Paper:

Identifying stroke types using machine learning method based on Multi-channel frequency domain electromagnetic signals

Shang Gao and Guohun Zhu

280-284 Sir Fred Schonell Dr, St Lucia QLD 4067
E-mail: shanegao1998@gmail.com
[Received 22/05/22; accepted 22/07/22]

Abstract. Stroke is an emergency and requires treatment within 3-6 hours. Therefore, rapid diagnosis of stroke subtypes is very important for timely clinical intervention. This study assesses four traditional machine learning methods with frequency domain features to classify stroke subtypes based on multichannel electromagnetic signals. Firstly, redundant data are removed from 256 channels scattering signals on each subject. Then 16 frequency domain features are extracted by a principle component analysis (PCA) method. Finally, features are forwarded into four classifiers to be built and used for stroke type prediction. The results indicate that the logistic regression classifier achieves better results than K-nearest neighbour, XGBoost and MLP neural network methods, achieving the highest accuracy (98%) while retaining one feature on each channel.

Keywords: Stroke; PCA; Machine Learning;Antenna array; Classification; Frequency Domain

1. Introduction

In the United States, approximately 795,000 people suffer from stroke each year, of which approximately 130,000 people die from it[1]. Stroke can cause brain tissue damage in different ranges and degrees, resulting in a variety of neuropsychiatric symptoms, and even life-threatening in severe cases. The length of time the brain is deprived of oxygen determines the degree of brain damage. That is, the less immediate treatment is, the more permanent or residual brain damage is left behind.

Subtypes of stroke can be divided into two main categories: Hemorrhagic(ICH) and Ischemic(IS). The main cause of ischemic stroke is cerebral vascular obstruction that prevents blood flow, which can be divided into cerebral thrombosis and cerebral embolism. Hemorrhagic stroke, also known as cerebral hemorrhage, is the rupture of a blood vessel in the brain, resulting in blood clots, which compress brain cells and affect blood supply. Strokes caused by blocked arteries account for about 85% of the total, and hemorrhagic strokes account for about 15% of the total[2]. Regardless of the type, a stroke is considered a fatal disease that requires urgent treatment. However, the key to rapid stroke diagnosis is not only to determine the location of the stroke, but also to accurately distinguish the type of stroke, so as to follow up with different treatment.

Generally, traditional medical stroke detection methods include CT and MRI. Nevertheless, some issues such as ionizing radiation and expensive time cost of them limit their practical use in most pre-clinical and non-clinical settings. For example, it should be noted that for relatively small ischemic strokes, it is generally not clear until 24 hours after the onset. In this case, the diagnosis should be confirmed by combining medical history, physical signs, head MRI, or CT review 24 hours or more later[3]. Furthermore, the disparities of regional development also lead to the imbalance in global health, as many people do not have access to abundant medical resources. Therefore, in recent years, researchers have been devoted to the development of imaging systems based on microwave imaging(MWI), as microwave technology has a series of advantages for being non-invasive, cost-effective, and transferable.

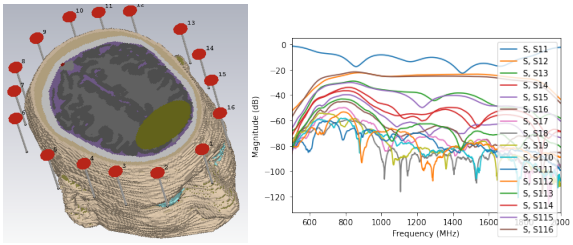
At present, feasible stroke identification methods for electromagnetic signals can be divided into two categories: microwave tomography (MWT) images or machine learning methods. MWT can reconstruct the complex permittivity image of the object by computing the inverse scattering, which is ill-posed, that is, the uniqueness and stability of the solution cannot be satisfied. In contrast, machine learning (ML) and deep learning (DL) algorithms circumvent traditional image formation, can quickly give prediction results, and has great potential to be deployed in mobile medical clinics, which have been explored in recent years[4]-[10]. A method to identify hemorrhagic stroke using stochastic deep neural network (DNN) was proposed in [8], but ischemic stroke was not considered in the experiment. Zhu et al. proposed a complex network (graph) method to identify two stroke subtypes and achieved 89% identification accuracy in mixed three groups with different signal-to-noise (SNR) ratios[5]. However, further research is still necessary to differentiate between hemorrhagic and ischemic stroke in medical applications.

This paper investigates stroke subtype classification performance with four rapid machine learning frameworks using frequency domain features. Firstly, for a simulated Vector Network Analyzer (VNA) consisting of

16 antenna arrays, the scattering parameter matrix is obtained and converted to dB values. After that, Principal Component Analysis (PCA) is used to extract the frequency domain features of different antenna pairs and combine them together. In the classification stage, four models with XGBoost (XGB), logistic regression (LR), and other classifiers are built with the extracted frequency features respectively. The best performance is evaluated by a 10-fold cross-validation.

2. Dataset Overview

In this study, the data are simulated from 10 digital brain models from electromagnetic analog signals generated by CST Microwave Studio. Those brain models have been used in [11], An antenna array of 16 dipole antennas is arranged around a real head model (the head model is scanned by MRI). Among them, each antenna acts as a transmitter and a receiver at the same time, with a frequency bandwidth of 0.5 GHz-2 GHz, a total of 751 frequency points, and sequentially transmits and captures scattered signal parameters[6]. The head model and simulated signals are demonstrated in Fig.1. Additionally, the simulation of stroke data for both labels is altered by changing the type of stroke clot (ICH or IS) and moving the clot location in the head model. Therefore, for each simulation case, there are 16 × 16 frequency-domain scattered signals, each containing 751 frequency points. Table.1 gives the specification and size of the dataset being used.



(a) (b)

Fig. 1. : (a) Emulated head model and 16 antennas, and (b) An example of frequency domain scattering parameters of port1.

Table 1. : The specification and size of dataset.

Type	ICH	IS	Total
Samples	83	82	165

3. Methodology

The flow chart for proposed stroke classification system is shown in Fig.2.

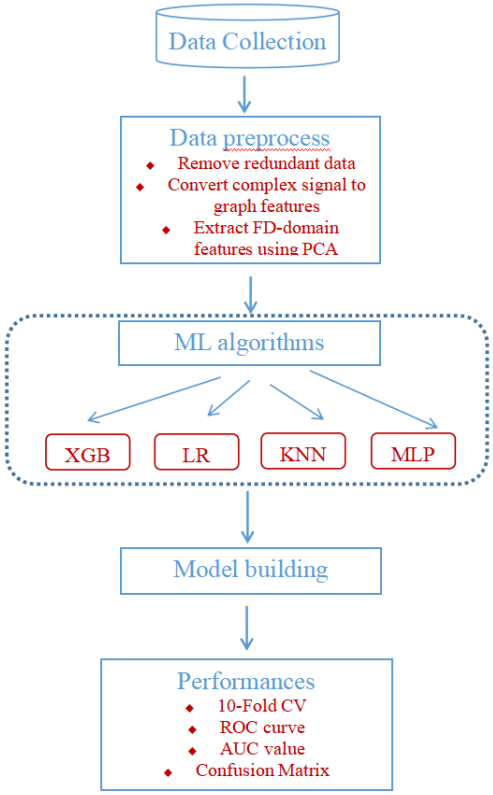


Fig. 2. : Stroke Classification procedures using multi-channel FD signals

3.1. Preprocess

Data preprocessing is required to remove redundant data from the dataset before model building. In the simulation case, the microwave signal transmission has no directionality for the reciprocal network, which is reflected in the S parameters as:

$$S_{ij} = S_{ji} \quad i,j = \{1,2,....16\} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

According to (1),for each simulation case, the upper triangular channel matrix (transmission gain), except the return loss of each antenna *Sii*, can be extracted. Thus, 120 channels can be retained from 16 × 16=256 signal channels. In addition, the transmission gain of each channel is used as the input of the subsequent procedures. It uses dB value as the unit for quantification, and the expression is as:

$$TG(S_{ij}) = 10\lg(S_{ij}) \quad i,j = \{1,2,....16\} \quad . \quad . \quad (2)$$

3.2. Feature selection

Principal component analysis(PCA) is a method of transforming a series of possibly linearly related variables into a set of linearly uncorrelated new variables using orthogonal transformation(principal components), so as to present the features of original dataset in a smaller dimension [12]. In this study, PCA is used separately for the

channels filtered out in preprocessing stage, in order to obtain the available frequency points, which can be combined for training in the next stage.

3.3. Machine Learning methods

3.3.1. XGBoost

XGBoost (XGB) is one of the implementations of Gradient Boosting Machine (GBM), and its basic idea is similar to GBDT. Different from the random forest giving each decision tree the same voting weight, the generation of the next decision tree in the XGB algorithm is related to the training and prediction of the previous decision tree (that is, the accuracy rate of the previous decision tree training is low. The samples are assigned higher learning weights to improve the model accuracy)[13]. Compared with other ensemble learning algorithms, on the one hand, XGBoost improves the robustness of the model by introducing regular terms and column sampling, and on the other hand, it adopts a parallelization strategy when selecting a split point for each tree, which greatly improves the running speed.

3.3.2. Logistic Regression

Logistic regression assumes that the data obeys Bernoulli distribution, and uses gradient descent to solve the parameters by maximizing the likelihood function to achieve the purpose of classifying the data into two categories[14]. Logistic regression is applicable because of the balance of positive and negative data samples obtained from simulation. Below Fig. 3 shows an image of the sigmoid function used in logistic regression.

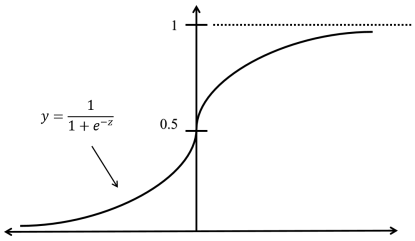


Fig. 3. : An example of the sigmoid function used for logistic regression classification, where the classification threshold is 0.5

3.3.3. K Nearest Neighbour

The k-nearest neighbor algorithm measures the distance between different features and classifies the sample points with close distances into the same category[15]. The key to KNN algorithm includes the selection of K nearest neighbors (K value) and the distance between point x and y in the multi-dimensional space. The distance between x and y uses the default Euclidean distance, given by equation(3). In this study, the value of K is implemented using *sklearn.model_selection.GridSearchCV* without manually specifying hyperparameters.

$$\mathrm{d}(\mathrm{x},\mathrm{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad \ldots\ldots\ldots \quad (3)$$

3.3.4. MLP Neural Network

A multilayer perceptron can learn an association (usually linearly inseparable) between features and objects for classification or regression purposes[16]. In this study, the MLP with one hidden layer is used, and the number of neurons in the hidden layer is determined with reference to the principles proposed by K. Gnana Sheela and S. N. Deepa[17]. An example of the network structure used is given in Fig.4 below, where the number of output units is 1.

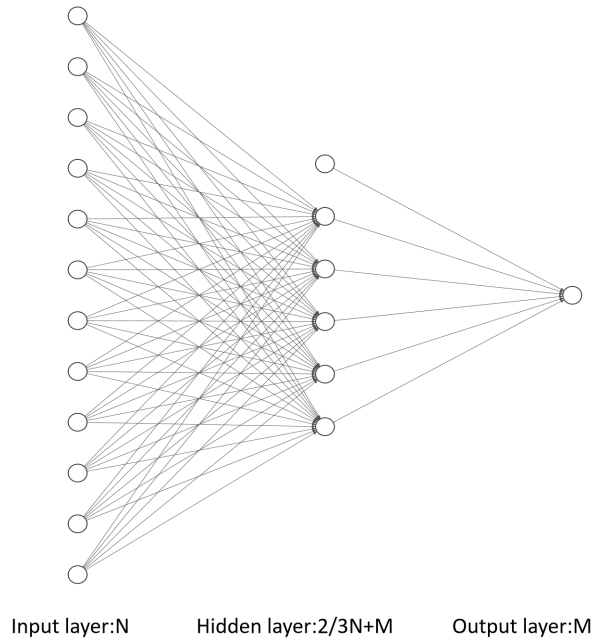


Fig. 4. : MLP neural network architecture diagram used in this study

4. Experiment result and Analysis

This experiment and result presentation are implemented in Python language using libraries including *sklearn*,*numpy*,*matplotlib*, etc.

4.1. Feature Extraction Results

For the 120-channel frequency-domain signals retained after data preprocessing, principal component analysis(PCA) is used to reduce feature dimensionality of each channel and filter out important frequency-domain features before being fed to training models. Feature ex-

traction result shows that majority the variance after dimensionality reduction (more than 90%) can be explained by the first and the second principal component independently for each channel. Fig.5. demonstrates the feature importance ratio of each channel, that is, the explained variance ratio of each feature.

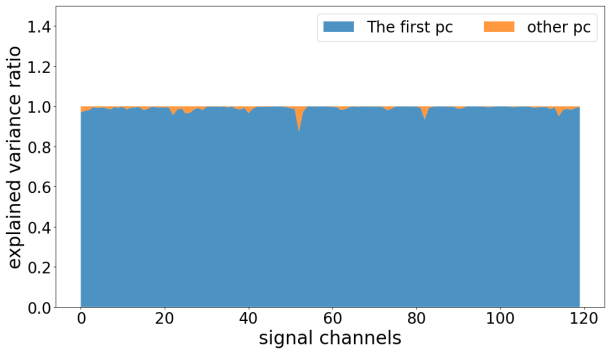


Fig. 5. : Feature importance ratio of selected channels

4.2. Classification result

In the classification stage, the feature extraction results are used as the input of the four classifiers.The dataset used is divided into training set and testing set with a ratio of 7:3. For each channel, one principal component and two principal components are retained for evaluation, and their 10-fold cross-validation scores on the training set are presented in Fig.8.

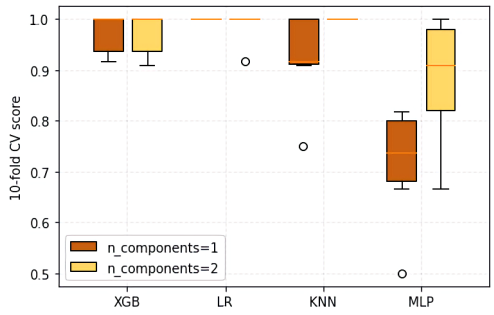


Fig. 6. : 10-Fold Cross Validation scores on training set

The performances of the four classifiers mentioned above on the test dataset are demonstrated in Table 2. The evaluation criteria are based on *accuracy_score* in *sklearn*. The experimental results indicate that, compared with other classifiers, the LR classifier achieves the highest recognition accuracy of 98% when each channel retains one principal component. In the cases where each channel retains two features, the classification performance of each classifier reaches reaches or even exceeds 80%, which proves the feasibility of adopted frequency-domain channel feature selection approach.

To further evaluate the results, the confusion matrices of these classifiers are illustrated in Fig.7. Receiver Operating Characteristic(ROC) curve and Area Under the Curve (AUC) of classifiers obtained by taking the number of feature retentions (per channel) of the best performing models on the test set are presented in Fig.8. The AUC values of the four classifiers all reach above 0.8, which indicates that proposed frequency domain feature selection method is effective. Specifically, compared to other classifiers, LR model shows the best classification performance (AUC=0.98) and is more sensitive to identify stroke types than others. MLP has the lowest sensitivity for identifying stroke, and has a higher false positive rate (FPR) than other classifiers, which can be attributed to the fact that the neural network only contains one hidden layer, thus it is too simple to identify deep patterns. In future work, possible performance improvements include increasing the number of hidden layers or using deep neural network(DNN).

Table 2. : Accuracy obtained using four classifiers, where *n_components* represent the number of principal components retained in each channel.

Classifer	feature=1	feature=2
XGB	0.52	0.88
LR	0.98	0.90
KNN	0.62	0.90
MLP	0.62	0.80

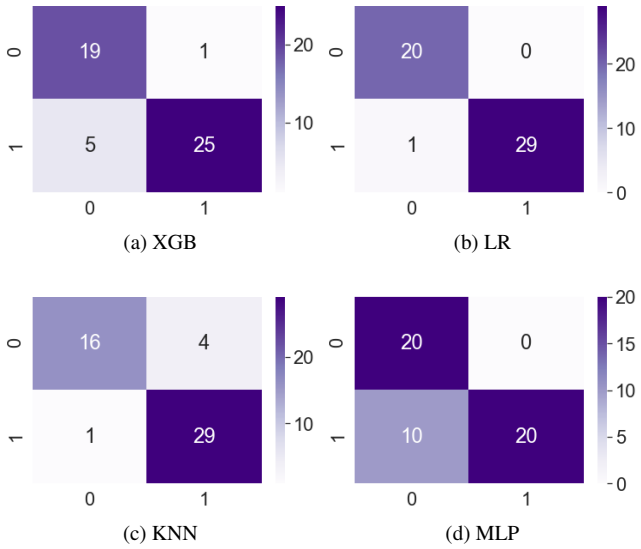


Fig. 7. : Confusion matrix of classifiers on test set

5. Conclusion

This study shows an effective brain stroke subtypes classification using multi-channel frequency-domain electromagnetic signals with machine learning algorithms. The proposed method achieves 98% classification

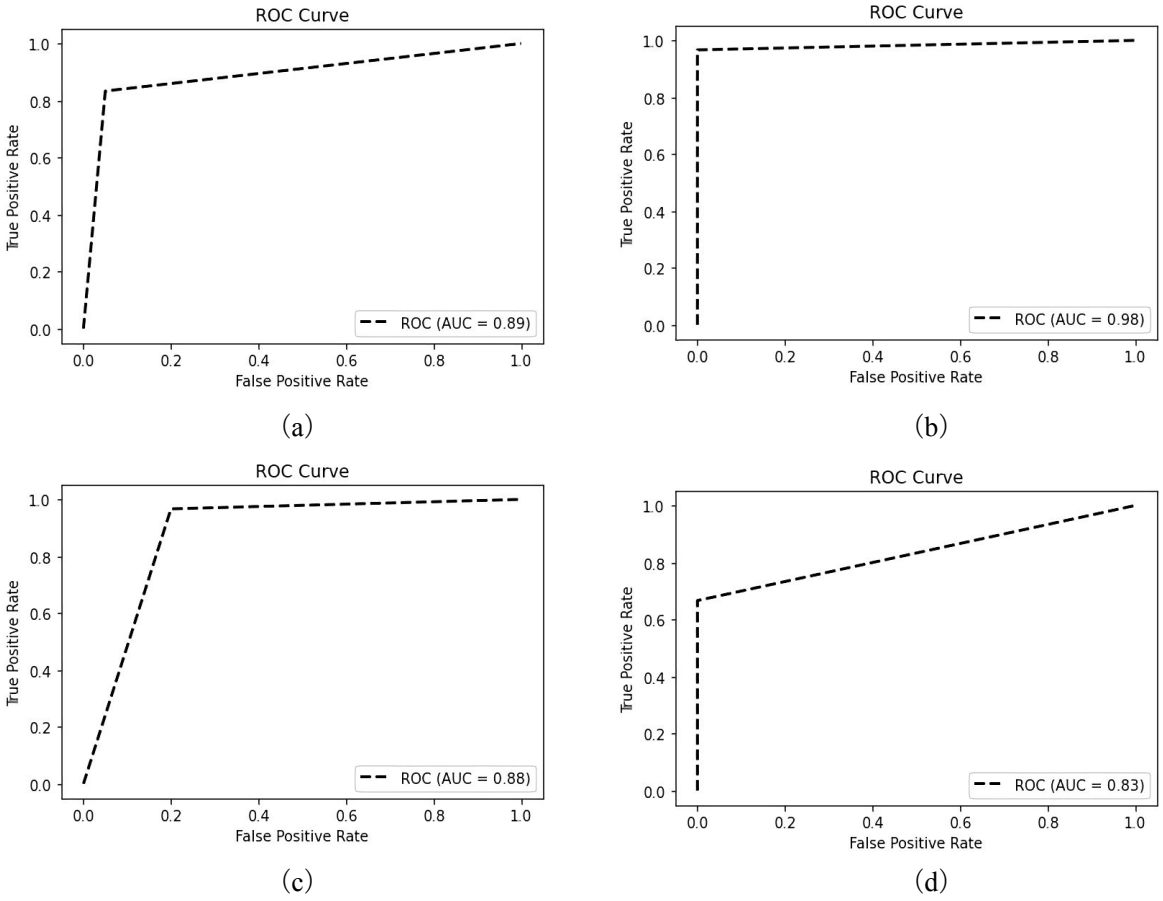


Fig. 8. : ROC curve and AUC of classifers

accuracy using Logistic Regression on the test set without the step of converting to the time domain, and has great potential to be deployed on mobile ambulances for rapid stroke diagnosis due to the portability and offline training nature.

However, this study still suffers from a limited data sample, as it usually takes a long time (on the order of a day) for CST studio to simulate a stroke data. Nonetheless, the proposed method remains promising for the urgent need for emergency diagnosis of stroke subtypes. To improve this problem, more human brain models of different sizes and stroke blood clot models will be considered in the next stage for training to improve the generality and practicality of building machine learning models. In addition, experimental data instead of simulated data will also be considered for training, since the noise in the process of experimental data collection will become a challenge in using machine learning to assist medical diagnosis.

References:

[1] Tsao CW, Aday AW, Almarzooq ZI, Alonso A, Beaton AZ, Bittencourt MS, et al. Heart Disease and Stroke Statistics—2022 Update: A Report From the American Heart Associationexternal icon. Circulation. 2022;145(8):e153–e639

[2] Stroke - Family Caregiver Alliance. (2022). Retrieved 20 May 2022, from https://www.caregiver.org/resource/stroke/

[3] Lin, M. P., and Liebeskind, D. S. (2016). Imaging of ischemic stroke. Continuum: Lifelong Learning in Neurology, 22(5), 1399.

[4] Xu, J., Chen, J., Yu, W., Zhang, H., Wang, F., Zhuang, W., ... and Chen, M. (2020). Noninvasive and portable stroke type discrimination and progress monitoring based on a multichannel microwave transmitting–receiving system. Scientific Reports, 10(1), 1-13.

[5] Li, J., Zhu, G., and Xi, M. Automating Stroke Subtype Classification from Electromagnetic Signals Using Principal Component Methods. The 7^{th} International Workshop on Advanced Computational Intelligence and Intelligent Informatics (IWACIII2021) Beijing, China, Oct.31-Nov.3, 2021

[6] G. Zhu, A. Bialkowski, L. Guo, B. Mohammed and A. Abbosh, "Stroke Classification in Simulated Electromagnetic Imaging Using Graph Approaches," in IEEE Journal of Electromagnetics, RF and Microwaves in Medicine and Biology, vol. 5, no. 1, pp. 46-53, March 2021, doi: 10.1109/JERM.2020.2995329.

[7] Jayachitra, S., and Prasanth, A. (2021). Multi-feature analysis for automated brain stroke classification using weighted Gaussian naïve Bayes classifier. Journal of Circuits, Systems and Computers, 30(10), 2150178.

[8] Alon, L., and Dehkharghani, S. (2021). A stroke detection and discrimination framework using broadband microwave scattering on stochastic models with deep learning. Scientific Reports, 11(1). doi: 10.1038/s41598-021-03043-y

[9] Salucci, M., Polo, A., and Vrba, J. (2021). Multi-step learning-by-examples strategy for real-time brain stroke microwave scattering data inversion. Electronics, 10(1), 95.

[10] Sailasya, G., and Kumari, G. L. A. (2021). Analyzing the performance of stroke prediction using ML classification algorithms. International Journal of Advanced Computer Science and Applications, 12(6).

[11] Guohun Zhu. (2019). DATASET for "Stroke Classification in Electromagnetic Imaging Using Graph Approaches". IEEE Dataport. https://dx.doi.org/10.21227/es43-bf10

[12] Abdi, H., and Williams, L. (2010). Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2(4), 433-459. doi: 10.1002/wics.101

[13] Chen, T., and Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

[14] Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley Sons.

[15] Peterson, L. E. (2009). K-nearest neighbor. Scholarpedia, 4(2), 1883.

[16] Pal, S. K., and Mitra, S. (1992). Multilayer perceptron, fuzzy sets, classifiaction.

[17] Sheela, K., Deepa, S. (2013). Review on Methods to Fix Number of Hidden Neurons in Neural Networks. Mathematical Problems In Engineering, 2013, 1-11. doi: 10.1155/2013/425740.