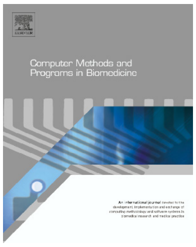

journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Different medical data mining approaches based prediction of ischemic stroke

Ahmet Kadir Arslan[a,*], *Cemil Colak*[a], *Mehmet Ediz Sarihan*[b]

[a] *Inonu University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Turkey*
[b] *Inonu University, Faculty of Medicine, Department of Emergency Medicine, Malatya, Turkey*

ARTICLE INFO

Article history:
Received 10 February 2016
Received in revised form
8 March 2016
Accepted 18 March 2016

Keywords:
Ischemic stroke
Medical data mining
Penalized logistic regression
Stochastic gradient boosting
Support vector machine

ABSTRACT

Aim: Medical data mining (also called knowledge discovery process in medicine) processes for extracting patterns from large datasets. In the current study, we intend to assess different medical data mining approaches to predict ischemic stroke.

Materials and methods: The collected dataset from Turgut Ozal Medical Centre, Inonu University, Malatya, Turkey, comprised the medical records of 80 patients and 112 healthy individuals with 17 predictors and a target variable. As data mining approaches, support vector machine (SVM), stochastic gradient boosting (SGB) and penalized logistic regression (PLR) were employed. 10-fold cross validation resampling method was utilized, and model performance evaluation metrics were accuracy, area under ROC curve (AUC), sensitivity, specificity, positive predictive value and negative predictive value. The grid search method was used for optimizing tuning parameters of the models.

Results: The accuracy values with 95% CI were 0.9789 (0.9470–0.9942) for SVM, 0.9737 (0.9397–0.9914) for SGB and 0.8947 (0.8421–0.9345) for PLR. The AUC values with 95% CI were 0.9783 (0.9569–0.9997) for SVM, 0.9757 (0.9543–0.9970) for SGB and 0.8953 (0.8510–0.9396) for PLR.

Conclusions: The results of the current study demonstrated that the SVM produced the best predictive performance compared to the other models according to the majority of evaluation metrics. SVM and SGB models explained in the current study could yield remarkable predictive performance in the classification of ischemic stroke.

1. Introduction

Ischemic stroke (IS) is associated with high mortality worldwide and is considered among the most important public health problems [1]. IS influences the management, diagnosis, and outcome. Treatments for acute IS should be made according to subtype of IS. Classification of subtypes for IS was arranged utilizing medical/clinical characteristics and the finding of supplementary clinical studies. The classification of Trial of Org in Acute Stroke Treatment (TOAST) defines five subtypes of IS: (1) big-artery atherosclerosis, (2) cardioembolism, (3) little-vein occlusion, (4) stroke of other identified etiology/causes, and (5) stroke of unidentified etiology/causes. The proposed rating system can determine etiologic diagnosis of IS in high proportions [2]. The important inference demonstrates the determination and prediction of causes and markers for the diagnosis and prevention of IS [1,2].

[*] *Corresponding author*. Tel.: +90 546 932 8246; fax: +90 422 3410036.
E-mail address: ahmetkadirarslan@gmail.com (A.K. Arslan).

http://dx.doi.org/10.1016/j.cmpb.2016.03.022
0169-2607/© 2016 Elsevier Ireland Ltd. All rights reserved.

Table 1 – The definition of the variables employed in the current study.

Variables	Abbreviation	Variable type	Definition	Role
Ischemic stroke	Is	Categorical	Present/absent	Target
Age (year)	–	Numerical	Natural number	Input
Gender	–	Categorical	Female/male	Input
Educational status	Es	Categorical	Elementary school/middle school/high school/university	Input
Marital status	Ms	Categorical	Single/married/widowed	Input
Alcohol consumption	Ac	Categorical	Present/absent	Input
White blood cell	Wbc	Numerical	Positive real number	Input
Hematocrit	Htc	Numerical	Positive real number	Input
Hemoglobin	Hb	Numerical	Positive real number	Input
Platelet	Plt	Numerical	Positive integer	Input
Glucose	Glc	Numerical	Positive integer	Input
Blood urea nitrogen	Bun	Numerical	Positive integer	Input
Creatinine	Cr	Numerical	Positive real number	Input
Sodium	Na	Numerical	Positive integer	Input
Potassium	K	Numerical	Positive real number	Input
Chlorine	Cl	Numerical	Positive integer	Input
Prothrombin time	Inr	Numerical	Positive real number	Input
Calcium	Ca	Numerical	Positive real number	Input

Data mining (also called knowledge discovery process) is a methodology for discovering hidden patterns from enormous datasets by using statistical approaches [3]. This methodology has many advantages compared to classical methods. For instance, in contrast to traditional statistical methods, data mining approaches require less presumptions in the classification and regression applications [4].

Alexopoulos et al. [5] applied inductive machine learning (ML) approaches in the diagnosis of stroke disease and used C4.5 algorithm by building a decision tree. These authors reported that inductive ML is a promising approach for computer-aided diagnosis of stroke. Linder et al. [6] used logistic regression (LR) and artificial neural networks (ANNs) for classifying acute ischemic stroke from the Database of German Stroke, and suggested that LR was the gold standard for the classification of acute ischemic stroke in comparison with ANNs, which may be employed as an alternative multivariate analysis. Khosla et al. [7] presented the comparison of the Cox proportional hazards model with a ML method for the prediction of stroke on the dataset of the Cardiovascular Health Study, and determined that combined with their suggested feature selection algorithm combined with support vector machine (SVM) achieved a higher area under the ROC curve when compared to the Cox proportional hazards model. In our previous study, ANNs and SVM were utilized to predict stroke disease using knowledge discovery process (KDP) approaches, and the results of the study determined that ANNs yielded more predictive performance as compared with SVM for the prediction of stroke and that the suggested ANNs might be beneficial for predictive purposes concerning stroke illness [3]. Additionally, there are some studies on ischemic stroke lesion segmentation using data mining or ML procedures [8–10].

The use of data mining approaches in many disciplines, especially in medicine, is increasing day by day. The medical application of data mining is called as medical data mining (MDM). Thence, MDM (also called knowledge discovery process in medicine) processes for extracting patterns from large datasets. In the current study, we intend to assess medical data mining approaches to predict ischemic stroke.

2. Material and methods

2.1. *Dataset*

This study which included 80 ischemic stroke patients (group I) and 112 healthy individuals (group II) was conducted in the department of emergency medicine, Turgut Ozal Medicine Center, Inonu University, Malatya, Turkey. Power analysis revealed that each group encapsulated minimum 68 individuals considering mean difference of creatinine for ischemic stroke patients and healthy individuals groups of 0.6, estimated standard deviations of 1.01 and 1.43, type I error (alpha) of 0.05 and type II error (beta) of 0.20. The definition of the variables that may associate with ischemic stroke [3,11,12] is summarized in Table 1.

2.2. *Preprocessing of the dataset*

In the current study, outliers were detected by local density cluster-based outlier factor [13]. This technique employs a cluster algorithm and allocates clusters into small and big ones. The outlier factor was calculated by dividing minimum sample distance to average cluster distance of all samples to the big cluster [14]. X-means was utilized as clustering algorithm in this technique. Also, z-transformation (standardization) was applied to the dataset.

2.3. *Support vector machines*

SVM is a supervised learning approach for classification and regression tasks and is utilized in order for linear/nonlinear classification problems with high-dimensional datasets [15]. To solve nonlinear classification problem, SVM maps the input sets to a high-dimensional space by applying various kernel functions [3]. A detailed explanation of SVM approach can be achieved in [16]. In this paper, SVM was employed with radial basis function (RBF) kernel function. SVM with RBF was applied by kernlab package [17] in R.

Table 2 – Tuning parameters and their scales of each models.

Model	Tuning parameters	Range	Number of combinations
SVM	C	$(2^{-5}–2^{15})$	399
	Sigma	$(2^{-15}–2^{3})$	
GBM	Interaction depth	(1–100)	3000
	n.trees[a]	(50–1500)	
	Shrinkage	0.1	
	n.minobsinnode[b]	20	
PLR	Lambda	$(10^{-4}–1)$	5
	cp[c]	AIC[d]	

[a] Total number of trees.
[b] Minimum number of observations in the trees terminal nodes.
[c] Complexity parameter.
[d] Akaike information criteria.

Table 3 – The performance metrics of each models with 95% CI.

Performance metric (95% CI)	Models		
	SVM	SGB	PLR
Accuracy	0.9789 (0.9740–0.9942)	0.9737 (0.9397–0.9914)	0.8947 (0.8421–0.9345)
AUC	0.9783 (0.9569–0.9997)	0.9757 (0.9543–0.9970)	0.8953 (0.8510–0.9396)
Sensitivity	0.9747 (0.9115–0.9969)	0.9512 (0.8797–0.9865)	0.8554 (0.7610–0.9230)
Specificity	0.9820 (0.9364–0.9978)	0.9907 (0.9494–0.9997)	0.9252 (0.8579–0.9671)
Positive predictive value	0.9747 (0.9115–0.9969)	0.9873 (0.9314–0.9996)	0.8987 (0.8101–0.9552)
Negative predictive value	0.9820 (0.9364–0.9978)	0.9640 (0.9103–0.9900)	0.8919 (0.8187–0.9428)

2.4. *Stochastic gradient boosting*

Boosting is an effective data mining ensemble meta-algorithm since it improves the prediction and classification performance of any learning approaches [18]. Stochastic gradient boosting (SGB) is a data mining approach presented by [19]. SGB is an important technique used for building prediction and classification tasks, and tunes predictive performance owing to implementation of preprocessing procedures. SGB was applied by gbm package [20] in R. More detailed definition of this method can be seen in [19].

2.5. *Penalized logistic regression*

The penalized log likelihood (PML) was maximized in a penalized logistic regression (PLR):

$$\mathrm{PML} = \log(L) - 0.5\lambda \sum ((s)_i \beta_i)^2$$

where L: the regular likelihood function, λ: a penalty factor, β_i: the predicted regression model coefficients and s_i: the scale factors. This prediction process reduces the coefficients of regression model toward zero, which increases the accuracy of novel predictions. The procedure of penalization decreases the estimated parameters and prevents overfitting difficulties [21]. PLR was implemented using stepPlr package [22] in R. More detailed information of this approach can be obtained from [23].

Table 4 – The variable importance values of the best classifier.

Variable	Importance[a]
Age	100.000
Cr	73.827
Cl	72.159
Bun	68.061
Glc	60.984
Hb	49.190
K	44.174
Gender	28.902
Inr	28.568
Htc	25.268
Ca	24.720
Wbc	23.100
Marital status	20.729
Alcohol consumption	19.145
Plt	15.678
Educational status	7.088
Na	0.000

[a] The values were transformed to scale of 0–100.

2.6. *Data partition, modeling and performance evaluation*

10-fold cross validation resampling method was employed to evaluate the model performance, and to obtain unbiased outputs and to take over fitting problem away. A seed number was determined randomly and used for each data partition process to ensure using the same training and the testing data in the modeling process. All of the modeling procedures were carried out under caret package [24] in R. The tuning parameters of the related classification models were optimized by grid search algorithm.

In the current study, the tuning parameters and their ranges of each models are tabulated in Table 2. The optimal values of the tuning parameters were identified based on the testing accuracy values that were calculated for each fold and

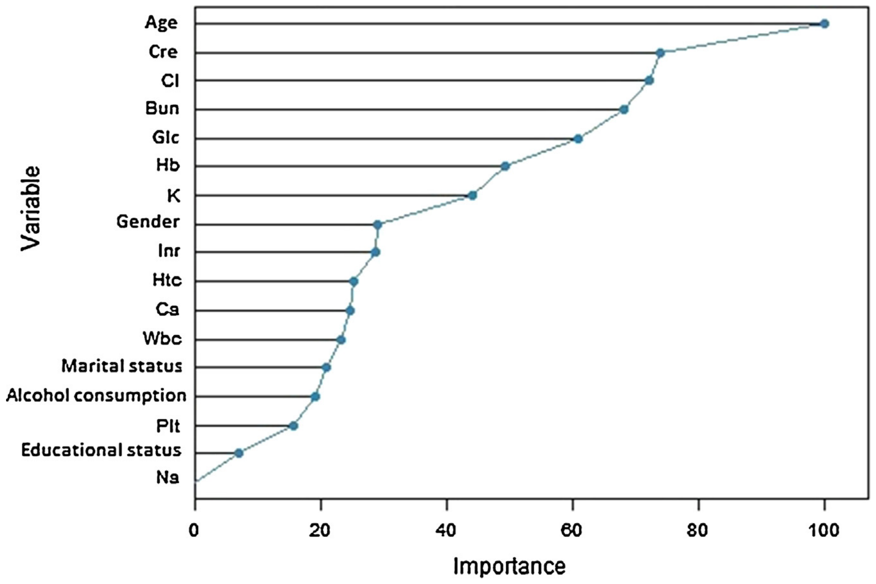

Fig. 1 – The model based-variable importance values of the best classifier.

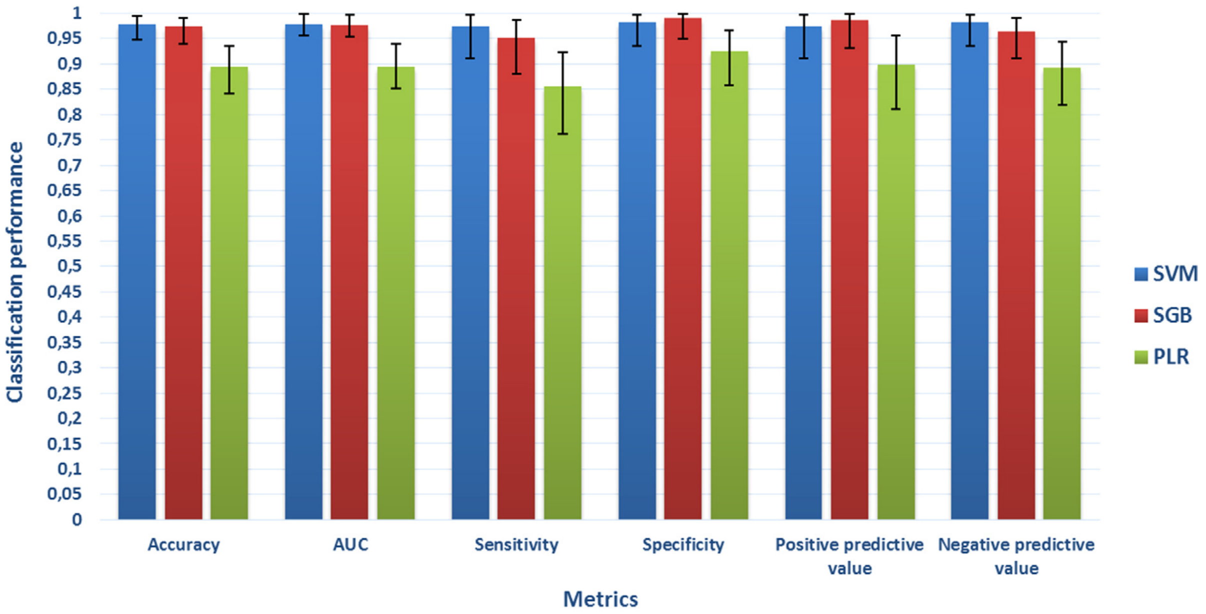

Fig. 2 – All the performance metrics of each model together with 95% CI values.

were averaged. After the optimal tuning parameters were discovered, the ultimate models were trained for the prediction.

In the current study, accuracy, area under receiver operating characteristic curve (AUC), sensitivity, specificity, positive predictive value and negative predictive value were utilized as model performance evaluation metrics [25]. These metrics are defined below:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{FN} + \text{TP}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}}$$

$$\text{Positive predictive value} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Negative predictive value} = \frac{\text{TN}}{\text{TN} + \text{FN}}$$

where TP is the number of true positives, TN is the number of true negatives, FN is the number of false negatives and FP is the number of false positives [26].

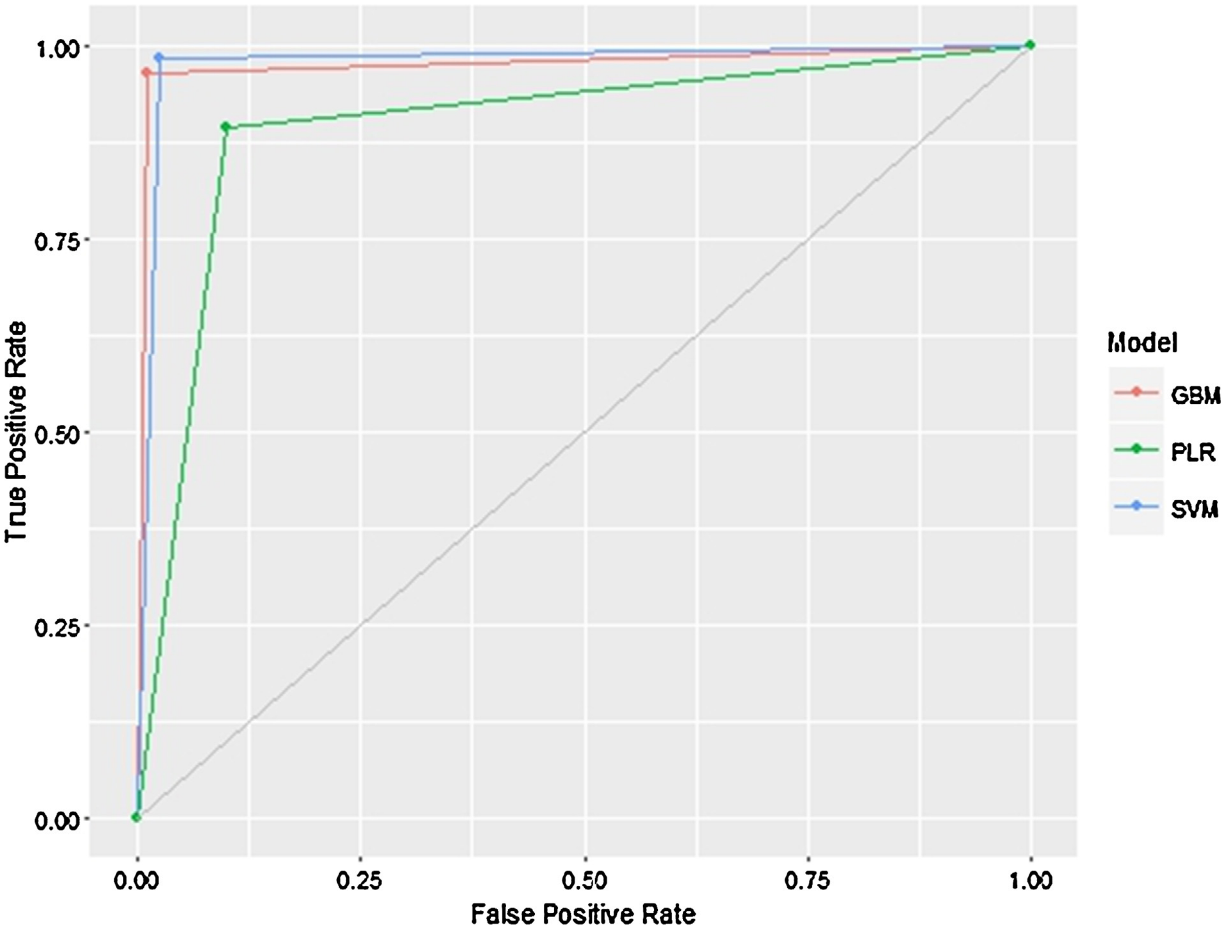

Fig. 3 – Illustration of the ROC curves of each model.

3. Results

Initially, the dataset was examined in terms of outliers. According to the outlier detection analysis, two observations were discarded from the data analysis. The remaining observations were 190 records used in the subsequent analysis. The study included 79 patients (41.6%) and 111 healthy persons (58.4%). The gender distribution of the study was 100 (52.6%) for men and 90 (47.4%) were women. The mean and standard deviation of age was 53.97 ± 21.38 years.

Table 3 presents the detailed results of the performance metrics of each models with 95% CI. The accuracy values with 95% CI were 0.9789 (0.9470–0.9942) for SVM, 0.9737 (0.9397–0.9914) for SGB and 0.8947 (0.8421–0.9345) for PLR. The AUC values with 95% CI were 0.9783 (0.9569–0.9997) for SVM, 0.9757 (0.9543–0.9970) for SGB and 0.8953 (0.8510–0.9396) for PLR.

Table 4 gives the model based-variable importance values of the best classifier (SVM) which was selected by the majority of evaluation metrics. The most related variables with the ischemic stroke were ranked by the importance values from higher to smaller.

The model based-variable importance values of the best classifier are plotted in Fig. 1. All the performance metrics of each model together with 95% CI values are illustrated in Fig. 2. Also, Fig. 3 illustrates the comparison of the ROC curves for each model.

4. Conclusions

MDM is one of the main application areas where performance metrics are very important to evaluate the predictions of the models [27]. In the current study, different data mining approaches were constructed and proposed for the prediction of ischemic stroke. For this purpose, SVM, SGB and PLR models were explained and were compared based on several predictive performance metrics: accuracy, AUC, sensitivity, specificity, positive predictive value and negative predictive value. When the values of accuracy and AUC were considered, SVM had the highest predictions as compared to GBM and PLR. While the values of accuracy, AUC, sensitivity and negative predictive value of SVM were slightly higher than GBM, all of the performance metric values for SVM and GBM were considerably higher than PLR. The values of all the performance metrics concerning the models were quite high and may be acceptable for the classification of ischemic stroke. The results achieved from the current study indicated that the constructed SVM showed a remarkable predictive performance in the majority of the performance metrics. For obtaining much more accurate and robust comparison results, comprehensive simulation study is necessary.

Finally, the current study and our previous study [3] revealed that, computer-aided approaches such as medical data mining or medical knowledge discovery are an effective instrument in the prediction of ischemic stroke and explore the hidden relationships and associations in the datasets.

Acknowledgment

We would like to thank to the RapidMiner Academia Team so much for providing RapidMiner Studio Academia Edition free licence key.

REFERENCES

[1] H.P. Adams, G. del Zoppo, M.J. Alberts, D.L. Bhatt, L. Brass, A. Furlan, R.L. Grubb, R.T. Higashida, E.C. Jauch, C. Kidwell, Guidelines for the early management of adults with ischemic stroke: a guideline from the American Heart Association/American Stroke Association Stroke Council, Clinical Cardiology Council, Cardiovascular Radiology and Intervention Council, and the Atherosclerotic Peripheral Vascular Disease and Quality of Care Outcomes in Research Interdisciplinary Working Groups: The American Academy of Neurology affirms the value of this guideline as an educational tool for neurologists, Circulation 115 (2007) e478–e534.

[2] H.P. Adams, B.H. Bendixen, L.J. Kappelle, J. Biller, B.B. Love, D.L. Gordon, E.r. 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 24 (1993) 35–41.

[3] C. Colak, E. Karaman, M.G. Turtay, Application of knowledge discovery process on the prediction of stroke, Comput. Methods Progr. Biomed. 119 (2015) 181–185.

[4] S. Sridhar, Improving diagnostic accuracy using agent-based distributed data mining system, Inf. Health Soc. Care 38 (2013) 182–195.

[5] E. Alexopoulos, G. Dounias, K. Vemmos, Medical diagnosis of stroke using inductive machine learning, Mach. Learn. Appl.: Mach. Learn. Med. Appl. (1999) 20–23.

[6] R. Linder, I. König, C. Weimar, H. Diener, S. Pöppl, A. Ziegler, Two models for outcome prediction, Methods Inf. Med. 45 (2006) 536–540.

[7] A. Khosla, Y. Cao, C.C.-Y. Lin, H.-K. Chiu, J. Hu, H. Lee, An integrated machine learning approach to stroke prediction, in: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2010, pp. 183–192.

[8] O. Maier, C. Schröder, N.D. Forkert, T. Martinetz, H. Handels, Classifiers for ischemic stroke lesion segmentation: a comparison study, PLOS ONE 10 (2015) e0145118.

[9] J.C. Griffis, J.B. Allendorfer, J.P. Szaflarski, Voxel-based Gaussian naïve Bayes classification of ischemic stroke lesions in individual T1-weighted MRI scans, J. Neurosci. Methods 257 (2016) 97–108.

[10] O. Maier, M. Wilms, H. Handels, Random forests with selected features for stroke lesion segmentation, in: Ischemic Stroke Lesion Segmentation, 2015, p. 17.

[11] P.A. Wolf, R.D. Abbott, W.B. Kannel, Atrial fibrillation as an independent risk factor for stroke: the Framingham study, Stroke 22 (1991) 983–988.

[12] P. Harmsen, A. Rosengren, A. Tsipogianni, L. Wilhelmsen, Risk factors for stroke in middle-aged men in Göteborg, Sweden, Stroke 21 (1990) 223–229.

[13] M. Amer, M. Goldstein, Nearest-neighbor and clustering based anomaly detection algorithms for rapidminer, in: Proc. of the 3rd RapidMiner Community Meeting and Conference (RCOMM 2012), 2012, pp. 1–12.

[14] M. Hofmann, R. Klinkenberg, RapidMiner: Data Mining Use Cases and Business Analytics Applications, CRC Press, 2013.

[15] R. Stoean, C. Stoean, M. Lupsor, H. Stefanescu, R. Badea, Evolutionary-driven support vector machines for determining the degree of liver fibrosis in chronic hepatitis C, Artif. Intell. Med. 51 (2011) 53–65.

[16] V.N. Vapnik, The Nature of Statistical Learning Theory. Statistics for Engineering and Information Science, Springer-Verlag, New York, 2000.

[17] A. Karatzoglou, A. Smola, K. Hornik, A. Zeileis, kernlab – An S4 Package for Kernel Methods in R, 2004.

[18] R.E. Schapire, The Boosting Approach to Machine Learning: An Overview, Nonlinear Estimation and Classification, Springer, 2003, pp. 149–171.

[19] J.H. Friedman, Stochastic gradient boosting, Comput. Stat. Data Anal. 38 (2002) 367–378.

[20] G. Ridgeway, gbm: Generalized Boosted Regression Models, R Package Version, vol. 1, 2006.

[21] A. Gastón, J.I. García-Viñas, Modelling species distributions with penalised logistic regressions: a comparison with maximum entropy models, Ecol. Model. 222 (2011) 2037–2041.

[22] M.Y. Park, T. Hastie, M.M.Y. Park, Package ‘stepPlr’, 2009.

[23] M.Y. Park, T. Hastie, Penalized logistic regression for detecting gene interactions, Biostatistics 9 (2008) 30–50.

[24] M. Kuhn, Building predictive models in R using the caret package, J. Stat. Softw. 28 (2008) 1–26.

[25] J. Friedman, T. Hastie, R. Tibshirani, The Elements of Statistical Learning, Springer Series in Statistics, Springer, Berlin, 2001.

[26] H. Chen, B. Yang, D. Liu, W. Liu, Y. Liu, X. Zhang, L. Hu, Using blood indexes to predict overweight statuses: an extreme learning machine-based approach, PLOS ONE 10 (2015) e0143003.

[27] T. Santhanam, M. Padmavathi, Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis, Proc. Comput. Sci. 47 (2015) 76–83.