IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 23, 2019, accepted December 30, 2019, date of publication January 6, 2020, date of current version January 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2964390

The Application of the Machine Learning Method in Electromyographic Data

TAO LIU[1], ZECHEN LI[1], YUQI TANG[2], DONGDONG YANG[2], SHUOGUO JIN[2], AND JUNWEN GUAN[1,3]

[1]Chengdu University of Information Technology, Chengdu 610103, China
[2]Sichuan Province Traditional Chinese Medicine Hospital, Chengdu 610075, China
[3]West China Hospital of Sichuan University, Chengdu 610041, China

Corresponding author: Zechen Li (307841726@qq.com)

This work was supported Scientific Research Projects of Sichuan Education Department under Grant 18ZA0111.

ABSTRACT This paper studies the application of machine learning in the analysis and diagnosis of electromyography data. Firstly, 2,352 electromyography examination reports have been recorded from Sichuan Provincial Hospital of Traditional Chinese Medicine for ten months. The data cleaning has been conducted based on the specific-designed inclusion criteria. Next, two data sets have been established, containing 575 facial motor nerve conduction study reports and 233 auditory brainstem response reports, respectively. And then, four machine learning algorithms including random forest, linear regression, support vector machine and logistic regression have been employed to the data sets. The performance comparisons of accuracy and recall rate among different algorithms indicate that the random forest algorithm has the optimal performance over the other two in both data sets. Moreover, the comparisons have been carried out in the cases with and without deviation standardization for each algorithm, and the results demonstrate that the deviation standardization has a certain effect on the accuracy improvement. Additionally, it is found that the random forest algorithm can present the ranking of the features in order of importance. Consequently, the random forest is proven to be an optimal algorithm for computer-aided diagnosis systems. Furthermore, it is worth mentioning that the feature ranking in order of importance can facilitate clinical diagnosis and has a certain clinical potential in diagnosis and diagnostic assessment.

INDEX TERMS Machine learning, electromyography, feature extraction, random forest, support vector machine.

I. INTRODUCTION

As a science for studying the bio-electrical activities of nerves and muscle cells, the nerve electrophysiology(EMG) has been employed clinically for nearly a century. Based on the nerve electrophysiology, a comprehensive clinical electrophysiological examination technology has been developed for years, including electroencephalography, electromyography, and evoked potentials. This examination technology has a clinical importance in the differential diagnosis of neurogenic diseases and myogenic diseases, as well as in the qualitative localization, pathological extent, and prognosis of peripheral neuropathy [1].

As an essential part of artificial intelligence, machine learning algorithms, such as traditional machine learning algorithms, deep learning algorithms, and reinforcement learning algorithms, have been extensively used in the medical field and played a vital role in the diagnosis and treatment of diseases.

Over the past few years, many efforts have been devoted to the machine learning application in the clinical diagnosis. In 2016, the Gulshan team from the University of California demonstrated in the JAMA magazine that the artificial intelligence could diagnose diabetic retinopathy from over 100,000 retinal fundus photographs. Compared with 54 ophthalmologists having US doctor licenses, the artificial intelligence method has higher sensitivity and specificity than the manual judgment [2]. In 2017, Golden proposed that deep learning can quickly scan pathological photos to diagnose breast cancer with lymph node metastasis. On the basis of the large-scale data set, these researches have achieved some success in the medical field. Though this method cannot completely replace the work of pathologists,

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

it would extremely enhance the diagnostic efficiency and lighten the burden of pathologists [3].

Over the past decades, in the scope of the open literature reporting the electromyography (EMG) data application by artificial intelligence and machine learning, there are a few researches obtaining achievements. From the birth of EMG in the 18th century to the end of the last century, EMG data were mainly used in qualitative analysis [4], [5]. In 1985, J.V. Basmajian and C.J. De Luca discussed all the latest advances in this rapidly developing field and offered accurate guidelines for the clinical application, thus assisting doctors in diagnosis [5]. In 1987, M.J. Aminoff published his book [6] and indicated the similar perspective.

Since the beginning of this century, scholars have devoted themselves to conduct quantitative analyses on EMG data, which has extremely promoted the development of EMG. In 2003, T.J. Doherty and D.W. Stashuk proposed the method and the preliminary scheme of EMG quantitative analyses [7]. Since then, EMG quantitative analyses have become a research hotspot. In 2005, S. Boe and D.W. Stashuk quantitatively estimated and analyzed the number of motor units of distal and proximal muscles of upper limbs, and the result was worth of reference for EMG-based gesture recognition [8]. In 2006, D.W. Stashuk and L. Pino et al presented the quantitative interpretation of EMG [9].

During the past decade, coupled with the vigorous development and the application of machine learning and deep learning, many efforts have been made to apply machine learning algorithms to the EMG data analysis. In 2013, Abdulhamit Subasi employed the support vector machine (SVM) method to classify the biceps muscle EMG data of 27 patients in the Neurology Department of the University of Gaziantep, which achieved good results for the computer-aided diagnosis. In his research, the method of combining particle swarm optimization (PSO) and SVM was proposed to improve the classification effect [10]. In 2014, Yousefi and Jamileh et al adopted traditional machine learning methods including Self-Organizing Feature Map (SOFM), Decision Trees, Bayesian Techniques, Artificial Neural Network (ANN), and Neuro-fuzzy system (NFS) to classify 57 participants, among whom there were 17 participants having symptoms of non-specific arm pain and the rest were healthy. By comparing the advantages and disadvantages among these algorithms, it was found that NFS had a better classification effect on EMG results [11]. In 2018, Phinyomark et al published a review paper about the significance of EMG in assisting clinical diagnosis and future direction in the big data era. In addition, the complete plan and assumption from the data collection to the data analysis were proposed in this paper [12].

Although the abovementioned researches have achieved great success in the medical field, there are still some researches worthy to be further explored.

First of all, these recent achievements have been mainly based on deep learning and the large-scale data set with abundant samples. However, numerous real investigations have shown that each type of examination could only produce hundreds of data during a whole year after classifying the EMG examinations, even in a well-known hospital in a big city. As a consequence, it is nearly impossible to obtain a large number of samples to satisfy the data quantity requirement of deep learning. Taking this reason into account, researches and clinical applications based on deep learning methods have been hindered. On the contrary, the traditional machine learning algorithms can obtain high accurate results even only based on small-scale data sets through manually selecting appropriate features [13]. Consequently, traditional machine learning algorithms are widely adopted in the field of the traditional Chinese medicine (TCM) diagnosis and treatment researches. These studies have played a significant role in the exploration of medical classification rules.

Second, AI-related researches on facial and head EMG data have been rarely reported, especially on the facial motor nerve conduction study (F-MNCS) and auditory brainstem response (ABR) data, thus having a tremendous research potential.

As for the term - ‘‘motor nerve conduction study’’, motor conduction velocity (MCV) is usually used as the professional academic expression in the scientific community. The output of this experiment equipment (MEB-9200K) is motor nerve conduction study. Therefore, this paper decided to respect the naming of equipment manufacturers in data processing and paper writing, and use motor nerve conduction study(MNCS).

In the present paper, the traditional machine learning algorithm using EMG data based on small-scale data set has been adopted to carry out the related researches on the clinical application. Firstly, 2,352 EMG examination reports are collected during ten months from Sichuan Provincial Hospital of Traditional Chinese Medicine. After designing and applying the inclusion criteria, 575 F-MNCS reports and 233 ABR reports are selected. Meanwhile, two data sets are established after data cleaning. Next, four most popular algorithms, including linear regression, logistic regression, support vector machine(SVM) and random forest, are applied to the two data sets, respectively. Furthermore, detailed comparisons and discussions are conducted on processed results of four algorithms, including the effect comparison in the cases with and without data standardization. Additionally, it is found that the random forest algorithm can present the ranking of the features in order of importance. Finally, it is concluded that random forest is an optimal algorithm for computer-aided diagnosis systems.

II. DATA

A. DATA COLLECTION

Firstly, 2352 EMG medical reports obtained from the Sichuan Provincial Hospital of TCM are collected under the confidentiality agreement and the authority approval. As mentioned above, it is nearly impossible to obtain a large number of data even in a well-known hospital. These over 2300 reports have been accumulated for about ten months. Only EMG raw data and images are presented in this paper. Table 1 shows

TABLE 1. Types of the obtained data.

Nerve Conduction	Motor Nerve Conduction Study (MNCS)
	Sensory Nerve Conduction Study (SNCS)
	Repetitive Stimulation
	F-wave
	H-reflex
	Blink Reflex
Somatosensory Evoked Potential	Somatosensory Evoked Potential (SEP)
	Short-latency Somatosensory Evoked Potential (SSEP)
	Electrocardiogram Triggered Somatosensory Evoked Potential (ECG-SSEP)
Auditory Evoked Potential	Auditory Brainstem Response (ABR)
	Middle Latency Responses (MLR)
	Slow Vertex Responses (SVR)
Visual Evoked Potential	Pattern Reversal Visual Evoked Potential (PR-VEP)
	Goggle Visual Evoked Potential (G-VEP)
	Flash Visual Evoked Potential (F-VEP)

the common EMG examination items involved in these reports.

The measurement equipment used in this paper is an electromyographic evoked potential inspection device MEB-9200K, which was produced by Nihon Kohden Corporation of Japan. The power consumption is 430W. This device features powerful extendibility and leads the heading position in EMG manufacturing fields. Up to eight separated examinations can be conducted at the same time using this device. This MEB-9200K device has been employed in Sichuan Provincial Hospital of TCM for many years.

B. EXAMINATION ITEMS SELECTION

Considering a large number of EMG examination items, only parts of these items are included in this study. Specifically, two types of EMG examination data are involved. F-MNCS and ABR are selected on account of the lower data dimension but a larger amount. In the clinical examination, the F-MNCS examination is usually used as an important auxiliary method to diagnose facial paralysis. Meanwhile, the ABR examination is usually employed to judge tinnitus for patients. Both examinations items are performed in the head. Some patients would have both of the examination items. The characteristics of each item will be discussed below.

1) FACIAL MOTOR NERVE CONDUCTION STUDY

Facial Motor Nerve Conduction Study(F-MNCS) is measured by electrical stimulations of the marginal mandibular branch and by evoked EMG of the mentalis muscle. The common F-MNCS value is 48.8 +/− 3.68 (means +/− SD) m/s [14]. The F-MNCS examination can provide a reliable reflection of the prognosis for facial palsy such as Bell's palsy(also named as idiopathic palsy) and other nerve-related disorders, as well as the degree of nerve damage. Thus, researches and applications on F-MNCS are popular all over the world. The common F-MNCS examination includes many data fields. Measurement devices made from different manufactures would have a slight impact on data fields. In the original reports, there are 19 data fields collected. Table 2 shows every data field and its corresponding description in F-MNCS examination. Considering the privacy of patients, their names are not mentioned.

TABLE 2. Data fields and descriptions for F-MNCS examination data.

Data field	Description
In/Out_Patient	Hospitalization status
Age	Patient's age
Sex	Patient's gender
Date	Date of electromyography
Do_ABR	Whether the ABR check was done meanwhile
Do_Blink	Whether the Blink check was done meanwhile
RT_L_Latency	Left latency of rami temporalis
RT_L_Amplitude	Left amplitude of rami temporalis
RT_L_Area	Left area of rami temporalis
BB_L_Latency	Left latency of buccal branch
BB_L_Area	Left area of buccal branch
BB_L_Amplitude	Left amplitude of buccal branch
RT_R_Latency	Right latency of rami temporalis
RT_R_Amplitude	Right amplitude of rami temporalis
RT_R_Area	Right area of rami temporalis
BB_R_Latency	Right latency of buccal branch
BB_R_Area	Right area of buccal branch
BB_R_Amplitude	Right amplitude of buccal branch
abnormal	inspection result

In Table 2, the data field *In/Out_Patient* represents whether the patient is hospitalized. The value of *age* is an integer. In the data field *sex*, the number 1 and 0 denote the man and woman, respectively. The data field *date* is expressed as yyyy/mm/dd. The unit of the incubation period is milliseconds or ms. The unit of the amplitude is mV. The unit of the area is mV/ms. As for the data field *abnormal*, the number 1 and 0 represent the conclusions of normal and abnormal originally drawn by the doctor who conducts the examination.

FIGURE 1. **EMG examination device MEB-9200K outline image.**

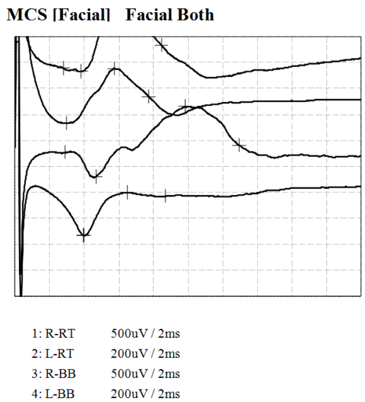


FIGURE 3. **Facial motor nerve conduction examination electromyography.**

The data field *Do_ABR* represents whether the ABR examination is performed at the same time. The data field *Do_Blink* represents whether the blink reflex examination is performed as well.

Figure 2 shows the mounting position of the electrode slices in our F-MNCS clinical examination. Figure 3 shows one typical EMG data obtained from a F-MNCS examination report. Since this is an original output figure from our EMG equipment, there is no x-axis and x-axis depicted in this figure. The horizontal axis and the vertical axis represent time and voltage, respectively. Every horizontal axis cell represents 2ms. The four waveforms from top to bottom are the right rami temporalis, left rami temporalis, the right-side buccal branch, and the left-side buccal branch, respectively. The indication *Facial Both* means both sides of the face are examined.

FIGURE 2. **Schematic of the position of the electrode slices in the facial motor nerve conduction examination.**

2) AUDITORY BRAINSTEM RESPONSE

The principle of ABR is to stimulate the auditory organs by a certain intensity of sound, and record a series of electrical activities generated by the auditory system in the cortex [15]. ABR is generally employed in newborn hearing screening, determination of organic and functional deafness, intraoperative monitoring of acoustic neuroma, and monitoring the effect of ototoxic drugs on hearing, etc. Moreover, ABR can also be used as an auxiliary examination for patients with clinical manifestations of facial paralysis, dizziness, headache, and tinnitus or hearing loss.

The common ABR examination report includes a great number of data fields, and Table 3 shows each data field and its corresponding meaning in our reports. The basic information of the patient in ABR is the same as the one in the F-MNCS examination in Table 2, while other examination contents and the data units are totally different. The units of the delay time and acoustic stimulus intensity are ms and dB, respectively. The units of interval time and amplitude are ms and μV, respectively. The data field *Do_MNCSF* represents whether the F-MNCS examination is performed at the same time.

Figure 4 illustrates the mounting position of the electrode slices in the ABR clinical examination, which is similar to the F-MNCS. One typical EMG data obtained from an ABR examination report containing the time-voltage curves is shown in Figure 5. The horizontal axis and the vertical axis represent time and voltage, respectively. Every horizontal axis cell represents 2ms. There are 8 waveforms depicted in this figure.

C. DATA INCLUSION CRITERIA

The inclusion criteria method is an ordinary method in medical data processing field. This method could be considered as a pre- or special data cleaning, focusing on data types or attributes instead of values. The data row would be included or excluded as a whole, instead of modifying its inner values. Nevertheless, considering that the data are all collected in clinical circumstance, some data would be influenced more or less by uncertain and unpredictable reasons. For instance, some patients would not conduct the comprehensive but only certain examination items. These data cannot be directly applied to the experimental training of machine learning without being cleaned in advance. To achieve this goal, some inclusion criteria are supposed to be designed conventionally in medical data processing. In view of the study objects

TABLE 3. The description of every data field of ABR examination data.

Data field	Description
In/Out_Patient	Hospitalization status
Age	Patient's age
Sex	Patient's gender
Do_MNCSF	Whether the facial MNCS check was conducted at the same time
Abnormal	Examination result
L_Smit	Left stimulus frequency
R_Smit	Right stimulus frequency
R_Latency_1	Right latency 1
R_Latency_2	Right latency 2
R_Latency_3	Right latency 3
R_Latency_4	Right latency 4
R_Latency_5	Right latency 5
R_Latency_A	Right latency A
R_Latency_B	Right latency B
L_Latency_1	Left latency 1
L_Latency_2	Left latency 2
L_Latency_3	Left latency 3
L_Latency_4	Left latency 4
L_Latency_5	Left latency 5
L_Latency_A	Left latency A
L_Latency_B	Left latency B
R_Interval_13	The interval on the right I-III
R_Interval_35	The interval on the right III-V
R_Interval_15	The interval on the right I-V
R_Amp_5A	The amplitudeon the right V-A
R_Amp_1B	The amplitudeon the right I-B
L_Interval_13	The interval on the left I-III
L_Interval_35	The interval on the left III-V
L_Interval_15	The interval on the left I-V
L_Amp_5A	The amplitudeon the left V-A
L_Amp_1B	The amplitudeon the left I-B

and circumstances, the corresponding inclusion criteria are designed as below. Based on these criteria, only about half of the 2352 raw reports are screened out.

1) The data of the reports should contain the complete information of the patients, such as age, gender, hospitalization, and reporting time.

2) The examination items, data and conclusions should be complete.

3) The patients should cooperate with the doctor actively during the examination.

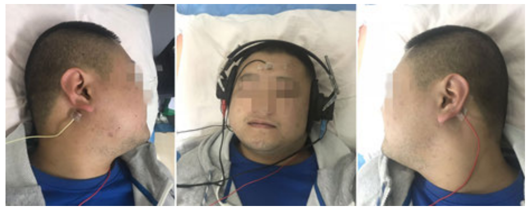


FIGURE 4. Schematic of the position of the electrode slices in the ABR examination.

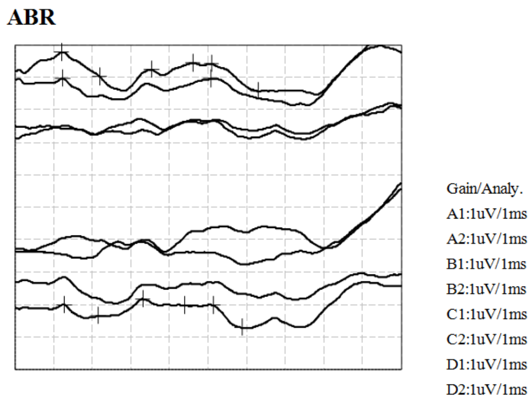


FIGURE 5. ABR examination electromyography.

D. DATA CLEANING

Data cleaning is a critical step in machine learning, since its methods and results would have a direct impact on the machine learning model performance and the final conclusion. Data cleaning would be depended on the specific task. The common methods of data cleaning include different sorts of mission-related preprocessing, data discarding or repair.

In real situations, some patients would only conduct left-side or right-side facial examination, buccal or temporal branch examination, instead of comprehensive examinations. As a consequence, some examinations would obtain the result like *NaN*(*Not a Number*). In this case, the *NaN* value processing method needs to be applied, such as deleting, mean-filling, and zero-filling. Since the data set employed in this paper is small-scale, the deleting method is obviously unsuitable. In the meantime, the mean-filling method may affect the feature selection of subsequent machine learning algorithms. Therefore, the zero-filling method is finally adopted for *NaN* value processing.

The data repetition is an inevitable issue that should be concerned beforehand, which is mostly due to the duplicated printing of EMG reports during the data collecting period. The duplicated reports would not appear adjacently in the data set tables, resulting in that the duplicated reports cannot be deleted immediately and manually. In this paper, the identifying/removing strategy is used. To be specified, if two or more rows with identical gender, age, examination time and first two data fields, only one would be remained.

Based on the above the inclusion criteria and data cleaning, 575 F-MNCS reports and 233 ABR examination reports are selected from the 2352 raw reports. Thus, data sets are

TABLE 4. F-MNCS and ABR normal and abnormal reports statistics.

	F-MNCS data set	ABR data set
0/normal/negative class	97	193
1/abnormal/positive class	478	40
total	575	233

established, which are organized in the form of data table for the purpose of future program visiting. Table 4 shows the statistics of the size and positive/negative distribution of the data sets.

E. DATA LABEL

The diagnosis conclusions conducted by the EMG examination doctors are considered as the standard conclusions or the labels. In general, the diagnosis conclusions of the EMG examination doctors can be roughly divided into four grades: the normal, the mild nerve damage, the moderate nerve damage and the severe nerve damage. In the actual treatment process, whether the patient is normal needs to be taken into consideration. Thus, in this paper, the diagnosis conclusions of doctors are classified into two categories: the normal and the abnormal, which can be directly obtained from the data field *Abnormal* in Table 2 and 3. It is particularly significant to note that label 0 denotes normal (without disease) or negative class, whereas 1 denotes abnormal or positive class. Consequently, this classification is a typical binary one. The data distribution for both labels is shown in Table 4.

III. METHOD

Both of the aforementioned data sets are used for the applications of four machine learning algorithms, including random forest, linear regression, SVM and logistic regression. In this work, the data standardization is implemented before applying the algorithms. In addition, the performances among four algorithms in cases with and without standardization are compared. The detailed comparisons on accuracy and recall rate among four algorithms are carried out.

A. DATA STANDARDIZATION

The data standardization is the foundation of machine learning. Both dimension and value of an indicator would make a great difference when it comes to evaluating an indicator. Without data processing, the results of the data analysis would be affected [16]. The common standardization methods include decimal calibration standardization, standard deviation standardization and deviation standardization.

The decimal calibration standardization method is to map attribute values to $[-1,1]$ by shifting decimal numbers of attribute values.This method is mainly used is to eliminate the influence of the units. Its conversion formula can be described as follows:

$$x^* = x/10^k \tag{1}$$

The standard deviation standardization is a method which can make the mean and the standard deviation of the data to be 0 and 1, respectively. This standardization method can be applied to eliminate the effect of units and the variation of variables. Its conversion formula can be expressed as:

$$x^* = (x - \bar{x})\big/\sigma \tag{2}$$

where x denotes the original data, $\bar{x}$ represents the mean, and σ is the standard deviation.

The deviation standardization is a method of linear mapping the original data to [0,1]. Its main purpose is to remove the impact of dimensions and range of the data, while maintaining the linear relationship among the original data. The specific definition can be written as follows:

$$x^* = (x{-}min)(max - min) \tag{3}$$

where max denotes the maximum of the sample data, min represents the minimum of the sample data.

Considering the multi-dimensional characteristics of the EMG data and the correlation among data, the deviation standardization is finally selected as the data standardization method in this paper.

B. LINEAR REGRESSION MODEL

The linear regression method describes the relationship among data with a straight line as accurate as possible, which can be expressed by the function [17]. The function of the linear regression model can be expressed as:

$$h_\theta = \theta_0 + \theta_1 x_1 + \ldots + \theta_n x_n \tag{4}$$

where the main parameters are $\theta = [\theta_0, \theta_1, \ldots, \theta_n]^T$.

Since the data fitting performance would vary from the type of the linear regression model, the selection of the model is particularly significant. In order to select a more accurate model to describe the linear relationship among data, a function reflecting the difference between the linear regression model and the real data is introduced. This function named as cost function is defined as follows:

$$J(\theta) = {}^1\!/_{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 \tag{5}$$

The meaning of the above-mentioned cost function is similar to that of L_2 distance or Euclidean distance. Since the smaller cost function value corresponds to the closer relationship and the better performance.

C. LOGISTIC REGRESSION MODEL

Logistic regression model is a common probabilistic nonlinear regression model [18] and its dependent variables can only be either 0 or 1. Assuming that there are p independent variables $X = [x_1, x_2, \ldots, x_p]$, the probability of y equivalent to 1 is denoted as $p = P(y = 1|X)$. The probability ratio (also known as the odds ratio) between y equivalent to 1 and 0 is $p/(1-p)$. The logistic regression conversion formula can be obtained by taking natural logarithm of the odds ratio, which can be defined as:

$$\log it(p) = \ln(\frac{p}{1-p}) \tag{6}$$

Based on the above derivation, the logistic regression model can be constructed by the logistic function and defined as:

$$\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon \tag{7}$$

D. SVM

As one of the classical classifiers, SVM(Support Vector Machine) was proposed by Vapnik et al. in 1995 [19]. In 2012, SVM was extensively applied by Guang-Bin Huang et al. in their studies [19], and was developed to the extreme.

However, there are some shortcomings of the SVM. For instance, the computation will be substantial if the sample size is large and the dimension of the kernel function is high; SVM is too sensitive to the missing data; kernel functions for nonlinear problems have no universal standard. Due to the above shortcomings, SVM has been gradually replaced by other methods.

The critical parameters of SVM are the kernel function, the kernel function coefficient gamma, and the error penalty coefficient C.

The frequently-used kernel functions are Gaussian function, polynomial kernel function, sigmoid kernel function, and linear kernel function. The higher the error penalty coefficient C is, the higher the accuracy of the training sample is. However, the model generalization ability would be lower.

E. RANDOM FOREST MODEL

The random forest model is based on the idea of bagging ensemble learning method. Many simple decision trees are integrated into a complex forest for final prediction.

At present, the dominant decision tree algorithms include ID3, C4.5 and CART. Firstly, The ID3 is regarded as the most classical algorithm. Its core is the selection of the proper attributes based on the information gain at any node in the tree [21]. Secondly, the improvement of C4.5 algorithm over ID3 is depending on information gain rate in node attribute selection instead of information gain. Consequently, C4.5 can solve the continuous attributes problem which cannot be solved by ID3 [22]. Thirdly, CART (classification and regression tree) proposed by Breiman et al in 1984. Is currently a quite popular decision tree algorithm [23]. In this paper, the CART algorithm is used as the decision tree algorithm for the construction of the random forest.

The difference among the above three algorithms is that the decision tree constructed by CART selects the optimal segmentation feature based on the Gini index rather than the information gain. Meanwhile, every branch of the decision tree is binary. The Gini index is theoretically similar to entropy. For a random variable X with K states and corresponding probabilities $p_1, p_2, \ldots, p_k$, its Gini index can be defined as follows:

$$Gini(X) = \sum_K p_k(1-p_k) = 2p(1-p)$$
$$X \sim Bernoulli(p) \tag{8}$$

Under the condition of given feature A, the Gini index for the data set D can be obtained by combining the above formulas, as shown below.

$$Gini(D, A) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \tag{9}$$

Similar to the principle of entropy, the greater the value of $Gini(D, A)$ is, the greater the sample uncertainty is. Taking this into account, the value of $Gini(D, A)$ should be as small as possible when selecting the feature A.

The case k = 1 represents a randomly selected attribute of a segment. In general, $k = \log_2 d$ is recommended. Figure 6 illustrates the random forest schematic diagram for CART algorithm.

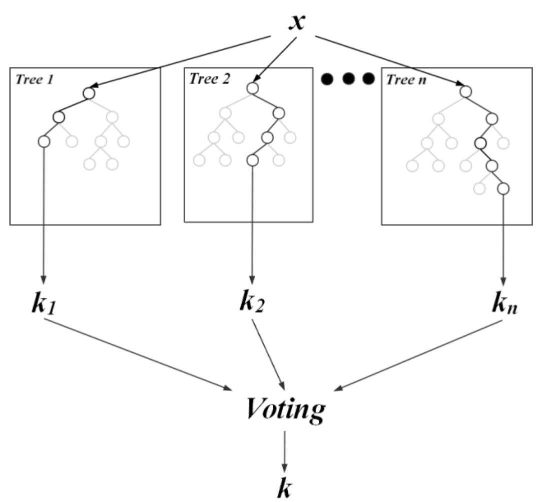


FIGURE 6. Schematic of random forest for CART algorithm.

F. SYSTEM DESIGN

According to the previous description, the whole process implemented in this paper can be divided into four steps: 1) data collection and cleaning; 2) data standardization; 3) implementation of four machine learning algorithms; 4) comparison. Figure 7 illustrates the flow chart of the data processing and verification.

The whole programming simulation is written with Python3.6 platform and sklearn toolkit. The sklearn (scikit-learn) is a python-based simple but efficient tool in data mining and data analysis, released in 2007 for the first time.

IV. RESULTS

As mentioned above, two data sets (F-MNCS and ABR) are finally constructed. Four machine learning algorithms will

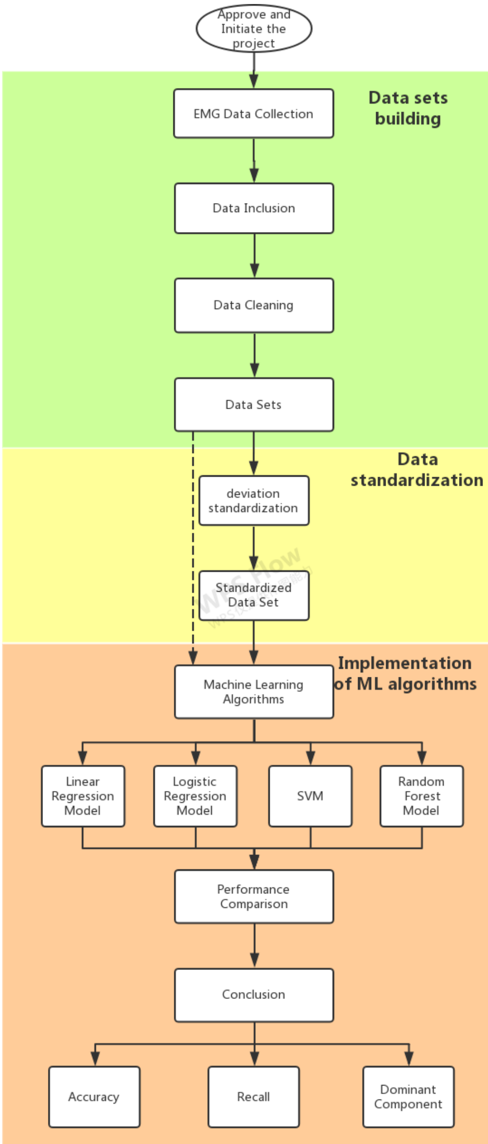


FIGURE 7. Flow chart of the data processing and verification.

be applied to these two data sets separately with subsequent comparison and discussion.

A. F-MNCS DATA

In term of F-MNCS examination, the data set size is 575, including 90 negative samples and 485 positive samples, as shown in Table 4. Four machine learning algorithms, including random forest, linear regression and logistic regression, are applied to the F-MNCS data set.

To all four models, the data set division strategy remains the same. There are 575 samples divided into half in the F-MNCS data set, namely, 287 samples in the training set and 288 samples in the test set. In linear regression, the training set is directly used to fit the regression coefficients, and the test set is used to calculate the accuracy without cross-validation. Nevertheless, in random forest and logistic regression, three times of cross-validation have been carried out, and the final accuracy is the average of cross-validation results.

First, after a great deal of parameter adjustment for the linear regression model, we have conducted non-playback test on the trained linear regression model with the test set of F-MNCS. The label *abnormal* in data set tables is regarded as the dependent variable, and all other data fields except *abnormal* are independent variables. As can be seen from Table 6, the accuracy based on linear regression model in the cases without and with deviation standardization is 0.8069 and 0.8313, respectively.

Then, similar to linear regression, the logistic regression model is established after the adjustment of model coefficients. And then, the non-playback and cross-validation tests are implemented on the test set of F-MNCS examinations. As shown in Table 6, the accuracy in the cases without and with deviation standardization is 0.8572 and 0.8624, respectively.

Subsequently, SVM is employed for data set processing. As mentioned above, there is no uniform standard for the kernel function of SVM. Based on this consideration, linear kernel function, Gaussian kernel function, and polynomial kernel function are used for performance comparison in this paper. Through parameter adjustment and experiments, the Gaussian kernel function with the optimal performance is finally selected. After balancing the generalization ability and accuracy of the model, the optimal solution is obtained with setting the penalty function C to 5 and kernel function coefficient gamma to 1. The average accuracy of SVM cross-validation data with and without standardization is 0.8348 and 0.8989, respectively.

Finally, we have built the random forest having 200 trees, each with 3-level depth. Three times of cross-validation are processed in the cases with and without deviation standardization. The accuracy and the final mean are shown in Table 5.

TABLE 5. Accuracy of random forest algorithm in cross-validation.

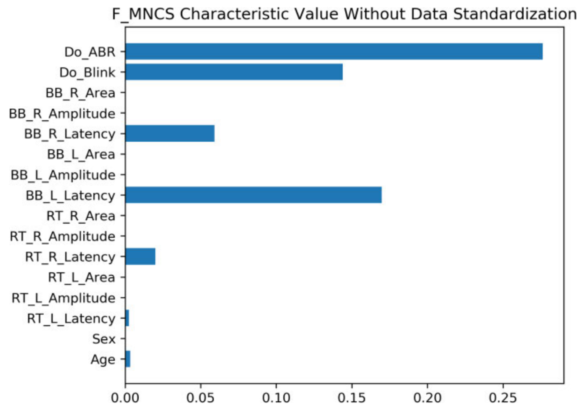
Numbers of cross-validation	Accuracy without deviation standardization	Accuracy with deviation standardization
1	0.9583	0.9687
2	0.9114	0.9166
3	0.9267	0.9476
Mean	0.9321	0.9443

In the cases with and without deviation standardization, the accuracy of four machine learning algorithms with F-MNCS data set is listed in Table 6. For a specific algorithm, the accuracy with deviation standardization is always a little bit higher than that without deviation standardization (presented in Table 6), indicating that the standardization is an effective method for performance improvement. Meanwhile,

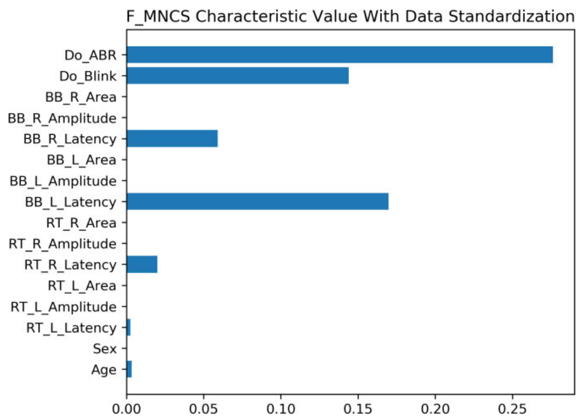
TABLE 6. Summary of the algorithm accuracy of F-MNCS.

Algorithm	Accuracy without deviation standardization	Accuracy with deviation standardization
Linear Regression	0.8069	0.8313
Logistic Regression	0.8572	0.8624
SVM	0.8348	0.8989
Random Forest	0.9321	0.9443

the results of the accuracy comparison demonstrate that the random forest algorithm has the optimal performance among the four algorithms. In addition, the accuracy of random forest is over 0.9, indicating that random forest should be adopted as the first and optimal choice in our research.



(a) without deviation standardization



(b) with deviation standardization

FIGURE 8. Comparison of the characteristic values of F-MNCS.

To verify the influence of deviation standardization on the original data relationship, the random forest is employed to extract characteristics of the test set in the cases with and without deviation standardization, respectively. The importance of each characteristic is depicted in Figure 8.

TABLE 7. Model performance comparisons of F-MNCS data.

	Logistic Regression	Random Forest	SVM
Accuracy	0.8624	0.9443	0.8989
Recall	0.4536	0.9278	0.8144
Precision	0.7096	1	0.9518
P-Value	0.0967		0.0865

TABLE 8. Summary of the algorithm accuracy of ABR examinations.

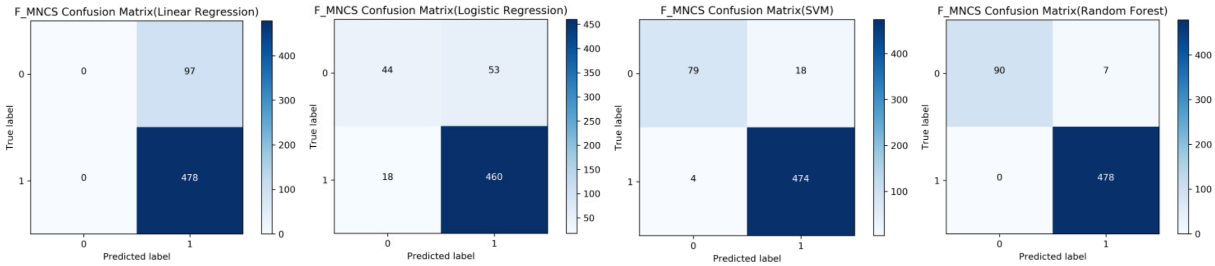
Algorithm	Accuracy without deviation standardization	Accuracy with deviation standardization
Linear Regression	0.1330	0.1416
Logistic Regression	0.8541	0.9915
SVM	0.8369	1
Random Forest	0.7802	0.9957

As can be seen from Figure 8, there is no significant difference in characteristic values and distributions extracted from random forest in the cases with and without the standardization. Combined with Table 4 , the accuracy of random forest algorithm has been improved significantly with deviation standardization, indicating that the deviation standardization is suitable for the data standardization process of F-MNCS.

Moreover, it can also be concluded from Figure 8 that gender and age are not critical factors in determining whether the patient is normal or not. In addition, the doctors would selectively add blink reflex and/or ABR examinations depending on individual differences of each patient, such as different diseases, different progression of the same disease, etc. Only considering EMG data rather than external factors, BB_L_Latency is the most significant factor or characteristic affecting the prediction or diagnosis results from EMG data, followed by BB_R_Latency. RT_R_Latence and RT_L_Latence also have certain impact on the diagnosis results, while the other examination items have almost no effect.

Besides accuracy, recall rate is also a very significant indicator to evaluate the performance of machine learning models. In order to visualize the recall rate, the confusion matrix is adopted in this paper. The confusion matrixes of the above-mentioned four algorithms with deviation standardization are illustrated in Figure 9.

In the above confusion matrix (shown in Figure 9), the y-label is the diagnosis made by the doctor in the original examination reports, and the x-label is the predicted value of the abovementioned four algorithms.



(a) Confusion matrix of linear regression (b) Confusion matrix of logistic regression (c) Confusion matrix of SVM (d) Confusion matrix of random forest

FIGURE 9. Confusion matrixes of four algorithms for F-MNCS.

TABLE 9. Cross-validation results of random forest and logical regression.

	Random Forest		Logical Regression		SVM	
Numbers of cross-validation	Accuracy without deviation standardization	Accuracy with deviation standardization	Accuracy without deviation standardization	Accuracy with deviation standardization	Accuracy without deviation standardization	Accuracy with deviation standardization
1	0.8462	0.9872	0.8481	0.9747	0.8354	1
2	0.9103	1	0.9091	1	0.8961	1
3	0.5844*	1	0.8052	1	0.7792	1
Mean	0.7803	0.9957	0.8541	0.9916	0.8369	1

As depicted in Figure 9, the data used in this paper is the unbalanced binary data. In the linear regression, all predicted values are set as 1 and more than 80% accuracy can be achieved. Therefore, more performance indicators are needed to determine the effectiveness of the model. Based on the analysis of the confusion matrix in Figure 9, a desirable result has not been achieved by applying the linear regression. As a consequence, the linear regression is excluded from the model comparisons in this paper. The model performance comparisons of logistic regression, SVM and random forest are shown in Table 7.

Generally, 0.05 is used as the threshold of P-Value. In the meantime, the data used in this paper is unbalanced binary data, and the data structure is too simple. Therefore, the P-Values shown in Table 7 indicate that there is a remarkable difference among the performance of logistic regression, SVM and random forest in F-MNCS data.

Through the confusion matrix visualization, we can conclude that the random forest has high accuracy and recall rate in predicting the F-MNCS data set.

B. ABR DATA

With respect to ABR examinations, the data set size is 233, including 193 negative samples and 40 positive samples (shown previously in Table 4). Four abovementioned machine learning algorithms are applied to the ABR data set. We have rebuilt all four models using new data set, where the random forest with 200 trees and 3-level depth for each tree.

The data set division strategy for the ABR data set is similar to that for the F-MNCS data set. The 233 samples are divided in half, that is, the training set with 116 samples and the test set with 117 samples. In all models, the final accuracy is the average of the three cross-validation results.

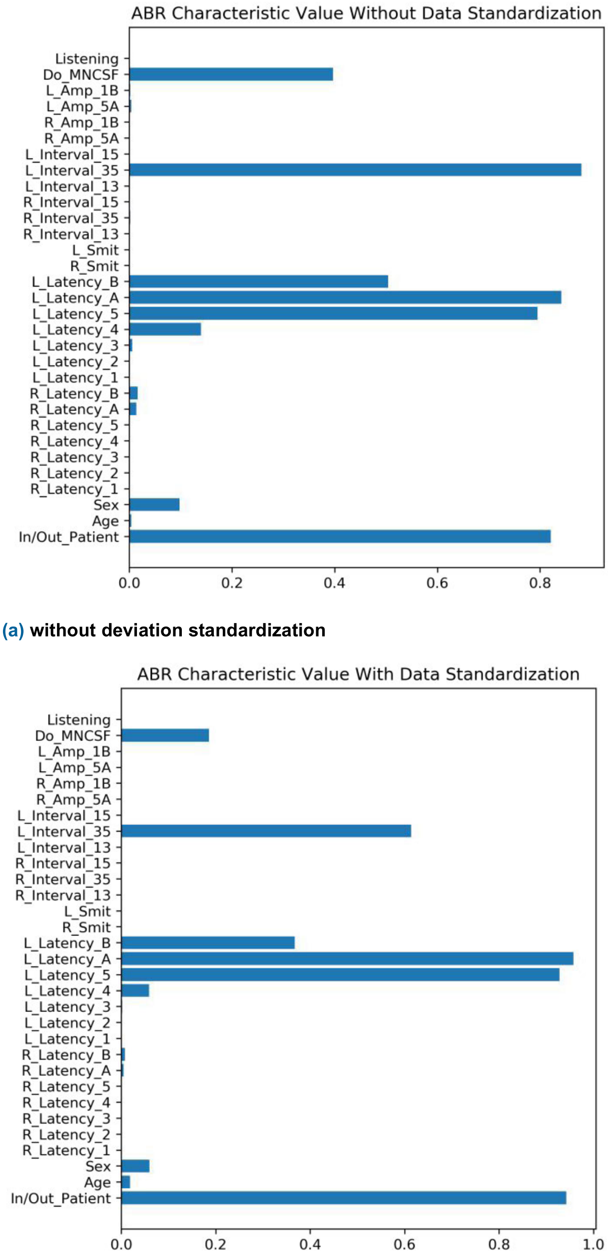
The accuracy in the cases with and without deviation standardization of four machine learning algorithms using ABR data set is summarized in Table 8. It can be found that the accuracy with deviation standardization is always a little bit higher than that without deviation standardization. The comparison results once again verify that the standardization is an effective method for the performance improvement. In the meantime, the accuracy comparison results among the four algorithms once again indicate that the performance of the random forest algorithm is evidently superior to that of the other two algorithms, thus selecting the random forest algorithm in our research.

As shown in Table 8, the accuracy of random forest is higher than that of logistic regression in the cases with data standardization. However, the result is opposite without data standardization. To explore the underlying reason, the accuracy of the two algorithms for several cross-validation tests has been further studied. The detailed cross-validation results of random forest and logical regression are presented in Table 9.

As can be seen from Table 10, the accuracy of random forest in the first two cross-validations is almost the same with that of logistic regression, while the accuracy of random

TABLE 10. Comparisons of model indicators for ABR data.

	Logistic Regression	Random Forest	SVM
Accuracy	0.9916	0.9957	1
Recall	1	0.9948	1
Precision	0.9747	0.9897	1
P-Value	0.5173	0.3739	



(a) without deviation standardization

(b) with deviation standardization

FIGURE 10. Feature extraction of ABR by random forest.

forest in the last cross-validation is significantly reduced. Through further analysis, it is considered that there are two reasons for this reduction. One is that the random values in the computer are not real or ideal random values but pseudo-random values or approximate values, which may lead to the over-concentration of some features in data segmentation, thus reducing the accuracy. The other reason is that the quantity of the ABR examination data is small, to be specific, only 233 data are available, resulting in inadequate training for the random forest model.

Similar to the previous study of F-MNCS, the feature extraction of ABR examination data is carried out to verify whether the deviation standardization would affect the correlation of the data. The characteristic values of ABR data set extracted from random forest are shown in Figure 10.

As illustrated in Figure 10, there are obvious differences in the characteristic values extracted from random forest in the cases with and without the deviation standardization. At the same time, the accuracy of random forest, SVM and logistic regression with standardization has achieved tremendous improvements. However, since the quantity of data is not enough to make the algorithm run effectively, the accuracy of the linear regression is relatively low. In summary, due to the small amount of data, there is a risk of overfitting for ABR data, even if using the traditional machine learning methods without massive data.

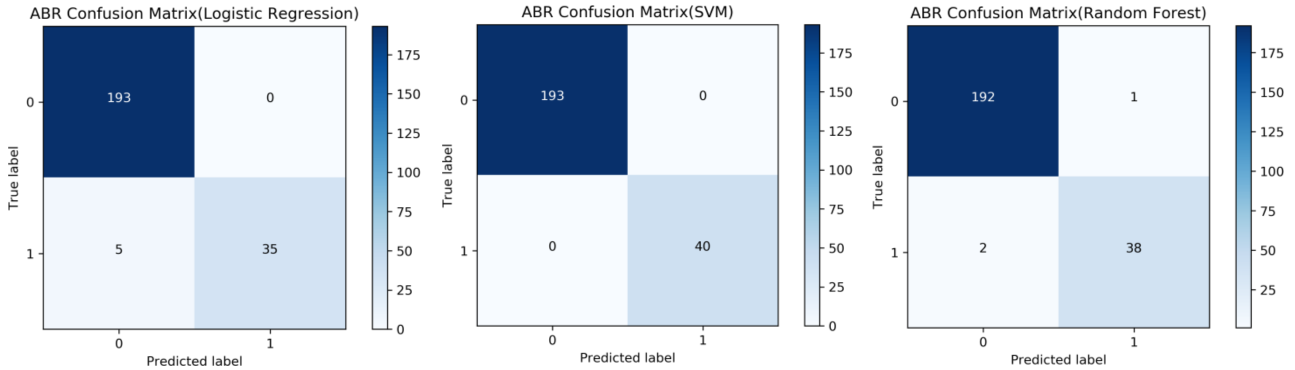
In addition, it can be found that the data fields *Age* and *Sex* are not significant to the results of ABR examinations, regardless of the external interference factors such as data fields *In/Out_Patient* (in or out patient) and *DO_MNCSF*(conducting both F-MNCS and ABR examinations). The most significant data fields in ABR examinations are L_Latency_5, L_Latency_A and L_Interval_35. Moreover, L_Latency_b and L_Latency_4 also have some certain influence on the results.

The confusion matrixes obtained by using logistic regression and random forest to predict ABR data sets are shown in Figure 11.

As shown in Figure 11, 193 patients are labeled as normal (also shown in Table 4), but 2 are misjudged as abnormal by random forest. Among the 40 patients judged as abnormal, only one of them is misjudged as normal by random forest. All predictions on normal are accurate, while there are 5 abnormal cases predicted to be normal by logistic regression. SVM has an excellent performance in ABR data and accurately predicts all data.

Table 10 presents the differences of the performance between logistic regression, SVM and random forest.

According to Table 10, the indicators of logistic regression, SVM and random forest in ABR data are close. The P-Values also far exceed the threshold of 0.05, indicating that there is no significant difference among the performance of logistic regression, SVM and random forest in ABR data. However, as mentioned in the previous cross-validation of ABR data, there is a risk of overfitting due to the small amount of



(a) Confusion matrix of logistic regression (b) Confusion matrix of SVM (c) Confusion matrix of random forest

FIGURE 11. Confusion matrixes of three algorithms for ABR.

ABR data. Therefore, based on the above results of F-MCNS and ABR, the final conclusion in this paper is that the random forest model is an ideal model for EMG data.

On the one hand, the small quantity of data leads to the inadequate training of the model. As a consequence, the random forest model obtained from ABR data training is not very reliable. It can be predicted that with the increase of the data set size, the model accuracy without standardization would be promoted, and the influence of the deviation standardization on the algorithm accuracy and feature values would be diminished.

On the other hand, the dimension would have a great impact on the model with small-scale data set. Consequently, the traditional machine learning algorithm is better than the DL method in the EMG data.

V. CONCLUSION

Based on the performance comparisons of four machine learning algorithms, it is can be found that the random forest algorithm is superior to the linear algorithm and logistic algorithm in EMG data, especially in F-MNCS and ABR data. Moreover, the data standardization such as derivation standardization is an effective method for performance improvement such as accuracy. Meanwhile, it is also found that the most significant influencing factor of the F-MNCS examination is BB_L_Latency, followed by BB_R_Latency. RT_R_Latence and RT_L_Latence also have some certain influence on the results, while the rest inspection indicators have little effect on the results. As for the ABR examination, the most important impact factors of ABR are L_Latency_5, L_Latency_A and L_Interval_35, followed by L_Latency_b and L_Latency_4.

The further research direction of this project will aim at identifying EMG data based on the waveform of EMG data and determining the values of various tests according to the characteristics of the waveform. At present, this process is manually done by the doctor on the basis of the characteristics of the waveform. Combined with the method applied in this paper and the further research, a complete system can be established. The final goal is to shorten time for the EMG examination, thus making full use of medical resources.

The results of this paper indicate that the application of machine learning in data mining and analysis provides an innovative way and direction for the development of the clinical diagnosis and the treatment technology. Simultaneously, the results also provide a significant reference for the diagnosis based on clinical data and the improvement of medical efficiency.

APPENDIX

Data set download link:

DOI: 10.21227/5007-5t85

ACKNOWLEDGMENT

(Tao Liu and Yuqi Tang contributed equally to this work.)

REFERENCES

[1] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: Review, opportunities and challenges," *Briefings Bioinformatics*, vol. 19, no. 6, pp. 1236–1246, Nov. 2018.

[2] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, vol. 316, no. 22, p. 2402, Dec. 2016.

[3] J. A. Golden, "Deep learning algorithms for detection of lymph node metastases from breast cancer: Helping artificial intelligence be seen," *JAMA*, vol. 318, no. 22, p. 2184, Dec. 2017.

[4] M. J. Aminoff, *Electromyography in Clinical Practice: Clinical and Electrodiagnostic Aspects of Neuromuscular Disease*, 2nd ed. New York, NY, USA: Churchill Livingstone, 1987.

[5] J. V. Basmajian, C. De Luca, and C. J. Deluca, *Muscles Alive: Their Functions Revealed by Electromyography*, 5th ed. Baltimore, MD, USA: William & Wilkins, 1985, pp. 1–20.

[6] M. J. Aminoff, *Electromyography in Clinical Practice: Clinical and Electrodiagnostic Aspects of Neuromuscular Disease*, 2nd ed. New York, NY, USA: Churchill Livingstone, 1987, pp. 3–15.

[7] T. J. Doherty and D. W. Stashuk, "Decomposition-based quantitative electromyography: Methods and initial normative data in five muscles," *Muscle Nerve*, vol. 28, no. 2, pp. 204–211, Aug. 2003.

[8] S. G. Boe, D. W. Stashuk, and T. J. Doherty, "Within-subject reliability of motor unit number estimates and quantitative motor unit analysis in a distal and proximal upper limb muscle," *Clin. Neurophysiol.*, vol. 117, no. 3, pp. 596–603, Mar. 2006, doi: 10.1016/j.clinph.2005.10.021.

[9] D. W. Stashuk, L. Pino, A. Hamilton-Wright, T. Doherty, and S. Boe, ''Interpretation of QEMG data,'' in *Proc. Gen. Meeting Amer. Assoc. Neuromuscular Electrodiagnostic Med. (AANEM)*, Phoenix, AZ, USA, 2007.

[10] A. Subasi, ''Classification of EMG signals using PSO optimized SVM for diagnosis of neuromuscular disorders,'' *Comput. Biol. Med.*, vol. 43, no. 5, pp. 576–586, Jun. 2013.

[11] J. Yousefi and A. Hamilton-Wright, ''Characterizing EMG data using machine-learning tools,'' *Comput. Biol. Med.*, vol. 51, pp. 1–13, Aug. 2014.

[12] A. Phinyomark and E. Scheme, ''EMG pattern recognition in the era of big data and deep learning,'' *Big Data Cogn. Comput.*, vol. 2, no. 3, p. 21, Aug. 2018.

[13] T. Ziemniak, ''Use of machine learning classification techniques to detect atypical behavior in medical applications,'' in *Proc. 6th Int. Conf. IT Secur. Incident Manage. IT Forensics*, May 2011.

[14] H. Tojima, ''Measurement of facial nerve conduction velocity and its application to patients with Bell's palsy,'' *Acta Oto-Laryngologica*, vol. 104, no. 446, pp. 36–41, Jan. 1987, doi: 10.3109/00016488709121839.

[15] E. Skoe and N. Kraus, ''Auditory brain stem response to complex sounds: A tutorial,'' *Ear Hearing*, vol. 31, no. 3, pp. 302–324, Jun. 2010.

[16] L. Liu and M. T. Özsu, Eds., ''Data Standardization,'' in *Encyclopedia of Database Systems*. Boston, MA, USA: Springer, 2009.

[17] F. O. Lorenz, J. Neter, W. Wasserman, and M. H. Kutner, ''Applied linear statistical models (3rd ed.),'' *J. Amer. Stat. Assoc.*, vol. 87, no. 419, p. 902, Sep. 1992.

[18] Y. Yu, ''Research of a new method for solving linear regression,'' in *Proc. Int. Conf. Transp. Logistics, Inf. Commun., Smart City (TLICSC)*, 2018, p. 5.

[19] C. Cortes and V. Vapnik, ''Support-vector networks,'' *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[20] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, ''Extreme learning machine for regression and multiclass classification,'' *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 2, pp. 513–529, 2012.

[21] Z. Wang, Y. Liu, and L. Liu, ''A new way to choose splitting attribute in ID3 algorithm,'' in *Proc. IEEE 2nd Inf. Technol., Netw., Electron. Autom. Control Conf. (ITNEC)*, Dec. 2017, p. 5.

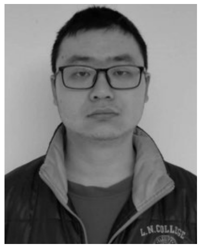
[22] S. Sathyadevan and R. R. Nair, ''Comparative analysis of decision tree algorithms: ID3, C4.5 and random forest,'' in *Proc. Int. Conf. Comput. Intell. Data Mining (ICCIDM)*, New Delhi, India, Dec. 2014.

[23] S. Hamze-Ziabari and T. Bakhshpoori, ''Improving the prediction of ground motion parameters based on an efficient bagging ensemble model of M5' and CART algorithms,'' *Appl. Soft Comput.*, vol. 68, pp. 147–161, Jul. 2018.

TAO LIU was born in Guizhou, China, in 1977. He received the B.S. and M.S. degrees in control theory and engineering from Chongqing University, Chongqing, China, in 2000 and 2003, respectively, and the Ph.D. degree in signal and information processing from the University of Electronic Science and Technology of China, Chengdu, Sichuan, China, in 2009.

From 2016 to 2017, he was supported by the Western China Talented Personnel Promotion Project of the China Scholarship Council to work as a Visiting Scholar with the University of Florida, Gainesville, FL, USA. He is currently a Teacher with the School of Electronic Engineering, Chengdu University of Information Technology. He has authored one book, over 20 articles, and three patents. His research interests include meteorological satellite remote sensing signal processing, radar signal processing, image processing, artificial intelligence, and software design.

Dr. Liu was a recipient of the Science Foundation of Education Department of Sichuan Province (award number 18ZA0111), in 2018.

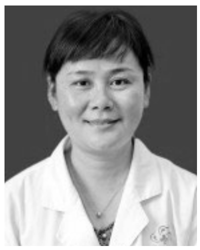


ZECHEN LI was born in Mianyang, Sichuan, China, in 1993. He received the B.E. degree from the Chengdu University of Information Technology, in 2015, where he is currently pursuing the master's degree.

His main research directions are machine learning and biomedical engineering. He won the Second-Class Scholarship at the school level, in 2015. His current research directions are traditional machine learning, biomedical engineering, and machine vision.

YUQI TANG born in Jiangyou, Sichuan, China, in January 1993. He received the bachelor's degree from the Chengdu University of Traditional Chinese Medicine, in 2016, where he is currently pursuing the master's degree in integrated traditional chinese and western medicine.

He got China's Licensed Doctor Qualification, in 2018.



DONGDONG YANG was born in Sichuan, China, in 1968. She received the Ph.D. degree from the Chengdu University of Traditional Chinese Medicine, in 2006.

She is currently the Chief Physician and Doctoral Supervisor of Hospital of the Chengdu University of Traditional Chinese (also known as Sichuan Hospital of Traditional Chinese Medicine). At present, the main research fields are epilepsy, depression, and facial neuritis.

SHUOGUO JIN was born in Henan, China, in 1981. She received the Ph.D. degree from the Chengdu University of Traditional Chinese Medicine.

She is currently the Attending Physician and Lecturer of the Hospital of Chengdu University of Traditional Chinese. She currently focuses on myasthenia gravis and peripheral neuropathy.

JUNWEN GUAN was born in Guangdong, China, in 1969. He received the Ph.D. degree from Sichuan University, in 2009.

He completed his postdoctoral program at West China Hospital, Sichuan University, in 2012, where he is currently the Chief Physician and a Master Tutor in neurosurgery. He concurrently serves as a Master Tutor of the Chengdu University of Information Technology.