Received March 9, 2018, accepted April 29, 2018, date of publication May 7, 2018, date of current version June 29, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2833442

Risk Detection of Stroke Using a Feature Selection and Classification Method

YONGLAI ZHANG[1], WENAI SONG[1], SHUAI LI[2], LIZHEN FU[1], AND SHIXIN LI[3]

[1]Software School, North University of China, Taiyuan 030051, China
[2]Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China
[3]North Automatic Control Technology Institute, Taiyuan 030006, China

Corresponding authors: Wenai Song (songwenai@nuc.edu.cn) and Shuai Li (lishuai@sia.cn)

This work was supported in part by the Natural Science Foundation of Shanxi under Grant 201601D102031 and in part by the National Natural Science Foundation of China under Grant 6160051296.

ABSTRACT Stroke places a heavy burden of care on global societies. Risk detection of stroke is a challenging and time-sensitive task across the world. This article investigated biomedical tests and electronic archives of 792 records that contained 398 records from the five years preceding the onset of stroke at a community hospital. The records included 28 features. We have proposed a new feature selection model that combines support vector machines with the glow-worm swarm optimization algorithm based on the standard deviation of the features. The results showed that the proposed model achieved 82.58% accuracy by means of the 18 features among the original data set. The new map thus represents an effective detection that can help to identify patients with an increased risk of stroke events.

INDEX TERMS Stroke, feature selection, classification, support vector machine, machine learning.

I. INTRODUCTION

Stroke is a common and disabling disease, and it places a heavy burden on society. Worldwide, the World Health Organization (WHO) estimates that almost 23.6 million people could die mostly from heart disease and stroke by 2030 [1]. According to the Department of Health Statistics, cerebrovascular disease has been ranked third in the top 10 fatal diseases since 2009 in Beijing. In 2015, cerebrovascular disease accounted for 21.7% of all causes of mortality in men and 21.68% in women. Stroke has significant morbid features, which include a high prevalence rate, high fatality rate, high disability rate and a high rate of recurrence. Most stroke-related research has been previously discussed [2]. In Beijing, the prevalence rates of stroke have been listed (Table 1), which showed that the prevalence rate in males were higher than that in females. The prevalence rate was determined to be 5.55% in the population of those 60-79 years of age in 2008, which increased to 7.4% by 2011. Additionally, the prevalence rate of female increased to 1.5% by 2014.

Automated techniques for stroke detection have been intensively researched in recent years, and many methods have been proposed and applied [3]–[9]. Data mining (DM) techniques, such as artificial neural networks (ANN), support vector machines (SVM), and intelligent algorithms, have recently been used in medical applications.

TABLE 1. Prevalence rate of stroke disease in Beijing.

Year	Age	Male	Female	Total
2008	18-39	0.23%	0.2%	0.21%
	40-59	1.74%	0.65%	1.14%
	60-79	7.35%	4.40%	5.55%
	All	1.73%	0.96%	1.3%
2011	18-39	0.1%	0%	0.1%
	40-59	1.6%	0.9%	1.2%
	60-79	10.2%	5.6%	7.4%
	All	1.8%	1.2%	1.5%
2014	18-79	1.4%	1.5%	1.4%

The wavelet-based image processing method has enhanced the ability to detect the subtlest signs of hypo-density, and the sensitivity of early stroke diagnosis has increased to 56.3% in comparison to 12.5% by previewing standard computerized tomography (CT) scans [3]. Most of this work in stroke falls into two broad categories of machine leaning: 1) rehabilitation of stroke patients and 2) stroke risk detection.

On the one hand, stroke patients showed higher mean distance errors compared to healthy individuals in semi-quantitative clinical tests of proprioception [4]. Cho *et al.* [5] proposed the effectiveness and possible use of the virtual reality (VR) rehabilitation system to recover the proprioception feedback of stroke patients in the upper limb. Mcclean *et al.* developed a model that combines Markov

models and discrete-event simulation for stroke patient care to reduce the burdens of this condition on society. After this group had accounted for patient heterogeneity and multiple care options, the model clustered patients with respect to their length of stay (LOS), which was based on data derived from the care of stroke patients at the Belfast City Hospital, UK [6]. Orthostatic hypertension (OH) is one of the catastrophic cardiovascular conditions found in elderly stroke patients. Hwang *et al.* used clinical data that included blood pressure measurements, the patient's basic clinical and physiological characteristics, and clinical symptoms that aimed to identify potential clinical factors that were associated with OH using Monte Carlo simulations. The simulations showed that the parameter estimates for the proposed model were robust with respect to the distribution assumption [7]. In contrast, effective early risk detection is of paramount importance, and automation of this assessment is highly desirable. Much effort has been devoted to solving this problem based on the technique of ultrasound imaging. Visual classifications of image segmentation based on non-invasive ultrasound plaque-image analysis have been associated with the risk of stroke. Reviewing ultrasound plaque-images was excellently reported by other groups [8]–[10].

Feature extraction creates new features based on transformations of the original dataset, while feature selection chooses distinguishing characteristics from existing features and does not construct new features [11]. Feature selection can also be categorized into unsupervised and supervised methods [12]. The strategy for feature selection is usually based on two basic components: 1) a search method and 2) an evaluation criterion [13]. Supervised feature selection methods include wrapper models and filter models. Wrapper models search in the subspace of feature sets and employ certain classifiers to evaluate the accuracy of the selected feature subsets [14]. Filter models use particular feature evaluation indices, including data variance, distance estimates, correlation estimates, relief algorithms, information entropy, and mutual information, to rank the features [15]. However, the wrapper approach can obtain more justifiable feature subsets for certain classifiers, although it requires high computational costs. The filter approach, which is based on the combination of individually acceptable features, uses the index of each single feature and does not necessarily lead to greater classification performance [16]. Thus, a hybrid model that combines the wrapper with the filter approach is proposed to balance the concerns with regard to accuracy and efficiency [17], [18].

The contributions of this study are twofold. First, we have focused on a supervised hybrid model for feature selection. In other words, this article presents a new hybrid feature selection model that is based on wrapper and filter models before the detection of the risk of stroke. The traditional factors, including family history of stroke, chronic diseases, smoking and work intensity, are impossible to quantify, and thus, it is challenging to determine how to detect the risk of stroke. Additionally, most prior research on stroke development risk detection models has used ultrasound imaging datasets. Second, the proposed model describes the importance of the qualitative order of each feature in detecting the risk of stroke. The current study was based on the dataset of the biomedical tests and the basic demographic characteristics.

TABLE 2. Biomedical test items.

Item	Normal range
α Hydroxybutyric dehydrogenase (α HBDH)	72-182
Gamma glutamyl transpeptidase (GGP)	7-32
Lactate dehydrogenase (LDH)	135-215
Low-density lipoprotein (LDL)	2.7-4.14
High-density lipoprotein (HDL)	0.7-2
Blood urea nitrogen (BUN)	3.2-7.1
Uric acid (UA)	Male: 149-476 Female: 89-434
Total cholesterol (TC)	3.5-6.1
Total bilirubin (TBIL)	3.4-20.5
Total protein (TP)	Male: 68-82 Female: 67-81
Triglyceride (TG)	0.48-1.7
Albumin (Alb)	34-55
Direct bilirubin (DBIL)	0.1-6
Alkaline phosphatase (ALP)	53-128
Serum phosphorus (PI)	0.97-1.62
Serum creatinine (SCr)	53-140
Creatine Kinase (CK)	Male: 24-195 Female: 24-170
Creatine Kinase Isoenzyme (CK-MB)	0-25
Glucose (Glu)	3.9-6.1
Alanine aminotransferase (ALT)	0-40
Aspartate aminotransferase (AST)	0-40
Apolipoprotein A1 (Apo-A1)	Male: 0.92-2.36 Female: 0.8-2.10
Apolipoprotein B (Apo-B)	Male: 0.42-1.14 Female: 0.42-1.26
Serum calcium (Ca)	2.03-2.54

The remainder of this paper is organized as follows. Section 2 describes our clinical datasets, the classification method, and the new feature selection model. Section 3 describes the results and discussion in which the application of the proposed model to the risk detection of stroke is analyzed and compared with some other existing state-of-the-art feature selection models using identical datasets. Finally, section 4 summarizes our conclusions.

II. METHODS

A. CLINICAL DATA

Data from biomedical tests (Table 2) and electronic archives of 792 patients in a community hospital in the city of Beijing were collected.

The data were derived from long-term historical records since March 2012 and tracking records for 398 patients over the 5 years preceding the onset of stroke, and all were anonymized prior to analysis. The dataset comprised 398 stroke risk records and 394 healthy individual records and contained 24 items of biomedical tests and four items of basic demographic characteristics (Table 3). There were 321 male and 471 female cases in the dataset. Compared with the data described in Table 2, the dataset described in Table 3 adds gender, age, height and BMI (Body Mass Index) as measured criteria.

TABLE 3. The dataset for stroke risk and healthy individuals[a].

Order	Feature	Description
1	α- HBD	Integer
2	GGP	Integer
3	LDH	Real
4	LDL	Real
5	HDL	Real
6	BUN	Real
7	UA	Integer
8	TC	Real
9	TBIL	Real
10	TP	Integer
11	TG	Real
12	Alb	Integer
13	DBIL	Real
14	ALP	Integer
15	PI	Real
16	SCr	Integer
17	CK	Integer
18	CK-MB	Integer
19	Glu	Real
20	ALT	Integer
21	AST	Integer
22	Apo-A1	Real
23	Apo-B	Real
24	Ca	Real
25	Gender	Male: 1 Female: 2
26	Age	Integer
27	Height	Real
28	BMI	Real
29	Label	Risk: -1 Normal: 1

[a] The first 24 items represent biomedical tests, the last item represents the classification label, and the others represent the demographic characteristics

B. SUPPORT VECTOR MACHINES

SVM is a theoretical machine learning classification technique that was adopted for structural risk minimization [19]. Our empirical analysis showed that SVM worked on the dataset with sound performance assessments. Therefore, we employed SVM for the benchmark classification algorithm to detect the accuracy rate of the feature subsets because of its stability. SVM was first presented at the Fifth Annual ACM Workshop on Computation Learning Theory (COLT). SVM preprocessing of the data represents patterns at a typically much higher level than the original feature space. With an appropriate non-linear mapping to the high-dimensional space, data from two categories can always be separated by a hyperplane.

To begin, we assume vectors "x" are the column vectors and that they have a dataset.

$$D = \{(x_1, y_1) \cdots (x_n, y_n)\}, x_n \in R^m, \quad y_n \in \{1, -1\} \quad (1)$$

When the data set "D" is linearly separable in high-dimensional space, SVM solves an optimization problem.

$$\begin{aligned} &\min_{\omega,b} \ \Phi(\omega) = \frac{1}{2} \|\omega\|^2 \\ &s.t.\ y_i(\omega^T x_i + b) \geq 1, \quad i = 1, 2, \cdots, n \end{aligned} \quad (2)$$

The hyperplane and decision function of the SVM classifier would be the following:

$$\omega \cdot x + b = 0 \quad (3)$$

$$f(x_i) = sgn(\omega_0 \cdot x_i + b_0) \quad (4)$$

Problem (2) is a differentiable convex problem with affine constraints, and therefore, this optimization problem is solved by Lagrange multipliers. We set the derivative of the Lagrange with respect to ω, α, and with b equal to zero, and then, we transform it into the Wolfe dual form using the Lagrange multipliers "α":

$$\begin{cases} \max\limits_{\alpha} \ \sum\limits_{i=1}^{n} \alpha_i - \frac{1}{2} \sum\limits_{i=1}^{n} \sum\limits_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i \cdot x_j) \\ s.t. \ \sum\limits_{i=1}^{n} \alpha_i y_i = 0 \\ \quad 0 < \alpha_i < C, i = 1, 2, \cdots, n \end{cases} \quad (5)$$

where C is the penalty parameter. C can be viewed as a way to control the fit.

Additionally, the solution "ω" has the following form:

$$w = \sum_{i=1}^{n} \alpha_i x_i y_i, \quad i = 1, 2, \cdots, n \quad (6)$$

We represent the dot products by a positive definite kernel or a Mercer kernel that is defined as $K(x_i, x_j)$. In this article, we have used the Gaussian kernel:

$$\Phi(x_i) \cdot \Phi(x_j) = K(x_i, x_j) = e^{-q\|x_i - x_j\|^2} \quad (7)$$

where $||\cdot||$ is the Euclidean norm. Parameter "q" represents the width of the kernel function.

In view of (6), and (7), we have the matrix dual optimization problem, which is defined as

$$\begin{aligned} &\max \ c^T \alpha - \frac{1}{2} \alpha^T H \alpha \\ &s.t.\ \alpha^T Y = 0, \alpha_i \geq 0 \quad (i = 1, 2, \cdots, n) \end{aligned} \quad (8)$$

where $c^T = [1, 1, \ldots, 1]1 \times n$, $Y = (y1, y2, \ldots, yn)^T$ and $H = (y_i y_j k(x_i, x_j))n \times n$ are a symmetric matrix with $k(x_i, x_j) = K_{ij}$.

Therefore, the Lagrange multipliers "α_i" and "w_0" are calculated by means of equations (8) and (6), and xi ($\alpha_i > 0$) are support vectors. The bias "b_0" can thus be calculated as

$$b_0 = -\frac{1}{2}[\max(w_0 \cdot x(1)) + \min(w_0 \cdot x(-1))] \quad (9)$$

With "w_0" and "b_0" calculated, the SVM decision function, in view of equation (4), can be given.

C. GLOW-WORM SWARM OPTIMIZATION

Optimization methodologies play an important role in training the SVM. Due to the different requirements on the training speed, memory constraints and the accuracy of optimizing variables, practitioners should choose different optimization algorithms. The GSO algorithm is a

new nature-inspired heuristic for optimization problems. Krishnan and Ghose [20] proposed this heuristic approach in 2005 as a derivative-free meta-heuristic algorithm that imitate the glow behavior of glow-worms [20]. The algorithm combines rules and randomness that are designed to imitate some natural phenomena. Each artificial glow-worm, namely, the agent, carries a fluorescent light in two-dimensional space, and has its own local decision range that depends on the number of neighbors. The agents are assumed to carry a luminescence quantity called luciferin with them. Briefly, the algorithm involves three phases:

The luciferin update phase, wherein the luciferin update rule is given by:

$$L_i(t) = (1-\rho)L_i(t-1) + \gamma \cdot J(X_i(t)) \tag{10}$$

where "$L_i(t)$" represents the luciferin level for the glow-worm "i" at time "t"; "ρ" is the luciferin decay constant $0 < \rho < 1$; "γ" is the luciferin update constant, and "J" represents the value of the objective function at a location of an agent at time "t".

Movement phase. For each glow-worm "i", the probability of moving toward a neighbor "j" is defined as

$$P_{ij}(t) = \frac{L_j(t) - L_i(t)}{\sum_{k \in N_i(t)} (L_k(t) - L_i(t))} \tag{11}$$

where "$N_i(t)$" is a set of the neighborhood of the glow-worm "i" at time "t" in the decision range domain. Then, the movement of the glow-worms can be stated as

$$X_i(t+1) = X_i(t) + s \cdot \frac{X_j(t) - X_i(t)}{\|X_j(t) - X_i(t)\|} + \alpha(rand - 0.5) \tag{12}$$

where "s" is the step size and $\alpha(rand - 0.5)$ is the disturbance term to avoid the local optima.

Decision range update. The decision range update rule is introduced as follows:

$$R_d^i(t+1) = \min\{r_s, \max\{0, R_d^i(t), \beta(n_t - |N_i(t)|)\}\} \tag{13}$$

where "r_s" represents the radial range of the luciferin sensor, "β" is a decision range update rate, and "n_t" is a parameter that is used to control the number of neighbors in the decision range domain. A brighter agent has a superior position. The higher luciferin intensity a neighbor has, the more attraction it gains within the local decision range. While the neighbor-density is low, the range is enlarged to locate more neighbors. The algorithm shares some common features with ant colony optimization (ACO) and particle swarm optimization (PSO), but it has an efficiently running algorithm, namely, the training speed of the SVM, which is a significant advantage.

D. A PROPOSED FEATURE SELECTION MODEL

The proposed model is a hybrid combined wrapped and filter model, which employs the SVM for the standard classifier to detect the classification accuracy of the feature subset. The basic idea of the proposed model is divided into four stages, as shown in Figure 1.

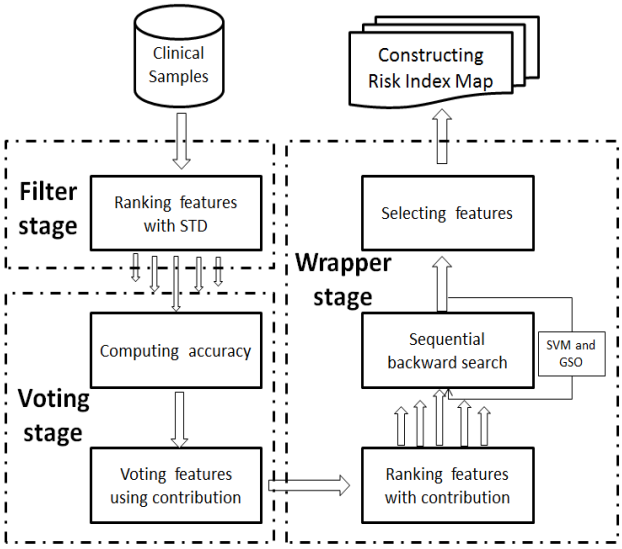


FIGURE 1. A proposed feature selection model with weighted voting.

- Filter stage

(1) Ranking features with STD. Data variance may be the simplest evaluation criteria for filter models. The variance of the signal feature reflects its representative power. We regarded the STD of each feature as positive weights, and then the features were ranked in descending order of weights.

- Voting stage

(2) Computing accuracy. Data samples with the first "d" highest ranked features are input into the SVM, respectively. The classification accuracy rates of the first "d" dimensions are then obtained.

(3) Voting features using contribution. The results of the values of STD are successively added to the corresponding classification. The accuracy rate indicated the contribution of each feature.

- Wrapper stage

(4) Ranking features with the contribution. Feature samples are ranked in descending order of contribution. We regarded the contribution of each feature as new weights.

(5) Sequential backward search. Data samples with the first "d (1<d<28)" highest ranked weights are input into the classifier, combining SVM with GSO. The procedure of the algorithm is to discard one feature that gives a negative growth in the value of the accuracy, which is deleted from the data set. The classification accuracy rates of the first "d" dimensions are again obtained, and the features are reranked according to the results.

(6) Selecting features. We select the first "n" features by means of the maximum of the classification accuracy rates.

- The last stage

(7) Constructing the risk index map using surface fitting technology based on the data set, including the first "n" features.

Algorithm 1 Algorithm of the Exhaustive Approach

Input: Training dataset $D = \{(x_1, y_1) \cdots (x_n, y_n)\}$, $x_n \in \mathbf{R}^m$, $y_n \in (1, -1)$, training class labels $\{y_1; y_2; \ldots; y_n\} \in \{1, -1\}$.

Output: Penalty parameter $C \in [0, 256]$, the width of the kernel function $q \in [0, 20]$, accuracy, S = set of feature select solutions.

Process:

1. Compute STDn of each feature and then rank the features in descending order of STD;

2. For each feature in x, do (d = 1, 2, . . . , n)

// the first "d" highest ranked features are input into the SVM, respectively.

for $q \in [0, 20]$//using GSO

for $C \in [0, 256]$

Compute α_i, $\omega = \sum_{i=1}^{n} \alpha_i x_i y_i$, $K(x_i, x_j)$

Accuracy = $P/(P + N)$

end

end

3. Compute the contribution of each feature: Cn = Ci + 1 − Ci (i = 1, 2, . . . n)

The implemented exhaustive algorithm for feature selection can be seen in Algorithm 1.

The optimization of the parameters q and C employs the GSO algorithm. The implemented exhaustive algorithm for GSO can be seen in Algorithm 2.

Algorithm 2 Algorithm of the Exhaustive Approach

Input: Training dataset $D = \{(x_1, y_1) \cdots (x_n, y_n)\}$, $x_n \in \mathbf{R}^m$, $y_n \in (1, -1)$, training class labels $\{y_1; y_2; \ldots; y_n\} \in \{1, -1\}$, $C \in [0, 256]$, $q \in [0, 20]$.

Output: Optimal classification accuracy

Process:

1. for each population size do

Randomly generate the initial solution;

Compute the value of the objective function J(t).

2. for each number of iterations do

for i=1 to population size

$$L_i(t) = (1 - \rho)L_i(t - 1) + \gamma \cdot J(X_i(t))$$

Compute

Compute Ni(t)

end

for $j \in N_i(t)$

$$P_{ij}(t) = \frac{L_j(t) - L_i(t)}{\sum_{k \in N_i(t)} (L_k(t) - L_i(t))}$$

Compute

$$X_i(t+1) = X_i(t) + s \cdot \frac{X_j(t) - X_i(t)}{\|X_j(t) - X_i(t)\|} + \alpha(rand - 0.5)$$

Compute

end

for i=1 to population size

$$R_d^i(t+1) = \min\{r_s, \max\{0, R_d^i(t), \beta(n_t - |N_i(t)|)\}\}$$

Compute

end

3. Get the optimal value of the objective function.

III. RESULTS AND DISCUSSION

The experimental data set included 792 samples with 29 features. There were 398 risk samples and 394 healthy samples. Table 4 described the first three steps of the feature selection model that is shown in Figure 1. We divided the data set into a test set and a validation set and used a ten-fold cross-validation strategy in the classifier. There are 300 patient samples and 300 healthy samples in the test set. There are 98 patient samples and 94 healthy samples in the validation set. The width "q" of the Gaussian kernel function and the penalty constant "C" need to be optimized during the process of the training phase of SVM. The penalty constant determines the trade-off between minimizing the training error and minimizing the model complexity. There is no systematic methodology for optimization of these parameters. In this study, the optimization of parameters employs the GSO algorithm. The initial GSO parameters are shown in Table 5. Decision function is the accuracy in the SVM classifier. The results of optimal parameters are listed in the third and fourth columns of Table 4. "Accuracy" indicates the classification accuracy of the validation set. "Contribution" indicates the difference between the accuracy and the previous contribution, "STD (0-1)" indicates the normalized result between 0 and 1, and "weight" is the sum of "STD (0-1)" and "Contribution (0-1)".

Table 6 describes the last three stages of the proposed model that is shown in Figure 1. In Table 6, "Weight" is reranked in descending order of the "Weight" that is shown in Table 4, "Contribution rate" indicates the weight divided by the sum of all weights, "Cumulative contribution" indicates the sum between the contribution and the previous contribution, and AUC (area under the curve) shows the SVM classifier. The maximum classification accuracy was 82.58%, which appears in the eighteenth line of Table 6, and the corresponding AUC is 0.8948. Therefore, we selected the first 18 dimension features for our optimal subset to enable us to detect the relative risk of stroke. Figure 2 shows the experimental results that contain accuracies as functions of the parameter C (0<C<400) and q (0<q<50) in the SVM models. In Figure 3, we employed receiver operating characteristic (ROC) metrics to evaluate the classifier featured true positive rate (TPR) on the y-axis and the false positive rate (FPR) on the x-axis. In these results, a larger area for AUC is usually more beneficial. In other words, the "steepness" of the ROC curves is also beneficial, since it is ideal to maximize the true positive rate while minimizing the false positive rate.

Additionally, ROC curves of the first nine dimensions of the dataset (Table 6) are described in Figure 3a, and the first 10–18 dimensions are described in Figure 3b. From Figure 3a, there is a steady increase in the AUC with the

TABLE 4. Weighting of stroke features based on standard deviations.

Order	Feature	STD [a]	C	q	Accuracy (%)	Contribution [b]	STD (0-1) [c]	Contribution (0-1)	Weight [d]
1	CK	0.2097	16	0.50	56.0707	-	1.0000	1.0000	2.0000
2	LDH	0.2073	64	0.50	57.0707	1.0000	0.9884	0.2987	1.2870
3	α-HBD	0.1914	128	8.00	58.5859	1.5152	0.9123	0.3667	1.2789
4	Height	0.1710	8	8.00	57.3232	-1.2627	0.8145	0.0000	0.8145
5	ALP	0.1521	2	4.00	58.8384	1.5152	0.7239	0.3667	1.0906
6	UA	0.1015	1	8.00	58.5859	-0.2525	0.4817	0.1333	0.6150
7	SCr	0.0874	16	4.00	61.8687	3.2828	0.4142	0.6000	1.0143
8	GGP	0.0845	16	2.00	61.6162	-0.2525	0.4000	0.1333	0.5334
9	TP	0.0775	2	8.00	61.6162	0.0000	0.3667	0.1667	0.5334
10	AGE	0.0755	64	1.00	67.9293	6.3131	0.3574	1.0000	1.3574
11	ALT	0.0661	128	1.00	67.5505	-0.3788	0.3120	0.1167	0.4287
12	AST	0.0580	64	1.00	67.9293	0.3788	0.2734	0.2167	0.4900
13	CK-MB	0.0490	128	0.15	69.1919	1.2626	0.2303	0.3333	0.5637
14	Alb	0.0468	128	0.05	69.5707	0.3788	0.2198	0.2167	0.4365
15	TBIL	0.0350	256	0.50	72.0960	2.5253	0.1633	0.5000	0.6633
16	BMI	0.0254	64	0.25	72.7273	0.6313	0.1171	0.2500	0.3671
17	Glu	0.0099	128	0.25	72.7273	0.0000	0.0430	0.1667	0.2097
18	DBIL	0.0085	64	0.50	73.1061	0.3788	0.0364	0.2167	0.2531
19	BUN	0.0071	64	0.50	72.9798	-0.1263	0.0296	0.1500	0.1796
20	TC	0.0061	64	0.50	73.2323	0.2525	0.0248	0.2000	0.2248
21	LDL	0.0050	128	1.00	72.9798	-0.2525	0.0197	0.1333	0.1530
22	TG	0.0042	128	1.00	72.8535	-0.1263	0.0156	0.1500	0.1656
23	Gender	0.0019	128	1.00	72.9798	0.1263	0.0049	0.1833	0.1882
24	Ca	0.0018	64	0.50	72.9798	0.0000	0.0040	0.1667	0.1707
25	Apo-A1	0.0017	128	1.00	73.2323	0.2525	0.0036	0.2000	0.2036
26	HDL	0.0016	128	1.00	73.1061	-0.1262	0.0032	0.1500	0.1533
27	Apo-B	0.0012	128	1.00	73.1061	0.0000	0.0013	0.1667	0.1680
28	PI	0.0009	128	1.00	72.9798	-0.1263	0.0000	0.1500	0.1500

[a] STD indicates the standard deviation; [b] Contribution indicates the difference between the accuracy and the previous contribution; [c] STD (0-1) indicates the normalized result between 0 and 1; [d] Indicates that the weight is the sum of STD (0-1) and the contribution (0-1).

TABLE 5. Parameters of the GSO algorithm.

Parameter	Value
Decision function	Max accuracy
Domain of C	0-256
Domain of q	0-20
Population size	20
Step size	0.05
Number of Iterations	500
Initial luciferin	5
Initial radial range	0.01
Luciferin update rate	0.6

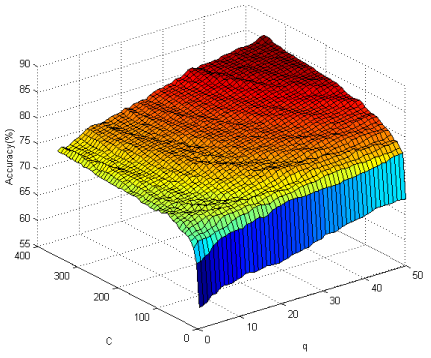


FIGURE 2. Accuracy of SVM with different values of parameter q and C.

instant influx of features. The first AUC peak appears in the ninth dimension in the last figure of Figure 3; meanwhile, the value of AUC was 0.8428. In addition, this value decreased, as shown in Figure 3b. The second peak appeared on the eighteenth dimension (AUC = 0.8948) in the last figure of Figure 3b. The optimal subset was selected to detect the risk of stroke, and the classification accuracy was determined as 82.58%.

As shown in Figure 4, the accuracy is used to compare the proposed model to the filter model. The first 16 features are the optimal subset that employs the filter model, and the first 18 features were the optimal subset that employed the proposed model. However, the proposed model achieved 82.58% accuracy compared to 73.23% for the filter model.

For comparison, we also conducted experiments on the same dataset using the wrapper model with SVM and ANN (artificial neural networks). Table 7 lists the results of the four models. Both SVM and ANN were limited to using the randomized search algorithm (the GSO algorithm) for searching an optimal subset [21]. The ANN architecture is composed

TABLE 6. Ordering of stroke features based on their weights.

Order	Feature	Weight	Contribution rate [a]	Cumulative contribution [b]	C	q	Accuracy (%)	AUC [c]
1	CK	2.0000	0.1274	0.1274	16	0.5	56.0707	0.5754
2	AGE	1.3574	0.0865	0.2139	128	0.25	62.2475	0.6661
3	LDH	1.2870	0.0820	0.2959	256	1	62.6263	0.6715
4	α-HBD	1.2789	0.0815	0.3774	256	1	64.6465	0.6994
5	ALP	1.0906	0.0695	0.4469	256	0.5	66.2879	0.7068
6	SCr	1.0143	0.0646	0.5116	16	1	67.2980	0.7155
7	Height	0.8145	0.0519	0.5635	128	1	73.7374	0.7941
8	TBIL	0.6633	0.0423	0.6057	256	0.25	71.9697	0.7800
9	UA	0.6150	0.0392	0.6449	256	1	77.5253	0.8428
10	CK-MB	0.5637	0.0359	0.6808	64	1	77.3990	0.8297
11	TP	0.5334	0.0340	0.7148	64	1	77.2727	0.8355
12	GGP	0.5334	0.0340	0.7488	128	0.5	75.2525	0.8238
13	AST	0.4900	0.0312	0.7800	64	1	78.5354	0.8430
14	Alb	0.4365	0.0278	0.8079	256	0.25	75.1263	0.8228
15	ALT	0.4287	0.0273	0.8352	256	0.5	78.0303	0.8539
16	BMI	0.3671	0.0234	0.8586	64	0.25	75.0000	0.8193
17	DBIL	0.2531	0.0161	0.8747	256	0.125	75.3788	0.8230
18	TC	0.2248	0.0143	0.8890	128	1	82.5758	0.8948
19	Glu	0.2097	0.0134	0.9024	64	0.5	77.5253	0.8462
20	Apo-A1	0.2036	0.0130	0.9154	64	0.5	77.7778	0.8468
21	Gender	0.1882	0.0120	0.9273	64	0.5	77.7778	0.8467
22	BUN	0.1796	0.0114	0.9388	64	0.5	77.6515	0.8475
23	Ca	0.1707	0.0109	0.9497	64	1	80.8081	0.8814
24	Apo-B	0.1680	0.0107	0.9604	64	1	80.8081	0.8817
25	TG	0.1656	0.0106	0.9709	64	1	80.4293	0.8822
26	HDL	0.1533	0.0098	0.9807	64	1	80.5556	0.8823
27	LDL	0.1530	0.0097	0.9904	128	1	75.5758	0.8346
28	PI	0.1500	0.0096	1.0000	128	1	72.9798	0.7921

[a] Contribution rate indicates the ratio, which is the weight divided by the sum of all weights; [b] Cumulative contribution indicates the sum between the contribution and the previous contribution; [c] AUC indicates the area under the ROC curves.

TABLE 7. Comparison of the models for the same dataset.

Method	Dimensions[a]	Accuracy	Times[b]
Proposed model	18	82.5758%	56
Filter model using STD	20	73.2323%	28
Wrapper model with SVM(without GSO)	19	82.12%	502
Wrapper model with ANN(without GSO)	21	78.63%	508

[a] Dimensions indicate the dimensions of the optimal subset; [b] Times indicate the mean number of searched subsets.

TABLE 8. The first 6 features' weighted voting.

Order	Feature	Weight	Contribution rate	Weight (0-1)[a]
1	CK	2.0000	0.1274	0.30
2	AGE	1.3574	0.0865	-
3	LDH	1.2870	0.0820	0.19
4	α-HBD	1.2789	0.0815	0.19
5	ALP	1.0906	0.0695	0.17
6	SCr	1.0143	0.0646	0.15

[a] Weight (0-1) indicates the normalized result except for AGE between 0 and 1 based on Weight.

of 30 neurons in hidden layers. In addition, the filter model using STD has the lowest accuracy because a combination of individually acceptable features does not necessarily lead to good accuracy. The wrapper model with SVM can obtain the significant feature subset that contains 19 dimension features, while it also requires a high computational cost (the mean number of the searched subsets is 502). A comparison of the results achieved by the wrapper model with ANN shows that ANN is not an effective algorithm for a dataset that contains a small sample size, as previously shown [22], [23]. In general, our proposed feature selection model provided the optimal feature subset with greater performance characteristics, a higher accuracy and lower computational costs.

Moreover, the weight described in Table 6 shows the important quality of features for detecting the risk of stroke. With the assistance of medical experts in stroke, the first 6 features, CK, AGE, LDH, α-HBD, ALP and SCr, were found to reflect the most important risk factors of stroke shown in Table 8. Based on Table 8, we found that the other five features belong to the category of enzymes, except for AGE, in the most important risk factors. This finding is a major discovery because we no longer need to collect data that includes family history of stroke, chronic diseases, smoking and work intensity to detect the risk of stroke. We only need a biomedical test, and we can detect the risk of stroke using our model. The weight (0-1) in Table 8 indicates the normalized result except for AGE, which is between 0 and 1 based on Weight.

Risk detection of stroke is shown in Figure 5. The risk index of the proposed model with different values of Age and Synthetic Value of Enzymes (SVE) expressed as 0.30*CK +

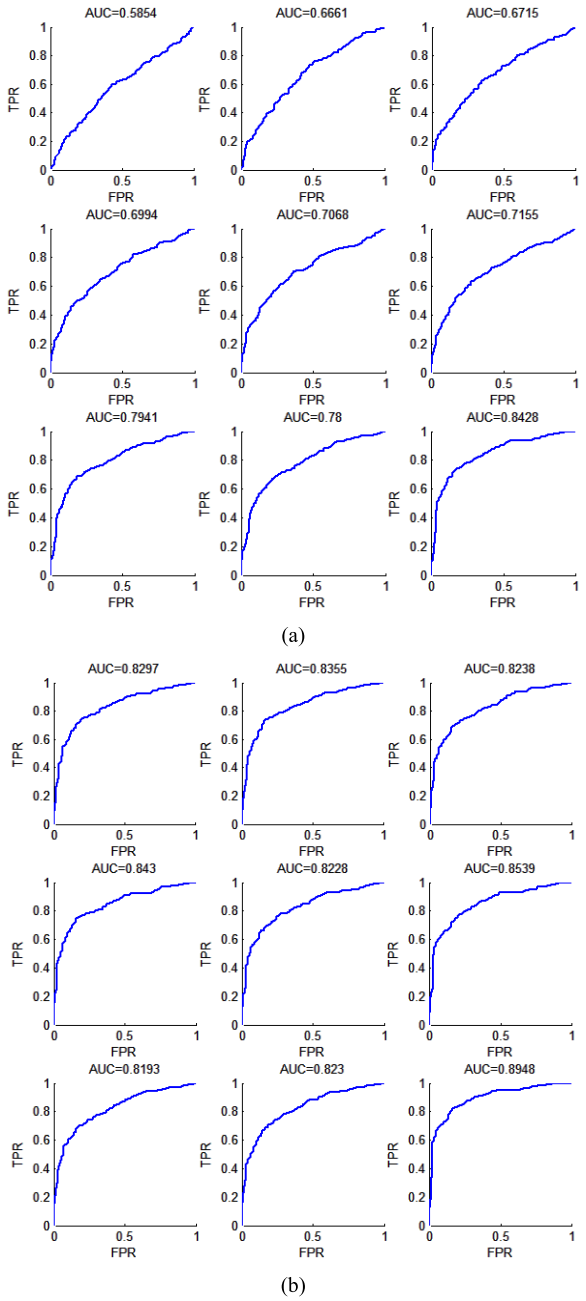


FIGURE 3. ROC curves for the first 18 dimensional datasets after feature selection: (a) represents the first nine dimensions of the feature subset; (b) represents the first 10 to 18 dimensions of the feature subset.

0.19*LDH + 0.19*α-HBD + 0.17*ALP + 0.15*SCr. In the expression, CK, LDH, α-HBD, ALP and SCr indicate the values of biomedical test items. The risk index is directly shown in the limited range ($20 < \text{Age} < 100, 0 < \text{SVE} < 300$). The higher the index is, the higher the risk becomes. From Figure 5, we find that the risk index of stoke is lower between 20 to 50 years of age, and the risk index becomes higher when SVE is between 115 to 145 at 45-65 years of age, while the risk index becomes higher when SVE is between 165 to 215 at 70-100 years of age. In other words,

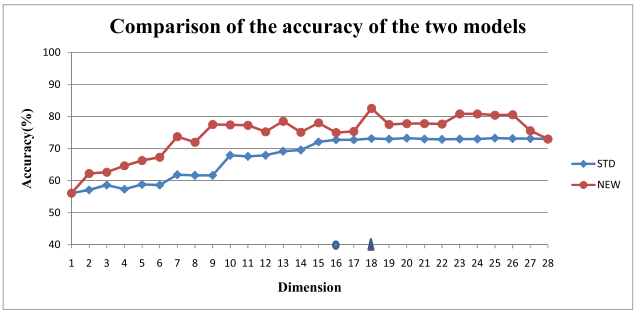


FIGURE 4. Comparison of the accuracy of the filter model using STD and the new model.

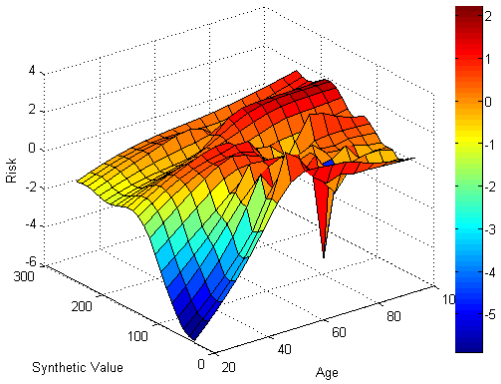


FIGURE 5. Risk index of the proposed model with different values of age and synthetic values expressed as 0.30*CK + 0.19*LDH + 0.19*α-HBD + 0.17*ALP + 0.15*SCr. CK, LDH, α-HBD, ALP and SCr indicate the values of biomedical test items.

the higher the value of SVE is, the higher the risk index becomes.

IV. CONCLUSION

In this study, we introduced the proposed feature selection model for detecting the risk of stroke based on a dataset that was obtained from biomedical tests and electronic archives on 792 patients at a community hospital in Beijing.

Our method applies the feature selection model to the medical field. This model combined STD, a filter-based variable, and SVM, which was remarkably effective in feature selection of stroke. Moreover, our model described an important quality of each feature for detecting the risk of stroke. The first 6 features, CK, AGE, LDH, α-HBD, ALP and SCr, were found to reflect the most important risk factors of stroke except for the traditional factors, which include family history of stroke, chronic diseases, smoking and work intensity. Our model shows its superior performance for detecting the relative risk of stroke based on the data set of biomedical tests. Improving the accuracy of this proposed novel approach will be accounted for in subsequent studies. The new map should thus have clinical utility in risk detection of stroke in the general population.

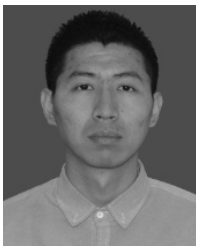
REFERENCES

[1] (May 2017). *World Health Organization, CVD*. [Online]. Available: http://www.who.int/mediacentre/fact-sheets/fs317/en/

[2] S. Zhang, Z. Liu, Y.-L. Liu, Y.-L. Wang, and X.-B. Cui, "Prevalence of stroke and associated risk factors among middle-aged and older farmers in western China," *Environ. Health Preventive Med.*, vol. 22, no. 1, p. 6, 2017.
[3] M. Breccia, M. Molica, I. Zacheo, A. Serrao, and G. Alimena, "Application of systematic coronary risk evaluation chart to identify chronic myeloid leukemia patients at risk of cardiovascular diseases during nilotinib treatment," *Ann. Hematol.*, vol. 94, no. 3, pp. 393–397, 2015.
[4] N. Leibowitz *et al.*, "Automated measurement of proprioception following stroke," *Disability Rehabil.*, vol. 30, no. 24, pp. 1829–1836, 2008.
[5] S. Cho *et al.*, "Development of virtual reality proprioceptive rehabilitation system for stroke patients," *Comput. Methods Programs Biomed.*, vol. 113, no. 1, pp. 258–265, 2014.
[6] S. McClean, M. Barton, L. Garg, and K. Fullerton, "A modeling framework that combines Markov models and discrete-event simulation for stroke patient care," *ACM Trans. Model. Comput. Simul.*, vol. 21, no. 4, 2011, Art. no. 25.
[7] Y.-T. Hwang, H.-Y. Tsai, Y.-J. Chang, H.-C. Kuo, and C.-C. Wang, "The joint model of the logistic model and linear random effect model—An application to predict orthostatic hypertension for subacute stroke patients," *Comput. Stat. Data Anal.*, vol. 55, no. 1, pp. 914–923, 2014.
[8] L. Saba *et al.*, "Web-based accurate measurements of carotid lumen diameter and stenosis severity: An ultrasound-based clinical tool for stroke risk assessment during multicenter clinical trials," *Comput. Biol. Med.*, vol. 91, pp. 306–317, Dec. 2017.
[9] I. M. B. Francischetti, A. Cajigas, M. Suhrland, J. M. Farinhas, and S. Khader, "Incidental primary mediastinal choriocarcinoma diagnosed by endobronchial ultrasound-guided fine needle aspiration in a patient presenting with transient ischemic attack and stroke," *Diagnostic Cytopathol.*, vol. 45, no. 8, pp. 738–743, 2017.
[10] H. Papadopoulos, E. Kyriacou, and A. Nicolaides, "Unbiased confidence measures for stroke risk estimation based on ultrasound carotid image analysis," *Neural Comput. Appl.*, vol. 28, no. 6, pp. 1209–1223, 2017.
[11] W. Zhuge, F. Nie, C. Hou, and D. Yi, "Unsupervised single and multiple views feature extraction with structured graph," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 10, pp. 2347–2359, Oct. 2017.
[12] S. Solorio-Fernández, J. F. Martínez-Trinidad, and J. A. Carrasco-Ochoa, "A new unsupervised spectral feature selection method for mixed data: A filter approach," *Pattern Recognit.*, vol. 72, pp. 314–326, Dec. 2017.
[13] U. Mlakar, I. Fister, J. Brest, and B. Potocnik, "Multi-objective differential evolution for feature selection in facial expression recognition systems," *Expert Syst. Appl.*, vol. 89, pp. 129–137, Dec. 2017.
[14] J. M. Cadenas, M. C. Garrido, and R. Martínez, "Feature subset selection filter-wrapper based on low quality data," *Expert Syst. Appl.*, vol. 40, no. 16, pp. 6241–6252, 2013.
[15] C. O. Sakar, O. Kursun, and F. Gurgen, "A feature selection method based on kernel canonical correlation analysis and the *minimum Redundancy–Maximum Relevance* filter method," *Expert Syst. Appl.*, vol. 39, no. 3, pp. 3432–3437, 2012.
[16] C. Yao, Y.-F. Liu, B. Jiang, J. Han, and J. Han, "LLE score: A new filter-based unsupervised feature selection method based on nonlinear manifold embedding and its application to image recognition," *IEEE Trans. Image Process.*, vol. 26, no. 11, pp. 5257–5269, Nov. 2017.
[17] M. Kang, M. R. Islam, J. Kim, J. M. Kim, and M. Pecht, "A hybrid feature selection scheme for reducing diagnostic performance deterioration caused by outliers in data-driven diagnostics," *IEEE Trans. Ind. Electron.*, vol. 63, no. 5, pp. 3299–3310, May 2016.
[18] M. Aladeemy, S. Tutun, and M. T. Khasawneh, "A new hybrid approach for feature selection and support vector machine model selection based on self-adaptive cohort intelligence," *Expert Syst. Appl.*, vol. 88, pp. 118–131, Dec. 2017.
[19] F. J. González-Serrano, Á. Navia-Vázquez, and A. Amor-Martín, "Training Support Vector Machines with privacy-protected data," *Pattern Recognit.*, vol. 72, pp. 93–107, Dec. 2017.
[20] K. N. Krishnanand and D. Ghose, "Detection of multiple source locations using a glowworm metaphor with applications to collective robotics," in *Proc. IEEE Swarm Intell. Symp.*, Jun. 2005, pp. 84–91.
[21] A. Deniz, H. E. Kiziloz, T. Dokeroglu, and A. Cosar, "Robust multiobjective evolutionary feature subset selection algorithm for binary classification using machine learning techniques," *Neurocomputing*, vol. 241, pp. 128–146, Jun. 2017.
[22] X. Zhou, Y. Zhang, M. Shi, H. Shi, and Z. Zheng, "Early detection of liver disease using data visualisation and classification method," *Biomed. Signal Process. Control*, vol. 11, pp. 27–35, May 2014.
[23] H. R. Pourghasemi, S. Yousefi, A. Kornejady, and A. Cerdà, "Performance assessment of individual and ensemble data-mining techniques for gully erosion modeling," *Sci. Total Environ.*, vol. 609, no. 3, pp. 764–775, 2017.

YONGLAI ZHANG received the B.S. degree from Shanghai Normal University, Shanghai, China, and the M.S. degree from Liaoning Technical University, Fuxin, China, in 2001 and 2007, respectively, all in electrical engineering, and the Ph.D. degree from the Big Data Analysis and Processing Group, Shenyang Institute of Automation, Chinese Academy of Sciences. His current research interests include pattern recognition, data mining, big data analysis, and medical informatics.

WENAI SONG received the B.S. and M.S degrees from the Taiyuan Mechanical Institute, Taiyuan, China, in 1985 and 1991, respectively, and the Ph.D. degree from the Beijing Institute of Technology. Her current research interests include signal and information system, data mining, and medical informatics.



SHUAI LI received the master's degree from Northeastern University, China, in 2013. He is currently a Scientific Researcher with the Big Data Analysis and Processing Group, Shenyang Institute of Automation, Chinese Academy of Sciences. He is the author of several refereed papers in pattern recognition, process monitoring, and machine vision. His current research interests include pattern recognition, big data analysis, and process monitoring. He is the Reviewer of the IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS.

LIZHEN FU received the B.S. and M.S degrees from Shanxi University, Taiyuan, China, in 2005 and 2008, respectively, and the Ph.D. degree from the Renming University of China. Her current research interests include database technology, pattern recognition, big data analysis, and medical informatics.

SHIXIN LI received the M.S. and Ph.D. degrees from the Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, China, in 2009 and 2014, respectively. He is currently a Scientific Researcher with the North Automatic Control Technology Institute. His current research interests include network communication, data mining, big data analysis, and medical informatics.