A Convenient Non-harm Cervical Spondylosis Intelligent Identity method based on Machine Learning

Nana Wang[1,2,4], Yunyou Huang[1,2,4], Li Cui[1,5], Fanda Fan[1,2], Xi Huang[1], Yi Rao[3], Jing Xiao[3], Jiahui Lu[1,2]

[1]*Institute of Computing Technology(ICT), Chinese Academy of Sciences(CAS), Beijing, China.*
[2]*University of Chinese Academy of Sciences, Beijing, China.*
[3]*Xiyuan Hospital, China Academy of Chinese Medical Sciences(CACMS), Beijing, China.*
[4]*These authors contributed equally to this work.*
[5]*e-mail:lcui@ict.ac.cn.*

Cervical spondylosis(CS), one of the most common orthopedic diseases, poses a serious financial burden on individuals and society. The early detection is an effective way to reduce the burden, since the earlier the CS is detected, the easier it is to treat, the higher the cure rate is, and the less the patient spend. However, the CS identification is mainly determined by the doctor's judgment based on the clinical symptoms and the morphological or structural changes of spine depending on expensive medical instruments in hospital, which may have potential harm to the human body, for example radiation. Owing to the development of the surface electromyography(sEMG, in short) technique and artificial intelligence, a convenient non-harm intelligent CS identification method EasiCNCSII is proposed in this paper, which consist of the sEMG data acquisition and CS identification. For the sEMG acquisition, a convenient method based on the relationship between muscle activity and CS is proposed. It spend less than 20 minutes with the simple instructor and can be easily performed in the non-hospital environment. For the CS identification, the EasiAI model based on machine learning is developed to intelligently identify CS and achieve a performance with 91.02% in accuracy, 97.14% in sensitivity, and 81.43% in specificity.

Cervical spondylosis(CS) is a degenerative disorder common which affect up to two-thirds of the population in their lifetime[1–4]. The CS seriously affect people's physical and

mental health and quality of life and increase the burden on individuals and society. What's more, it leads to the loss of human-related functions and is accompanied by depression, anxiety and other psychological damage. The main complaint of CS is neck pain which is reported by approximate 30-50% from the patients[5]. Meanwhile, the global point prevalence of neck pain was 4.9% and the neck pain ranked 4th highest in terms of disability as measured by YLDs, and 21st in terms of overall burden in the Global Burden of Disease 2010 Study[6]. It is critical for burden lighten to early identify CS. As the earlier the disease is discovered, the easier it is to treat, the higher the cure rate is, and the less the patient spend. CS is a chronic 'wear and tear' degenerative process of the cervical spine that initially is the vertebral bodies and intervertebral disks degeneration in the neck, and can develop into disk ruptures and herniation, osteophyte, compression of the spinal cord, or cervical spondylotic myelopathy(the most important complication of degenerative disease of the cervical spine)[7–9]. As cervical degeneration worsens, clinical manifestations become more obvious, and the difficulty and cost of treatment increases. In detail, for the patients with the vertebral bodies and intervertebral disks degeneration in the neck, non-operative treatment continues to play an important role in treatment[2]. For the patients with cervical spondylotic myelopathy, surgical treatment have been conventional means but may lead to significant problems including adjacent level[2,10].

Due to the complex of the pathogenesis and clinical symptoms of CS, the identification of CS is a sophisticated and complicated work. A large number of auxiliary inspection methods are used to assist the identification. The following methods[11–16] included in the domestic and foreign objective examination of CS are clinical examinations, spinal angiography, vertebral artery angiography, X-ray, computed tomography(CT), and magnetic resonance imaging(MRI), etc. Most of them depend on expensive medical instruments in hospital to directly observe the physical changes in the spine and ancillary structures. However, the physical changes above can cause chronic dysfunction as well as pain[17], leading to anomalous pattern of muscle activity[18–24]. And when the muscles are activated in the activity Motor Unit Action Potential Trains(MUAPTs) are generated by motor units, superimposed on the surface of the skin and form a non-stationary week signal which can be acquired by sEMG device and generate electromyography. The relevant works[18–24] demonstrated that there are differences in sEMG signals between population with cervical musculoskeletal disorders or neck pain and the healthy. Thus, it provides an chance that we can explore the relationship

between muscle activity and CS with a a non-invasive means that analyzing the sEMG signals between CS sufferer and healthy. Benefit from the development of sensors technology, sEMG device become more portal and more cheaper, promoting the sEMG technology to become a competitive choice for the convenient CS identification. The sEMG have attracted a lot of attention in muscle function assessment[25,26], muscle activity assessment[20,21,23], rehabilitation effect tracking[27] and rehabilitation guidance[28]. Meanwhile, the development of machine learning has made great progress in medical[29–32]. With the convenient of sEMG technology and the development of artificial intelligence, the convenient, no-harm, intelligent CS identification method can be considered.

The data acquired by portal sEMG device present a huge challenge to the identification of CS. The high-dimensional sEMG data can cause dimensional disaster which decrease computational efficiency, increase memory storage requirements, and cause overfit. Faced with the high-dimensional data, feature extraction and feature selection, which are effective means of data preprocessing, have the advantages of improving model performance, increasing computational efficiency, decreasing memory storage requirements, giving model better readability and interpretability, and building better generalization model[33]. Besides, the data shows weak availability of faulty, redundant, insufficient, sparse distribution since the data acquisition is susceptible to external factors in non-lab environments by portal sEMG device. So a powerful machine learning model should be considered. Ensemble learning containing a number of weak learners can not only learn linear and complex nonlinear function but also boost weak learners which are slightly better than random guess to strong learners which can make very accurate predictions[34]. The gradient boosting regression tree (GBRT, in short), one of powerful ensemble learning, has been successfully used in classification task[35–38]. And it is a competitive choice for the classification task on limited weakly available data.

In this work, we propose a new convenient, non-harm and intelligent method EasiC-NCSII to identify CS based on sEMG and machine learning as the Figure 1 shown. The method mainly consist of data collection and CS identification. For data acquisition, we proposed the convenient, time-saving data acquisition method. The subject only need to independently complete a set of simple movements according to instruction(see Supplementary for the instruction of data acquisition), spending less than 20 minutes, after connecting

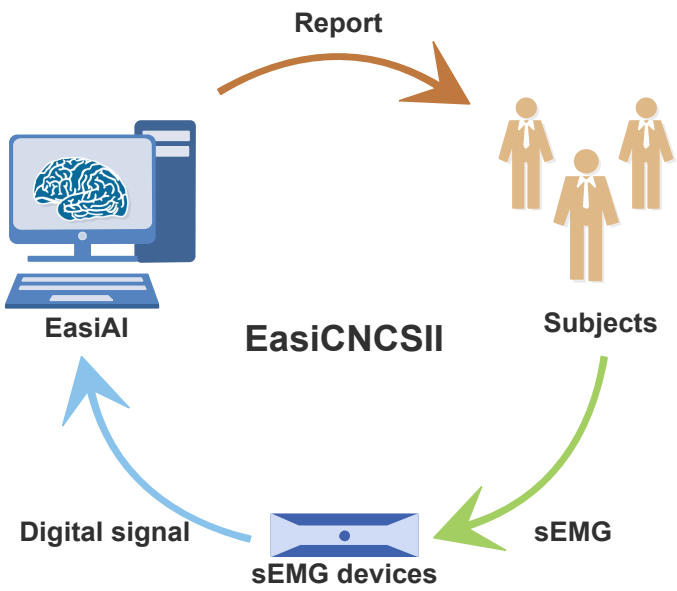


Figure 1: The CS identify based on sEMG and machine learning. Firstly, subjects perform a set of movements according to instructions. Meanwhile, the portable sEMG device acquire sEMG signals of the subject and send it to the smart terminal. Then, the EasiAI, an intelligent CS identification model, predict the state of subject according to it's sEMG signals and return the report which shown that whether the subject suffering from CS.

the portal sEMG device to the laptop and subject's neck muscles. The relevant data is uploaded to the intelligent processing terminal while being collected by the sEMG device. For CS identification, the EasiAI model based on 3-tier architecture, was developed to identify CS. The EasiAI consist of feature extraction, feature selection and intelligent CS identify algorithm. For feature extraction, we utilize parameter model, nonlinear feature, feature combination methods to extract 11 types of features from sEMG data as well as common time-domain, frequency-domain. Besides, the 5 types of enlightening combination features are builded. Most of the features are proved to be significantly associated with the CS by Pearson($p \leq 0.05$). For feature selection, EasiRF, a embedded feature selection method, was developed to select the most relevant features and improve the performance of the CS identification. The easyRF is validated effective compared with traditional feature selection algorithms. For the intelligent CS identify algorithm, the GBRT is developed to the CS identification, and to the best of our knowledge the GBRT is firstly used to identify CS. The EasiAI achieve a opt-perfect performance with 91.02% in accuracy, 97.14% in sensitivity, and 81.43% in specificity compared with 4 kinds of machine learning classification model.

Result

Data acquisition A total of 179 subjects are recruited, of which 109 are patient and 70 are heathy. With a set of movements repetitively completed three times by each subject, the 537 samples are obtained, of which 327 are patient and 210 are healthy.

The data acquisition method involve 6 muscles and 7 movements. The muscles above consist of the the left sternocleidomastoid(M_1), left upper trapezius(M_2), left cervical erector spinae (M_3), right cervical erector spinae(M_4), the right upper trapezius(M_5), the right sternocleidomastoid(M_6) (see Supplementary for the selection of muscles and movements and Figure S_1 for the location of surface electrodes). The movements above include bow(A_1), head backwards(A_2), left flexion(A_3), right flexion(A_4), left rotation(A_5), right rotation(A_6), hands up(A_7)(see Supplementary for the selection of muscles and movements and Figure S_2 for the instruction of the movements). Each movement is held for 20 seconds and repeated 3 times with 5 seconds rest between each repetition.

After connecting the sEMG device to the subject's neck muscles above as well as the

laptop, the data acquisition begins. The subject performed the 7 movements according to instruction(see Supplementary for the instruction of data acquisition). After data acquisition, we obtain 7 multiple high-dimensional time series data from a sample, each of which is generated from the movement A_i and represented as S_i. The S_i is expressed as Equation 1. The $S_{i,j}$ denotes the sEMG signals collected from the muscles M_j activated by movement A_i, expressed as Equation 2.

$$S_i = \left[\begin{array}{cccc} S_{i,1}, & S_{i,2}, & ... & S_{i,j} \end{array} \right]^{\mathrm{T}}, (1 \leq i \leq 7, 1 \leq j \leq 6) \tag{1}$$

$$S_{i,j} = \left[\begin{array}{cccc} p_1, & p_2, & ... & p_n \end{array} \right] \tag{2}$$

Here, $S_{i,j}$ converted from the electrical signal are high dimensional time series data. The p_n is a value of $S_{i,j}$.

Feature extraction The 5 methods: time-domain, frequency-domain, time-frequency, parametric model and nonlinear feature analysis are used to extracted features from the sEMG signal which are high-dimensional time series data(details on the calculation are shown in Supplementary Table S_2). With the methods above, we extracted 63 features from $S_{i,j}$, of which 11 are extracted in methods of time-domain, 14 are computed in methods of frequency-domain, 23 are computed in methods of time-frequency, 14 are extracted in parametric model AR and 1 nonlinear feature are extracted. Among 63 features above, the 5 features including the root mean square (RMS), median frequency(MF), mean power frequency (MPF), the average electromyogram (AEMG), and the integrated electromyogram (IEMG) are common features in clinical research. Besides, we build 45 features from S_i, of which 2 called as cervical flexion-relaxation ratio(FRR)[39] is commonly used in clinical research and can only be extracted from the data of S_1, and the others are new features inspired by relevant research and pathology(see Supplementary for Feature extraction).

The $423(45+6\times63)$ features are extracted from the A_1 and $421(43+6\times63)$ features are respectively extracted from the other 6 movements. Thus, $2949(423 + 6 \times 421)$ features are

extracted from the raw sEMG signal generated by all the movements above[1]. The Pearson correlation analysis indicated that 1789 features are significantly associated with the CS ($p \leq 0.05$, the feature number distribution is shown in Supplementary Figure S_4]). The 63 features from $S_{i,j}$ and 45 features from S_i are divided into 11 types: TF, FF, WL, WLP, AR, EY, FRR, DU, ACI, UN, SYM (details are shown in Supplementary Table S_1-S_2), facilitating statistical analysis.

Data preparation The data set consist of 537 samples and the number of features in each sample is 2949. We split the data into 346 training samples(training set), 86 validation samples(validation set), and 108 test samples(test set). Meanwhile, it is ensured that samples belonging to the same subject are only distributed in one of the three sets above.

Feature Selection We developed a feature selection algorithm EasiRF based on random forest(RF) to select the most relevant features and improve the performance of the CS identification. We use the 537 samples to generate 7 different data sets, each of which is represented as $D_i (0 \leq i \leq 7)$. The RF(with different number of trees) is iteratively used to select the top 25 most important features from D_i. And put the 25 selected features of each iteration into the feature set S_i until the feature number of the S_i is not growing. As shown in Figure 2, The selected feature number tends to be stable on each data set D_i when the iteration number reaches 40. The final feature set including 282 features are generated after merging the all the S_i above.

In order to validate the effectiveness of EasiRF, the Fisher Score(FS)[40], Conditional Infomax Feature Extraction(CIFE)[41], Multi-Cluster Feature Selection(MCFS)[42], f-score are respectively selected from 4 kinds of traditional feature selection algorithms[33] as well as EasiRF. With the metrics of accuracy, sensitivity, specificity, FNR and FPR (see Supplementary for The metric of model performance), we compare the performance of the GBRT on test set with the feature selection algorithms above, using five-folds cross-validation. The large the accuracy, sensitivity and specificity are, the smaller the FNR and FPR are, the better performance of GBRT is, the more efficient the feature selection algorithm is. As shown in table 1, the mean accuracy of the model is 86.54% without feature selection. In spite of the slight drop in mean accuracy with the MCFS and CIFE, the mean accuracy of

[1]The movements include A_1, A_2, A_3, A_4, A_5, A_6, A_7

other algorithms is over 87.10%. What's more, the mean accuracy with the EasiRF is the highest with the value of 91.02%. The mean sensitivity of the model is 92.51% without feature selection. All the mean sensitivity of the model are over 92.51% with feature selection algorithms. What's more the mean sensitivity with the EasiRF is the highest with value of 97.14%. The mean specificity is 77.14% without feature selection. And the mean specificities with feature selection algorithms are less than 77.14%. However the mean specificity with the EasiRF is 81.43% and over 77.14%. The mean FNR of the model is 7.49% without feature selection. All the mean FNR with the feature selection algorithms are less than 7.49%. Besides, the mean FNR with the EasiRF is the smallest with the value of 2.86%. The mean FPR of the model is 22.86% without feature selection. And the mean FPR with feature selection algorithms are larger than 22.86%. However the FPR with EasiRF are the smallest with 18.57%. Compared with the feature selection algorithms above, the EasiRF perform best and are validated most effective. The comparison of more feature selection algorithms are shown in Supplementary Tables S_3.

the train and test of the EasiAI We train the EasiAI on the 346 training samples and validate on 86 validation samples(validation set). We use AUC(the area under the sensitivity and specificity curve) to illustrate the performance of the classifier with larger AUC values indicating higher classification accuracy, sensitivity and specificity across a range of threshold choices[29,43]. The performance of the classifier is mainly affected by model parameters, especially the number of weak classifiers. Thus we first assessed that the number of trees(weak classifiers) included in the EasiAI were enough to obtain the highest AUC on validation set. As shown in Supplementary Figure S_5, the higher the number of trees is, the higher the AUC on the validation set is. Furthermore, the AUC on the validation set achieved highest when the tree number is 535. The other parameters of the EasiAI model are shown in Supplementary Table S_4. The final model is generated by training EasiAI with the parameters in Supplementary Table S_4 on the 432 training samples consisting of the 346 training samples and 86 validation samples.

To validate the performance of the final EasiAI, we tested the model on test set, using five-fold cross-validation. The metrics are accuracy, sensitivity, specificity, FNR, FPR as well as AUC. The large accuracy, sensitivity, specificity and AUC are, the smaller FNR and FPR are, the better the model classification performance is. The AUC, the mean accuracy,

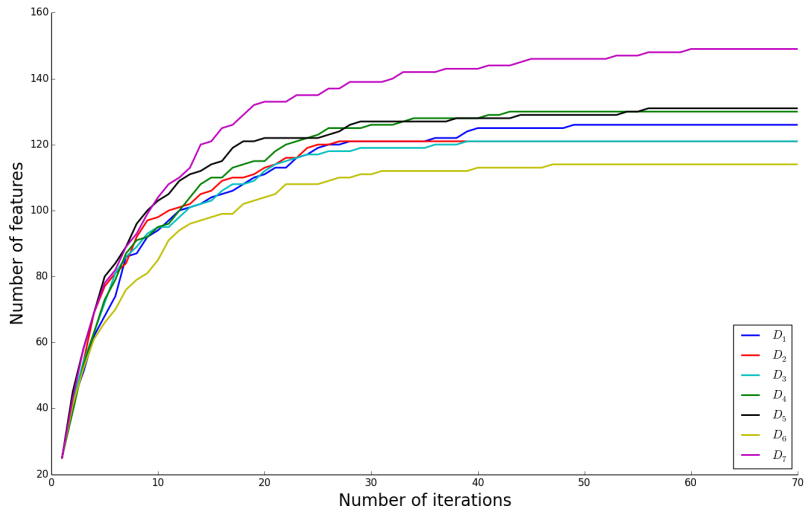


Figure 2: The feature selection. Different colors indicate different data sets. The x-axis represents the number of iterations. The y-axis represents the number of selected features. When the number of iterations is less than 20, the number of features increases exponentially. When the number of iterations is greater than 20, the number of features increases slowly and tends to be stable.

Table 1: The comparison of performance with different feature selection algorithms

	Accuracy	Sensitivity	Specificity	FNR	FPR
Non	86.54%	92.51%	77.14%	7.49%	22.86%
FS	87.10%	95.28%	74.29%	4.72%	25.71%
CIFE	86.00%	92.60%	75.71%	7.40%	24.29%
MCFS4	83.16%	92.60%	68.57%	7.45%	31.43%
F-score	87.10%	95.28%	74.29%	4.72%	25.71%
EasiRF	91.02%	97.14%	81.43%	2.86%	18.57%

the mean sensitivity, the mean specificity respectively are 0.95, 91.02%, 97.14% and 81.43%. The mean FNR, the rate of missed diagnosis, is 2.86%. The mean FPR, the misdiagnosis rates, is 18.57%. Overall, the accuracy of model is higher than 90% and the missed diagnosis rate of our model is less than 3%.

Comparing with support vector machines(SVM), Logistic regression(LR), NativeBayes(NB), random forests(RF) The identification of CS is a task of classification. Faced with limited data, we chose the following four models that are often used, which include support vector machines(SVM), Logistic regression(LR), Native Bayes(NB), and random forests(RF). We validate the effectiveness of the EasiAI by comparing the EasiAI with SVM, LR, NB and RF, on test set in the same classification task, using five-fold cross-validation. As shown in table 3, the highest mean accuracy is 91.02% achieved by the EasiAI, and the lowest mean accuracy is 82.10%. The highest mean sensitivity is 97.14% achieved by the EasiAI, and the lowest mean sensitivity is 88.92%. The highest mean specificity is 82.86% and the lowest mean specificity is 70%. The mean specificity of the EasiAI is 81.43%, just smaller than the highest. The highest mean FNR(false negative rate) is 11.08% and the smallest value is 2.86% achieved by the EasiAI. The highest mean FPR(false positive rate) is 30% and the lowest value is 17.14%. The mean FPR of the EasiAI is 18.57%, just larger than the smallest one. Besides, the curves of sensitivities and specificities of models above are shown in Figure 3. Most of the sensitivity and specificity point falls below the red curve of EasiAI, especially between point a and b. In short the EasiAI achieve a opt-perfect performance compared with RF, SVM and LR, NB with the metrics above.

Analysis of learned knowledge about CS The EasiAI achieved high accuracy in identifying CS, so we believe that the features in final model play an important role in the classification and there are significant differences in features between CS suffers and the healthy. The significant differences in feature indicated the features, which are extracted from the raw signal generated from the muscle contraction induced by movements, have unseen relation with the CS, so we believe that the informative knowledge about the CS can be learned from the relationship between features and muscles activated by movements. In

Table 2: The performance of machine learning on test train

	Accuracy	Sensitivity	Specificity	FNR	FPR
EasiAI	91.02%	97.14%	81.43%	2.86%	18.57%
RF	87.65%	93.42%	78.57%	6.60%	21.43%
SVM	86.00%	89.78%	80.00%	10.22%	20.00%
LR	86.57%	88.92%	82.86%	11.08%	17.14%
NB	82.10%	89.83%	70.00%	10.17%	30.00%

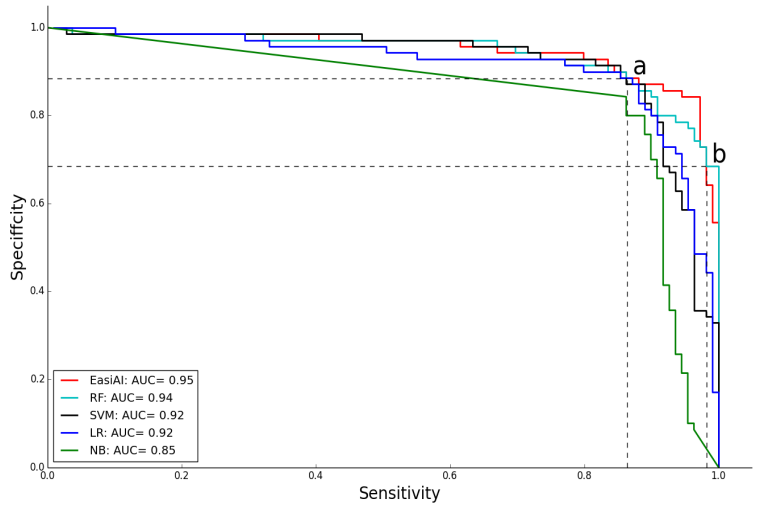


Figure 3: The CS identify performance of the EasiAI, RF, SVM, LR, NB. The x-axis is the sensitivity. The y-axis is the specificity. Different color denotes different machine learning models. The curves of different colors consist of sensitivity and specificity points of different models. The closer the sensitivity and specificity points are to the upper left corner, The large the sensitivity and specificity are, the better the model performs. All the sensitivity and specificity points of the NB are the farthest from the upper left corner and fall below the curves of other models. The two curves of LR and NB are intertwined between curve of SVM and curve of RF whose most of sensitivity and specificity points fall below the curve of the RF. The two curves of RF and EasiAI are intertwined and most of the RF's sensitivity and specificity points fall below the ones of EasiAI, especially between point a and b.

order to further analyze the relationship above, the number[2] and importance[3] of the features on the muscles being activated by movements are plotted as a unit of the heat map as shown in Figure 4. The features on A_1(bow head) are the most in number in Figure 4 (a). The features on A_1(bow head) are most important in Figure 4 (b), followed by the ones on A_7(hands up). The features on M_3(left CE) involved in A_1, M_4(right CE) involved in A_1, M_2(left UT) involved in A_7 and M_5(right UT) involved in A_7 are the most in number and most important in Figure 4. Thus, it is inferred that the CE activated by the movement of bowing and the UT activated by the movement of hands up may be more closely related to CS and there are significant difference in CE and UT between CS suffers and the healthy. Besides, we found that the number and importance of features are asymmetric distribution between the left and right muscles(movements) compared the A_3, A_4 and the A_5, A_6 as well as the M_2, M_3 and the M_4, M_5. It indicated that the asymmetry of muscle distribution may be related to the asymmetry of movements distribution.

In order to further understand difference in muscles as well as asymmetry, the feature number distribution on muscles, which the feature number distribution on muscles are called as A-distribution, and the importance distribution on muscles, which the feature number distribution on muscles are called as B-distribution, are plotted as Figure 5. The M_4(left CE), M_3(right CE) and M_5(right UT) rank the top 3 in the two different distributions. The features on UT (M_2 and M_5) and CE (M_3 and M_4) take up 85.47% in the A-distribution and 74.60% in the B-distribution. So the CE and UT play an important role in CS identify and is more closely related to CS, especially CE. The features on right muscles are more than ones on left but slightly less important than the ones left[4]. It is inferred that all the muscles play an important role in identifying CS and the left muscles are slightly more important

[2]Feature number shows the number of features generated from muscles or movements. The more the features that are distributed on the muscle(movement) are, the stronger muscle(movement) have the ability to identify CS, the more the muscle(movement) have differences between the healthy and the CS suffer, the more the muscle(movement) is related to CS.

[3]The feature importance shows the feature contribution to the task above during the training model. The more important the features that are distributed on the muscle(movement) are, the muscle(movement) have the ability to identify CS, the more the muscle(movement) have differences between the healthy and the CS suffer, the more the muscle(movement) is related to CS.

[4]The features on right muscle(M_4, M_5, M_6) take up 56.06% in the A-distribution and take up 49.13% in the B-distribution.

than the right ones. Besides, the asymmetry of the feature distribution in the left and right muscles may be caused by that the population are accustomed to using the unilateral muscles. Humans more frequently use their right hands at work and in life, for example mouse with your right hand, chopsticks with your right hand. The right hand activities will lead to the right arm and right scapula movement, activating the right-side muscles more frequently than left. The pattern is found in A-distribution and B-distribution that the CE rank first, followed by UT and SCM in two different distributions. The higher the ranking of the muscle is, the more significant the muscle have difference between the CS suffers and the healthy is, the more the muscle have relationship with CS. Thus, it is inferred that the pattern of feature distribution may be related to the amount of muscles activities and the muscle with more activity rank higher and are more closely related to CS. The activity active muscles, and moderate muscle activation is conducive to muscle health. However, the muscles excessive use of muscle can cause muscle strain which affect the muscle contraction as well as sEMG signals. The signals above have a more difference from the healthy, making the muscles rank higher. The most activities appeared in population above are bow and the hand and arms movements, the posture of bow for long time lead to that CE are stretched for a long time, and the hand and arms movements for long time will constantly activate the UT. The overuse of muscles will lead to muscle strain, accelerating the occurrence and development of CS.

In order to further analyze the difference in movements as well as asymmetry, the feature number distribution on movements, which the feature number distribution on muscles are called as C-distribution, and the features importance distribution on movements, which the feature number distribution on muscles are called as D-distribution, are plotted as Figure 6. The A_1(bow) rank first in two distributions above, followed by A_7(hands up) in D-distribution, which is consist with the conclusion above that the features on the CE involved in A_1 and UT involved in A_7 are most in two distributions in Figure 4-5. It has been strengthened that the muscles with more activity ranked higher. There is an asymmetry distribution compared left-side movements(A_3 and A_5)[5] and right-side movements(A_4 and A_6)[6] in the two distributions. The features involving right-side movements are more

[5]The features on right flexion and right rotation take up 25.89% in the C-distribution and 22.39% in the D-distribution.

[6]The features on left flexion and left rotation take up 29.79% in the C-distribution and 27.78% in the

than ones in the left-side. The asymmetric distribution of features in movement indicated that unbalanced use of muscles cause the asymmetry of the feature distribution in the left and right movements. It can be explained that the more muscle activity, the closer the relationship with CS.

In short, it is demonstrated that the the sEMG generated on CE and UT have significant differences between CS suffer and the healthy which is consist with the previous research[18–24]. It is uncertain whether it is caused by bone problem or muscle problem. However, the conclusion focus on the CE involved in A_1 and UT involved in A_7. so we inferred that the overuse of CE and UT for long time is an cause of CS. The muscles and bones work together to maintain the balance of power in exercise. The excessive use of CE and UT lead to the strain of the CE and UT, followed by dynamic imbalance of muscle strength during exercise. The imbalance above increase the burden on spine, aggravating spinal degeneration and the occurrence of CE. In detail, the inference is that improper posture cause muscle strain, accelerate cervical degeneration and lead to CS which is consist with the dysfunction of the spine stabilizing system[17]. It may provides a suggestion for sedentary population to prevent CS by avoiding excessive use of the CE and UT, strengthen the UT and CE protection.

Discussion The CS identification methods mainly include clinical symptoms examination and imaging examination since the diagnosis of CS is determined by clinical symptoms and imaging information[16,44]. Currently, the clinical symptoms examination are performed by experts or doctors in the form of the inquiry. With the standardized development of the inquiry, scale is introduced to assist the examination of symptom information. The Tanaka Yasushi cervical spondylosis symptom Scale are used to assess the CS including the ability of working and living, physical signs, hand functions and pain, etc. The higher the score is, the more likely the user is suffering from CS. The images examination[7] mainly depend on observation of the physical changes in spine and its subsidiary structure by imaging instruments to identify CS. The EasiCNCSII depend on detecting abnormal sEMG signal associated with muscle activity to identify CS. Here, we explore the practicality of the EasiCNCSII in different aspects by comparing sEMG with the inquiry, imaging method as

D-distribution.

[7]The imaging methods include spinal angiography, vertebral artery angiography, X-ray, computed tomography(CT), and magnetic resonance imaging(MRI), etc.

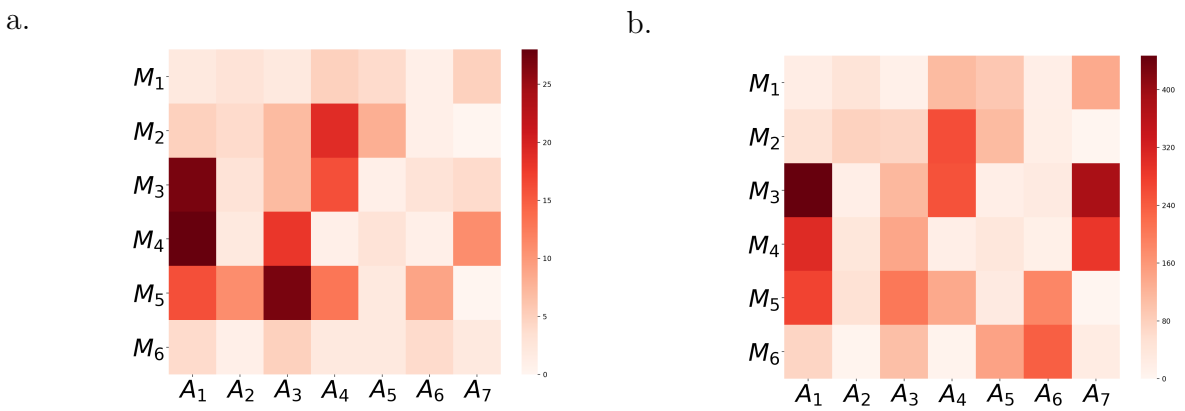


Figure 4: The heat map. The M_1, M_2, M_3, M_4, M_5, M_6 respectively denotes left SCM, left UT, left CE, right CE, right UT and right SCM. The A_1, A_2, A_3, A_4, A_5, A_6 and A_7 respectively denotes low head, head backwards, left flexion, right flexion, left rotation and right rotation. (a), The number distribution of features in muscles and movements. A square represent the number of feature on M_i activated by A_i. The darker the square's color is, the more features the square have. (b), The importance distribution of features in muscles and movements. A square represent the importance of feature on M_i activated by A_i. The darker the square's color is, the more important features square have.

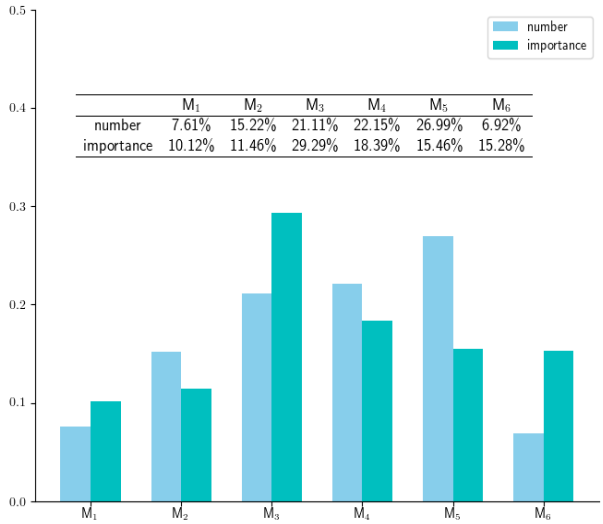
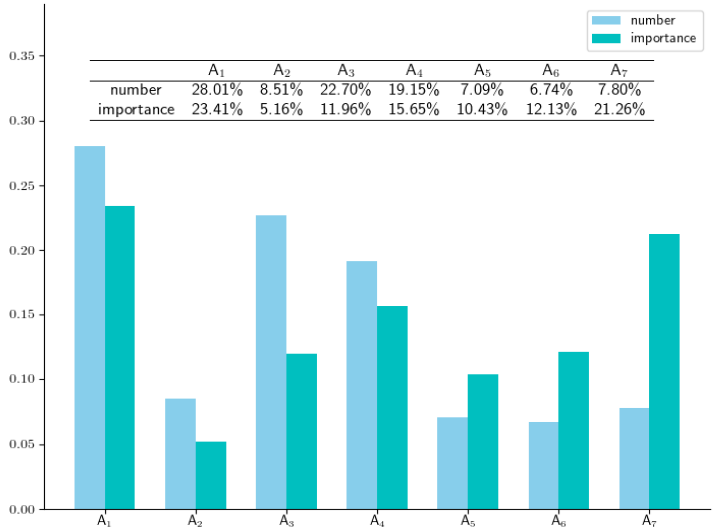


Figure 5: The feature distribution on muscle. The M_1, M_2, M_3, M_4, M_5, M_6 respectively denotes left SCM, left UT, left CE, right CE, right UT and right SCM. (a), The light bars show the distribution of feature numbers. The right UT rank first, followed by right CE, left CE, left UT, left SCM, and right SCM. The right UT, right CE and left CE take up 70.25%. (b), The dark bars shows the distribution of feature importance. The left CE rank first, followed by right CE, right UT, right SCM, left UT and left SCM. The left CE, right CE and right UT take up 66.90%.



	A_1	A_2	A_3	A_4	A_5	A_6	A_7
number	28.01%	8.51%	22.70%	19.15%	7.09%	6.74%	7.80%
importance	23.41%	5.16%	11.96%	15.65%	10.43%	12.13%	21.26%

Figure 6: The feature distribution on movement. The A_1, A_2, A_3, A_4, A_5, A_6 and A_7 respectively denotes low head, head backwards, left flexion, right flexion, left rotation and right rotation. (a), The light bars denotes the distribution of numbers. The movement of head backwards rank first, followed by left flexion, right flexion, low head, hand up, left rotation, right rotation. (b), The dark bars shows the distribution of feature importance. The movement of head backwards rank first, followed by hands up, right rotation, left flexion, left rotation, low head.

shown in table 3. The inquiry method is the easiest, fastest and repeatable. However, the method is suitable for population with obvious symptoms, for example severe pain, since information is mainly determined by suffer's subjective feelings and judgment. And it needs the help of the doctor and auxiliary equipment to accurately identify CS. The imaging method is the essential CS examination currently. However, its cost is relatively high and it is time-consuming considering time spent to go to the hospital and wait for the results which requires the intervention of doctors or experts. What's more, it can not be used frequently with concern on health since the frequent use of imaging instruments can put a strain on the health, for example the radiation. Compared with inquiry and imagine method, the sEMG examination used in EasiCNCSII is an best choice to identify CS outside the clinic with the advantage of easy use, low cost, no-harm. Due to the intelligent algorithm, users can get results quickly after examination. What's more, combined with mobile application, the EasiCNCSII potentially provide low-cost universal access to indispensable identify care outside the clinic, and even promote the development of telemedicine, especially in areas short of medical resources.

To our best knowledge, previous research on CS based on the sEMG and machine learning are relatively few, so there is a lot room for improvement. Traditional classification or regression algorithms can achieve good performance when fed with a wealthy of high quality data. However, the amount of data is limited and the data dimension is high. Besides, data acquisition is vulnerable to the environment so that there are some poor quality data collected by portal sEMG. The EasiAI can handle these influences by boosting multiple weak learners, hence enabling EasiAI to reach higher prediction accuracies. Compared with the RF, SVM, LR and NB, the EasiAI can achieve better performance with the metric of not only accuracy but sensitivity, FNR and AUC as well. The higher accuracy and smaller FNR allows us to identify complex crowd with low missed diagnosis rate. Because it is hardly to fully understand the relation between CS and the activity of the deep and shallow muscles with the limited data, the data-driven machining learning can achieve more perfect performance with more accumulated data, and can accelerate our understanding of the principles behind CS identify to guide treatment. We are looking forward to more convenient and intelligent applications of machine learning based on sensors technology in CS studies.

Methods

Cohort description The study population are made up of CS outpatient from Beijing Xiyuan Hospital Physiotherapy Department and the healthy from Institute of Computing Technology(ICT) and the hospital above which were recruited as volunteer by email and advertising. A total of 179 objects participated in the experiment, of which 109 were patient and 70 were heathy. The former are outpatients that have received a clinical diagnosis of CS, which was in accordance with 2012 ICD-9-CM Diagnosis Code 721(721.0 Cervical spondylosis without myelopathy, 721.1 Cervical spondylosis with myelopathy) and the diagnostic criteria of diagnosis and treatment for CS issued by China Rehabilitation Medicine association, by doctors. The latter are healthy diagnosed by the rich experienced clinicians. The exclusion criteria for both the groups includes cervical vertebral trauma, cervical vertebral surgery, congenital spinal deformity, syringomyelia, amyotrophic lateral sclerosis, spinal cord tumor, spinal cord injury, adhesive arachnoiditis, the fascitis, cervical injury, tumor or infection, pregnant, breast-feeding, menstruation. All the subjects were given a clear explanation of tests approved the research ethics board of Beijing Xiyuan Hospital before taking part in this research.

EasiCNCSII In this paper, we proposed a method EasiPNACSI to identify CS. As shown figure 7, the method include two parts: data acquisition and the EasiAI model. For data acquisition, subjects connect sEMG device to the SCM, UT and CE of neck muscles and complete 7 neck collection movements. The data are sent to the EasiAI. The EasiAI is a three layer data processing model: feature extraction, feature selection algorithm and machine learning classification algorithm. Feature extraction algorithm includes 9 kinds of methods: time domain, frequency domain, time frequency domain, AR, wavelet, wavelet packet, entropy, feature combination. Feature selection algorithm is developed to selected the most relevant features by iterative RF. The gradient tree boosting is developed to identify CS, achieving the good performance on limited data set with relative small computing overhead. The lightweight algorithms can be integrated into the user end, quickly returning report to subjects without concerns on privacy.

The EasiRF algorithm The EasiRF based on the RF is a stochastic model since the samples are randomly selected to generate the tree and the features are randomly selected to be splitting rule. Firstly, we generate 7 different data sets and iteratively use random forest

Table 3: The comparison of the Inquiry, Images, sEMG

	Iquiry	Images	sEMG
Time	dozens of minutes	dozens of minutes	about 20 minutes
Cost	low	high	middle
Harmful	no	yes	no
Location	no limite	hospital	no limite
Convenience	high	low	middle
Frequent	yes	no	yes
Doctor/professional	yes	yes	no
The time of getting the results	a few minutes	days	several seconds

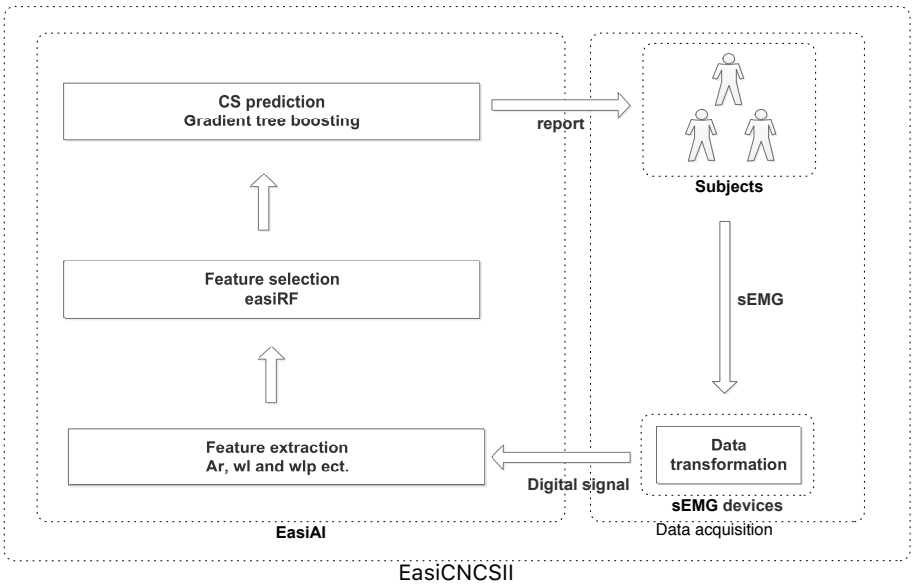


Figure 7: The EasiCNCSII method. The left side of figure is the CS identify model EasiAI consisting of feature extraction, feature selection, classification model. The right side of figure is data acquisition. The data collected by sEMG is automatically transmitted to the intelligent terminal equipped with EasiAI. The report generated by the EasiAI is sent to the subjects.

to train on each data set. Secondly, set the tree number of RF model different, select the top 25 most important features and merge the features of each iteration until the feature number of the merged set is not growing in each iteration. Finally, the final feature set is generated by merging 7 feature sets from 7 data sets above. As shown in algorithm 1, the feature selection algorithm EasiRF based on RF is developed to get the relative features.

Algorithm 1 The EasiRF algorithm

Input: the data set S

Output: the set of features $FeaSet$

```
k=7, FeaSet = ∅
S = {S_1, S_2, S_3, ..., S_7}
for j = 1 to 7 do
    DS = S - S_j
    FeaSet_j = ∅
    Num = 0
    while | FeaSet_j | is not growing do
        FeaSet_t = ∅
        Setting the number of the RF with (50 + Num * 5).
        Run the RF with DS, output the features and its importances.
        Put the 25 the most important features into FeaSet_t
        FeaSet_j = FeaSet_j ∪ FeaSet_t
    end while
    FeaSet = FeaSet ∪ FeaSet_j
end for
return FeaSet
```

Experimental set-up The python (version 2.7.13) and matlab (version 2016r) were used to implement features extraction from sEMG data. The python (version 3.6.0) and xgboot (version 0.6) were used to implement the esaiAI model. The RF, NB, LR and SVM model are implemented by the scikit-learn(0.19.1) and the python (version 3.6.0). The laptop and sEMG device which used to collect the sEMG signal are provided by Efficient Autonomous Sensitive and Intelligent sensor Networks group (this research does not develop the hardware). The EasiAI is deployed in the laptop with the CPU of i7, the memory of 8GB and operating system of 64-bit.

Code availability The code of EasyAI will be available at Github after being arranged. If necessary, contact wangnana@ict.ac.cn for the EasiAI currently.

Data availability The raw sEMG data supporting the findings of this study involves user privacy, we do not have the right to make it public currently. The data set consist of 537 data samples will be public at Github as mentioned in the manuscript after being arranged. Each sample consist of 2949 features which are extracted in the methods mentioned in our paper. If necessary, contact wangnana@ict.ac.cn for the data set currently.

1. Côté, P., Cassidy, J. D. & Carroll, L. The saskatchewan health and back pain survey: the prevalence of neck pain and related disability in saskatchewan adults. *Spine* **23**, 1689–1698 (1998).

2. Todd, A. G. Cervical spine: degenerative conditions. *Current reviews in musculoskeletal medicine* **4**, 168 (2011).

3. Jebri, B. *et al.* Detection of degenerative change in lateral projection cervical spine x-ray images. In *Medical Imaging 2015: Computer-Aided Diagnosis*, vol. 9414, 941404 (International Society for Optics and Photonics, 2015).

4. Singh, S., Kumar, D. & Kumar, S. Risk factors in cervical spondylosis. *Journal of Clinical Orthopaedics and Trauma* **5**, 221–226 (2014).

5. Hoggjohnson, S., Velde, G. V. D., Carroll, L. J., Holm, L. W. & Cassidy, J. D. The burden and determinants of neck pain in the general population: Results of the bone and joint decade 2000–2010 task force on neck pain and its associated disorders. best evidence on the burden and determinants of neck. *European Spine Journal* **17**, 39–51 (2008).

6. Hoy, D. *et al.* The global burden of low back pain: estimates from the global burden of disease 2010 study. *Annals of the rheumatic diseases* **73**, 968–974 (2014).

7. Matz, P. G. *et al.* Joint section on disorders of the spine and peripheral nerves of the american association of neurological surgeons and congress of neurological surgeons. *J Neurosurg Spine* **11**, 157–169 (2009).

8. Kotil, K. & Bilge, T. Prospective study of anterior cervical microforaminotomy for cervical radiculopathy. *Journal of Clinical Neuroscience* **15**, 749–756 (2008).

9. Cai, Z. *et al.* Trend of the incidence of cervical spondylosis: decrease with aging in the elderly and increase with aging in the young and the adults. *Int J Clin Exp Med* **9**, 14329–14336 (2016).

10. Quinn, J. C., Kiely, P. D., Lebl, D. R. & Hughes, A. P. Anterior surgical treatment of cervical spondylotic myelopathy. *HSS Journal®* **11**, 15–25 (2015).

11. Xie, X., Yin, S., Gao, H. & Kaynak, O. Asymptotic stability and stabilisation of uncertain delta operator systems with time-varying delays. *Control Theory and Applications Iet* **7**, 1071–1078 (2013).

12. Shen, Y., Ding, S. X., Haghani, A., Hao, H. & Zhang, P. A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark tennessee eastman process. *Journal of Process Control* **22**, 1567–1581 (2012).

13. Liu, M., Cui, L., Guan, Y., Li, B. & Du, H. Single-fiber electromyography in amyotrophic lateral sclerosis and cervical spondylosis. *Muscle and nerve* **48**, 137–139 (2013).

14. Borden, S. Abnormal magnetic-resonance scans of the cervical spine in asymptomatic subjects. *J Bone Joint Surg.* **72**, 1178–1184 (1990).

15. Lehto, I. *et al.* Age-related MRI changes at 0.1 T in cervical discs in asymptomatic subjects. *Neuroradiology* **36**, 49–53 (1994).

16. Yu, X., Liu, M., Meng, L. & Xiang, L. Classifying cervical spondylosis based on x-ray quantitative diagnosis. *Neurocomputing* **165**, 222–227 (2015).

17. Panjabi, M. M. The stabilizing system of the spine. part i. function, dysfunction, adaptation, and enhancement. *Clinical Spine Surgery* **5**, 383–389 (1992).

18. Airaksinen, M. K. *et al.* Wireless on-line electromyography in recording neck muscle function: a pilot study. *Pathophysiology* **12**, 303–306 (2005).

19. Falla, D. Unravelling the complexity of muscle impairment in chronic neck pain. *Manual therapy* **9**, 125–133 (2004).

20. Falla, D., Bilenkij, G. & Jull, G. Patients with chronic neck pain demonstrate altered patterns of muscle activation during performance of a functional upper limb task. *Spine* **29**, 1436–1440 (2004).

21. Johnston, V., Jull, G., Souvlis, T. & Jimmieson, N. L. Neck movement and muscle activity characteristics in female office workers with neck pain. *Spine* **33**, 555–563 (2008).

22. Madeleine, P., Xie, Y., Szeto, G. P. & Samani, A. Effects of chronic neck–shoulder pain on normalized mutual information analysis of surface electromyography during functional tasks. *Clinical Neurophysiology* **127**, 3110–3117 (2016).

23. Johnston, V., Jull, G., Darnell, R., Jimmieson, N. & Souvlis, T. Alterations in cervical muscle activity in functional and stressful tasks in female office workers with neck pain. *European journal of applied physiology* **103**, 253–264 (2008).

24. Falla, D., Farina, D., Dahl, M. K. & Graven-Nielsen, T. Muscle pain induces task-dependent changes in cervical agonist/antagonist activity. *Journal of Applied Physiology* **102**, 601–609 (2007).

25. Yang, L.-f. & Kang, B. Study on human neck muscles' comfort of different height levels based on semg method. In *Proceedings of the 6th International Asia Conference on Industrial Engineering and Management Innovation*, 563–574 (Springer, 2016).

26. Venugopal, G., Navaneethakrishna, M. & Ramakrishnan, S. Extraction and analysis of multiple time window features associated with muscle fatigue conditions using semg signals. *Expert Systems with Applications* **41**, 2652–2659 (2014).

27. YANG, H.-s. *et al.* Influence of different acupoint combinations on immediate effect of surface electromyography of patients with cervical spondylosis. *World Journal of Acupuncture-Moxibustion* **26**, 7–13 (2016).

28. Lambelet, C., Lyu, M., Woolley, D., Gassert, R. & Wenderoth, N. The ewrist–a wearable wrist exoskeleton with semg-based force control for stroke rehabilitation. In *Rehabilitation Robotics (ICORR), 2017 International Conference on*, 726–733 (IEEE, 2017).

29. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115 (2017).

30. Lezcano-Valverde, J. M. *et al.* Development and validation of a multivariate predictive model for rheumatoid arthritis mortality using a machine learning approach. *Scientific reports* **7**, 10189 (2017).

31. Zhou, X. X. *et al.* pdeep: Predicting MS/MS spectra of peptides with deep learning. *Analytical Chemistry* (2017).

32. Subasi, A., Alickovic, E. & Kevric, J. Diagnosis of chronic kidney disease by using random forest. In *CMBEBIH 2017*, 589–594 (Springer, 2017).

33. Li, J. *et al.* Feature selection: A data perspective. *ACM Computing Surveys (CSUR)* **50**, 94 (2017).

34. Zhou, Z.-H. Ensemble learning. *Encyclopedia of biometrics* 411–416 (2015).

35. Li, P. Robust logitboost and adaptive base class (abc) logitboost. *arXiv preprint arXiv:1203.3491* (2012).

36. Burges, C. J. From ranknet to lambdarank to lambdamart: An overview. *Learning* **11**, 81 (2010).

37. Bennett, J., Lanning, S. *et al.* The netflix prize. In *Proceedings of KDD cup and workshop*, vol. 2007, 35 (New York, NY, USA, 2007).

38. He, X. *et al.* Practical lessons from predicting clicks on ads at facebook. In *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*, 1–9 (ACM, 2014).

39. Shin, S.-j., An, D.-h., Oh, J.-s. & Yoo, W.-g. Changes in pressure pain in the upper trapezius muscle, cervical range of motion, and the cervical flexion–relaxation ratio after overhead work. *Industrial health* **50**, 509–515 (2012).

40. Duda, R. O., Hart, P. E. & Stork, D. G. *Pattern classification* (John Wiley & Sons, 2012).

41. Brown, G., Pocock, A., Zhao, M.-J. & Luján, M. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *Journal of machine learning research* **13**, 27–66 (2012).

42. Cai, D., Zhang, C. & He, X. Unsupervised feature selection for multi-cluster data. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 333–342 (ACM, 2010).

43. Baldi, P., Sadowski, P. & Whiteson, D. Searching for exotic particles in high-energy physics with deep learning. *Nature communications* **5**, 4308 (2014).

44. Binder, A. I. Cervical spondylosis and neck pain. *BMJ: British Medical Journal* **334**, 527 (2007).

Acknowledgements The authors thank all medical workers in Beijing Xiyuan Hospital rehabilitation physiotherapy for providing the outpatients and Research Center for Ubiquitous Computing Systems(in short CUbiCS) for the providing the healthy. Besides, the authors also thank all subjects for providing valuable data.

Author contributions N.-N.W and Y.-Y.H designed the experiment and model, collected the data, wrote the codes, analyzed the data and wrote the manuscript. L.C. directed the project and revised the manuscript. F.D.F wrote the code of mode. X.H guided the data preparation. Y.R. and J.X. provided outpatients and places for data acquisition. J.H.L participated in the data collection.

Competing Interests The authors declare that they have no competing financial interests.

Correspondence Correspondence and requests for materials should be addressed to Institute of Computing Technology, Chinese Academy of Sciences, No. 6 South Road of Academy of Sciences, Haidian District, Beijing 100089, China. (email: lcui@ict.ac.cn).