IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON FEATURE REPRESENTATION AND LEARNING METHODS WITH APPLICATIONS IN LARGE-SCALE BIOLOGICAL SEQUENCE ANALYSIS

Received June 4, 2020, accepted June 21, 2020, date of publication June 25, 2020, date of current version July 7, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3004977

Automated Ischemic Stroke Subtyping Based on Machine Learning Approach

GANG FANG[1,3], PENG XU[1,2], AND WENBIN LIU[1], (Member, IEEE)

[1]Institute of Computing Science and Technology, Guangzhou University, Guangzhou 510006, China
[2]School of Computer Science of Information Technology, Qiannan Normal University for Nationalities, Duyun 558000, China
[3]Departments of Neurology, Guangdong Province Traditional Chinese Medical Hospital, Guangzhou 510120, China

Corresponding author: Gang Fang (gangf@gzhu.edu.cn)

This work was supported by the National Science Foundation of China under Grant 61972107.

ABSTRACT Ischemic stroke subtyping was not only highly valuable for effective intervention and treatment, but also important to the prognosis of ischemic stroke. The manual adjudication of disease classification was time-consuming, error-prone, and limits scaling to large datasets. In this study, an integrated machine learning approach was used to classify the subtype of ischemic stroke on The International Stroke Trial (IST) dataset. We considered the common problems of feature selection and prediction in medical datasets. Firstly, the importances of features were ranked by the Shapiro-Wilk algorithm and 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 important to ischemic stroke subtyping. Furthermore, the importances of selected features were determined by Extra-Trees-Classifier. Finally, the selected features were used by Extra-Trees-Classifier and a simple deep learning model to classify the ischemic stroke subtype on IST dataset. It was suggested that the described method could classify ischemic stroke subtype accurately. And the result showed that the machine learning approaches outperformed human professionals.

INDEX TERMS Machine learning, ischemic stroke subtype, feature selection, IST.

I. INTRODUCTION

Stroke had become a major cause of disability worldwide. It was predicted that by 2030, there could be almost 70 million stroke survivors, and more than 200 million disability-adjusted life-years (DALYs) lost from stroke each year [1]. Stroke burden in high-income countries was very serious, and the burden of stroke increases rapidly in low-income and middle-income countries in recent years with the rapid development of social economy [2]. In ischemic stroke (IS), subtype classification was critical for management and outcome prediction. Numerous medical studies and data analyses had been conducted to classify IS subtype. Classification of ischemic stroke subtype required synthesis of historical, examination, laboratory, electrocardiographic, and imaging data to infer a mechanism and assign causal, etiologic, or phenotypic classification. A few different subtype schemas had been proposed including the Trial of Org10172 in Acute Stroke (TOAST) classification [3], Causative Classification System (CCS) [4], Oxfordshire Community Stroke Project (OCSP) [5], [6], and Atherosclerosis, Small-vessel disease, Cardioembolism and Other causes (ASCO) system [7]. All these classifications possessed their own advantages and weaknesses. For an example, the TOAST system had become the most widely used in recent literature, most often in studies that did not investigate the efficacy of new acute stroke treatments, such as genetic association studies, evaluations of new potential risk factors or causes of stroke, epidemiologic studies, etc. It had the weaknesses of flawing the medical decision-making process, causing major biases, etc. [8]. On the other hand, the OCSP system had the apparent advantages. Patients were easy to classify into groups based on clinical grounds and CT scanning. The outcome of stroke was driven strongly by the severity of the stroke, which was well reflected in this classification. This system had been seldom used in the 21st century to investigate potential risk factors or causes of stroke. But compared to other systems, such as ASCO, CCS, it was easy to control and can be reliably used

The associate editor coordinating the review of this manuscript and approving it for publication was Quan Zou.

TABLE 1. OCSP classification.

Classification	Diagnosis
Lacunar syndrome (LACS)	One of the 4 classic clinical lacunar syndromes. Patients with faciobrachial or brachiocrural deficits are included, but more restricted deficits are not.
Total anterior circulation syndrome (TACS)	Combination of new higher cerebral dysfunction (e.g. dysphasia, dyscalculia, visuospatial disorders), homonymous visual field defect, and ipsilateral motor and/or sensory deficit of at least 2 areas of the face, arm, and leg. If the conscious level is impaired and formal testing of higher cerebral function or the visual fields is not possible, a deficit is assumed.
Partial anterior circulation syndrome (PACS)	Only 2 of the 3 components of the TACS syndrome, with higher dysfunction alone, or with a motor/sensory deficit more restricted than those classified as LACS (e.g. confined to 1 limb, or to the face and hand but not the whole arm).
Posterior circulation syndrome (POCS)	Any of the following: ipsilateral cranial nerve palsy with contralateral motor and/or sensory deficit, bilateral motor and/or sensory deficit, disorder of conjugate eye movement, cerebellar dysfunction without ipsilateral long-tract deficit (i.e. ataxic hemiparesis), or isolated homonymous visual field defect.

in emergent situation. The OCSP system provided a simple assessment of stroke severity, with total anterior circulation syndrome having the worst prognosis [5]. The classification was based on clinical findings only. Computed tomography (CT) scanning was the best investigational test performed, but assessment of extra and intracranial arteries and precise cardiac work-up were not available [8]. In TABLE 1, the detailed OCSP classification was presented.

II. BACKGROUND AND DEVELOPMENT

A stroke subtype classification should be useful both in daily clinical practice and in epidemiological and genetic studies, randomized acute clinical trials, and prevention studies of various types (e.g. including the hemorrhagic aspects). The OCSP classification could be easily used to assess IS severity and predict the prognosis [9]. But the manual IS subtype classification was time-consuming, error-prone, professional dependent, and limits scaling to large datasets. Now, machine learning algorithms are capable of identifying features highly related to stroke occurrence efficiently from the huge set of features [10]; therefore, we believe machine learning can be used to overcome these limitations. In this study, we tested the hypothesis that an integrated machine learning method performed on features identification of structured medical data could identify OCSP subtype with high accuracy compared to manually determined OCSP subtyping performed by board-certified stroke neurologists.

Thus far, there had been a few studies on machine learning methods in processing censored medical data that outperformed traditional statistical methods. Kattan [11] 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 were expected to effectively deal with a large number of features. Recently, Stephen *et al* [12] and JoonNyung *et al* [13] presented modern machine learning based model for prediction of stroke risk and prognosis. In their work, random forest, gradient boosting machines and deep neural network were used and the accuracy of prediction was significantly increased. Ravi *et al* [14] had tested that advanced machine learning methods performed on unstructured textual data in the electronic health record (HER) can identify TOAST subtype with high concordance and inter-rater reliability. With the rapid development of machine learning method and theory in recent years, more powerful and effective integrated methods were developed [15], [16], and [17]. In this paper, Recursive Feature Elimination with Cross-Validation (RFECV) was used to select features that can automatically subtype IS.

III. MATERIAL

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 [18]. The patients in this trial were treated more than 20 years ago, and many have died. Patients and hospitals were identified only by an anonymous code; there were 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 presented no material risk to confidentiality of study participants.

The dataset included 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 IS subtypes were classified as one of the OCSP categories (TABLE 1): TACS, PACS, POCS and LACS. Nineteen thousand four hundred and thirty five patients from 467 hospitals in 36 countries were randomized within 48 hours of symptoms onset. In this dataset 984 patients were involved in the pilot phase and 1815 patients in the regular phase were finally not diagnosed as IS. The patients of pilot phase and not diagnosed as IS were excluded in this study, and then 16636 entries were kept. The data of these 16636 patients were used to select robust features for automatically IS subtyping.

IV. WORKFLOW

Given an external estimator that assigns weights to features (e.g., the coefficients of a linear model), recursive feature elimination (RFE) was to select features by recursively

considering smaller and smaller sets of features. First, the estimator was trained on the initial set of features and the importance of each feature was obtained either through a coef_attribute or through a feature_importances_attribute. Then, the least important features were pruned from current set of features. That procedure was recursively repeated on the pruned set until the desired number of features to select was eventually reached. RFECV performed RFE in a cross-validation loop to find the optimal number of features [19]. The integrated machine learning approach of 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 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 randomization were selected. Some features, such as time, date information and comments, were deleted manually (these features apparently were not related to the IS subtyping). The feature of OCSP deficit subtypes (STYPE) was kept as the target of the dataset. Then, twenty-two features were kept. The importances of these features were ranked by the Shapiro-Wilk algorithm and Pearson correlations between features were analyzed. The Shapiro-Wilk algorithm was utilized to assess the normality of the distribution of instances with respect to the feature, and was improved by Royston to process large data [20], [21]. In order to overcome the time-consuming problem of RFECV, advised by nerve physician and considering the results of Shapiro-Wilk ranking, 8 features which were related and important to IS subtyping were selected firstly. Now, we wanted to know which features would be more important to IS subtyping in the selected 8 features. 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 RFECV with its estimators respectively. After this, the selected features were ranked by Extra-Trees-Classifier which performed better than other estimators. Thirdly, the selected features were used by Extra-Trees-classifier and a simple deep neural network to subtype IS. And these two classifiers were compared with board-certified stroke neurologists to test their effectiveness.

V. METHODS

Initially, according to Shapiro-Wilk ranking and Pearson Correlation analysis, the features of continuous variables (Delay between stroke and randomization in hours (RDELAY), age (AGE), Systolic blood pressure at randomization (RSBP)) were closer to normal distribution than other features with respect to STYPE (FIGURE 2). But this analysis could not indicate which features were important to IS subtyping. In order to track the clue of important feature to IS subtyping, all the features of discrete variables were dummied. Then

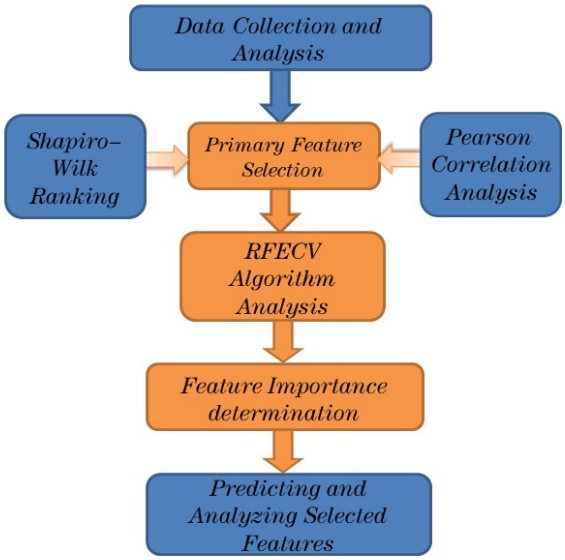

FIGURE 1. Workflow of the method.

Shapiro-Wilk ranking and Pearson Correlation analysis were carried out, the results showed that some dummied features get the same rank between binary-state variable (FIGURE 3). These features included sex (SEX), Symptoms noted on waking (RSLEEP), Atrial fibrillation (RATRIAL) and CT before randomization (RCT) etc. (FIGURE 3 (a)). This indicated that the binary-state variable of dummied feature (whether the feature present or not) exerted same influence on the feature of STYPE. It implied that these features were less important to IS subtyping. And the other features except neurological deficits features in the dataset were directly related to IS severity [9].

According to above analyses and referred to TABLE 1, it was suggested that the features of 8 neurological deficits exerted the most influence on IS subtyping. In other words, the OCSP IS subtype was based on these neurological deficits (RDEF1, RDEF2…RDEF8. Readers can be referred to the APPENDIX explaining IST_variables.). These 8 features were used by different classifiers incorporated in RFECV algorithm to subtype IS in the study. The results showed that Random-Forest and Extra-Trees Classifiers outperformed others and attained the highest accuracy 0.989 by selecting all 8 neurological deficits (FIGURE 4 and 5). When 5 features of neurological deficit selected, both two classifiers could attain accuracy above 0.95 (FIGURE 4 and 5). The shaded area in FIGURE 4, 5, 6, 7 and 8 represented the variability of cross-validation, one standard deviation above and below the mean accuracy score drawn by the curve.

Because the Extra-Trees-classifier outperformed other classifiers, the importances of 8 neurological deficits features crucial to IS subtyping were ranked by it (FIGURE 9). The feature importances were assessed by computing the differences of out of bag errors in every decision tree of the Extra-Trees-classifier. The importance of each feature was determined by formula (1). In the formula, errOOB1 means

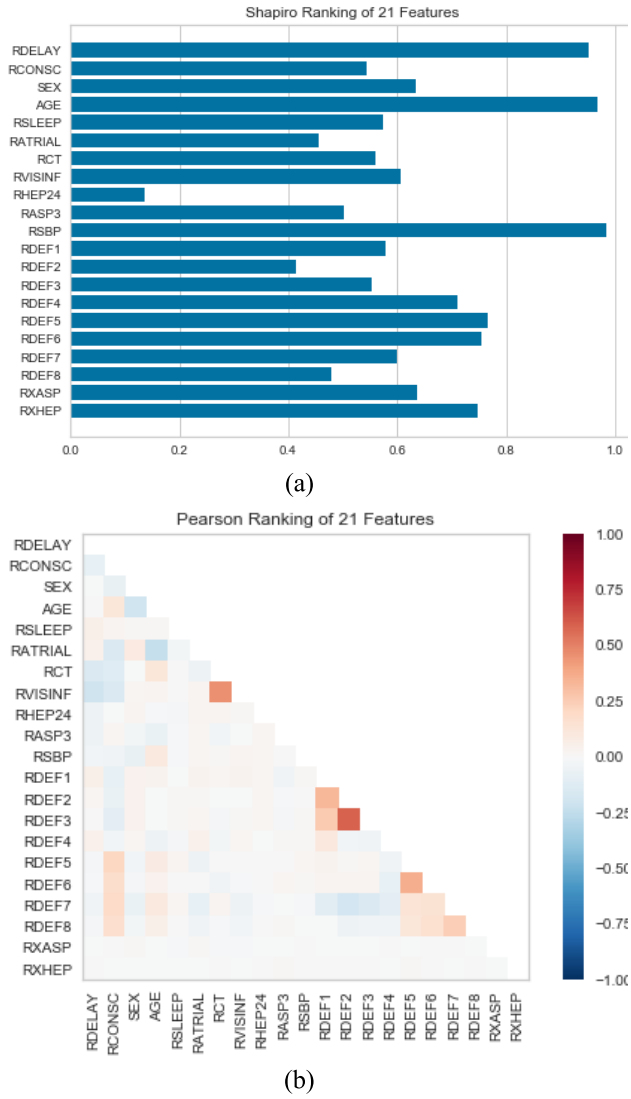

FIGURE 2. Primary Shapiro-Wilk ranking (a) and Pearson Correlation analysis (b) of initial 21 features (except STYPE).

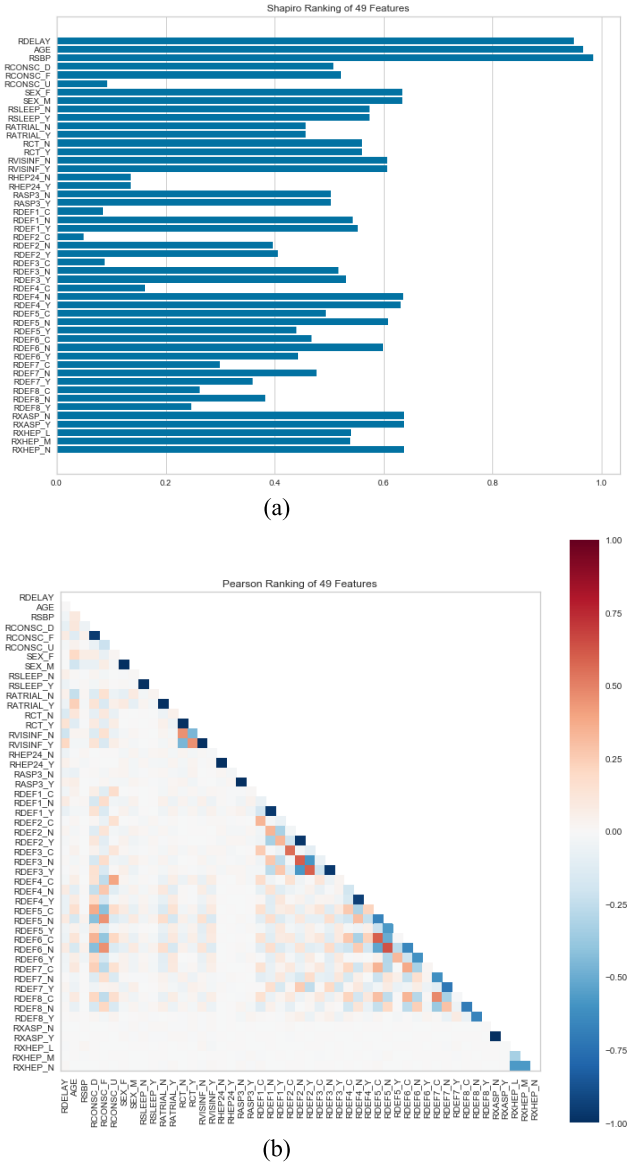

FIGURE 3. Shapiro-Wilk ranking (a) and Pearson Correlation analysis (b) of dummied 21 features (except STYPE).

error of out of bag data in *ith* decision tree and errOOB2 means error of out of bag data with noises in feature K. '*n*' is the number of decision trees in Extra-Trees-classifier.

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

VI. RESULTS

In FIGURE 9 the X-axis was the feature importance determined by formula (1), and the Y-axis was the name of these 8 selected features. The results showed that RDEF5 (Hemianopia), RDEF7 (Brainstem/cerebellar signs), RDEF4 (Dysphasia), RDEF6 (Visuospatial disorder) and RDEF2 (Arm/hand deficit) were more important. When subtyping IS in an emergent situation, less number of neurological deficits was always needed. Considering the feature correlations, importances and the analyzed results presented in FIGURE 2 (b), 9, 4 and 5, these deficits, including RDEF2 (Arm/hand deficit), RDEF4 (Dysphasia), RDEF5 (Hemianopia), RDEF6 (Visuospatial disorder) and RDEF7 (Brainstem/cerebellar signs), were kept for IS subtyping in next step. The features RDEF1 (Face deficit), RDEF3 (Leg/foot deficit, which was highly correlated to RDEF2 in FIGURE 2 (b)) and RDEF8 (Other deficit) were eliminated.

According to previous results, Extra-Trees and Random-Forest classifiers performed better than others. The Extra-Trees-classifier was used to automatically subtype IS (The Random-Forest-classifier worked in a similar way with it [22], [23]). To avoid over-fitting, a 10-fold cross validation was performed and the classifier attained a mean accuracy of 0.950 within test dataset (FIGURE 10 and 11). Furthermore, a fully connected neural network with 4 hidden layers was constructed. The structure and parameters of this neural network

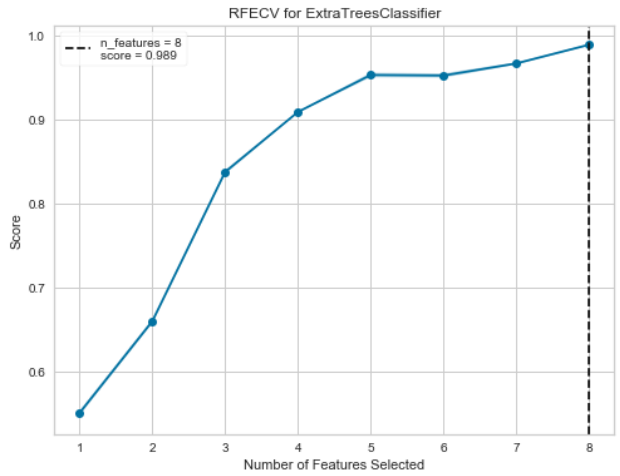

FIGURE 4. Performance of RFECV to select features important to IS subtyping with Extra-Trees-Classifier.

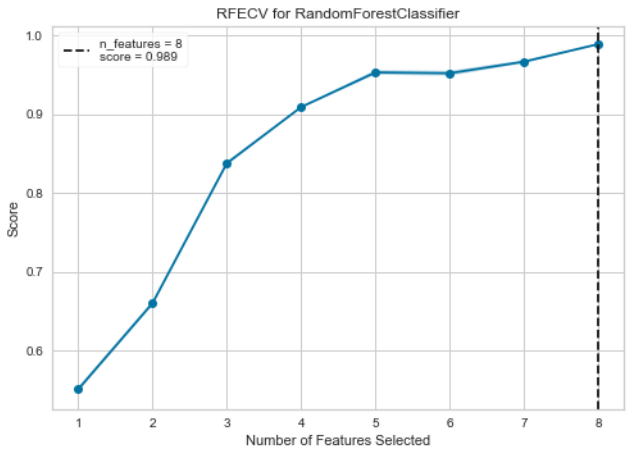

FIGURE 5. Performance of RFECV to select features important to IS subtyping with Random-Forest-Classifier.

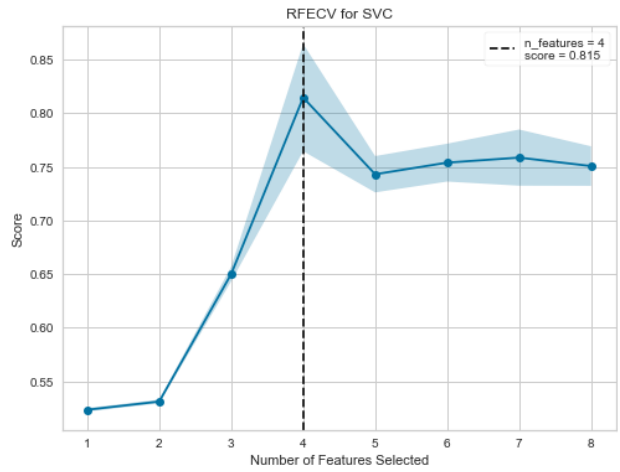

FIGURE 6. Performance of RFECV to select features important to IS subtyping with Linear SVC.

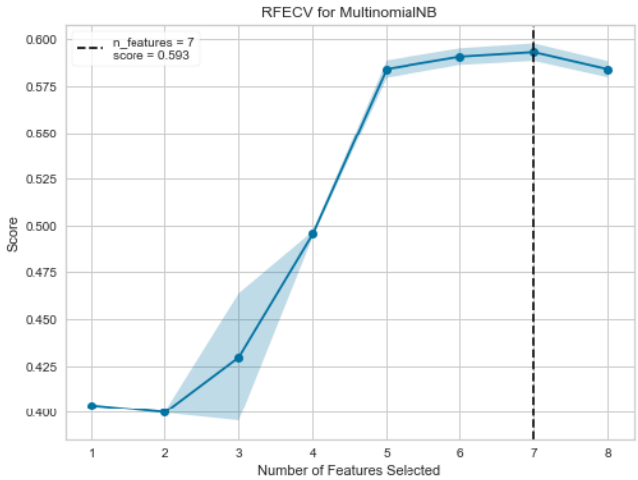

FIGURE 7. Performance of RFECV to select features important to IS subtyping with Multinomial-Naïve-Bayes-Classifier.

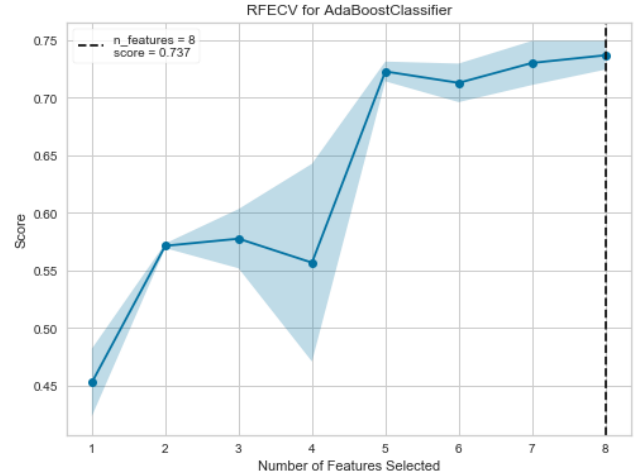

FIGURE 8. Performance of RFECV to select features important to IS subtyping with AdaBoost-Classifier.

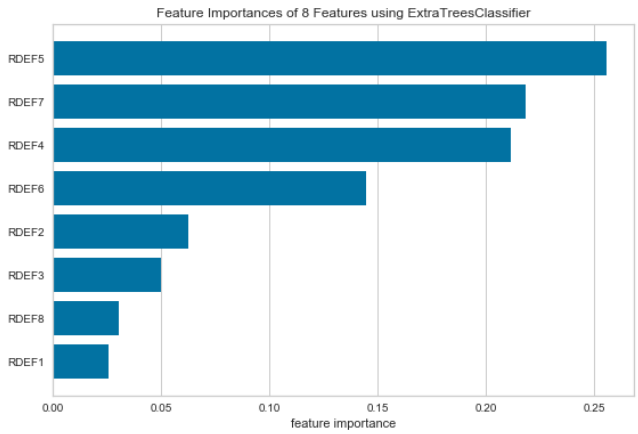

FIGURE 9. Feature importance of the 8 features selected by Extra-Trees-Classifier.

had been optimized by Gridsearch strategy with cross validation. This neural network was called deep learning because its layers were more than 4. The simple deep learning model consisted of 70, 40, 70, 40 nodes in each hidden layer with '*tanh*' as its activation function (other parameters: *alpha* = 1.0, *random_state* = 62). This model was performed to

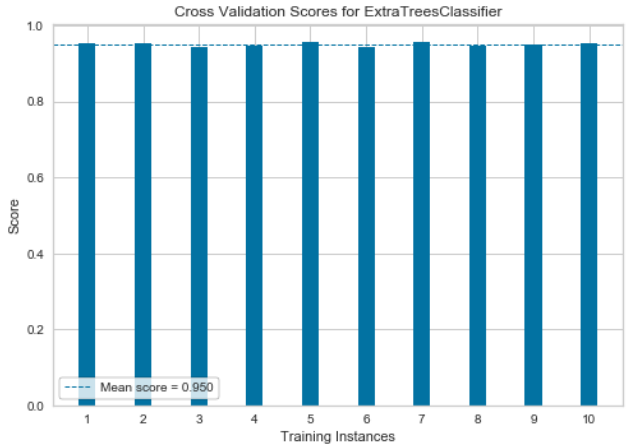

FIGURE 10. Cross validation result of Extra-Trees-Classifier.

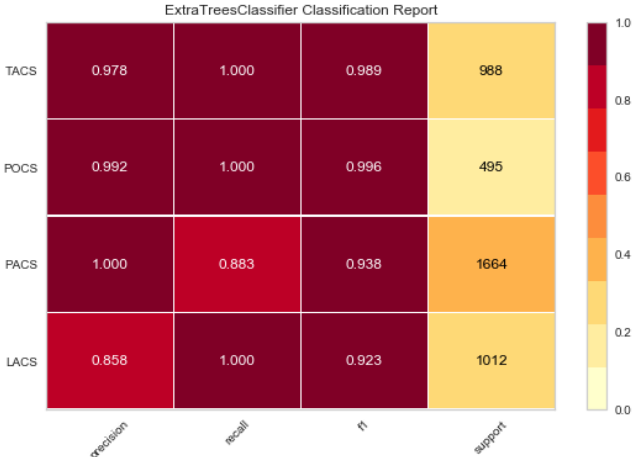

FIGURE 11. Classification report of Extra-Trees-Classifier(The final test dataset contain 4159 samples).

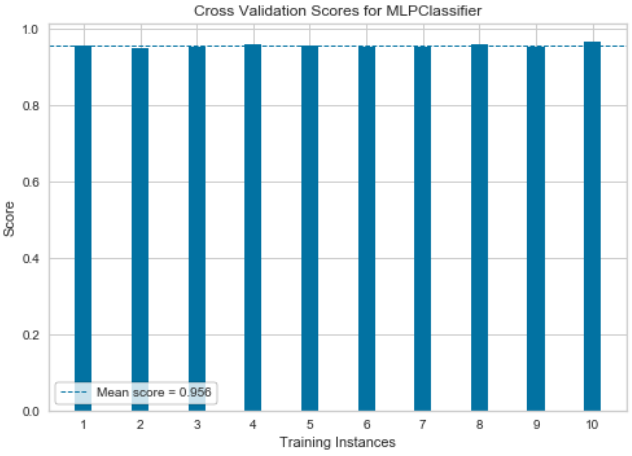

FIGURE 12. Cross validation result of neural network.

subtype IS with the 5 selected features. A 10-fold cross validation was also carried out and a mean accuracy of 0.956 within test dataset was attained (FIGURE 12 and 13). We randomly selected 400 patients from IST dataset, then Extra-Trees-Classifier and the simple deep learning model automatically subtype IS by attaining accuracy of 0.954 and 0.960 respectively. But four well-trained nerve physicians

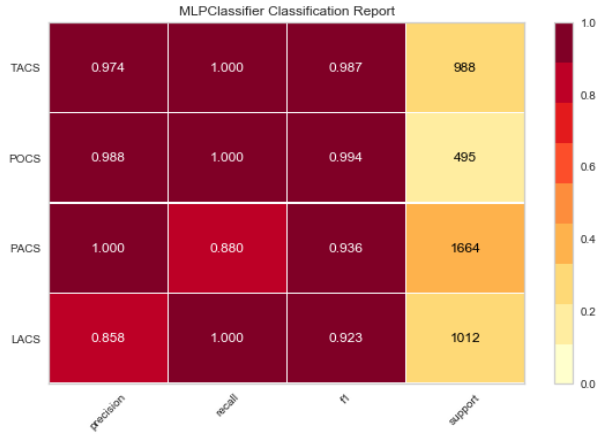

FIGURE 13. Classification report of neural network (The final test dataset contain 4159 samples).

subtyping these 400 patients with 5 selected features attained the accuracy of 0.86 (Chi-square test was performed between deep learning model and neurologists, resulted in $\chi^2 = 10.249$ and $p = 0.017$. The hypothesis test between Extra-Trees-Classifier and neurologists resulted in $\chi^2 = 9.984$ and $p = 0.019$.). The performance evaluation indicators were given by following formulas. In the formulas TP, TN, FP and FN was for true positive, true negative, false positive and false negative respectively. Support was the number of actual occurrences of the class in the specified dataset.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$precision = \frac{TP}{TP + FP} \tag{3}$$

$$recall = \frac{TP}{TP + FN} \tag{4}$$

$$f1\ score = \frac{2 * precision * recall}{precision + recall} \tag{5}$$

VII. CONCLUSION

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 value, data preprocessing was not carried out. Even if data preprocessing was carried out with standardization, normalization, and et al, the classifiers, such as linear SVC, Multinomial-Naïve-Bayes and AdaBoost didn't perform better. The RFECV method worked well in other fields, such as image processing, financial data analyzing, and was already used in medical research [24], [25]. The classifiers used in the study; except ExtraTrees, RandomForest and the simple deep learning model, didn't work well (with highest accuracy of 0.815) to subtype ischemic stroke (IS) with 8 neurological deficits. But the simple deep learning model and ExtraTrees could subtype IS accurately with only 5 selected neurological deficits. The first main reason was that the OCSP subtype was based on these neurological deficits,

and the OCSP was a primitive and simple classification system. The two classifiers worked well in the study would fail to subtype IS in other advanced classification system, such as TOAST, CCS and ASCO. The second one was that the newly developed machine learning methods (such as deep learning) could process medical data better [26], [27], even the features collected in 1990's were not very mature. In the study the dataset was firstly analyzed by Shapiro-Wilk algorithm and Pearson Correlation. Compared to RFECV result, 5 neurological deficits were selected to subtype IS at last. The last result showed that these 5 deficits could be used by classifiers to subtype IS accurately. It was also suggested that these 5 deficits can be used in emergent situation to subtype IS according to OCSP system and assess IS severity. The result also showed that machine learning approaches outperformed human professionals by subtyping IS.

In this study OCSP IS subtype system was used. Today, this system was seldom used to subtype and classify IS. But the system had the advantages of easily to use and assessing IS severity instantly in emergency. In the study we just used features in early IST, next step some new features would be collected to subtype IS according to other advanced IS classification system. And more sophisticated machine learning approach would be used to investigate new potential risk factors or causes of stroke.

APPENDIX

Appendix included IST variable (feature) names and interpretations.

ACKNOWLEDGMENT

The authors would like to thank the Yellowbrick project founded by B. Bengfort and R. Bilbro upon which this study was conducted.

REFERENCES

[1] V. L. Feigin, M. H. Forouzanfar, R. Krishnamurthi, G. A. Mensah, M. Connor, D. A. Bennett, A. E. Moran, R. L. Sacco, L. Anderson, T. Truelsen, M. O'Donnell, N. Venketasubramanian, S. Barker-Collo, C. M. M. Lawes, W. Wang, Y. Shinohara, E. Witt, M. Ezzati, M. Naghavi, and C. Murray, "Global and regional burden of stroke during 1990–2010: Findings from the global burden of disease study 2010," *Lancet*, vol. 383, pp. 245–255, Jun. 2014.

[2] A. S. Kim, E. Cahill, and N. T. Cheng, "Global stroke belt: Geographic variation in stroke burden worldwide," *Stroke*, vol. 46, no. 12, pp. 3564–3570, Dec. 2015.

[3] H. P. Adams, B. H. Bendixen, L. J. Kappelle, J. Biller, B. B. Love, D. L. Gordon, and E. E. Marsh, "Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of org 10172 in acute stroke treatment," *Stroke*, vol. 24, no. 1, pp. 35–41, Jan. 1993.

[4] H. Ay, T. Benner, E. M. Arsava, K. L. Furie, A. B. Singhal, M. B. Jensen, C. Ayata, A. Towfighi, E. E. Smith, J. Y. Chong, W. J. Koroshetz, and A. G. Sorensen, "A computerized algorithm for etiologic classification of ischemic stroke: The causative classification of stroke system," *Stroke*, vol. 38, no. 11, pp. 2979–2984, Nov. 2007.

[5] J. Bamford, P. Sandercock, M. Dennis, C. Warlow, and J. Burn, "Classification and natural history of clinically identifiable subtypes of cerebral infarction," *Lancet*, vol. 337, no. 8756, pp. 1521–1526, Jun. 1991

[6] R. I. Lindley, C. P. Warlow, J. M. Wardlaw, M. S. Dennis, J. Slattery, and P. A. Sandercock, "Interobserver reliability of a clinical classification of acute cerebral infarction.," *Stroke*, vol. 24, no. 12, pp. 1801–1804, Dec. 1993.

[7] P. Amarenco, J. Bogousslavsky, L. R. Caplan, G. A. Donnan, M. E. Wolf, and M. G. Hennerici, "The ASCOD phenotyping of ischemic stroke (updated ASCO phenotyping)," *Cerebrovascular Diseases*, vol. 36, no. 1, pp. 1–5, 2013.

[8] P. Amarenco, J. Bogousslavsky, L. R. Caplan, G. A. Donnan, and M. G. Hennerici, "Classification of stroke subtypes," *Cerebrovascular Diseases*, vol. 27, no. 5, pp. 493–501, 2009.

[9] S. Ricci, S. Lewis, and P. Sandercock, "Previous use of aspirin and baseline stroke severity: An analysis of 17 850 patients in the international stroke trial," *Stroke*, vol. 37, no. 7, pp. 1737–1740, Jul. 2006.

[10] A. Khosla, Y. Cao, C. C.-Y. Lin, H.-K. Chiu, J. Hu, and H. Lee, "An integrated machine learning approach to stroke prediction," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Washington, DC, USA, Jul. 2010, pp. 25–28.

[11] M. W. Kattan, "Comparison of cox regression with other methods for determining prediction models and nomograms," *J. Urol.*, vol. 170, pp. S6–S10, Dec. 2003.

[12] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?" *PLoS ONE*, vol. 12, no. 4, Apr. 2017, Art. no. e0174944.

[13] J. Heo, J. G. Yoon, H. Park, Y. D. Kim, H. S. Nam, and J. H. Heo, "Machine learning-based model for prediction of outcomes in acute stroke," *Stroke*, vol. 50, no. 5, pp. 1263–1265, May 2019.

[14] R. Garg, E. Oh, A. Naidech, K. Kording, and S. Prabhakaran, "Automating ischemic stroke subtype classification using machine learning and natural language processing," *J. Stroke Cerebrovascular Diseases*, vol. 28, no. 7, pp. 2045–2051, Jul. 2019.

[15] P. Xu, G. Zhao, Z. Kou, G. Fang, and W. Liu, "Classification of cancers based on a comprehensive pathway activity inferred by genes and their interactions," *IEEE Access*, vol. 8, pp. 30515–30521, 2020.

[16] W. Liu, D. Li, and H. Han, "Manifold learning analysis for allele-skewed DNA modification SNPs for psychiatric disorders," *IEEE Access*, vol. 8, pp. 33023–33038, 2020.

[17] X.-L. Qiang, P. Xu, G. Fang, W.-B. Liu, and Z. Kou, "Using the spike protein feature to predict infection risk and monitor the evolutionary dynamic of coronavirus," *Infectious Diseases Poverty*, vol. 9, no. 1, pp. 1–8, Dec. 2020, doi: 10.1186/s40249-020-00649-8.

[18] P. A. Sandercock, M. Niewada, and A. Członkowska, "The international stroke trial database," *Trials*, vol. 12, p. 101, Dec. 2011.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[20] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, nos. 3–4, pp. 591–611, Dec. 1965.

[21] J. P. Royston, "An extension of Shapiro and Wilk's *W* tests for normality to large samples," *Appl. Statist.*, vol. 31, pp. 115–124, Jun. 1982.

[22] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[23] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, Apr. 2006.

[24] F. Akhtar, J. Li, Y. Pei, Y. Xu, A. Rajput, and Q. Wang, "Optimal features subset selection for large for gestational age classification using gridsearch based recursive feature elimination with cross-validation scheme," in *Frontier Computing* (Lecture Notes in Electrical Engineering), Kyushu, Japan, vol. 551. Berlin, Germany: Springer, Feb. 2020, pp. 63–71.

[25] E.-M. Cui, F. Lin, Q. Li, R.-G. Li, X.-M. Chen, Z.-S. Liu, and W.-S. Long, "Differentiation of renal angiomyolipoma without visible fat from renal cell carcinoma by machine learning based on whole-tumor computed tomography texture features," *Acta Radiol.*, vol. 60, no. 11, pp. 1543–1552, Nov. 2019.

[26] Y. Ge, Q. Wang, L. Wang, H. Wu, C. Peng, J. Wang, Y. Xu, G. Xiong, Y. Zhang, and Y. Yi, "Predicting post-stroke pneumonia using deep neural network approaches," *Int. J. Med. Inform.*, vol. 132, pp. 1–8, Jan. 2019.

[27] A. Hilbert, L. A. Ramos, H. J. A. van Os, S. D. Olabarriaga, M. L. Tolhuisen, M. J. H. Wermer, R. S. Barros, I. van der Schaaf, D. Dippel, Y. B. W. E. M. Roos, W. H. van Zwam, A. J. Yoo, B. J. Emmer, G. J. L. à Nijeholt, A. H. Zwinderman, G. J. Strijkers, C. B. L. M. Majoie, and H. A. Marquering, "Data-efficient deep learning of radiological image data for outcome prediction after endovascular treatment of patients with acute ischemic stroke," *Comput. Biol. Med.*, vol. 115, pp. 1–7, Nov. 2019.