Comparison of Machine Learning Models and NIHSS for Prognosis Prediction in Stroke Patients

Yeonwoo Noh
College of Medicine, Gachon Univ.
Seoul, Republic of Korea
nyw0207@gachon.ac.kr

Yun-Young Chang
School of Computing, Gachon Univ.
Seoul, Republic of Korea
jangyyoung88@gachon.ac.kr

Sewon Jeon
College of Medicine, Gachon Univ.
Seoul, Republic of Korea
sewon1001@gachon.ac.kr

Minwoo Lee
College of Medicine, Hallym University
Seoul, Republic of Korea
minwoo.lee.md@gmail.com

Wonjong Noh
College of Information Science, Hallym University
Chuncheon, Republic of Korea
wonjong.noh@hallym.ac.kr

***Abstract*—In this study, we evaluated ten machine learning models to predict stroke prognosis and compared them with the NIHSS scoring system using metrics like AUC, precision, recall, and F1-score. The top performers were DNN, XGBoost, and LightGBM, excelling in AUC, recall, and F1-score. Given the importance of recall in hospital settings, our findings suggest that machine learning models, particularly DNN, are more effective than NIHSS for predicting the prognosis of stroke patients.**

***Keywords*—Machine learning, NIHSS, prognosis, stroke.**

I. Introduction

Stroke is a neurological deficit caused by damage to blood vessels in the central nervous system [1]. Stroke is the second leading cause of death worldwide, and is a dangerous disease with a low long-term survival rate. According to a study in Sweden, 135 of 400 (33%) first-time stroke patients died within three years of their stroke [2]. Therefore, predicting the prognosis of stroke patients and appropriately treating them is critical.

To predict the prognosis of stroke patients, various scoring systems such as Cincinnati Prehospital Stroke Severity Scale [3], National Institutes of Health Stroke Scale (NIHSS) [4], and PLAN score (derived from preadmission comorbidities, level of consciousness, age, and neurologic deficit) [5] have been created. Among them, NIHSS considers the most variables, comprising 15 variables. However, predicting the prognosis of patients by relying solely on this score is difficult because NIHSS does not consider all variables associated with stroke prognosis.

However, machine learning (ML) models have the advantage of considering all patient characteristics because they can be trained on unlimited variables. Accordingly, ML is currently used to predict the prognosis of cancer, traumatic brain injury, and Alzheimer's disease [6]–[8].

In this study, we trained various ML models using more variables than conventional scoring systems. This study aimed to predict the prognosis of patients with stroke more effectively than the existing NIHSS scoring system using ML models.

II. Method

A. Data

In this study, we used data of stroke patients from Hallym University Sacred Heart Hospital and Hallym University Chuncheon Sacred Heart Hospital. The patient data comprised 56 variables, including age, sex, time of admission, blood tests, medical history, and medication history, for patients hospitalized between 2010 and 2023. Among these, 29 variables were used to train the model, including age, time from onset to arrival, body mass index, initial NIHSS score, pre-stroke modified Rankin Scale (mRS), medication history, specific treatment, blood test values, blood pressure, medical history, and smoking status [9]. Then, to address multicollinearity among variables, we converted low-density lipoprotein into categorical data ranging from 0 to 4 based on 100, 130, 160, and $190\,\mathrm{mg\,dL^{-1}}$. High-density lipoprotein was converted into categorical data ranging from 0 to 2 based on 40 and $60\,\mathrm{mg\,dL^{-1}}$. Systolic and diastolic blood pressures were converted to categorical data from 0 to 5 based on the Focused Update of the 2018 KSH Guideline from the Korean Society of Hypertension [10].

Before using the data, we removed data from patients with missing values and used those from 7190 stroke patients. (Fig.1) We used 80% (n = 5752) of the above data as training data and the remaining 20% (n = 1338) as test data. The data were divided such that the ratio of negative to positive patients in the training and test sets was the same.

For this study, we considered mRS scores of 0, 1, or 2 as negative data with a favorable prognosis and mRS scores of 3, 4, 5, or 6 as positive data with a poor prognosis. Accordingly, we entered 0 as the PoorOutcome variable for patients with mRS scores of 0, 1, or 2, and 1 for the remainder. In contrast, the NIHSS score was used as the control for all models. At this time, patients with a good prognosis were predicted if the score was five or less, and those with a poor prognosis if the score was six or more [11].

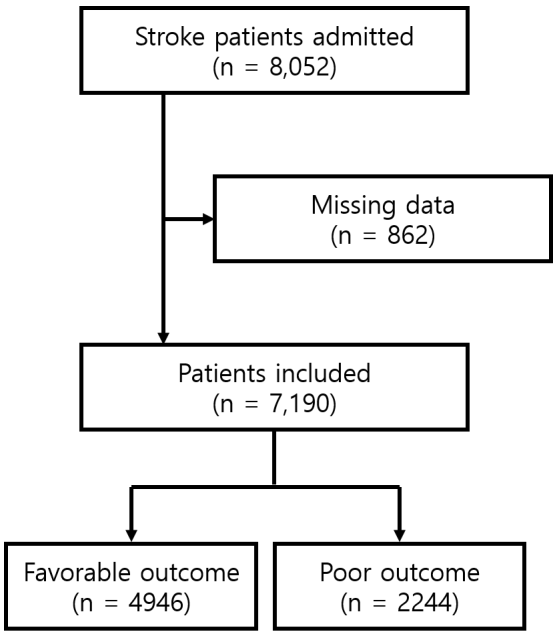

Fig. 1. Flow chart illustrating patient selection.

B. Machine Learning Algorithms

A total of ten ML algorithms were used: logistic regression, random forest classifier, support vector classifier (SVC), deep neural network (DNN), extreme gradient boosting (XGBoost), histogram-based gradient boosting, adaptive boost classifier (AdaBoost), light gradient boosting (LightGBM), TabNet, and ghost batch normalization TabNet (GBNTabNet).

- Logistic regression: An ML algorithm used for classification, where the linear combination of input features is transformed into probabilities using a sigmoid function.
- Random forest classifier [12]: An ensemble learning method that constructs multiple decision trees and combines their predictions. A total of 300 decision trees were used in this study.
- SVC [13]: A type of supervised learning algorithm that determines the optimal hyperplane to separate classes in the feature space, aiming to maximize the margin between different classes.
- DNN: An artificial neural network with multiple hidden layers between input and output layers. Two hidden layers with 64 neural network units were used. The weights in the first layer are initialized using a Glorot uniform initializer. Both hidden layers include batch normalization, ReLU activation function, and dropout at rates of 0.3 and 0.5. The output layer comprises a single neuron with a sigmoid activation function. The model is compiled using the RMSprop optimizer with a learning rate of 0.00001 and momentum of 0.96. The binary cross-entropy loss function is used, along with the area under the receiver operating characteristic (ROC) curve (AUC) metric, to evaluate the performance.
- XGBoost [14]: An advanced ensemble learning method that constructs multiple decision trees sequentially with optimization and regularization. For XGBoost, 300 decision trees with a maximum depth of 3 and a learning rate of 0.05 were used.
- Histogram-based gradient boosting [15]: A variant of gradient boosting that uses histogram-based algorithms to accelerate the training process by discretizing continuous feature values into bins. For histogram-based gradient boosting, 300 decision trees with a maximum depth of 3 and a learning rate of 0.05 were used.
- AdaBoost [16]: An ensemble learning technique that sequentially combines multiple weak classifiers, adjusting the weights of misclassified samples. For AdaBoost, 300 decision trees with a maximum depth of three and a learning rate of 0.05 were used.
- LightGBM [17]: A gradient boosting framework that uses gradient-based one-sided sampling and exclusive feature bundling. For LightGBM, 300 decision trees with a maximum depth of 3 and a learning rate of 0.05 were used.
- TabNet [18]: A deep learning model for tabular learning that uses a sequential attention mechanism, instance-wise feature selection, and visualization of selection masks. For TabNet, a batch size of 64 and a virtual batch size of 32 were used.
- GBNTabNet [19]: A TabNet with ghost batch normalization that normalizes data based on smaller mini-batches within each larger batch. For GBNTabNet, a batch size of 64, a virtual batch size of 32, and a ghost batch size of 16 were used.

III. Results

In this study, we compared the performance of ML algorithms with the NIHSS score in terms of AUC, precision, recall, and F1-score.

- AUC: Area under the ROC curve. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The AUC provides a single metric that summarizes the ability of the model to distinguish between two classes. A model with an AUC close to 1.0 is considered to have good discriminatory power.
- Precision: A metric that measures the proportion of correct positive predictions made by the model. The formula for precision is as follows:

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

- Recall: A metric that measures the proportion of actual positive cases correctly identified by the model. The formula for recall is as follows:

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

We compared recall because of its medical importance: predicting a positive patient with a poor prognosis as having a good prognosis is life-threatening. The higher

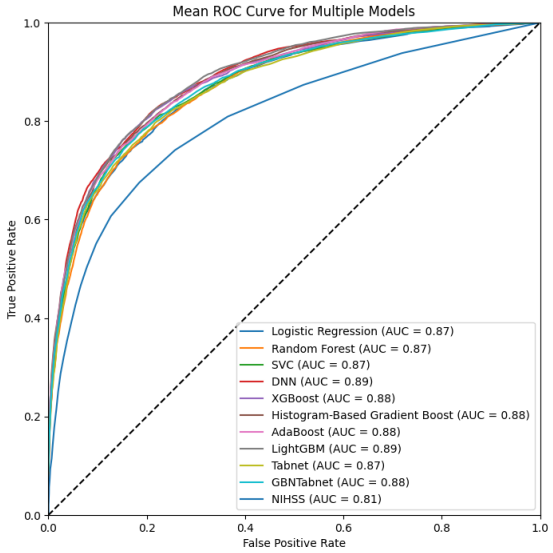

Fig. 2. Mean receiver operating characteristic(ROC) curve for NIHSS and multiple models. ROC curves for all models are above NIHSS. DNN and LightGBM have the largest area under the curve (AUC), but there is not much difference between the 10 models.

the recall of the ML model, the greater the number of patients with a poor prognosis that could be predicted. Therefore, recall is an important indicator in determining the models that can be used in hospital settings.

- F1-score: A metric used to evaluate the performance of a binary classification model by balancing precision and recall. This is the harmonic mean of precision and recall. The formula for the F1 score is:

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

A. Comparison of the Models

First, as summarized in TABLE I, the AUC of the ML models was higher than that of the NIHSS. Comparing the AUC of each model, the top three performing models were LightGBM, DNN, and XGBoost, which outperformed NIHSS by up to 9.901%. However, the AUC values for LightGBM, DNN, and XGBoost were only 2.069% higher than logistic regression, which was the lowest. In other words, no significant difference was observed in the AUC of the models. The ROC curve in Fig.2 shows that the ML models performed better than the NIHSS; however, no significant difference existed between them.

Second, the precision of all models except the DNN was higher than that of the NIHSS. Comparing the precision of each model, the top three performing models were random forest classifier, SVC, and GBNTabNet, which outperformed NIHSS by up to 26.608%.

Third, the recall of the models excluding logistic regression, random forest classifier, and SVC was higher than that of the NIHSS. Comparing the recall of each model, the top three performing models were DNN, XGBoost, and LightGBM, which outperformed the NIHSS by up to 28.007%.

Finally, the F1-scores of all models, except the random forest classifier, were higher than that of the NIHSS. Comparing the F1 scores of each model, the top three performing models were DNN, XGBoost, and LightGBM, which outperformed NIHSS by up to 11.975%.

IV. Discussion

This study showed that using medical data, ML models can predict stroke patient prognoses. The AUC of all models were higher than that of the NIHSS. Except for the logistic regression and random forest classifiers, the remaining eight models outperformed the NIHSS in terms of recall and F1- score. This was because, unlike the NIHSS score, they considered more variables associated with stroke prognosis without limiting the number of variables. Additionally, the top three models, DNN, LightGBM, and XGBoost, exhibited significant dominance in AUC, recall, and F1-score compared with the NIHSS score.

Logistic regression, random forest classifier, and GBNTabNet, which had the lowest recall and F1-score, had the highest precision. This implies that the three models were trained to be conservative in their positive predictions to avoid false positives. However, several patients with poor prognoses have not been identified using these models. This is an undesirable characteristic of the model in hospital settings.

This study concluded that DNN is the most appropriate model for predicting stroke prognosis among other ML models. DNN had the highest AUC, recall, and F1 scores compared to the other models. In addition, DNN exhibited a significant dominance in recall, which is an important metric for determining whether a model is practical for patients in practice. Therefore, if the DNN model in this study is used as a stroke prognosis prediction model, 77.7% of the patients with poor prognosis can be diagnosed in advance based on clinical data. Diagnosing patients with poor prognosis in advance, which is insufficient for the NIHSS score, will contribute toward improving the survival rate of these patients. The potential of the DNN model to achieve these outcomes should instill confidence in its effectiveness and the potential to significantly improve patient outcomes.

However, the models used in this study have several limitations. First, they can only classify patient prognoses into positive or negative categories. In addition, patients who died because of poor prognosis should have been used as essential data for training the models. However, these were removed during data preprocessing because of missing clinical data. A clear need for future research in this field exists. If multi-classification models are trained with the inclusion of clinical data from deceased patients, they could potentially surpass the models presented in this paper. These models can provide more detailed prognostic information, further enhancing the ability to predict and manage stroke outcomes.

Acknowledgment

This research was supported by the [Bio&Medical Technology Development Program] of the National Research Founda-

TABLE I
COMPARISON OF THE NIHSS AND MACHINE LEARNING MODELS

Machine learning model	AUC	Precision	Recall	F1-score
NIHSS	0.808	0.684	0.607	0.643
Logistic Regression	0.870 [0.861-0.879]	0.791 [0.761-0.822]	0.597 [0.581-0.612]	0.680 [0.667-0.693]
Random Forest Classifier	0.870 [0.864-0.877]	0.866 [0.850-0.906]	0.413 [0.389-0.437]	0.559 [0.536-0.582]
SVC	0.874 [0.868-0.881]	0.798 [0.764-0.826]	0.594 [0.575-0.613]	0.681 [0.666-0.696]
DNN	0.887 [0.887-0.888]	0.670 [0.661-0.680]	0.777 [0.772-0.783]	0.720 [0.716-0.723]
XGBoost	0.884 [0.879-0.890]	0.782 [0.756-0.807]	0.657 [0.642-0.693]	0.714 [0.703-0.725]
histogram-based gradient boosting	0.884 [0.877-0.891]	0.780 [0.744-0.802]	0.646 [0.636-0.656]	0.706 [0.696-0.716]
AdaBoost	0.883 [0.873-0.893]	0.787 [0.753-0.812]	0.643 [0.620-0.666]	0.707 [0.691-0.724]
LightGBM	0.888 [0.883-0.894]	0.781 [0.760-0.798]	0.653 [0.638-0.668]	0.711 [0.700-0.721]
TabNet	0.871 [0.864-0.879]	0.786 [0.771-0.818]	0.631 [0.611-0.650]	0.699 [0.686-0.713]
GBNTabNet	0.875 [0.866-0.885]	0.797 [0.776-0.824]	0.621 [0.602-0.640]	0.698 [0.683-0.713]

tion (NRF) funded by the Korean government (MSIT) (No. RS-2023-00223501).

REFERENCES

[1] S. J. Murphy and D. J. Werring, "Stroke: causes and clinical features," *Medicine*, vol. 48, no. 9, pp. 561–566, 2020.

[2] J. Aked, H. Delavaran, and A. G. Lindgren, "Survival, causes of death and recurrence up to 3 years after stroke: A population-based study," *European journal of neurology*, vol. 28, no. 12, pp. 4060–4068, 2021.

[3] R. U. Kothari, A. Pancioli, T. Liu, T. Brott, and J. Broderick, "Cincinnati prehospital stroke scale: reproducibility and validity," *Annals of emergency medicine*, vol. 33, no. 4, pp. 373–378, 1999.

[4] L. K. Kwah and J. Diong, "National institutes of health stroke scale (nihss)," *Journal of physiotherapy*, 2014.

[5] M. J. O'Donnell, J. Fang, C. D'Uva, G. Saposnik, L. Gould, E. McGrath, M. K. Kapral, I. of the Registry of the Canadian Stroke Network *et al.*, "The plan score: a bedside prediction rule for death and severe disability following acute ischemic stroke," *Archives of internal medicine*, vol. 172, no. 20, pp. 1548–1556, 2012.

[6] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and structural biotechnology journal*, vol. 13, pp. 8–17, 2015.

[7] H. Khalili, M. Rismani, M. A. Nematollahi, M. S. Masoudi, A. Asadollahi, R. Taheri, H. Pourmontaseri, A. Valibeygi, M. Roshanzamir, R. Alizadehsani *et al.*, "Prognosis prediction in traumatic brain injury patients using machine learning algorithms," *Scientific reports*, vol. 13, no. 1, p. 960, 2023.

[8] B. Y. Kasula, "A machine learning approach for differential diagnosis and prognostic prediction in alzheimer's disease," *International Journal of Sustainable Development in Computing Science*, vol. 5, no. 4, pp. 1–8, 2023.

[9] M. Won-Young Jung, M. Gun-Han Lim, M. Hyung-Gyun Oh, M. Seung-Heon Lee, and J.-G. N. M.D., "The longterm prognostic factors after acute cerebral infartion," *Journal of the Korean Neurological Association*, vol. 13, no. 4, pp. 806–814, 1995.

[10] H. L. Kim, E. M. Lee, S. Y. Ahn, and et al., "The 2022 focused update of the 2018 korean hypertension society guidelines for the management of hypertension," *Clinical Hypertension*, vol. 29, p. 11, 2023.

[11] D. Schlegel, S. J. Kolb, J. M. Luciano, J. M. Tovar, B. L. Cucchiara, D. S. Liebeskind, and S. E. Kasner, "Utility of the nih stroke scale as a predictor of hospital disposition," *Stroke*, vol. 34, no. 1, pp. 134–137, 2003.

[12] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[13] S. Suthaharan and S. Suthaharan, "Support vector machine," *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, pp. 207–235, 2016.

[14] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[15] A. Guryanov, "Histogram-based algorithm for building gradient boosting ensembles of piecewise linear decision trees," in *Analysis of Images, Social Networks and Texts: 8th International Conference, AIST 2019, Kazan, Russia, July 17–19, 2019, Revised Selected Papers 8*. Springer, 2019, pp. 39–50.

[16] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[17] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.

[18] S. Ö. Arik and T. Pfister, "Tabnet: Attentive interpretable tabular learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 8, 2021, pp. 6679–6687.

[19] E. Hoffer, I. Hubara, and D. Soudry, "Train longer, generalize better: closing the generalization gap in large batch training of neural networks," *Advances in neural information processing systems*, vol. 30, 2017.