such that plugging

these estimates into the model for $p(X)$ yields a number close to one for all individuals who had a stroke and a number close to zero for individuals who did not. This intuition can be formalized using a mathematical equation called a likelihood function:

$$\mathcal{L}(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i)).$$

The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to maximize this likelihood function. Measure of accuracy of the coefficient estimates is by computing their standard errors:

$$Z\text{-statistic} \quad \text{associated} \quad \text{with} \quad \beta_1 = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}.$$

2.2.5 Making Predictions

Once the coefficients have been estimated, predictions can be made:

$$\hat{p}(x) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x}}.$$

2.3 Decision Tree

According to Anshul [1], this algorithmic model utilizes conditional control statements and is non-parametric, supervised learning useful for both classification and regression tasks. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules, and each leaf node represents the outcome. The algorithm works as follows:

1. Start with the root node.
2. Select the best attribute to split the data based on a criterion such as information gain or Gini index.

3. Create a new internal node with the selected attribute.
4. Split the data into subsets based on the attribute values.
5. For each subset, repeat steps 2-4 recursively until a stopping criterion is met (e.g., all instances belong to the same class or a maximum depth is reached).
6. Assign a class label to each leaf node based on the majority class of the instances in that node.
7. The decision tree is now ready for predictions.

2.4 Random Forest

According to Biau and Scornet [2], random forest is composed of multiple decision trees (classified regression trees).

The model randomly drafts N training subsets $M = \{m_1, m_2, \ldots, m_n\}$ based on bootstraps. The probability P of each sample not being drawn is calculated as:

$$P = \left(1 - \frac{1}{N}\right)^N.$$

The N decision trees $T = \{T_1, T_2, \ldots, T_n\}$ are developed on their corresponding training subsets. The Gini Index is calculated as:

$$G(M) = \sum_{k=1}^{k} P_k(1 - P_k) = 1 - \sum_{k=1}^{k} P_k^2$$

where:
- M is the independent training subset.
- The probability that the sample belongs to the k-th category is denoted as P_k.

Incorporating random forest in this research is aimed at leveraging the ensemble learning approach to improve prediction accuracy and robustness while gaining insights into the important features that contribute to stroke outcomes. This can enhance the understanding of stroke management and strategies, potentially leading to improved patient care and decision-making in the field of medicine.

2.5 Support Vector Machines

Vapnik [12] developed Support Vector Machines (SVM) to tackle the issue of binary classification by constructing the hyper-plane to separate the positive class and the negative class. The Gaussian Kernel function of SVM is given by:

$$K(x_i, x_j) = e^{-(x_i - x_j)^2}.$$

Incorporating support vector machines in this research is aimed at developing a predictive model that can effectively handle non-linear relationships, high-dimensional data, and outliers, while potentially providing interpretable insights through the analysis of support vectors. SVMs offer a robust and well-established approach to predicting stroke outcome, contributing to the field of ischemic stroke management and enhancing clinical decision-making.

3 Results

3.1 Logistic Regression Analysis

Table 1 evaluates the statistical significance and reliability of the estimated coefficients in the logistic regression model. A non-significant (i.e., large) p-value, greater than 0.05, suggests insufficient evidence to prove that the variables have a significant impact on the log-odds of the outcome.

- **Age:** The coefficient for age is 0.012882, indicating that for each one-unit increase in age, the log-odds of the outcome increase by 0.012882. However, this coefficient is not statistically significant at the 0.05 significance level.

- **Gender:** The coefficient of gender is 0.060505, suggesting that being in the gender category (male or female) is associated with an increase in the log-odds of the outcome by 0.060505. However, this coefficient is also not statistically significant (p-value = 0.857).

- **Smoking, Heart Rate, Chest Pain, Cholesterol, Blood Pressure, and Blood Sugar:** The coefficients for these values are not statistically significant (all p-values > 0.05). Therefore, there is insufficient evidence to suggest that these variables have a significant impact on the log-odds of the outcome.

Table 1: Assessing Coefficient Estimates, Standard Errors, Z-Values, and P-Values.

Variable	Estimate	Std. Error	Z-Value	P-Value
Intercept	-0.692868	1.171708	-0.591	0.554
Age	0.012882	0.011316	1.138	0.255
Gender	0.060505	0.334619	0.181	0.857
Smoking	0.692868	0.355815	0.011	0.991
Heart Rate	0.004296	0.006356	0.676	0.499
Chest Pain	0.011730	0.060741	0.193	0.847
Cholesterol	0.000109	0.002199	0.050	0.960
Blood Pressure	-0.000342	0.005059	-0.068	0.946
Blood Sugar	0.003742	0.003168	1.181	0.237

Source: Authors' Computation, 2023

3.2 Model Performance

The Confusion matrix below provides a summary of the predictions made by the model against the actual true values of the target variable.

Predicted Classes	
0	1
62	118

Based on the confusion matrix, the logistic regression model predicted 62 instances as class 0 (no stroke) and 118 instances as class 1 (stroke). These numbers represent the count of observations classified into each class based on the modelâs predictions.

3.3 Analysis of Decision Tree Model

Test for Correlation

Table 2 below focuses on the relationship between the predictor variables and the target variable i.e. it examined how changes in the predictor variables are associated with the target variable. The decision tree has a total of 420 observations (n = 420). The root node is the starting point of the tree. The first split occurs based on the variable "Age". If the age is less than 45.5, we move to node 2; otherwise, we move to node 3. Node 2 represents individuals with an age less than 45.5. It contains 119 observations. The deviance for this node is 27.462180, and the predicted value (Yval) is 0.6386555. The lower the deviance, the better the prediction accuracy. In this case, the deviance is relatively low, indicating a reasonably good fit. At node 2, there are two further splits based on blood sugar and heart rate. If blood sugar is less than 136.5, we move to node 4. This node contains 76 observations. The deviance is 18.671050, and the predicted outcome value is 0.5657895. Again, this deviance is relatively low, suggesting a decent fit. At node 4, there is another split based on "Heart Rate".

If "Heart Rate" is less than 55.5, we move to node 8. This node represents individuals with low heart rates (21 observations). The deviance is 4.666667, and the predicted outcome value is 0.3333333. At this point, gender becomes a splitting criterion. If Gender is greater than or equal to 0.5, we reach node 16. This terminal node represents individuals with a predicted outcome value of 0.1538462. Terminal nodes are the endpoints of the decision tree, where no further splits occur. If "Gender" is less than 0.5, we reach node 17. This terminal node represents individuals with a predicted outcome value of 0.6250000. If "Heart Rate" is greater than or equal to 55.5, we move to node 9. This node represents individuals with higher heart rates (55 observations). The deviance is 12.436360, and the predicted outcome value is 0.6545455. This node is a terminal node, and no further splits occur. If Blood Sugar is greater than or equal to 136.5, we move to node 5. This node represents individuals with higher blood sugar levels (43 observations). The deviance is 7.674419, and the predicted outcome value is 0.7674419. This node is also a terminal node. Node 3 represents individuals with an age greater than or equal to 45.5. It contains 301 observations. The deviance for this node is 59.661130, and the predicted outcome value is 0.7275748. This node is a terminal node.

Table 2: Relationship between Predictor Variables and Target Variable (Functional Outcome After Ischemic Stroke).

Node	Observations (n)	Deviance	Predicted Outcome Value (Y value)
1) Root	420	87.797620	0.7023810
2) Age < 45.5	119	27.462180	0.6386555
4) Blood Sugar < 136.5	76	18.671050	0.5657895
8) Heart Rate < 55.5	21	4.666667	0.3333333
16) Gender ≥ 0.5	13	1.692308	0.1538462 *
17) Gender < 0.5	8	1.875000	0.6250000 *
9) Heart Rate ≥ 55.5	55	12.436360	0.6545455 *
5) Blood Sugar ≥ 136.5	43	7.674419	0.7674419 *
3) Age ≥ 45.5	301	59.661130	0.7275748 *

* denotes terminal node Source: Authors' Computation, 2023

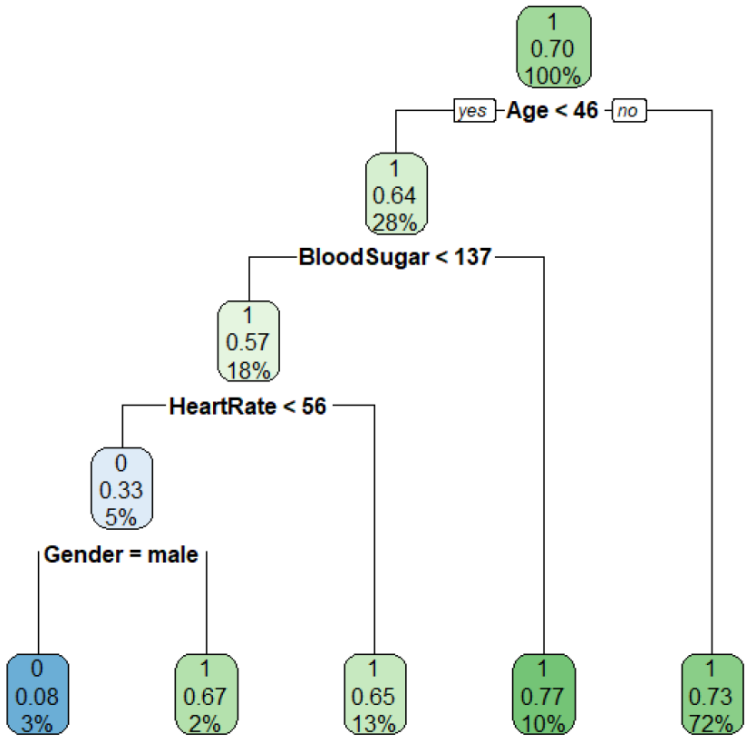

Figure 1: Decision Tree Plot

3.4 Model Performance

3.4.1 Confusion Matrix

The confusion matrix is a summary of the model's predictions compared to the actual true values of the target variable.

	Predicted Classes	
	2	1
	60	117

Based on the confusion matrix:

- The model correctly predicted a positive outcome (functional outcome after stroke = 1) in 117 instances (True Positives).
- The model correctly predicted a negative outcome (functional outcome after stroke = 0) in 2 instances (True Negatives).
- The model incorrectly predicted a positive outcome (functional outcome after stroke = 1) when the actual outcome was negative (functional outcome after stroke = 0) in 1 instance (False Positives).
- The model incorrectly predicted a negative outcome (functional outcome after stroke = 0) when the actual outcome was positive (functional outcome after stroke = 1) in 60 instances (False Negatives).

3.5 Analysis of Random Forest Regression Model

3.5.1 Confusion Matrix

The confusion matrix provides an overview of the model's predictions compared to the actual true values of the target variable.

		Predicted	
		0	1
Actual	0	1	0
	1	61	118

Based on the confusion matrix:

- The model correctly predicted a positive outcome (functional outcome after stroke = 1) in 118 instances (True Positives).
- The model correctly predicted a negative outcome (functional outcome after stroke = 0) in 1 instance (True Negatives).

- The model did not make any false positive predictions (False Positives), meaning it did not incorrectly classify a negative outcome as positive.
- The model incorrectly predicted a negative outcome (functional outcome after stroke = 0) when the actual outcome was positive (functional outcome after stroke = 1) in 61 instances (False Negatives).

3.6 Analysis of Support Vector Machine

3.6.1 Confusion Matrix

The confusion matrix provides valuable information about the performance of the classification model.

		Predictions	
		0	1
Actual	0	0	62
	1	0	118

Based on the predictions:

- The model performed well in predicting positive functional outcomes (functional outcome 1) after ischemic stroke, achieving 118 correct predictions.
- However, it failed to identify any negative functional outcomes (functional outcome 0) correctly, with 62 false positive predictions.

Figure 2 shows the scatter plot illustrating the relationship between age and heart rate for individuals who have had a stroke (red) and those who have not (blue). Additionally:

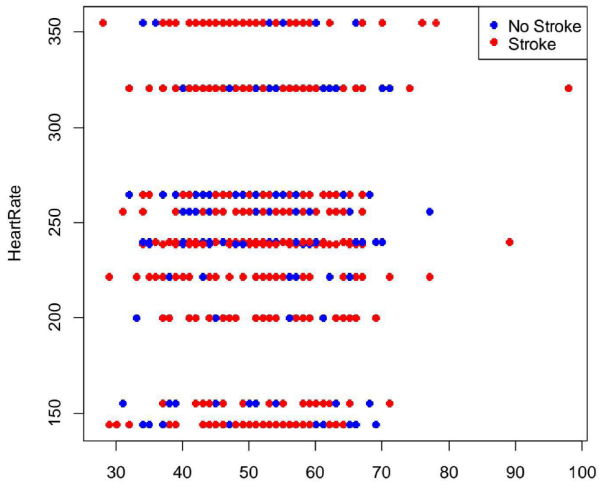

Figure 2: SVM Scatter Plot for Age and Heart Rate.

- The scatter plot displays the decision boundary of the SVM classifier, a curved line that separates the red and blue data points.

- The SVM classifier seeks to find the optimal decision boundary that maximizes the margin and minimizes the classification error.

- Most of the red data points are positioned above the decision boundary, while most of the blue data points are below it, indicating that the SVM classifier can correctly predict the functional outcome for most individuals based on their age and heart rate.

- However, there are some data points that fall on the wrong side of the decision boundary (outliers), suggesting that age and heart rate alone are not sufficient to predict the functional outcome for some individuals. Other factors, such as comorbidities, medications, lifestyle, etc., may also play a role.

3.7 Performance Metrics of Logistic Regression and Machine Learning Models

In this section, we compare the performance of three supervised machine learning models and the Logistic Regression model to determine which model fits the data better.

The results showed that the Logistic Regression model achieved the highest accuracy of 90%. This implies that it correctly predicted the functional outcome in 90% of cases. However, it exhibited the lowest precision at 50%, indicating a higher likelihood of false positive predictions. On the other hand, the SVM model demonstrated the highest precision at 71.3%, suggesting a lower rate of false positive predictions. Additionally, it achieved the highest F1-Score of 77.5%, indicating a better balance between precision and recall compared to other models. The Logistic Regression model exhibited perfect recall (sensitivity) of 100%, meaning it correctly identified all positive functional outcomes. The Random Forest model also achieved high recall at 93.2%. The decision tree model exhibited relatively lower performance across all metrics, indicating that it might not be as suitable for this specific prediction task compared to other models. Considering these findings, the choice of theâbestâ model depends on the specific priorities and requirements of the analysis. If correctly identifying positive functional outcomes is crucial, the Logistic Regression model with perfect recall may be preferred. However, if a balanced performance between precision and recall is important, the SVM model with the highest F1-Score could be considered. In summary, to predict functional outcome after 36 ischemic stroke, the most crucial metric to focus on is recall (sensitivity) which measures the ability of the model to correctly identify positive cases (functional outcome) out of all actual positive cases in the dataset. The Logistic Regression model with perfect recall is an excellent option. The SVM model achieved a relatively high recall of 84.92% and also demonstrated a good balance between precision and recall, as indicated by its higher F1-Score of 77.54%. This suggests that the SVM model is also a viable option for predicting functional outcomes after stroke, especially in balancing the model sensitivity and

precision.

Table 3: Performance metrics of the models.

Model	Accuracy	Precision	F1-Score	Recall (Sensitivity)
Decision Tree	0.6611	0.6667	0.0615	0.0323
Random Forest	0.6333	0.6548	0.7692	0.9322
Logistic Regression	0.9000	0.5000	0.6667	1.0000
SVM	0.6865	0.7133	0.7754	0.8492

4 Discussion

This research employed binary logistic regression and machine learning models, including Decision Tree, Random Forest, and Support Vector Machine (SVM), to predict functional outcomes after ischemic stroke. Additionally, the models were evaluated to identify the best-fit model for the dataset. In the analysis, four models were assessed for predicting functional outcomes after ischemic stroke: Decision Tree, Random Forest, Logistic Regression, and Support Vector Machine (SVM). Model performance was evaluated using metrics such as Accuracy, Precision, F1-Score, and Recall (Sensitivity). The results revealed that the Logistic Regression model achieved the highest accuracy of 90%, indicating correct predictions in 90% of cases. However, it had the lowest precision at 50%, implying a higher likelihood of false positive predictions. In contrast, the SVM model demonstrated the highest precision at 71.3%, indicating a lower rate of false positive predictions. Moreover, it achieved the highest F1-Score of 77.5%, signifying a better balance between precision and recall compared to other models. The Logistic Regression model exhibited perfect recall (sensitivity) of 100%, correctly identifying all positive functional outcomes. The Random Forest model also achieved high recall at 93.2%. The Decision Tree model showed moderate accuracy of 66.11%, but it had relatively low precision (66%), an F1-Score of

6.15%, and recall (sensitivity) of 3.2%, indicating a high number of false positives and false negatives. This makes it less reliable for classification tasks. The choice of the âbestâ model depends on specific priorities and requirements, with Logistic Regression being preferred for correctly identifying positive outcomes and SVM for a balanced performance between precision and recall.

5 Conclusion

In conclusion, based on the findings, the Support Vector Machine (SVM) model appears to be a promising approach for predicting functional outcomes after ischemic stroke. It exhibited high precision and a balanced F1-Score, suggesting accurate predictions with a lower rate of false positives.

References

[1] Saini, Anshul (2021). Decision Tree Algorithm - A Complete Guide. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/

[2] Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*, 197-227. https://doi.org/10.1007/s11749-016-0481-7

[3] Choi, Y.-A., Park, S., Jun, J.-A., Ho, C. M. B., Pyo, C.-S., Lee, H., & Yu, J. (2021). Machine-learning-based elderly stroke monitoring system using electroencephalography vital signals. *Applied Sciences*, *11*(4), 1761. https://doi.org/10.3390/app11041761

[4] Chang, W., Liu, Y., Xiao, Y., Yuan, X., Xu, X., Zhang, S., & Zhou, S. (2019). A machine-learning-based prediction method for hypertension outcomes based on medical data. *Diagnostics*, *9*(4), 178. https://doi.org/10.3390/diagnostics9040178

[5] Hanna, K. L., & Rowe, F. J. (2017). Health inequalities associated with post-stroke visual impairment in the United Kingdom and Ireland: A systematic review.

Neuro-Ophthalmology, *41*(3), 117-136. https://doi.org/10.1080/01658107.2017.1279640

[6] Wang, L. (2023). Logistic regression for stroke prediction: an evaluation of its accuracy and validity. *Highlights in Science, Engineering and Technology*, *39*, 1086-1092. https://doi.org/10.54097/hset.v39i.6712

[7] Mirzaikamrani, S. (2020). Predictive modeling and classification for Stroke using machine learning methods. http://www.diva-portal.se/smash/get/diva2:1430021/FULLTEXT01.pdf

[8] Restrepo, L. (2004). Handbook of stroke prevention in clinical practice. *Texas Heart Institute Journal*, *31*(4), 460.

[9] Volinsky, C. T., & Raftery, A. E. (2000). Bayesian information criterion for censored survival models. *Biometrics*, *56*(1), 256-262. https://doi.org/10.1111/j.0006-341x.2000.00256.x

[10] Wu, O., Cloonan, L., Mocking, S. J. T., Bouts, M. J. R. J., Copen, W. A., Cougo-Pinto, P. T., Fitzpatrick, K., Kanakis, A., Schaefer, P. W., Rosand, J., & others. (2015). Role of acute lesion topography in initial ischemic stroke severity and long-term functional outcomes. *Stroke*, *46*(9), 2438-2444. https://doi.org/10.1161/strokeaha.115.009643

[11] Yang, L., Liu, Q., Zhao, Q., Zhu, X., & Wang, L. (2020). Machine learning is a valid method for predicting prehospital delay after acute ischemic stroke. *Brain and Behavior*, *10*(10), e01794. https://doi.org/10.1002/brb3.1794

[12] Vapnik, V. (1999). The nature of statistical learning theory. *Springer Science & Business Media*.