Automating Stroke Subtype Classification from Electromagnetic Signals Using Principal Component Methods

Jinrui Li[1] , Guohun Zhu[1], Min Xi[2]

[1]School of ITEE, The University of Queensland, Brisbane QLD 4000, AU
[2]Shandong institute of medicine and health, Jinan 250022, China
E-mail: g.zhu@uq.edu.au

Abstract. Stroke is an emergency. Automatic stroke classification based on electromagnetic images needs to establish a large correlation matrix array which leads to low computational performance. The paper presents a principal component analysis (PCA) to extract the efficient features to conduct a fast stroke subtypes classification. Firstly, it is shown that stroke classification using a single feature was failed. By discovering the data pattern associated with correlations, the second attempt has proposed a novel PCA feature extraction method, the result has achieved 99% accuracy after the application of PCA and SVM on the extracted features.

Keywords: Stroke, PCA, Electromagnetic image

1. INTRODUCTION

Stroke is medical emergency and is considered as one of the most high-risk cause of acute death [1]. The economic cost of stroke is extraordinarily high, according to a data analysis report conducted by Deloitte, the economic cost of stroke in Australia exceeded $6.2 billion in 2020, with a further $26.0 billion in lost wellbeing, caused by short or long-term disability, and premature death [2]. On the other hand, the human cost is more terrific. One stroke occurs in Australia every 19 minutes, more than 8000 Australians would die because of stroke in 2020, more and more people below 54 years old are suffering the stroke [2].

There are mainly two types of stroke, haemorrhagic (ICH) and ischemic (IS). An ischemic stroke is caused by a blockage cutting off the blood supply to the brain. A haemorrhagic stroke is caused by a bleeding in or around the brain. [3] While both are sharing a very similar symptom, the treatment can be fairly different. However, it is difficult to known whether all clots in patient's vessel are open or not during treatment because CT cannot scan the brain frequency due to radiation exposure. The key technique of this fellowship is to identify the multiple stroke lesions and investigate biomarkers for effective stroke therapies using electromagnetic image (EM) scanners. EM is an emerging technology which promises to provide a mobile, and rapid scan patient's head based on the dielectric properties of the tissue.

To successfully control the disaster caused by stroke, the early diagnosis is extremely important. The higher chance patients will recover if they receive the diagnosis results earlier. Currently, stroke-related deaths account for a large number due to the time spent on diagnosing and moving patients to available instruments. The diagnosis of stroke is basically relying on the brain image. Computed Tomography (CT) scan and Magnetic Resonant Imaging (MRI) are two commonly used methods. However, there exists a large space to improve for these two methods. CT scan might be harmful for human brain due to the radiation exposure. Some studies have shown that CT scan might increase the risk of getting cancer mostly occurring at chest, abdomen and pelvis [4] [5]. Moreover, both CT and MRI are not portable, which means they are not allowed to be placed on the ambulance or on the road. As a result, researchers are focusing on new stroke detection approaches: Microwave system, which has been believed by the researchers that it is the most promising tool for reforming medical diagnosis, and it is currently leading the research path [6] [7].

Microwave imaging (MWI) is a promising tool for improving medical diagnosis because it has the strong potential to compensate for the limitations of CT and MRI. It is a technique aimed at sensing a given scene by means of interrogating microwaves [8].

There are mainly three methods to develop microwave head application: tomography, radar imaging and machine learning. Tomography is to quantitatively reconstruct the spatial representation of tissue dielectric distribution. Although the equation will be reduced iteratively, it is still likely to produce multiple solutions. Radar imaging is sensible to skin reflections removal although it is considered as a feasible method for prehospital use [7]. In this situation, machine learning might be the most promising strategy. However, since the current machine learning algorithm does not allow for the development of image systems such as CT or MWI, a more sophisticated model would be needed in the future.

This paper presents a principal component method to identify the stroke subtype by extract the principal component from multiple electromagnetic signals. The scatter parameter data is collected from 16 antenna arrays, one antenna transmits a signal, all antennas receive it, and then one by one. Firstly, all received signals from one subject was transferred into one dimension feature to conduct the classification. The results are poor. Then a correlation matrix is built for each simulated data. Unlike the previously proposed graph degree mutual information systems [9], this novel model applies the PCA to extract the electromagnetic signal directly and select the major component to conduct classification. The results have shown high performance results.

2. DATASET DESCRIPTION

The electromagnetic data used for analysis and classification in this paper are from the 10 digital human brain models [24], where each human brain has five ICH and five IS targets. During the simulation, the human brain was surrounded by 16 antennas. Each antenna emits a radio frequency signals and all other antennas received the signals, which the emitting and receiving signals are reflecting signals but others are transmitted signals. The previous length of signals are 4096 sampling points [9], the signals used in this study are resampling as 751 points because the realistic signals in clinical are 751. The frequency starts from 0.5GHz and ended on 2.0GHz. The recordings were storage in touchstone format. All signals were sampled at 256 samples per second with 16-bit resolution.

3. METHODOLOGY

The flow chart for the stroke subtypes classification system is shown in Figure 1.

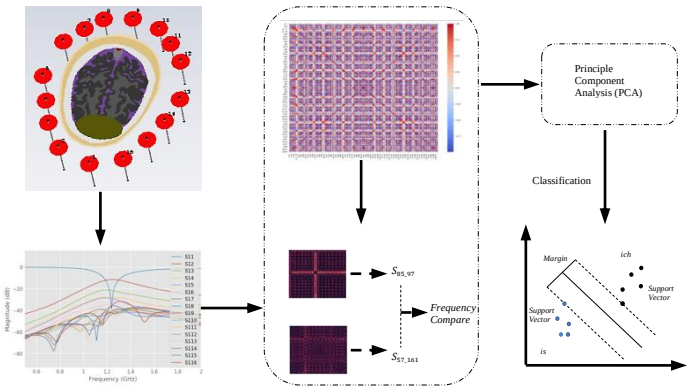

Figure.1 Flow chart of stroke classification

3.1. Inverse of Fast Fourier transform (IFFT) and zero padding

The fast Fourier transform (FFT) is a method for efficiently computing the discrete Fourier transform (DFT) of a time series (discrete data samples) [10]. IFFT, however, is the inverse of FFT, indicating that it could transform frequency into time series, which is defined as [11].

$$X(n) = \frac{1}{N} * \sum_{k=0}^{N-1} X(k) * e^{i*2*pi*n*k/N} \quad (1)$$

where X(k) is the frequency domain samples from the raw signals, Y(n) is the output time series. It is notice that the N is the FFT size, which is equal to the 2^n. In this paper, it should be 2048.

Due to the input raw signals is started from 0.5 Ghz and only 751 points, it needs to pad zero into the raw data to fill the frequency points from 0 to 0.5Ghz.

3.2. Correlation filter method

The new feature extraction method which is designed for the specific pattern of this dataset is proposed by the work.

To be specific, the difference of correlations was calculated firstly to obtain n*n number of matrixes:

$$\begin{pmatrix} 0 & dis(r_{s12\_s11}) & \cdots & dis(r_{s1616\_s11}) \\ dis(r_{s11\_s12}) & 0 & \cdots & \vdots \\ \vdots & \vdots & \cdots & dis(r_{s1616\_s1515}) \\ dis(r_{s11\_s1616}) & dis(r_{s12_{s1616}}) & \cdots & 0 \end{pmatrix} \quad (2)$$

Where r represents correlation coefficient, difference of correlation is calculated as $ABS(r(ICH_m) - r(IS_n))$.

The candidate features (S_{ij_mn} , where $i, j, m, n \in [0 \ldots .16]$.) were extracted from each matrix which can significantly distinguish ICH and IS by ranking each dis(r) and it appears these features have significantly similar pattern. For example, S_{85_97}, S_{86_96} and S_{216_89} may be retrieved by ranking the calculation result from one of the matrixes (as shown in the following table) and this order can be found in many other matrixes, which indicates that the extracted features may have a strong correlation, and this has been proved by the further calculation. Based on this reason, the frequency is calculated to extract the final selections and Principal Component Analysis is applied before classification.

$r1$	$r2$	$dis(r1\_r2)$
S_{85}	S_{97}	0.019713
S_{86}	S_{96}	0.019561
S_{216}	S_{89}	0.019499

3.3. Principle Component Analysis

Principle Component Analysis (PCA) is a multivariate technique for analysing a data table in which measurements are represented by multiple interrelated quantitative dependent variables and extracting the most important information from the table [12]. It is a common technique to be used in dimension reduction and feature extraction.

In this work, PCA is particularly useful when extracting the important information after applying the proposed filter method due to the existence of multi-collinearity between the features/variables.

3.4. Support Vector Machines

Support Vector Machine (SVM) is a computer algorithm that learns by example to assign labels to objects [13]. Support Vector Machine is using a set of kernel functions. The kernel function is to require data as input and transform it into the desire form to find the proper Support Vector classifier. These functions include linear, nonlinear, polynomial, radial basis function (RBF), etc.

RBF is one of the most preferred kernel functions in SVM as it generally performs good on non-linear data or large dataset [14].

The work prioritized the Support Vector Machine with RBF kernel function to construct the classification model.

4. RESULTS

The algorithms were implemented using Python embedded with various libraries, including Pandas, Scipy, Seaborn, Numpy, Sklearn, etc.

4.1. The classification based on single features

The main issue was that the time series data structure was destroyed by converting them from d*k*N to k*N. More data processing on time series could be performed before feature extraction. The final result of the first attempt was fairly disappointing with the highest 50.1% of accuracy.

4.2. Feature Selection

The correlation matrix was initially used to select good features by distinguishing those that are closely related to the class but not to other features. The filtered features were then fed into training models to determine the importance of each aspect. The scores were compared on Light GBM and the average of importance score of the other three models. However, the findings of four different models used to calculate the score of feature value were varied, suggesting that the importance of features is sensible to different models.

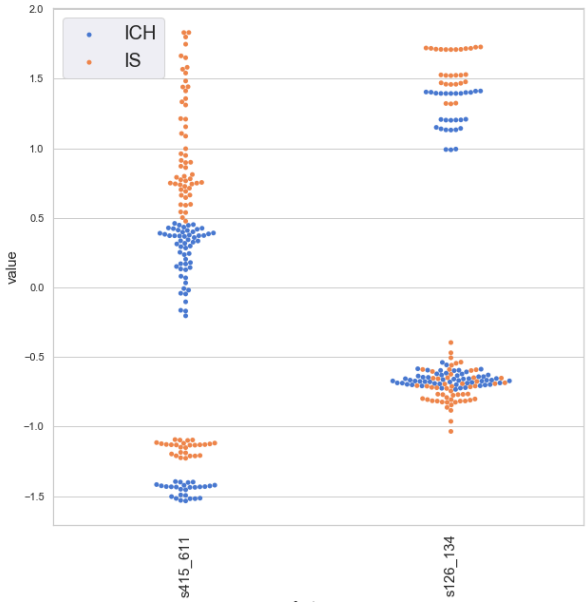

Figure. 2 The swarm plot of example extracted features.

4.3. Classification results

After the feature extraction proposed in the second attempt, PCA and SVM had been applied on the selected feature after using this approach and the performance was significantly improved. The model had then been applied on the dataset containing 56 ICH and 56 IS samples as well as the combination of the datasets totaling 166 samples, the result of PCA 5 components+ SVM on 21 features had also achieved 99% accuracy.

Features	Approach	Result
17 sets	PCA 10 components + SVM rbf	88%
17 sets	PCA 5 components + SVM rbf	99%
21 sets	PCA 5 components + SVM rbf	99%
25 sets	PCA 5 components+ SVM rbf	99%

5. DISCUSSION

5.1. Building the model on single original features is nearly impossible

The process of feature selection in the first attempt also had provided a positive indication of that it is hard to build the model on original features, i.e., S11, S12, …, S1616. The scores from feature importance did provide the useful information about the features that could play the important role in classifying stroke types; however, the "good features", such as S116, S115, did not reveal any major distinctions between ich and is. (Figure 3)

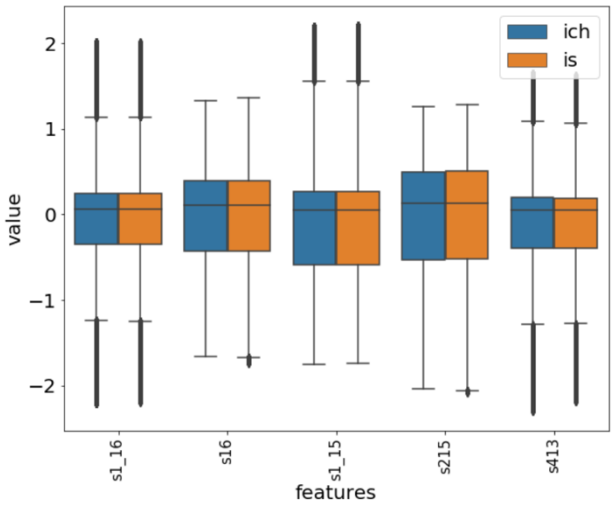

Figure. 3 Box plot of example "good feature" from first attempt

5.2. Feasibility of the proposed correlation filter method

The model on the representation of the missed outcome was changed in the second attempt. The first phase involved reprocessing the time series data and searching for hidden data patterns between features. While some correlation between features was discovered, the disparity between two targets appeared to be notable (Figure 4).

Moreover, some of the correlations which could identify ICH and IS significantly appeared frequently across the sample. By ranking those correlations according to their significance (differences between targets), it even appeared the order of each rank tended to

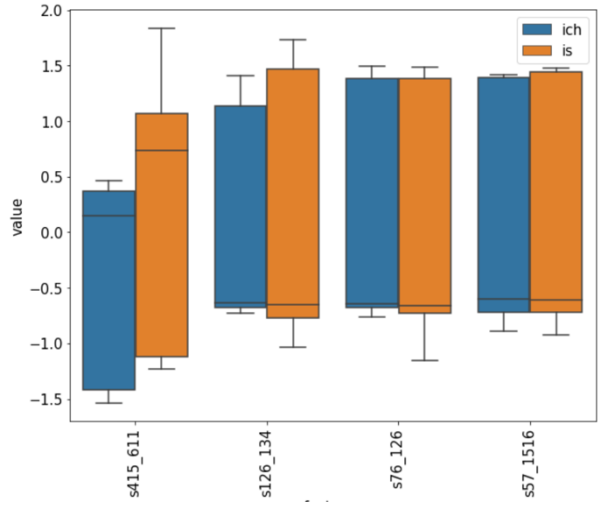

Figure. 4 Box plot of example features by using proposed feature selection method

be incredibly similar, which had indicated even some of the correlations are strongly correlated with each other. Based on this finding, this work proposed the new feature extraction method.

6. CONCLUSION

The work has applied the principal component analysis and correlation matrix to discover the data pattern difference between two stroke subtypes: ICH and IS. It is verified whether one major component of a dimension feature could classify the two types of stroke or not. It is shown that the correlation pattern of simulation data may vary from that of experimental data, implying that the feature extraction process dependent on correlation may not be able to provide important features.

Using the proposal principal component feature extraction method from the correlation matrix, the result achieves well performance on the simulated data. Thus, it is potentially to be applied in realistic clinical stroke medical diagnosis.

REFERENCES:

[1] Persson, M., Fhager, A., Trefná, H. D., Yu, Y., McKelvey, T., Pegenius, G., ... & Elam, M. (2014). Microwave-based stroke diagnosis making global prehospital thrombolytic treatment possible. IEEE Transactions on Biomedical Engineering, 61(11), 2806-2817.) C. D. Jones, A.B. Smith, and E.F. Roberts, Book Title, Publisher, Location, Date.

[2] Simone, C. (2020). The economic impact of stroke in Australia. https://www2.deloitte.com/au/en/pages/economics/articles/economic-impact-stroke-australia.html

[3] Stroke Association. (n.d.). Type of Stroke. https://www.stroke.org.uk/what-is-stroke/types-of-strokeMotorola, Inc.: DSP56002 Users Manual, 1993.

[4] Harvard Medical School. (2013, March). Do CT scans cause cancer? https://www.health.harvard.edu/staying-healthy/do-ct-scans-cause-cancer

[5] Science Daily. (2018, July 19). CT scans may increase risk of brain cancer, study suggests. https://www.sciencedaily.com/releases/2018/07/180719085337.htm#:~:text=A%20new%20study%20suggests%20that,the%20risk%20of%20brain%20tumors.&text=A%20new%20study%20in%20the,the%20risk%20of%20brain%20tumors

[6] Munawar Qureshi, A., Mustansar, Z., & Mustafa, S. (2018). Finite-element analysis of microwave scattering from a three-dimensional human head model for brain stroke detection. Royal Society open science, 5(7), 180319.

[7] Fhager, A., Candefjord, S., Elam, M., & Persson, M. (2018). Microwave diagnostics ahead: Saving time and the lives of trauma and stroke patients. IEEE Microwave Magazine, 19(3), 78-90.

[8] Pastorino, M. (2010). Microwave imaging (Vol. 208). John Wiley & Sons.

[9] Zhu, G., Bialkowski, A., Guo, L., Mohammed, B., & Abbosh, A. (2020). Stroke classification in simulated electromagnetic imaging using graph approaches. IEEE Journal of Electromagnetics, RF and Microwaves in Medicine and Biology, 5(1), 46-53.

[10] Cochran, W. T., Cooley, J. W., Favin, D. L., Helms, H. D., Kaenel, R. A., Lang, W. W., ... & Welch, P. D. (1967). What is the fast Fourier transform?. Proceedings of the IEEE, 55(10), 1664-1674.

[11] RF Wireless World. (n.d.). IFFT vs FFT-Difference between IFFT and FFT. https://www.rfwireless-world.com/Terminology/IFFT-vs-FFT.html

[12] Abdi, H., & Williams, L. J. (2010). Principal component analysis. Wiley interdisciplinary reviews: computational statistics, 2(4), 433-459

[13] Boser, B.E., Guyon, I.M. & Vapnik, V.N. A training algorithm for optimal margin classifiers. in 5th Annual ACM Workshop on COLT (ed. Haussler, D.) 144–152 (ACM Press, Pittsburgh, PA, 1992).

[14] Prajapati, G. L., & Patle, A. (2010, November). On performing classification using SVM with radial basis and polynomial kernel functions. In 2010 3rd International Conference on Emerging Trends in Engineering and Technology (pp. 512-515). IEEE.