การทำนายโรคหลอดเลือดสมองโดยใช้การเรียนรู้ของเครื่อง

สากล พัชรปัญญวัฒน์ [1,*] และ จันตรี ผลประเสริฐ[1]

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาการทำนายความเสี่ยงต่อการเกิดโรคหลอดเลือดสมองในวัยผู้ใหญ่โดยใช้การเรียนรู้ของเครื่อง การศึกษานี้เราต้องการตรวจสอบประสิทธิภาพของโมเดลการเรียนรู้ของเครื่องในสามโมเดลซึ่งประกอบไปด้วยโมเดล Logistic Regression (LR), Random Forest (RF) และ Support vector machine (SVM) เราใช้ชุดข้อมูลการดูแลสุขภาพที่มีอยู่ในชุดข้อมูลของ Kaggle[9] ซึ่งมีข้อมูลผู้ป่วย 5,110 คนและเราเลือกผู้ป่วยเหลือเพียง 4,254 คน ที่เป็นผู้ป่วยวัยผู้ใหญ่ที่มีอายุ 18 ปีขึ้นไป เมทริกซ์ความสับสนใช้สำหรับการสรุปประสิทธิภาพของโมเดลการจำแนกประเภท และ AUC (Area Under The Curve) แสดงถึงระดับของความสามารถในการแยกตัวออกจากกันได้ ซึ่งค่า AUC ที่สูงกว่า แสดงว่าโมเดลมีความสามารถในการแยกแยะระหว่างผู้ป่วยโรคหลอดเลือดสมองและผู้ป่วยที่ไม่มีโรคหลอดเลือดสมองได้ดีกว่า จากการทดลองครั้งนี้ RF เป็นโมเดลที่มีประสิทธิภาพที่ดีที่สุดด้วย accuracy 0.94, precision 0.93, recall 0.95, f1-score 0.94 และค่า AUC เท่ากับ 0.94 และสามอันดับสูงสุดของความสำคัญของฟีเจอร์ของ Random Forest ที่มีลำดับตามความสำคัญจากมากไปน้อยคือ age มีค่า 0.38, avg_glucose_level มีค่า 0.20 และ bmi มีค่าเท่ากับ 0.05 ตามลำดับ

คำสำคัญ : โรคหลอดเลือดสมอง การเรียนรู้ของเครื่อง ประสิทธิภาพของโมเดล

[1] Department of Computer Science, Faculty of Science, Srinakharinwirot University
Bangkok, Thailand, 10110

[*] Corresponding author email: Sakol.pat@g.swu.ac.th

STROKE PREDICTION USING MACHINE LEARNING

Sakol Patcharapanyawat_[1,*] and Chantri Polprasert [1]

The purpose of this research is to study the prediction of stroke risk in adults using machine learning.
This research investigates the performance of three machine learning models that include Logistic Regression (LR), Random Forest (RF), and the Support Vector Machine (SVM). This research uses healthcare datasets that are available in the Kaggle dataset [9], which contains data on 5,110 cases, and this research selected only 4,254 cases that are adults 18 years of age or older [10]. A confusion matrix is used for summarizing the performance of a classification model and the AUC (Area Under The Curve) represents the degree of separability. The higher the AUC, the better the model is at distinguishing between patients with a stroke and those without. From the experiment, RF achieves the best performance with an accuracy of 0.94, precision of 0.93, recall of 0.95, f1-score of 0.94, and an AUC of 0.94. The three top-rankings of Random Forest Feature Importance by importance are the age of 0.38, the average_glucose_level of 0.20, and bmi of 0.05, respectively.

Keywords: Stroke, Machine learning, Model performance

[1] Department of Computer Science, Faculty of Science, Srinakharinwirot University
Bangkok, Thailand, 10110

[*] Corresponding author email: Sakol.pat@g.swu.ac.th

I.INTRODUCTION

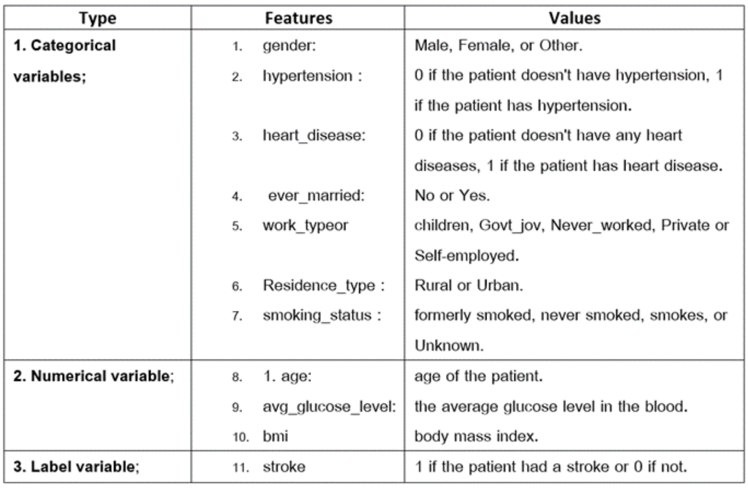
Currently, Cerebrovascular disease (stroke), or paralysis, is one of many health problems worldwide due to its high rate of death and disability. According to the World Health Organization (WHO), in 2020, more than 80 million people suffered from a stroke, about 5.5 million of whom are dead, and the survivors tend to have disabilities. It also saw an increase of 14.5 million new cases per year, of which one in four were aged 25 and over [1]. Therefore, it is necessary to develop a comprehensive care system for stroke patients. Patients who are at risk of having a stroke or are in the acute stroke phase must receive immediate preventive care or treatment, which helps to reduce mortality, complications, disability, and treatment costs.

At present, the application of Artificial Intelligence (AI) technology in the assessment of stroke risk can achieve favorable results. Previous research showed that AI algorithms can be used for the early diagnosis of atrial fibrillation using normal sinus rhythm electrocardiographs, which allows for early intervention to reduce stroke risk [2]. Han et al. [3] applied machine learning to develop a classification model for predicting short-term probabilities of stroke that yields higher performance than the traditional CHA2DS2-VASc score. Lekadir et al. [4] showed the potential of using Convolutional Neural Network (CNN) for automatic characterization of carotid plaque composition (lipid core, fibrous cap, and calcium) in ultrasound, which is correlated with risk stratification in ischemic stroke. Chilamkurthy et al. [5] applied deep learning algorithms to automatically identify abnormalities in head computed tomography scans. The network performed well in detecting intracranial hemorrhage (AUC, 0.94) and calvarial fractures (AUC, 0.92). In addition, Titano et al [6] demonstrated the effectiveness of CNN through a double-blinded randomized controlled trial, which showed that the deep learning-based system could detect acute neurological events in cranial imaging 150 times faster than radiologists. Therefore, in this era of big data and medical services, it has become an important matter that is always related. It also helps to support the doctor in making a decision to diagnose the symptoms of the disease early and aids in planning the treatment. The goal of this study is to use machine learning (ML) to predict stroke risk by analyzing healthcare stroke datasets with three ML models: LR, RF, and SVM, which have shown the best performance for analyzing the healthcare dataset [7].

II. Methodology

A. Dataset

This is the data from the Kaggle dataset [9] in CSV format, which is 317 KB. There are 11 features and 5,110 records in our database. The samples used in this investigation were drawn solely from the records of adult patients aged 18 and up, leaving just 4,254 records. After that, the patients were separated into two groups: those with risk factors for stroke and those who had already had a stroke.



Type	Features	Values
1. Categorical variables;	1. gender:	Male, Female, or Other.
	2. hypertension :	0 if the patient doesn't have hypertension, 1 if the patient has hypertension.
	3. heart_disease:	0 if the patient doesn't have any heart diseases, 1 if the patient has heart disease.
	4. ever_married:	No or Yes.
	5. work_typeor	children, Govt_jov, Never_worked, Private or Self-employed.
	6. Residence_type :	Rural or Urban.
	7. smoking_status :	formerly smoked, never smoked, smokes, or Unknown.
2. Numerical variable;	8. 1. age:	age of the patient.
	9. avg_glucose_level:	the average glucose level in the blood.
	10. bmi	body mass index.
3. Label variable;	11. stroke	1 if the patient had a stroke or 0 if not.

Table 1. The dataset type, feature, and values.

From Table 1. The primary variables consist of 10 variables, which were divided into 7 categorical and 3 numerical variables, and the dependent variable was a stroke.

B. WORKFLOW

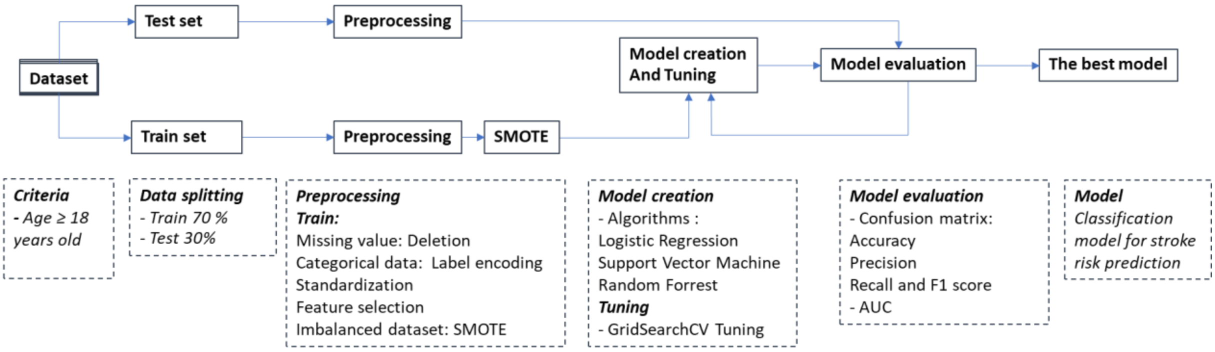


Figure 1. Block diagram of the proposed ML model.

IMPLEMENTATION STEPS

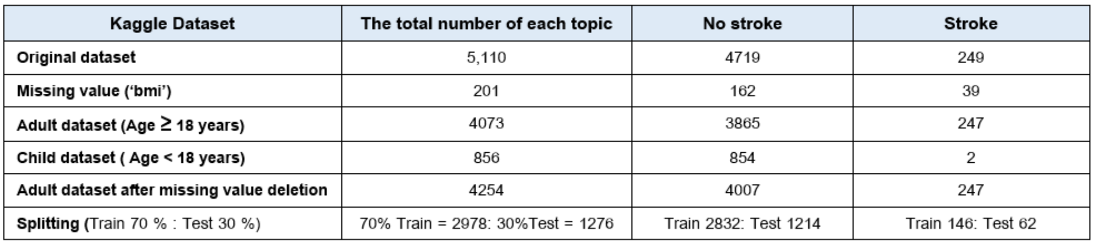
From Figure 1. The details of each step are as follows.

1. Data

The adult patients are 18 years of age or older [10].

2. Data splitting

Data splitting and then dividing the data into the training set and test set by test set, we will just place it without having to interfere or sneak it. The data is divided into 2 sets—namely, train 70%, test set 30% shown in Table 2.



Kaggle Dataset	The total number of each topic	No stroke	Stroke
Original dataset	5,110	4719	249
Missing value ('bmi')	201	162	39
Adult dataset (Age ≥ 18 years)	4073	3865	247
Child dataset (Age < 18 years)	856	854	2
Adult dataset after missing value deletion	4254	4007	247
Splitting (Train 70 % : Test 30 %)	70% Train = 2978: 30%Test = 1276	Train 2832: Test 1214	Train 146: Test 62

Table 2. The dataset is divided into 2 sets is train 70%, test 30 %.

From Table 2. The data we used in this study, have dealt with missing values, this is data of 4254 adult patients aged 18 years and over, comprising 4007 people with strokes and 247 non-strokes.

3. Data Pre-processing

There is no outlier data. The bmi missing values are dropped 181 rows. This study performs feature engineering by standardizing avg_glucose_level to make the feature's data is at a similar scale.

4. Model Selection and creation

This study selects three ML algorithms, SVM), RF, and LR due to having shown so far the best performance for analyzing the healthcare data to create predictive models.

5. Model evaluation

This study evaluates the performance of our predictive models using a confusion matrix, accuracy, sensitivity, precision, F1 score.

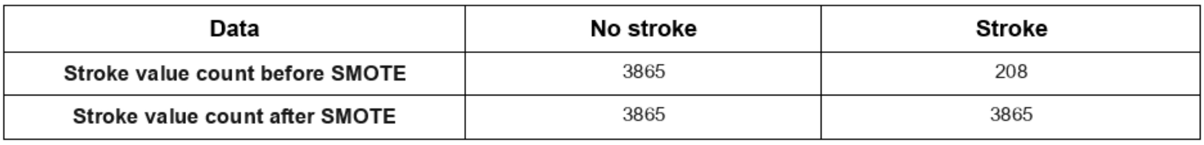
C: Handling with imbalance dataset

Imbalance dataset

Working with imbalanced datasets presents the challenge that most machine learning techniques will ignore, and thus perform poorly on, the minority class, even though performance on the minority class is typically the most important.

Synthetic Minority Oversampling Technique (SMOTE) for handling with imbalanced dataset

By definition, SMOTE is an oversampling technique that generates synthetic samples from the minority class. It is used to obtain a synthetically class-balanced or nearly class-balanced training set, which is then used to train the classifier.



Data	No stroke	Stroke
Stroke value count before SMOTE	3865	208
Stroke value count after SMOTE	3865	3865

Table 3. The stroke value count before and after using SMOTE technique.

From Table 3. The dataset is imbalanced data that contain class 0 (No stroke) 3866 records and class 1 (Stroke) 208 records, so after using the SMOTE technique we got a balanced dataset with stroke value count of class 0 (No stroke) 3866 and class 1 (Stroke) 3866 records equally.

D. Feature Selection

Correlation is a measure of the linear relationship between two or more variables. Through correlation, we can predict one variable from the other.

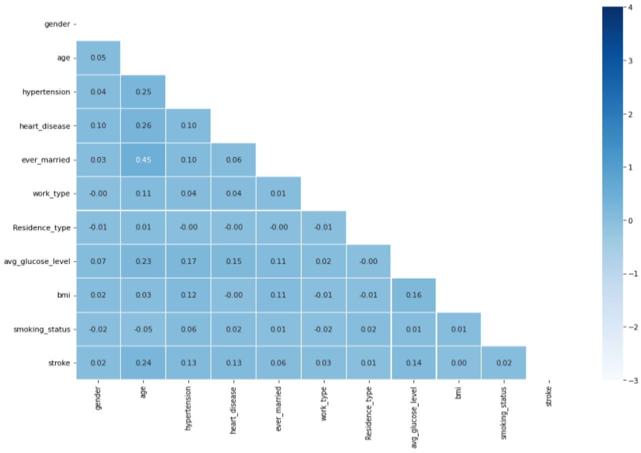


Figure 2. Correlation matrix of the features.

This study needs to set an absolute value, say 0.5, as the threshold for selecting the variables. If we find that the predictor variables are correlated among themselves, we can drop the variable that has a lower correlation coefficient value with the target variable. So, from Figure 2., we found the correlation coefficient of the predictor variables is not correlated among themselves. We still keep them for creating models.

III. Results and Discussion

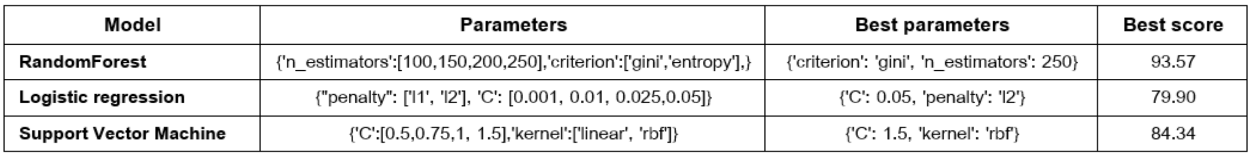
Training and ML algorithm comparison: Different ML algorithms were used for the training process, whose performance is shown in Table 4.

Model performance with imbalanced dataset				Model performance with balanced dataset after using SMOTE			
Models	**precision**	**recall**	**F1-score**	**Models**	**precision**	**recall**	**F1-score**
RandomForest (Accuracy = 0.95)				**RandomForest** **(Accuracy = 0.79)**			
0	0.95	1.00	0.97	**0**	0.84	0.71	0.77
1	0.00	0.00	0.00	**1**	0.75	0.86	0.80
Logistic regression **(Accuracy = 0.95)**				**Logistic regression** **(Accuracy = 0.79)**			
0	0.95	1.00	0.97	**0**	0.80	0.77	0.79
1	0.00	0.00	0.00	**1**	0.78	0.81	0.80
Support Vector Machine **(Accuracy = 0.95)**				**Support Vector Machine** **(Accuracy = 0.83)**			
0	0.95	1.00	0.97	**0**	0.85	0.80	0.83
1	0.00	0.00	0.00	**1**	0.81	0.86	0.84

Table 4. Comparison of model performance. Between models derived from an imbalanced dataset and balanced dataset (shows the performance of logistic regression models, Support vector machine, and random forest).

From Figure 6. Using SMOTE techniques to manipulate the imbalanced dataset can enhance the model's efficiency in predicting stroke likelihood. Although the accuracy of models from the balanced dataset is lower than that of those from the imbalanced dataset, with very high recall values, models from the balanced dataset have better performance because they can fit better with other data based on the assumptions set.

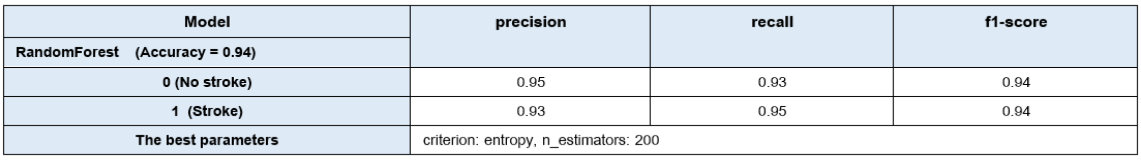
Hyperparameters Tuning (Using GridSearchCV)



Model	Parameters	Best parameters	Best score
RandomForest	{'n_estimators':[100,150,200,250],'criterion':['gini','entropy'],}	{'criterion': 'gini', 'n_estimators': 250}	93.57
Logistic regression	{"penalty": ['l1', 'l2'], 'C': [0.001, 0.01, 0.025,0.05]}	{'C': 0.05, 'penalty': 'l2'}	79.90
Support Vector Machine	{'C':[0.5,0.75,1, 1.5],'kernel':['linear', 'rbf']}	{'C': 1.5, 'kernel': 'rbf'}	84.34

Table 5. The best parameters of each model after hyperparameter tuning by GridSearchCV.

GridSearchCV is a module of the Sklearn model_selection package that is used for hyperparameter tuning. From Table 5. The Random Forest model is the best performing model for stroke risk prediction. We define the hyperparameter as shown below for the random forest classifier model. The parameters are tuned randomly, and the performance results are checked as Table 6. below.



Model	precision	recall	f1-score
RandomForest (Accuracy = 0.94)			
0 (No stroke)	0.95	0.93	0.94
1 (Stroke)	0.93	0.95	0.94
The best parameters	criterion: entropy, n_estimators: 200		

Table 6. The best parameters are {'criterion': 'entropy', 'n_estimators': 200} and the best performance of the random forest model after hyperparameter tuning.

After hyperparameter tuning, we found the Random Forest model to be the best performing model for stroke risk prediction in this study, as the Confusion matrix, and AUC scores.

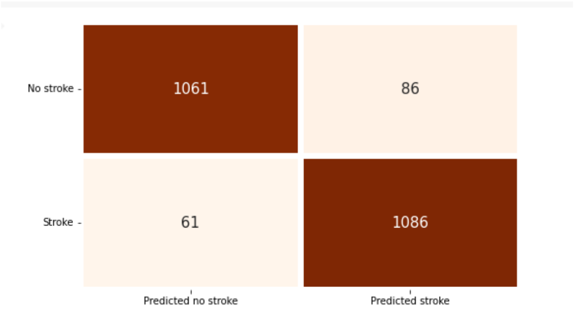


Figure 3. A confusion matrix of the RF model.

From Figure 3. The RF classifier model is favorable performance, showing the model predicted a total of 61 false negatives and 86 false positives, with an accuracy of 0.94 and an f1 score of 0.94. The AUC of 0.94. The higher the AUC is the better model for correctly classifying instances.

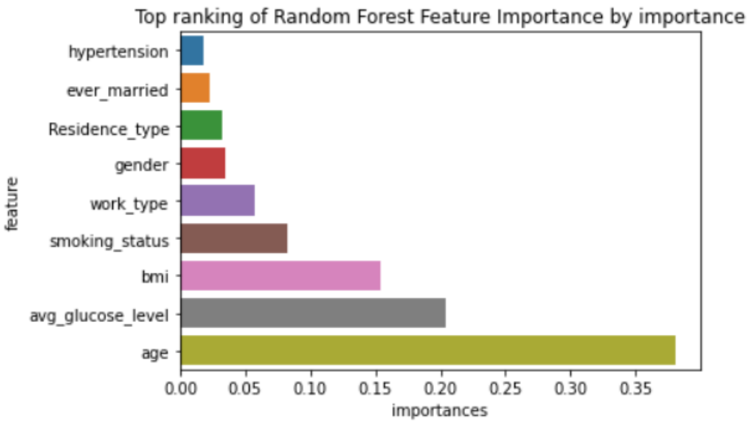


Figure 4. Feature importance of the Random forest model, in descending order.

Feature Importance of the Random forest model

Feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. The higher the value the more important the feature.
From Figure 4. the important feature in descending order of value is age, avg_glucose_level, bmi, smoking_status, work_type, gender, Residence_type, ever_married and hypertension with the importance value of 0.381, 0.204, 0.154, 0.083, 0.057, 0.035, 0.033, 0.023, 0.018, 0.014 respectively.

IV. DISCUSSION

From this study, the RF classification outperforms other methods tested with a classification with an accuracy of 0.94, a recall of 0.95, a precision of 0.93, an f1-score of 0.94, and an AUC of 0.94. When using cross-validation metrics to predict stroke likelihood, the RF method is more efficient than other methods. The feature importance of the RF model indicates the relationship of each feature to the likelihood of getting strokes in the prediction that a feature has greater feature importance means that feature has an association with a greater likelihood of getting a stroke. This study's finding is the features are relevant to the probability of getting a stroke in descending order are as follows: age, avg_glucose_level, bmi, smoking_status, work_type, gender, Residence_type, ever_married and hypertension with an importance value of 0.381, 0.204, 0.154, 0.083, 0.057, 0.035, 0.033, 0.018, and 0.014, respectively. Based on what we know about medicine, risk factors are behaviors or traits that make you more likely to develop a disease or condition. Some risk factors for stroke that cannot be changed, which we call uncontrollable risk factors, are age, gender, and race. However, there are many risk factors that can be treated, modified, or controlled that can help to reduce your risk of a stroke by making some healthy lifestyle choices. Whether it’s your diet, activity level, smoking, or drinking, it’s never too late to make a change.

V. Conclusions

This study uses only data from a single source. If this study can validate the model with data from other sources (multi-center validation), it will help to determine how well the model is adapting to other datasets. The artificial intelligence system provides efficiency and accuracy that are on par with experts. We think artificial intelligence will help, but not yet replace the doctor. because we still don't know how well artificial intelligence systems can perform in healthcare settings. use of artificial intelligence. In hospitals, this may raise new ethical and legal questions. Physicians may face liability issues when artificial intelligence recommendations are followed. Especially when these recommendations are inconsistent with the judgment of the physician or based on clinical knowledge and instinct. Furthermore, artificial intelligence systems are unable to explain the social and cultural contexts that influence patient care. Therefore, these barriers and pitfalls should be carefully considered when using the technology, Artificial Intelligence Systems vs. Clinical Settings.

Acknowledgments

I would like to thank the Graduate School of Srinakharinwirot University for the support and guidance in presenting the research results of the graduate studies

References

[1] Hfocus. (Thu, 2020-10-29). Know quickly, survive! Cerebrovascular disease is the leading cause of the elderly.handicapped-death.RetrievedFebuary19, 2022, from https://www.hfocus.org/content/2020/10/20381

[2] Attia ZI, Noseworthy PA, Lopez-Jimenez F, Asirvatham SJ, Deshmukh AJ, Gersh BJ, Carter RE, Yao X, Rabinstein AA, Erickson BJ, et al.. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction.Lancet. 2019; 394:861–867. doi: 10.1016/S0140-6736(19)31721-0CrossrefMedlineGoogle Scholar

[3] Han L, Askari M, Altman RB, Schmitt SK, Fan J, Bentley JP, Narayan SM, Turakhia MP. Atrial fibrillation burden signature and near-term prediction of stroke: a machine learning analysis.Circ Cardiovasc Qual Outcomes. 2019; 12:e005595. doi: 10.1161/CIRCOUTCOMES.118.005595LinkGoogle Scholar

[4] Lekadir K, Galimzianova A, Betriu A, Del Mar Vila M, Igual L, Rubin DL, Fernandez E, Radeva P, Napel S. A convolutional neural network for automatic characterization of plaque composition in carotid ultrasound.IEEE J Biomed Health Inform. 2017; 21:48–55. doi: 10.1109/JBHI.2016.2631401CrossrefMedlineGoogle Scholar

[5] Chilamkurthy S, Ghosh R, Tanamala S, Biviji M, Campeau N, Venugopal V, et al. Deep learning algorithms for detection of critical findings in head CT scans a retrospective study. The Lancet. 2018;392.

[6] Titano JJ, Badgeley M, Scheffler J, Pain M, Su A, Cai M, et al. Automated deep-neural-network surveillance of cranial images for acute neurologic events. Nat Med. 2018;24(9):1337-41.

[7] A. Site, J. Nurmi and E. S. Lohan, "Systematic Review on Machine-Learning Algorithms Used in Wearable-Based eHealth Data Analysis," in *IEEE Access*, vol. 9, pp. 112221-112235, 2021, doi: 10.1109/ACCESS.2021.3103268.

[8] Tavares, Jose-A. (2021). Stroke prediction through Data Science and Machine Learning Algorithms. 10.13140/RG.2.2.33027.43040.

[9] Fedesoriano.(2021-Jan-27),StrokePredictionDataset.Version1.RetrievedFebuary19,2022,from https://www.kaggle.com/fedesoriano/stroke-prediction-dataset/metadata

[10] Wikipedia, the free encyclopedia. Adult. Retrieved19 February 2022, from https://en.wikipedia.org/wiki/Adult