Computational Biology and Chemistry 88 (2020) 107316

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/cbac

Research Article

A machine learning approach to select features important to stroke prognosis

Gang Fang[a,*], Wenbin Liu[a], Lixin Wang[b]

[a] Institute of Computing Science and Technology, Guangzhou University, Guangzhou 510006, China
[b] Departments of Neurology, Guangdong Province Traditional Chinese Medical Hospital, Guangzhou 510120, China

ARTICLE INFO

Keywords:
Machine learning
Ischemic stroke
Feature Selection
IST

ABSTRACT

Ischemic stroke is a common neurological disorder, and is still the principal cause of serious long-term disability in the world. Selection of features related to stroke prognosis is highly valuable for effective intervention and treatment. In this study, an integrated machine learning approach was used to select the features as prognosis factors of stroke on The International Stroke Trial (IST) dataset. We considered the common problems of feature selection and prediction in medical datasets. Firstly, the importance of features was ranked by the Shapiro-Wilk algorithm and the Pearson correlations between features were analyzed. Then, we used Recursive Feature Elimination with Cross-Validation (RFECV), which incorporated linear SVC, Random-Forest-Classifier, Extra-Trees-Classifier, AdaBoost-Classifier, and Multinomial-Naïve-Bayes-Classifier as estimator respectively, to select robust features. Furthermore, the importance of selected features was determined by Random-Forest-Classifier and Shapiro-Wilk algorithm. Finally, twenty-three selected features were used by SVC, MLP, Random-Forest, and AdaBoost-Classifier to predict the RVISINF (Infarct visible on CT) of acute stroke on IST dataset. It was suggested that the selected features could be used to infer the long-term prognosis of acute stroke at a high accuracy, and it also could be used to extract factors related to RVISINF, which is associated with large artery occlusion (LAO) in ischemic stroke patient.

1. Introduction

Stroke has become the second leading cause of death worldwide. It is predicted that by 2030, there could be almost 12 million stroke deaths, 70 million stroke survivors, and more than 200 million disability-adjusted life-years (DALYs) lost from stroke each year (Feigin et al., 2014). Stroke burden in high-income countries is very heavy, and the burden of stroke increases rapidly in low-income and middle-income countries in recent years with the rapid development of social economy (Kim et al., 2015). Stroke prognosis prediction can contribute significantly to its effective intervention and treatment. Numerous medical studies and data analyses have been conducted to identify effective predictors of stroke and its prognosis. The Framingham Study (Dawber et al., 1951; Wolf et al., 1991) reported a list of stroke risk factors including age, systolic blood pressure, and the use of anti-hypertensive therapy, diabetes mellitus, cigarette smoking, prior cardiovascular disease, atrial fibrillation, and left ventricular hypertrophy by electrocardiogram. Furthermore, more other studies (Longstreth et al., 2001; McGinn et al., 2008; Manolio et al., 1996) have led to the discovery of more risk factors such as creatinine level, time to walk 15 feet, and others. Most previous prediction models have adopted features (risk factors) that are verified by clinical trials or selected manually by medical experts. For example, Lumley et al. (2002) built a 5-year stroke prediction model based on the Cardiovascular Health Study (Fried et al., 1991) dataset using a set of 16 manually selected features (given in (Manolio et al., 1996)) from a total of roughly one thousand features. With a large number of features in current medical datasets, it is a very tough task to identify and verify each risk factor manually. Now, machine learning algorithms are capable of identifying features highly related to stroke occurrence efficiently from the huge set of features; therefore, we believe machine learning can be used to improve the prediction accuracy of stroke risk and its prognosis and discover new prognosis factors.

The Cox proportional hazards model has ever been one of the most commonly used statistical methods in medical research (Bender et al., 2005). It has been extensively studied (Akazawa and Nakamura, 1991) and applied to the prediction of various diseases including stroke (Liang et al., 1990). However, the performance of the original Cox model depends heavily on the quality of the pre-selected features. To address this problem, several approaches have been proposed (Park and Hastie,

* Corresponding author.
E-mail address: gangf@gzhu.edu.cn (G. Fang).

https://doi.org/10.1016/j.compbiolchem.2020.107316
Received 22 May 2020; Accepted 21 June 2020
Available online 23 June 2020

2007). Thus far, there have been a few studies on machine learning methods in making predictions on censored medical data that outperformed traditional statistical methods. Kattan (2003) compared Cox proportional hazards regression with several machine learning methods (neural networks and tree-based methods) based on three urological datasets. However, Kattan's study focused on datasets with only five features, while machine learning algorithms are expected to effectively deal with a large number of features. In 2010 Aditya et al. (Khosla et al., 2010) presented a machine learning approach for stroke risk prediction. They investigated machine learning algorithms to improve the prediction accuracy and conducted extensive comparisons between their results and those with the traditional statistical methods. But they didn't carry out prognosis factor analyzing. With the rapid development of machine learning method and theory in recent years, more powerful and effective integrated methods have been developed (Han and Liu, 2019). These methods not only outperformed the traditional statistical methods, but also found new problems and solved them. Recently, Stephen et al. (Weng et al., 2017) and JoonNyung et al. (Heo et al., 2019) presented modern machine learning based model for prediction of stroke risk and prognosis. In their work, random forest, gradient boosting machines and particularly the deep neural network were used and the accuracy of prediction was significantly increased. In this paper, improved Recursive Feature Elimination with Cross-Validation (RFECV) was used to select features which would influence and determine the prognosis of stroke.

2. Material and methods

2.1. Data collection

The dataset analyzed in this paper was downloaded from The International Stroke Trial (IST) website. IST was conducted between 1991 and 1996 (including the pilot phase between 1991 and 1993). It was a large, prospective, randomized controlled trial, with 100 % complete baseline data and over 99 % complete follow-up data. The aim of the trial was to establish whether early administration of aspirin, heparin, both or neither influenced the clinical course of acute ischemic stroke (Sandercock et al., 2011). The patients in this trial were treated more than 20 years ago, and many have died. Patients and hospitals are identified only by an anonymous code; there are no identifying data such as name, address or social security numbers; patient age has been rounded to the nearest whole number. In our opinion, usage of the dataset clearly presents no material risk to confidentiality of study participants.

The dataset includes the following baseline data: age, gender, time from onset to randomization, presence or absence of atrial fibrillation (AF), aspirin administration within 3 days prior to randomization, systolic blood pressure at randomization, level of consciousness and neurological deficit. The deficits were classified as one of the Oxfordshire Community Stroke Project (OCSP) categories: total anterior circulation syndrome (TACS), partial anterior circulation syndrome (PACS), posterior circulation syndrome (POCS) and lacunar syndrome (LACS). Nineteen thousand four hundred and thirty five patients from 467 hospitals in 36 countries were randomized within 48 h of symptoms onset, of whom 13,020 had a CT before randomization, 5569 were first scanned after randomization and 846 were not scanned at all. We deleted entries with missing data, and then 18,128 entries were left. The data of these 18,128 patients were used to select robust features for stroke prognosis prediction.

2.2. Workflow

Given an external estimator that assigns weights to features (e.g., the coefficients of a linear model), recursive feature elimination (RFE) is to select features by recursively considering smaller and smaller sets of features. First, the estimator is trained on the initial set of features and the importance of each feature is obtained either through a coef_attribute or through a feature_importances_attribute. Then, the least important features are pruned from current set of features. That procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached. RFECV performs RFE in a cross-validation loop to find the optimal number of features (Pedregosa et al., 2011). The integrated machine learning approach RFECV used in the study adopted linear SVC, Random-Forest-Classifier, Extra-Trees-Classifier, AdaBoost-Classifier, and Multinomial-Naïve-Bayes-Classifier as its estimator respectively. In this study, a step by step Recursive Feature Elimination (RFE) algorithm was carried out with automatic tuning of the number of features selected with cross-validation.

Firstly, features collected at the beginning of and on 14 days of randomization were used. Some features, such as date information and comments, were deleted manually (these features apparently are not related to the prognosis of stroke). The features of 6th month outcome were kept as the target of the dataset. Some features collected on 6th month overlapped with 6th month outcome were removed manually. Then, fifty-four features were kept. Now, we want to know which features are important to 6th month outcome (long-term prognosis) of acute ischemic stroke.

Secondly, an integrated machine learning approach of RFECV was constructed. Linear SVC, Random-Forest-Classifier, Extra-Trees-Classifier, AdaBoost-Classifier, and Multinomial-Naïve-Bayes-Classifier were given as external estimators. Feature selections were carried out by primary RFECV with its estimators respectively. These estimators were evaluated and the best one would be kept in next RFECV. Shapiro-Wilk algorithm and Pearson correlation analysis were carried out to assess the importance of features. The least important features were eliminated, and then features which are related and important to 6th month outcome (long-term prognosis) were selected by RFECV firstly. Next steps, features related and important to RVISINF (Infarct visible on CT) were selected. After these, the selected features were ranked by Random-Forest-Classifier which performed better than other estimators.

Thirdly, the selected features were used by classifiers to predict RVISINF (Infarct visible on CT) of acute ischemic stroke on IST dataset (Fig. 1).

3. Methods and results analyzing

3.1. Initial RFECV

Initially, the machine learning approach of RFECV selected some features that important to 6th month outcome (abbreviate OCCODE)

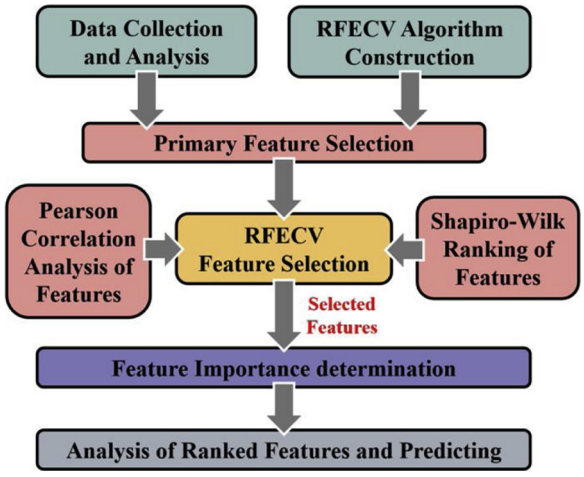

Fig. 1. Workflow of the method.

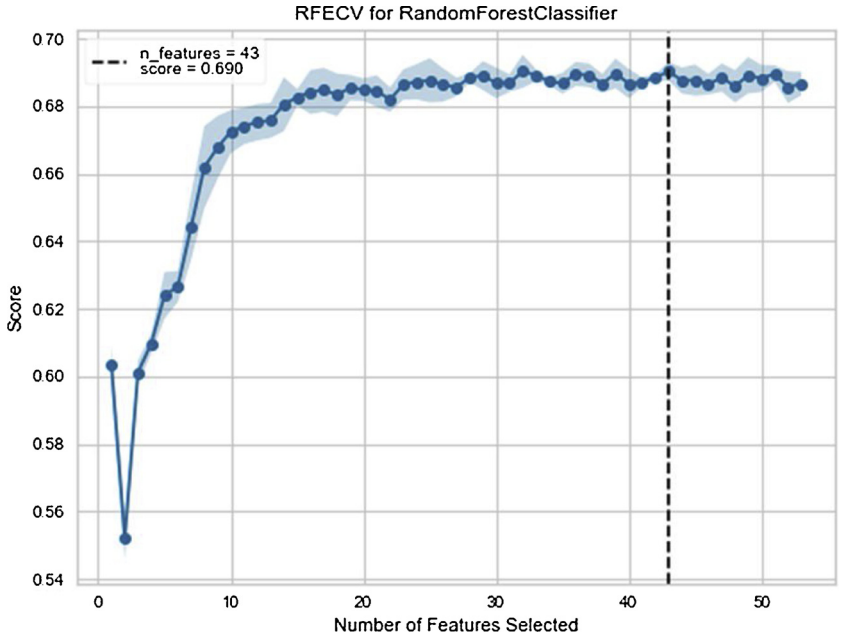

Fig. 2. Performance of RFECV for RandomForestClassifier to select features important to OCCODE.

from all 53 features by using different estimators. The RandomForestClassifier outperformed others by selecting 43 features, and got the highest f1 score of 0.690 (Fig. 2 and Supplementary Figures Fig. 1). Compared with other feature selection methods, RFECV also outperformed them by attaining the highest f1 score —— 0.690.

The initial RFECV excluded DSCH (Non trial subcutaneous heparin), DIVH (Non trial intravenous heparin), DCAREND (Carotid surgery), DTHROMB (Thrombolysis), DMAJNCH (Major non-cerebral hemorrhage), DDIAGHA (Final diagnosis of ischemic stroke), DDIAGUN (Final diagnosis of Indeterminate stroke), DRSISC (Ischemic recurrent stroke within 14 days), DRSH (Hemorrhagic recurrent stroke within 14 days), and DPE (Pulmonary embolism within 14 days). The selected features included HOSPNUM (Hospital number), RDELAY (Delay between stroke and randomization in hours), RCONSC (Conscious state at randomization), SEX (sex), AGE (age), RSBP (Systolic blood pressure at randomization), RSLEEP (Symptoms noted on waking), CNTRYNUM (Country code), RATRIAL (Atrial fibrillation), and et al. But the performance of this coarse RFECV method is not satisfactory (its f1 score is only 0.690, Fig. 2).

In order to improve the RFECV method, the feature importance was firstly analyzed by Shapiro-Wilk algorithm (Shapiro and Wilk, 1965). The algorithm was improved by Royston to process large data (Royston, 1982). It was utilized to assess the normality of the distribution of instances with respect to the feature. All 53 features except OCCODE were ranked by the algorithm (Fig. 3). Considering the initial RFECV analyzing, all the results showed that CNTRYNUM is one of the most important feature related to the 6th month outcome. It was suggested that different country (developing and developed country in 1990's) made the difference of outcome. In that time a developed country always provided better medical facilities and made different outcomes. To eliminate this bias, 5998 entries from 4 major English speaking countries (UK, Ireland, US, Canada) were included in the next feature selection.

3.2. Refined RFECV

Firstly, the feature importance of this smaller dataset was ranked by Shapiro-Wilk algorithm. The result showed that the feature CNTRYNUM was less important in the dataset (Fig. 4). Then, the RFECV (with default parameters) algorithm with RandomForestClassifier was carried out to select features important to OCCODE. Forty-nine features were selected from all 53 features by attaining the f1 score of 0.792 (Fig. 5).

The result was not very informative for selecting too many features. To refine the result, the correlation between features was analyzed by calculating their Pearson correlation coefficient (Fig. 6). Combining with the Shapiro ranking of features in this dataset, the results showed that DTHROMB, DCAREND, DHAEMD and DGORM were the least four important features and had strong colinearity correlation (Figs. 4 and 6). These four features were eliminated (The primary RFECV algorithm of this dataset eliminated DTHROMB, DCAREND, DHAEMD and DRSH). Then RFECV (set the parameter *step* = 2) was carried out to reselect important features. The result showed that 35 features were selected by attaining the f1 score of 0.794 (Fig. 7).

The thirty-five selected features included HOSPNUM, RDELAY, RCONSC, SEX, AGE, RSLEEP, RATRIAL, RCT, RVISINF, RASP3, RSBP, RDEF1, RDEF2, RDEF3, RDEF4, RDEF5, RDEF6, RDEF7, RDEF8, STYPE, RXASP, RXHEP, DASP14, DASPLT, DLH14, DMH14, ONDRUG, DOAC, DCAA, DDIAGISC, DDIAGUN, DPLACE, FPLACE, CNTRYNUM, and TD. The result showed that the feature CNTRYNUM was still important to OCCODE. In order to eliminate all the influence of this feature, all 5713 samples (patients) from UK were used to select features important to stroke prognosis. The another important reason for only using samples from UK was that The International Stroke Trial (IST) was leaded and conducted by University of Edinburgh and Western General Hospital from UK. All the samples were firmly controlled and the data were accurately recorded. Firstly, the features were ranked by Shapiro-Wilk algorithm and the Pearson correlation between them was analyzed (Supplementary **Fig. 2**). In the dataset, DTHROMB, DCAREND, DHAEMD and DGORM were also the least four important features and had strong colinearity correlation (Supplementary **Fig. 2**). These four features were deleted, then RFECV (set the parameter *step* = 1) was carried out. The result showed that 30 features were selected by attaining the f1 score of 0.809 (Fig. 8). The shaded area in Figs. 7 and 8 represented the variability of cross-validation, one standard deviation above and below the mean accuracy score drawn by the curve.

3.3. Feature importance determination and prediction

The thirty selected features in the dataset included HOSPNUM,

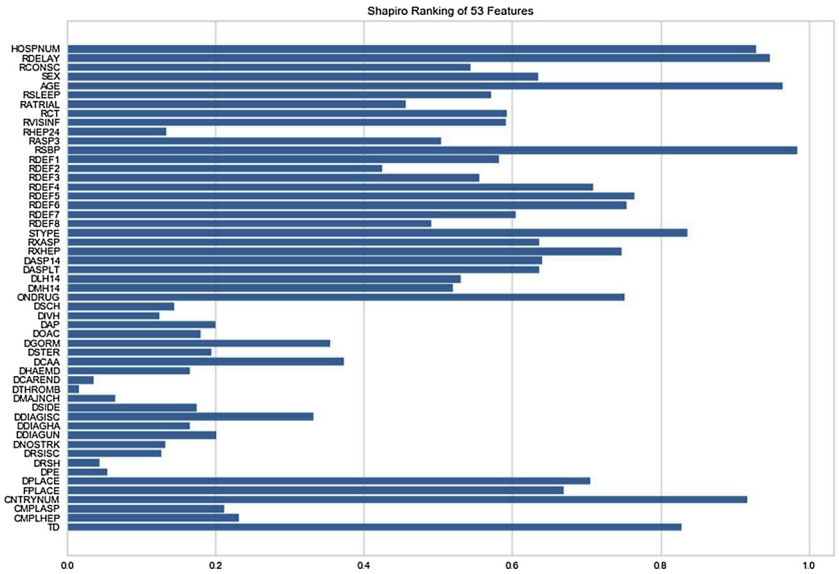

Fig. 3. Importance of features ranked by Shapiro-Wilk algorithm.

RDELAY, RCONSC, SEX, AGE, RSLEEP, RATRIAL, RCT, RVISINF, RASP3, RSBP, RDEF1, RDEF2, RDEF3, RDEF4, RDEF5, RDEF6, RDEF7, STYPE, RXASP, RXHEP, DASP14, DASPLT, DLH14, DMH14, ONDRUG, DDIAGISC, DPLACE, FPLACE, and TD. The importance of these features were firstly ranked by Shapiro-Wilk algorithm and then by RandomForestClassifier (Figs. 9 and 10). When ranked by RandomForestClassifier, the feature importance was assessed by computing the differences of out of bag errors in every decision tree of the Random Forest. The importance of each feature is determined by formula (1). In the formula, errOOB1 means error of out of bag data in *ith* decision tree and errOOB2 means error of out of bag data with noises in feature M. 'n' is the number of decision trees in Random Forest.

$$\text{Feature M importance} = \sum_{i=1}^{n} \frac{(errOOB2 - errOOB1)}{n} \quad (1)$$

The Shapiro-Wilk algorithm assesses the normality of the distribution of instances with respect to the feature. Because the RandomForestClassifier was utilized to predict the 6th month outcome of stroke in this paper, the feature importance ranked by RandomForestClassifier was used to analyze the results.

In Fig. 10, the results showed that TD (Time of death or censoring in days), FPLACE (Place of residence at 6 month follow-up), ONDRUG (Estimate of time in days on trial treatment), AGE (age), HOSPNUM (Hospital number), RDELAY (Delay between stroke and randomization in hours), RSBP (Systolic blood pressure at randomization) and DPLACE (Discharge destination) were most important. TD, FPLACE, ONDRUG, HOSPNUM, RDELAY, and DPLACE were features associated with medical facilities. It was suggested that active treatment and better medical condition (different hospital) made the difference of outcome. RSBP was another most important feature. It was suggested that systolic blood pressure controlling was important to stroke prognosis. It was not surprised that AGE was selected (the older the worse healthy condition and worse outcome). In other selected features, DASPLT (Discharged on long term aspirin), RASP3 (Aspirin within 3 days prior to randomization), DASP14 (Aspirin given for 14 days or till death or discharge), and RXASP (Trial aspirin allocated) were features associated with aspirin using. It was suggested that aspirin usage was important to ischemic stroke prognosis. RXHEP (Trial heparin allocated), DLH14 (Low dose heparin given for 14 days or till death/discharge), and DMH14

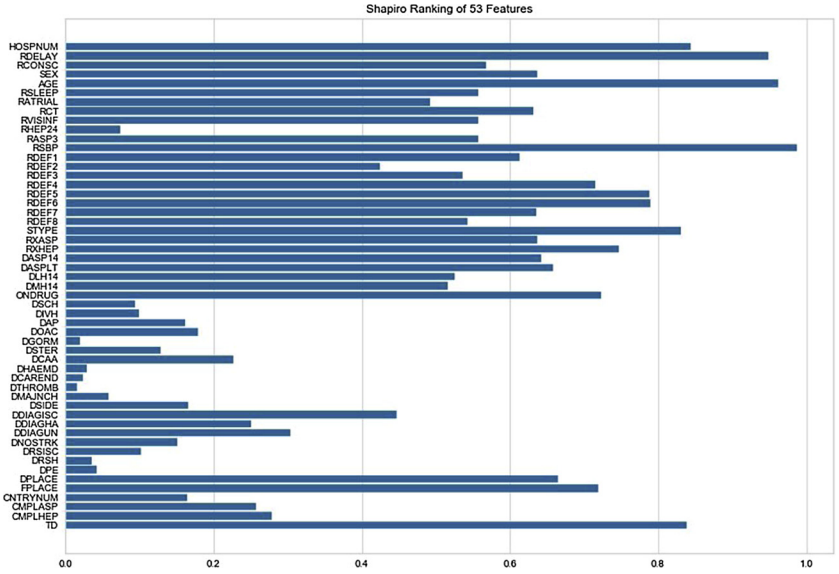

Fig. 4. Importance of features ranked by Shapiro-Wilk algorithm in 4 major English speaking countries dataset.

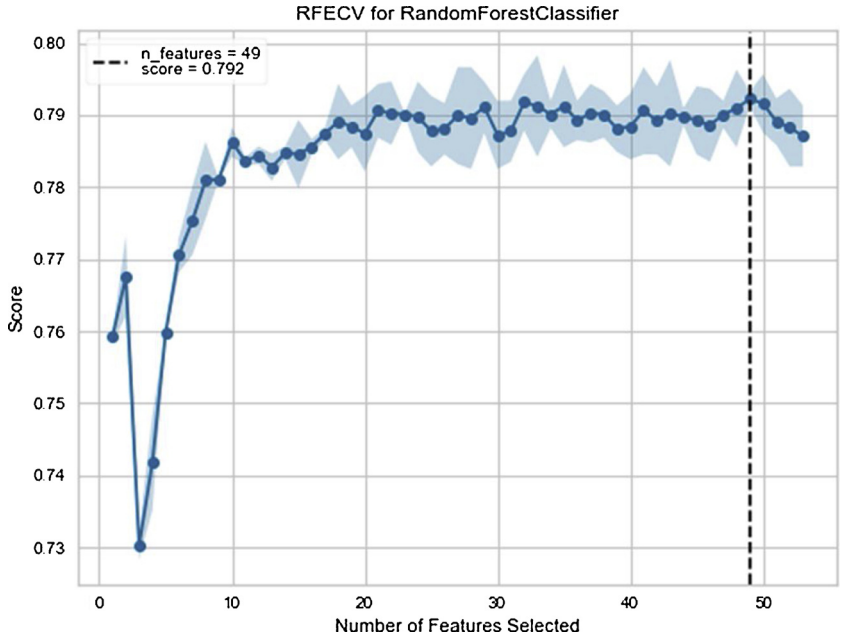

Fig. 5. Primary RFECV for RandomForestClassifier to select features important to OCCODE in 4 major English speaking countries dataset.

(Medium dose heparin given for 14 days or till death/discharge) were features associated with heparin using. But RHEP24 (Heparin within 24 h prior to randomization) was excluded at the primary RFECV selection. It was suggested that properly heparin using was important to ischemic stroke prognosis. STYPE (Stroke subtype), RDEF1 (Face deficit)……RDEF7 (Brainstem/cerebellar signs), RCONSC (Conscious state at randomization), RVISINF (Infarct visible on CT), RSLEEP (Symptoms noted on waking), and RATRIAL (Atrial fibrillation) were all clinical signs collected at the beginning of the trial. These were all selected by the algorithm as the important features. DDIAGISC (Ischemic stroke) was the final diagnosis of initial event, and it also influenced the OCCODE. SEX (sex) was another testified feature important to OCCODE (Li et al., 2016). RCT (CT before randomization) was another feature associated with medical facilities. Considering the excluded features, DPE was always believed to be a major reason led to high mortality. But in this study it was excluded from the major factors that related to 6th month outcome of acute ischemic stroke. The reason of this could be intensive care of most randomized patients, especially in UK. But

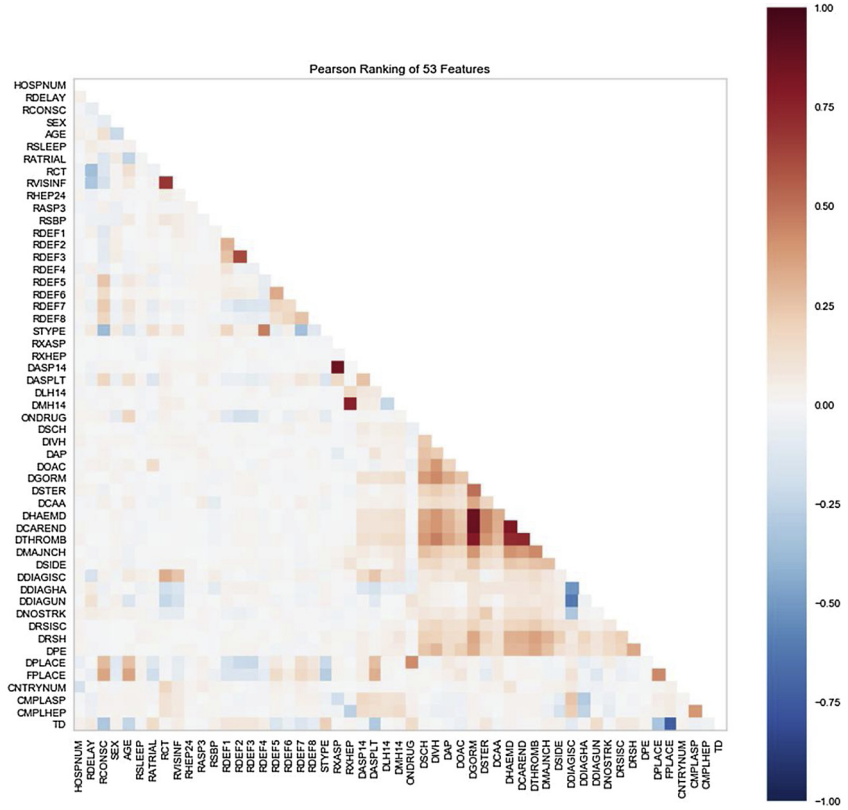

Fig. 6. Pearson correlations between features in 4 major English speaking countries dataset.

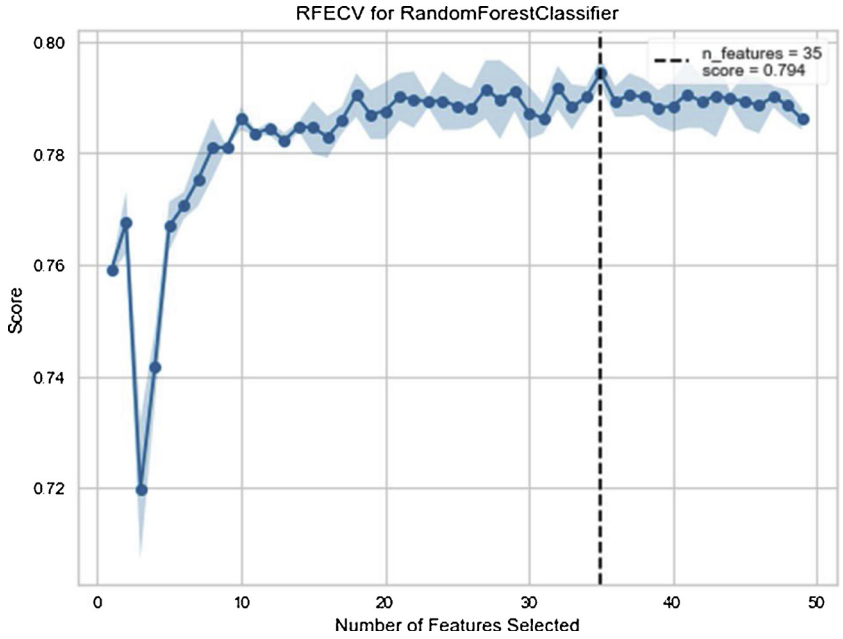

Fig. 7. RFECV for RandomForestClassifier to refine the primary result in 4 major English speaking countries dataset.

pulmonary embolism is still a serious complication and has more chance led to death. Today, DCAREND (Carotid surgery) and DTHROMB (Thrombolysis) are considered as effective therapies to cure ischemic stroke. But the two features were excluded; the reason of this was that these therapies were seldom carried out in 1990's.

Now, to identify candidates for thrombectomy are of utmost importance in acute stroke. No prognostic tool has yet gained any widespread use, which can predict large artery occlusion (LAO) (Cooray et al., 2018). In this study, the feature of RVISINF was associated with LAO. Factors related to it can be extracted by machine learning method. Firstly, features directly related and less related to RVISINF were deleted manually. For an example, according to Pearson correlations analyzing RCT (CT before randomization) was directly related to RVISINF and HOSPNUM was less related to it (Fig. 6 and Supplementary Figures Fig. 2). Forty-five features were left to be processed for selecting features to predict RVISINF that related to LAO. RFECV with RandomForestClassifier which outperformed other estimators was performed to select these features (Fig. 11). Twenty-three selected features included RDELAY, RCONSC, SEX, AGE, RSLEEP, RATRIAL, RASP3, RSBP, RDEF1, RDEF3, RDEF4, RDEF5, RDEF6, RDEF7, RDEF8, STYPE, RXASP, RXHEP, DASP14, DASPLT, DLH14, DMH14, and DDIAGISC. Then these features were ranked by Random Forest and can be used to extract factors to predict LAO (Fig. 12). At last, different classifiers were used to predict RVISINF in UK dataset, and ROC curves and AUC (areas under the curves) were presented in Supplementary file (Supplementary Figures Fig. 3, Fig. 4, Fig. 5, Fig. 6).

In these 23 selected features, DDIAGISC (Ischemic stroke) was ranked higher than before (Figs. 12 and 10). It meant that DDIAGISC was important to RVISINF that related to LAO. RDEF2 (Arm/hand deficit) was excluded and RDEF8 (Other deficit) was kept in these

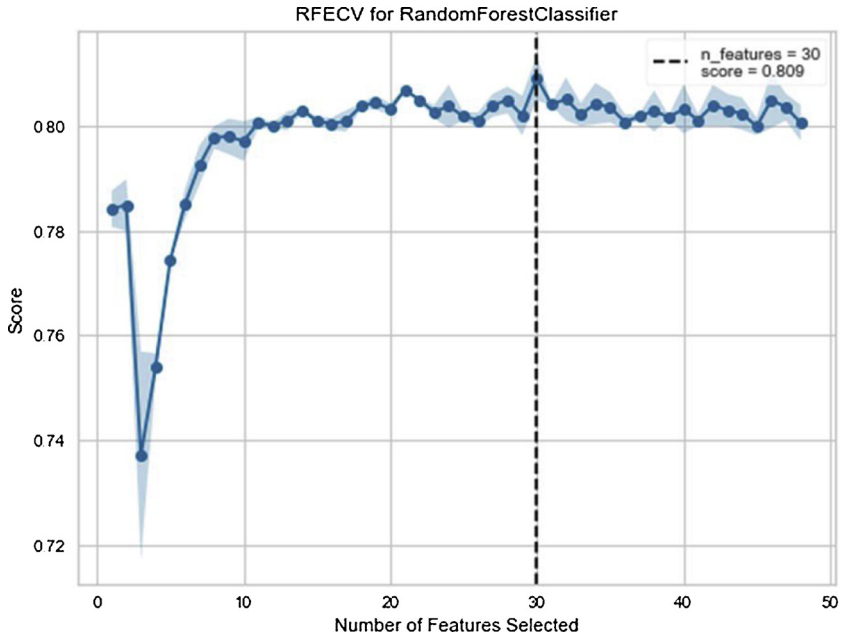

Fig. 8. RFECV for RandomForestClassifier to select the important features in UK dataset.

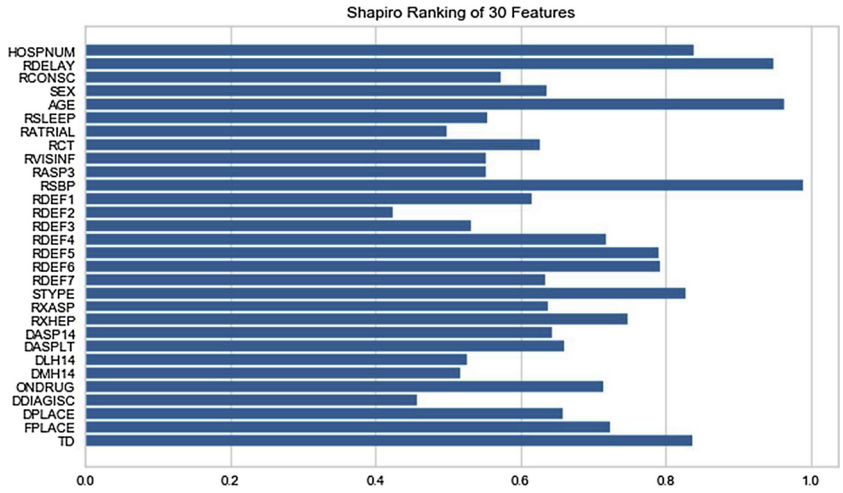

Fig. 9. Feature importance of the 30 selected features ranked by Shapiro-Wilk algorithm.

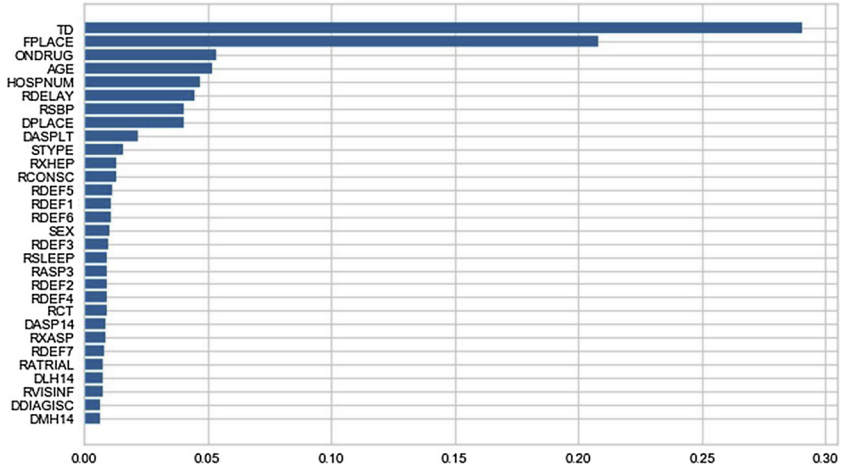

Fig. 10. Feature importance of the 30 selected features ranked by RandomForestClassifier.

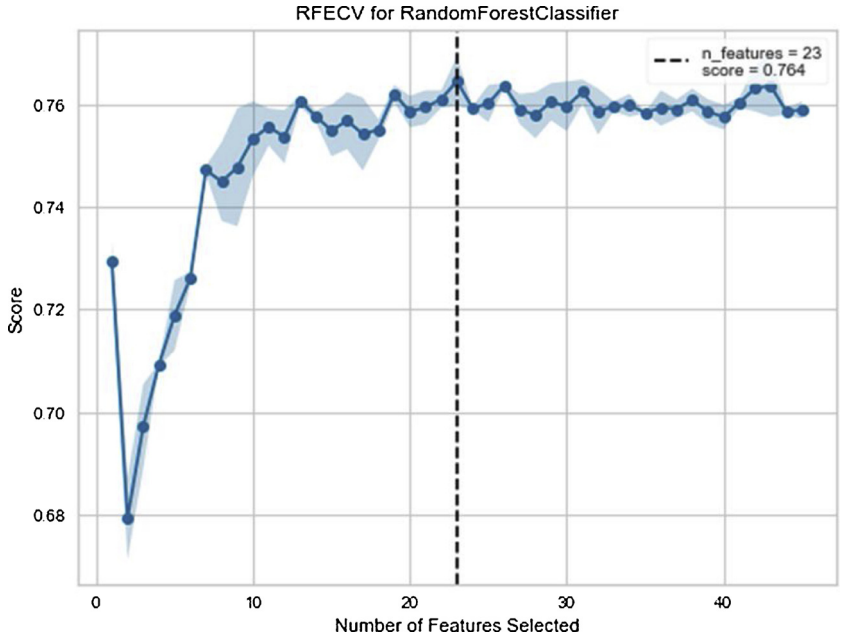

Fig. 11. RFECV for RandomForestClassifier to select features related to RVISINF in UK dataset.

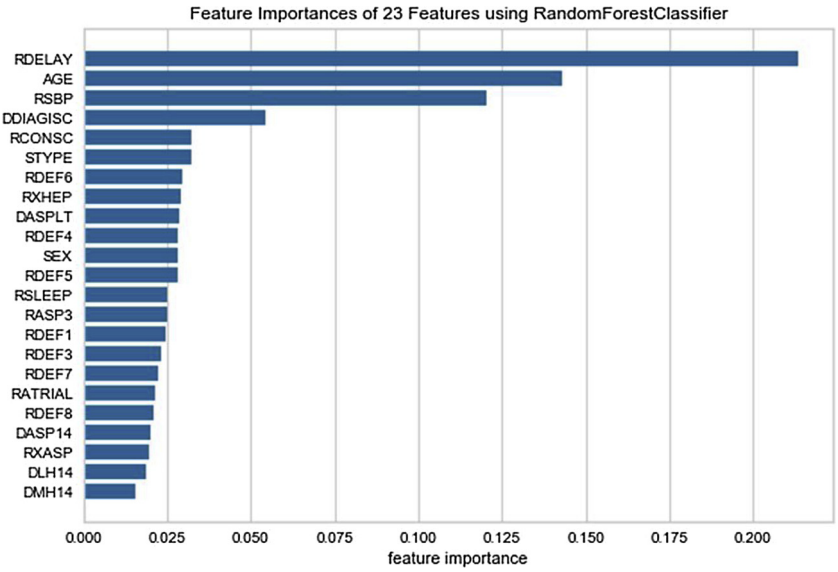

Fig. 12. Feature importance of the 23 selected features related to RVISINF ranked by RandomForestClassifier.

features. Compared with the selected features for OCCODE, it was suggested that neurological deficits exerted different influences related to RVISINF and 6th month outcome. In 23 selected features, RDELAY (Delay between stroke and randomization in hours) was ranked as the highest feature. It meant that delay between stroke and randomization would cause more infarct in congested area of brain when ischemic stroke happened. DASPLT (Discharged on long term aspirin), RASP3 (Aspirin within 3 days prior to randomization), DASP14 (Aspirin given for 14 days or till death or discharge), and RXASP (Trial aspirin allocated) were selected. RXHEP (Trial heparin allocated), DLH14 (Low dose heparin given for 14 days or till death/discharge), and DMH14 (Medium dose heparin given for 14 days or till death/discharge) were selected too. It was suggested that both aspirin and heparin usage influenced RVISINF.

4. Discussion

In this study, IST dataset was used. It was a large, prospective, randomized controlled trial, with 100 % complete baseline data and over 99 % complete follow-up data. When collecting data, we just deleted entries with missing data without imputing the missing data in the dataset. Because the dataset mostly consisted of discrete data, data preprocessing was not carried out. Even if data preprocessing was carried out with standardization, normalization, and et al., the classifiers did not perform better. The RFECV method work well in other fields, such as image processing, financial data analyzing, but was seldom used in medical research. All the classifiers used in the study did not work well (with highest accuracy of 0.690) to predict the 6th month outcome of all acute ischemic stroke data. The first main reason is that the features collected in 1990's are not mature. Some other features should be included in new study. The second one is that the data type is not proper to the study, and continuous variable should be adopted for data preprocessing. In this way prediction accuracy can be increased and the performance of classifiers can be improved. In this study, the results showed that TD, FPLACE, ONDRUG, AGE, HOSPNUM, RDELAY, RSBP and DPLACE were most important, and the TD, FPLACE, ONDRUG, HOSPNUM, RDELAY, DPLACE were features associated with medical facilities. It was suggested that active treatment and better medical condition made the difference of outcome. And these features were not mentioned as key features to stroke prognosis in previous studies (to the best of our knowledge). On the other hand, the Shapiro ranking of features showed that DTHROMB, DCAREND, DHAEMD and DGORM were the least four important features (Fig. 4). It was suggested that these four features were far from normal distribution and their data were skewed or uniform. They would exert the least influence on the outcome and were deleted first. In 23 selected features related to RVISINF that was associated with large artery occlusion (LAO) in ischemic stroke patient, RDELAY was ranked as the highest feature. It meant that delay between stroke onset and intervention would cause more infarct in congested area of brain when ischemic stroke happened. It was also suggested that these 23 selected features can be used to infer clinical symptoms and signs that can diagnose LAO.

Furthermore, an integrated machine learning approach of feature selection was presented in this study. Firstly, Shapiro-Wilk algorithm and the Pearson correlations between features were used to eliminate the least important features and highly correlated features. Then, step by step RFECV were carried out to select features important to stroke prognosis. The RFECV method performed better in feature selection for OCCODE prediction in UK dataset. The reason is that all the samples in UK were firmly controlled and the data were accurately recorded. In this study we just used features in early IST, the results coincided with current medical studies of acute ischemic stroke. Next step, some new features and variables would be collected to enhance the performance of the machine learning approach.

Funding

The study is supported by National Natural Science Foundation of China with grant number 61972107.

Declaration of Competing Interest

The authors declare that there is not conflict of interest in the paper and in the study.

Acknowledgements

Authors acknowledge the Yellowbrick project founded by Benjamin Bengfort and Rebecca Bilbro upon which this study was conducted.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.compbiolchem.2020.107316.

References

Akazawa, K., Nakamura, T., 1991. Simulation Program for Estimating Statistical Power of Cox's Proportional Hazards Model Assuming No Specific Distribution for the Survival Time. Elsevier, Ireland.

Bender, R., Augustin, T., Blettner, M., 2005. Generating survival times to simulate Cox proportional hazards models. Stat. Med. 24, 1713–1723.

Cooray, C., Mazya, M.V., Bottai, M., et al., 2018. Are you suffering from a large arterial occlusion? Please raise your arm!. Stroke Vasc. Neurol. 3, e000165.

Dawber, T.R., Meadors, G.F., Moore, F.E., 1951. Epidemiological approaches to heart disease: the Framingham study. Am. J. Public Health Nations Health 41, 279–286.

Feigin, V.L., Forouzanfar, M.H., Krishnamurthi, R., Mensah, G.A., Connor, M., Bennett, D.A., Moran, A.E., Sacco, R.L., Anderson, L., Truelsen, T., O'Donnell, M., Venketasubramanian, N., Barker-Collo, S., Lawes, C.M.M., Wang, W., Shinohara, Y., Witt, E., Ezzati, M., Naghavi, M., Murray, C., 2014. Global and regional burden of stroke during 1990–2010: findings from the Global Burden of Disease Study 2010. Lancet 383, 245–255.

Fried, L.P., Borhani, N.O., Enright, P., Furberg, C.D., Gardin, J.M., Kronmal, R.A., Kuller, L.H., Manolio, T.A., Mittelmark, M.B., Newman, A., O'Leary, D.H., Psaty, B., Rautaharju, P., Tracy, R.P., Weiler, P.G., 1991. The cardiovascular health study: design and rationale. Ann. Epidemiol. 1, 263–276.

Han, H., Liu, W., 2019. The coming era of artificial intelligence in biological data science. BMC Bioinformatics 20, 712.

Heo, Joon Nyung, Yoon, Jihoon G., Park, Hyungjong, et al., 2019. Machine learning-based model for prediction of outcomes in acute stroke. Stroke 50, 1263–1265.

Kattan, M.W., 2003. Comparison of cox regression with other methods for determining prediction models and nomograms. J. Urol. 170, S6–S10.

Khosla, Aditya, Cao, Yu, Lin, Cliff Chiung-Yu, et al., 2010. An integrated machine learning approach to stroke prediction. In: Conference: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, DC, USA. pp. 25–28 July.

Kim, A.S., Cahill, E., Cheng, N.T., 2015. Global stroke belt: geographic variation in stroke burden worldwide. Stroke 46, 3564–3570.

Li, Wen-Xing, et al., 2016. Integrated analysis of ischemic stroke datasets revealed sex and age difference in anti-stroke targets. PeerJ 4, e2470.

Liang, K.-Y., Self, S.G., Liu, X., 1990. The Cox proportional hazards model with change point: an epidemiologic application. Biometrics 46, 783–793.

Longstreth, W.T., Bernick, Jr.C., Fitzpatrick, A., Cushman, M., Knepper, L., Lima, J., Furberg, C., 2001. Frequency and predictors of stroke death in 5,888 participants in the Cardiovascular Health Study. Neurology 56, 368–375.

Lumley, T., Kronmal, R.A., Cushman, M., Manolio, T.A., Goldstein, S., 2002. A stroke prediction score in the elderly: validation and web-based application. J. Clin. Epidemiol. 55, 129–136.

Manolio, T.A., Kronmal, R.A., Burke, G.L., O'Leary, D.H., Price, T.R., 1996. Short-term predictors of incident stroke in older adults: the Cardiovascular Health Study. Stroke 27, 1479–1486.

McGinn, A.P., Kaplan, R.C., Verghese, J., Rosenbaum, D.M., Psaty, B.M., Baird, A.E., Lynch, J.K., Wolf, P.A., Kooperberg, C., Larson, J.C., Wassertheil-Smoller, S., 2008. Walking speed and risk of incident ischemic stroke among postmenopausal women. Stroke 39, 1233–1239.

Park, M.-Y., Hastie, T., 2007. An L1 regularization-path algorithm for generalized linear models. JRSSB 69, 659–677.

Pedregosa, F., Varoquaux, G., Gramfort, A., et al., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Royston, J.P., 1982. An extension of Shapiro and Wilk's W tests for normality to large samples. Appl. Stat. 31, 115–124.

Sandercock, et al., 2011. The international stroke trial database. Trials 12, 101.

Shapiro, S.S., Wilk, M.B., 1965. An analysis of variance test for normality (Complete samples). Biometrika 52, 591–611.

Weng, Stephen F., Reps, Jenna, Kai, Joe, et al., 2017. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One 12, e0174944.

Wolf, P.A., D'Agostino, R.B., Belanger, A.J., Kannel, W.B., 1991. Probability of stroke: a risk profile from the framingham study. Stroke 22, 312–318.