Stroke Factor Analysis based on Granger Causality Test

Tiancheng Shi *

Department of Literature and Science, University of Illinois Urbana-Champaign, Illinois, United States

* Corresponding author email: ts15@illinois.edu

Abstract. Stroke is a common brain disease that happens when brain vessels get blocked. As one of top-rated reason for human death, stroke not only leads to death but also causes huge economic damage and occupies a lot of medical resources. Every year 610,000 people died in US because of stroke. The economy loss is up to 30 billion dollars. This paper tries to predict and find the relationship between stroke and other related variables. The dataset from Kaggle collects 5110 participants' data about stroke with other health or habitats' relative variables like gender, age, bmi, average glucose level, whether have hypertension, smoking status, to a total of 12 variables. With the use of the VAR model and Granger Causality test, 10 different pairs of VAR models are created and compared. The result of the final test shows that 9 variables have Granger Causality with stroke while only 1 attribute is not.

Keywords: Stroke Factor Analysis; Granger Causality Test; Vector Auto Regression.

1. Introduction

A stroke happens when the vessel used for transferring oxygen to the brain has been blocked. This causes the brain to start to have problems with getting oxygen and blood, which will cause the cells in the brain to die and start to affect human lives.

There are two types of stroke, one is Ischemic stroke and the other is Hemorrhagic stroke. Most patients get Ischemic stroke. According to CDC, "An ischemic stroke occurs when blood clots or other particles block the blood vessels to the brain." While Hemorrhagic stroke, the blood vessel is damaged by the pressure caused by blood or ruptures leaked by the brain.

As a reason for human death, stroke acts an important role in human society now. From CDC, 610,000 people died because of stroke in the US each year, and in the world every 3.5 minutes there is one person who died because of a stroke. Stroke also causes a huge economic effect. Each year, $15 billion to $30 billion is spent on stroke in the US, When the patients' lost earnings are taken into account [1].

With the development of Artificial Intelligence, many studies are researching disease prediction by using techniques like machine learning. Stroke prediction is also a popular topic for now. In addition, using big data as a method to predict stroke statistically could also provide hints to physicians for finding the real reason why stroke happens and how to avoid it. The following authors have done stroke prediction with different machine learning techniques.

Ferdib-Al-Islam et al. [2] predict the stroke with the use of SMOTE and machine learning based on the dataset from Kaggle. His team did 99.07 percent in accuracy, 99.0 percent in precision, and 99.0 percent in recall.

Premisha [3] has tested the accuracy of DT, NB, ANN, and SVM these base classifiers, and RF, XGBoost, LightGBM, AdaBoost, and CatBoost these bagging and boosting techniques.

Dhiman [4] also predicts stroke with the Kaggle datasets with her teammates. XgBoost, K-nearest neighbor, Support vector machine, Random Forest, and Decision tree are applied in their research with a best of 98.45 percent accuracy in the SVM classifier.

Razzaq et al. [5] did the stroke prediction with other machine learning techniques. In their study, XgBoost has 94 percent accuracy, 91% in terms of precision, and 94% for F1-score as perspectives.

In R.K. Kavitha [6] and her colleagues' work, maximum of 98% accuracy was obtained by using the Decision Tree classifier which is also based on the Kaggle dataset.

Azizul Hakim [7] trained a self-modified bagging techniques by using extra trees algorithms. This gave them a 96% accuracy in stroke prediction.

In Puri and his teams' work [8], they use linear, quadratic, and cubic SVM method to predict the happening of the stroke, with a high accuracy of 95.2%, 95.2% and 94.7% respectively.

Jayashree also works in stroke prediction [9]. This study uses SVM, KNN, RF, DT, and LR classifiers by using Relief algorithm with feature selection has a best 88% accuracy.

In Li's [10] design of stroke prediction on ischemic stroke by using variable induced method with 0.7222 in c-test, 0.7000 in p-test and 0.04 on development of coefficients.

Though this study has done many works on stroke prediction, neither of them has done a prediction for the causality of stroke. This study will use the Vector Auto regression model to predict the relation between stroke with other variables. And use the Granger Causality test to find whether there exists Granger Causality between those variables to stroke.

2. Methodology

In this paper, unrestricted models will be used to predict the relationship between stroke with different variables. The variables in the dataset are assumed as independent since the relation between the variables cannot be determined. The dataset contains 12 variables including "stroke" and "id". Therefore 10 pairs of Vector Auto Regression models will predict the relation with stroke. Moreover, this research will use granger causality to test the causality between these variables with stroke.

2.1 VAR Model

The Auto Regression model is a method used to predict how x is compared to x that in the past. While Vector Auto Regression is better, it could test multiple variables at the same time instead of one. This study will use different variables to predict stroke with each attribute in the dataset, where x is stroke, and y is other attributes.

Since each VAR pair only have two models, so the formula is a VAR(1) model(1):

$$\begin{bmatrix} y_t \\ x_t \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} a_{1,1} & a_{2,1} \\ a_{1,2} & a_{2,2} \end{bmatrix} \begin{bmatrix} y_{t-1} \\ x_{t-1} \end{bmatrix} + \begin{bmatrix} e_{1,t} \\ e_2, t \end{bmatrix} \quad (1)$$

Since there are 10 variables to test for the stroke, 10 pairs of VAR models will be made.

2.2 Granger Causality Test

As a statistical hypothesis test the Granger Causality test is used for judging whether one-time series is useful in forecasting another variable. In this paper Granger Causality test used for getting the casuality of stroke for the VAR model above. The formula for the Granger Causality(2):

$$y_t = a_0 + a_1 y_{t-1} + b_1 x_{t-1} + error_t \quad (2)$$

All 10 pairs of VAR models will do the Granger Causality. Because there is no true time series for stroke with those variables, the lag value will take 1. The hypothesis for the test for each model will be:

H_0: The variable has Granger Causality toward stroke.

H_A: The variable does not have Granger Causality toward stroke.

A Wald Test will be used to compare each pair of the model. The F-statistics will be used to determine the hypothesis.

3. Experiment

In this part, the specific steps of this study will be stated.

3.1 Dataset Information and Description

This study used a dataset which is based on the stroke prediction dataset in Kaggle, which is a CSV file containing 5110 observations and 12 attributes. The Major attributes are in the following Table1.

Table 1. Stroke prediction dataset description

Feature	Description	Data type
Id(int)	the identifier for the observer	Numerical
gender(char)	"Male", "Female" or "Other" represents the gender of the observer	Categorical
age(int)	the age of the observer	Numerical
hypertension(int)	0 or 1 whether 0 for no hypertension in the observer, 1 for there is hypertension in the observer	Categorical
heart_disease(int)	0 or 1 whether 0 for no heart disease in the observer, 1 for there is heart disease in the observer	Categorical
ever_married(char)	"No" for no marriage and "Yes" for having marriage	Categorical
work_type(char)	"children" for the observer is now children, "Govt_job" for the observer working in government, "Never_worked" for the observer never work, "Private" for observer working in private work, or "Self-employed" for observer working for themselves	Categorical
Residence_type(char)	"Rural" for the observer lives in the rural area and "Urban" for the observer lives in the urban area	Categorical
avg_glucose_level(double)	the average blood glucose level of the observer	Numerical
bmi(double)	the BMI of the observer	Numerical
smoking_status(char)	"formerly smoked" represents the observer used to smoke, "never smoked" represents the observer never smoke, "smokes" represents the observer is smoking now and "Unknown" represents observer does not want to show their smoking status	Categorical
stroke(int)	1 or 0, whether 1 for having a stroke, 0 for no	Categorical

3.2 Pre-Processing of the Data

Since the attributes gender, work type, ever married, smoking status, and residence type in the dataset are all categorical attributes that collect with strings, which cannot test using a math model, in this study, they are turned into numeric data. The detailed information is in the Table2 below.

Table 2. Changed data description

Feature	**Description**	**Data type**
gender(char)	1 represents "Male", 0 represents "Female" and 2 represents "Other"	Categorical
ever_married(char)	0 represents "No" and 1 represents "Yes"	Categorical
work_type(char)	0 represents "children", 1 represents "Govt_job", 2 represents "Never_worked", 3 represents "Private" and 4 represents "Self-employed"	Categorical
Residence_type(char)	0 represents "Rural" and 1 represents "Urban"	Categorical
smoking_status(char)	0 represents "formerly smoked", 1 represents "never smoked", 2 represents "smokes" and 3 represents "Unknown"	Categorical

In addition, there are few empty data in variable BMI. They are considered as 0 when calculating.

3.3 VAR Models and Granger Causality Test Implementation

After processing the dataset, different pairs of VAR have been run. They are gender-stroke, age-stroke, hypertension-stroke, heart_disease-stroke, ever-married-stroke, work_type-stroke, residence_type-stroke, avr_glucose_level-stroke, bmi-stroke, and smoking_status-stroke. Then, the same as in methodology, a Granger Causality test takes place with a wald test as a method for determining the hypothesis.

4. Result and Analysis

The granger causality test for all attributes with stroke when lag equals 1 is below in Table3.

Table 3. Granger Causality of different variables with stroke

Attributes	Lag-value	df	F-statistics	p-value
Hypertension	1	-1	0.0171	0.8959
Heart disease	1	-1	2.5111	0.1131
Ever married	1	-1	0.0415	0.8385
Residence type	1	-1	4.1282	0.04223
Average glucose level	1	-1	2.5639	0.1094
BMI	1	-1	0.3049	0.5808
Age	1	-1	1.0553	0.3043
Smoking Status	1	-1	0.1658	0.6839
Gender	1	-1	0.176	0.6748

From the analysis above, when the lag value is 1, the p-value of heart disease, hypertension, ever-married, BMI, average glucose level, age, gender, and smoking status are all greater than 0.05, which cannot reject the null. Therefore all attributes except residence type are granger causes of stroke. The residence type has a p-value less than 0.05. So in the extent of 95% confidence, residence type is not the granger cause of stroke.

5. Conclusion

Stroke as an important disease to human beings has been a problem for people for thousands of years. This study tries to use statistics way to find why stroke happens. Though this is not the first study to use the statistical method to find the reason for stroke but is the first study to test causality. From the data set in Kaggle with 5110 observations and 12 attributes, 10 different pairs of models have been created. the final result is surprising for all attributes: heart disease, hypertension, ever married, BMI, average glucose level, age, gender, and smoking status have the granger causes of stroke.

Since the granger causality has self-limitations providing the true casualty, all these attributes could not represent the true casualty of stroke. In the future, this work could give insight for any research that has an interest in stroke to find the true casualty in each attribute with stroke.

References

[1] Dobkin B. The economic impact of stroke. Neurology, 1995, 45:6-9.

[2] Ferdib-Al-Islam, M. Ghosh, An Enhanced Stroke Prediction Scheme Using Smote and Machine Learning Techniques. International Conference on Computing Communication and Networking Technologies, 2021,12: 1-6.

[3] P. Premisha, S. Prasanth, M. Kanagarathnam, K. Banujan. An Ensemble Machine Learning Approach for Stroke Prediction. International Research Conference on Smart Computing and Systems Engineering, 2022, 165-170.

[4] H. Dhiman, R. Kumar, P. Rani. A Hybrid Model for Early Prediction of Stroke Disease. International Conference on Reliability, Infocom Technologies and Optimization, 2022, 10:1-6.

[5] N. Razzaq, N. Masood, S. Nawaz, N. Anjum et al. Stroke Prediction in Elderly Persons using Remote Health Monitoring. IEEE International Conference on Electronics, Circuits and Systems, 2022, 29:1-4.

[6] R. K. Kavitha, W. Jaisingh, S. R. Sujithra. Applying Machine Learning Techniques for Stroke Prediction in Patients. International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation,2021, 2-5.

[7] Md. Azizul Hakim, Md. Zahid Hasan, Md. Mahabur Alam, Md. Mehadi Hasan, et al. An Efficient Modified Bagging Method for Early Prediction of Brain Stroke. International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering, 2019, 1-4.

[8] Harshita Puri, Jhanavi Chaudhary, Kulkarni Rakshit Raghavendra, Rhea Mantri et al. Prediction of Heart Stroke Using Support Vector Machine Algorithm. International Conference on Smart Computing and Communications, 2021,8:21-26.

[9] Jayashree K, N Krishna Prakash. Prediction of Brain Stroke using Machine Learning with Relief Algorithm. International Conference on Edge Computing and Applications, 2022, 1255-1260.

[10] X. Li, X. Yao, Y. Liang, L. Tang. Research on the Construction of Multivariate-Induced ischemic stroke prediction model based on medical big data. IEEE Advanced Information Technology, Electronic and Automation Control Conference, 2022, 6: 983-987.