frontiers | Frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 11 July 2022
doi: 10.3389/fbioe.2022.935481

A New Strategy for Identification of Coal Miners With Abnormal Physical Signs Based on EN-mRMR

Mengran Zhou, Kai Bian, Feng Hu and Wenhao Lai*

School of Electrical and Information Engineering, Anhui University of Science and Technology, Huainan, China

Coal miners' occupational health is a key part of production safety in the coal mine. Accurate identification of abnormal physical signs is the key to preventing occupational diseases and improving miners' working environment. There are many problems when evaluating the physical health status of miners manually, such as too many sign parameters, low diagnostic efficiency, missed diagnosis, and misdiagnosis. To solve these problems, the machine learning algorithm is used to identify miners with abnormal signs. We proposed a feature screening strategy of integrating elastic net (EN) and Max-Relevance and Min-Redundancy (mRMR) to establish the model to identify abnormal signs and obtain the key physical signs. First, the raw 21 physical signs were expanded to 25 by feature construction technology. Then, the EN was used to delete redundant physical signs. Finally, the mRMR combined with the support vector classification of intelligent optimization algorithm by Gravitational Search Algorithm (GSA-SVC) is applied to further simplify the rest of 12 relatively important physical signs and obtain the optimal model with data of six physical signs. At this time, the accuracy, precision, recall, specificity, G-mean, and MCC of the test set were 97.50%, 97.78%, 97.78%, 97.14%, 0.98, and 0.95. The experimental results show that the proposed strategy improves the model performance with the smallest features and realizes the accurate identification of abnormal coal miners. The conclusion could provide reference evidence for intelligent classification and assessment of occupational health in the early stage.

Keywords: coal miners, occupational health, accurate identification, machine learning, feature screening, intelligent optimization

OPEN ACCESS

Edited by:
Bing Wang,
Anhui University of Technology, China

Reviewed by:
Wenzheng Bao,
Xuzhou University of Technology, China
Teng Li,
Anhui University, China

****Correspondence:***
Kai Bian
kbian92@163.com

Specialty section:
This article was submitted to Preclinical Cell and Gene Therapy, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *04 May 2022*
Accepted: *06 June 2022*
Published: *11 July 2022*

Citation:
Zhou M, Bian K, Hu F and Lai W (2022) A New Strategy for Identification of Coal Miners With Abnormal Physical Signs Based on EN-mRMR. Front. Bioeng. Biotechnol. 10:935481. doi: 10.3389/fbioe.2022.935481

INTRODUCTION

As most coal mines are buried underground, underground mining is a very important mining method, and the technical difficulty and risk of underground mining are high (Blondeel and Van de Graaf, 2018; Liu et al., 2019). Because the underground environment and equipment of coal mines are restricted, the health status of underground miners cannot be ignored (Xie et al., 2020). With the continuous increase in coal mining depth, the underground geological conditions are complex, the mining conditions are difficult, the working environment is poor (Paul et al., 2020), and the possibility of miners suffering from occupational diseases has also increased significantly (Lu et al., 2020). Therefore, how to accurately identify coal miners with abnormal physical signs and make early judgments on miners' physical health is an important premise for the prevention and treatment of miners' occupational diseases (Hanoa et al., 2011; Volobaev et al., 2016).

The possible health hazard factors in the working environment of the coal mine mainly include dust (Perret et al., 2017), chemical poisons (Pone et al., 2007), and harmful physical conditions (Takacs et al., 2015), which may affect the health of coal miners. Various physical signs of the human body are interdependent, and the change of each physical sign will not be carried out independently, which is a comprehensive organism (Zachurzok-Buczyńska et al., 2011). When the basic physical signs of the human body are abnormal, the physical status of the human body must be changed. However, if the physical status of the human body is abnormal, one of the physical sign parameters may not change, and these abnormal parameters will be the precursor of occupational disease (Ackermann, 2004; Zhu et al., 2014). Therefore, an accurate assessment of the health status of the human body was made only by comprehensively analyzing a variety of physical signs (Pucciarelli et al., 2019). In the traditional approaches, the evaluation of miners' health status is primarily performed by experienced doctors by integrating the signs information based on the physical examination report (Wu et al., 2019a). This method is not only time-consuming and laborious, especially for doctors with insufficient diagnostic experience, but also with a higher chance of misdiagnosis and missed diagnosis because of subjective (Mackenzie Ross, 2016). The doctors may only consider a single disease that is related to a single factor without paying attention to the correlation between diagnostic results and different signs.

In recent years, artificial intelligence algorithms have been applied to the intelligent aided analysis and evaluation of physical examination data, which provides a lot of theoretical basis for health management and disease prevention (Hoogendoorn et al., 2016; Grzywalski et al., 2019; Koshimizu et al., 2020; Yang et al., 2020). For example, Wu et al. (2019b) put forward the learning vector quantization and Fisher-SVM to assess the risk of hypertension in steel workers and achieved a good evaluation effect. Lee et al. (2020) developed the prediction model based on XGBoost to assess the risk of metabolic syndrome with body weight control. Maxwell et al. (2017) used a Deep Neural Networks (DNN) algorithm based on deep learning for multi-label classification and prediction of chronic diseases. Galarraga et al. (2017) combined principal component analysis (PCA) and multiple linear regressions to process physical examination and surgery data for the prediction of postoperative gait in cerebral palsy. However, at present, all the attributes of physical examination data were analyzed by some of these machine learning algorithms without considering the redundancy of attribute parameters. The parameter adjustment of the predictive model based on the deep learning method is complex, a large number of samples need to be iterated a certain number of times to achieve the targeted accuracy, and the efficiency is low. The feature extraction methods including PCA change the raw data structure, and the dimensionality reduction results in features without physical meaning, which is not interpretable.

Elastic net (EN) is a feature screening method based on regulation (Teisseyre, 2017), which can effectively solve the over-fitting problem, reduce the relevant features, and remove them from the model. For example, Kocsmár et al. (2020) used the elastic net feature selection method to improve the parameter sample size ratio and found the influence factor in periampullary adenocarcinomas. Fukushima et al. (2019) proposed an improved elastic net to realize the accurate prediction of treatment response for multiple sclerosis (MS) patients. Bravo-Merodio et al. (2019) combined the Elastic Net with different supervised learning methods to assess the quality of clinical biomarkers. Watts et al. (2021) used the elastic net to select important clinical variables from risk factor data to predict different types of criminal offenses. Max-Relevance and Min-Redundancy (mRMR) is a feature selection method that uses the dependence of features and tags and the correlation between features (Bose et al., 2019). Wei et al. (2020) found useful texture features in CT images of COVID-19 by mRMR technology and the prediction model showed a good predictive performance. Aghaeipoor and Javidi (2020) put forward the algorithm based on the mRMR framework to improve the accuracy of the estimation method for real-world regression datasets. Özyurt (2020) applied a feature selection algorithm based on CNN–mRMR and an extreme learning Machine (ELM) classifier to achieve a higher classification accuracy for white blood cell detection. Bose et al. (2019) chose attributes with scores greater than zero and a linear model for identifying the key information in nursing documentation. There are many applications of mRMR in biomedicine. Zhang et al. (2021) used the mRMR and IFS-SVM to classify the microbiota biomarkers with orthologous gene annotation. Wu et al. (2020) combined mRMR with SVM for classifying the osteoarthritis hip samples and osteoarthritis knee samples evaluated with LOOCV. Cai et al. (2018) adopted the mRMR and the variance inflation factor regression algorithm to identify their interacting schizophrenia genes in brains. Chen et al. (2019) used the mRMR and recurrent neural network to select the discriminate features for classifying widely expressed genes. The gravity search algorithm (GSA) is a new swarm intelligence optimization algorithm based on the law of universal gravitation and the interaction between particles (Mosa, 2019). Han et al. (2015) combined the enhanced GSA algorithm with BP neural network to segment the image. Goswami and Chakraborty (2015) employed GSA and fireworks algorithm (FWA) to optimize the parameters of the Ultrasonic machining (USM) process. Guha et al. (2020) developed a new method called Clustering-based Population in Binary GSA (CPBGSA), which improved the classification performance of the model for UCI datasets.

The present work is focused on the development of an analytical method for identification of coal miners with abnormal physical signs based on EN-mRMR and intelligent optimization. Firstly, the physical examination data of coal miners are collected, and the samples are randomly divided into the training set and the test set. Then, the elastic net is used for the initial screening of raw data to obtain important physical signs. The identification model of coal miners with abnormal signs is established by mRMR combined with the GSA-SVC algorithm. The features of the data from preliminarily selected physical signs are further simplified, and the key feature subset is selected to obtain the optimal

identification model. Finally, the conclusion and future work plan of this paper are summarized.

MATERIALS AND METHODS

Collection of Physical Examination Data of Coal Miners

With the assistance from the research platform of the Huaihe Energy Hospital for the Prevention and Treatment of Occupational Disease, Biosensing and Comprehensive Health Laboratory of School of electrical and information engineering, and Key Laboratory of Industrial Dust Control and Occupational Health, Ministry of Education, and Anhui University of Science and Technology. The occupational health data of coal miners in the Huainan mine area in 2020 were taken as the research object, and the data set of physical signs were constructed. These coal miners came from six departments, including electromechanical, transportation, fully mechanized mining, development, drivage, and coal preparation plant. The data set contains 320 samples, which are mainly composed of three parts. The first part is the coal miner's basic information, including age (AGE), length of service (LS), and toxic length of service (TLS). The second part is the coal miner's physical parameters, including heart rate (HR), systolic blood pressure (SBP), diastolic blood pressure (DBP), height (HGT), weight (WGT), alanine aminotransferase (ALT), triglyceride (TG), cholesterol (CHOL), glucose (GLU), the predicted value (Pred), measured value, ratio of the measured value to the predicted value (%) of forced vital capacity (FVC), and forced expiratory volume in one second (FEV1) and FEV1/FVC. SBP and DBP are the evaluating indicator of cardiac function. The ALT is the evaluating indicator of liver function. TG and CHOL are the evaluating indicators of blood lipid. The GLU is the evaluating indicator of blood glucose. FVC and FEV1 are the evaluating indicators of pulmonary function. The third part is the final evaluation result, including two examination conclusions: "abnormal physical signs that need further clinical examination" and "normality of the items currently examined". There are 141 coal miners with normal physical signs and 179 coal miners with abnormal physical signs. The proportion of samples with normal and abnormal physical signs of coal miners is about 1:1.27. There is little difference between the number of coal miners with normal and abnormal physical signs. According to the examination results, the category of coal miners is marked.

The hardware conditions of the computer used in the experiment with the Intel ninth-generation core i7-9700, the 3.0GHz eight-core processor, the NVIDIA RTX2070 graphics card (8GB Video memory), the 16G Kingston memory module, etc. The algorithm simulation runs on the MATLAB R2021a (MathWorks, United States) platform.

Elastic Net

Elastic network (EN) is a new embedded feature selection method (Zou and Hastie, 2005). Based on the lasso algorithm, which can select more representative feature variables. The EN uses both L1 and L2 as regularization terms, L1 regularization makes the model become sparse, and L2 makes the model parameters closer to zero. When the model parameters are limited or normalized, some parameters can shrink toward zero. The L1 controls the number of features so that a few features are more important and can be used for feature selection. L2 cannot control the number of features, but it can prevent the model from over-fitting to a feature.

The EN combines the characteristics of the L1 regularized model based on the lasso and the L2 regularized model based on the ridge, which not only ensures the sparsity of the model but also inherits the stability of the ridge.

$$L = \mathrm{X}\omega + E, \tag{1}$$

where $X = [x_1, x_2, \cdots, x_m]^T$ $(X \in R^{m\times n})$ is the attribute variable, $L = [l_1, l_2, \cdots, l_m]^T$ $(L \in R^{m\times 1})$ is the result label, $E \in R^{m\times 1}$ is the random error, and $\omega = [\omega_1, \omega_2, \cdots, \omega_n]^T$ $(\omega \in R^{n\times 1})$ is the regression coefficient vector.

Parameter α can be adjusted according to **Eq. 1** to achieve sparse dimensionality reduction of target variables.

$$Q(\omega) = \arg\min\{\|L - X\omega\|^2 + \lambda_1|\omega| + \lambda_2\|\omega\|_2\}, \tag{2}$$

where λ_1 and λ_2 are penalty coefficients, let $\alpha = \frac{\lambda_1}{\lambda_1+\lambda_2}$, $\lambda = \lambda_1 + \lambda_2$, and we obtain the following formula:

$$Q(\omega) = \arg\min\{\|L - X\omega\|^2 + \lambda[\alpha|\omega| + (1-\alpha)\|\omega\|_2]\}. \tag{3}$$

In the optimization function $Q(\omega)$, the value of α is strictly between zero and one.

Max-Relevance and Min-Redundancy

The Max-Relevance and Min-Redundancy (mRMR) is a filter feature selection method based on mutual information (Hanchuan Peng et al., 2005), which can select features according to the maximum statistical dependence criterion, and has the advantages of high speed and robustness. The algorithm minimizes the redundancy of feature subsets, maximizes the correlation between feature subsets and response variables, and finds several features with the greatest correlation and the least redundancy with each other from the feature space.

The mRMR is defined as follows:

$$\max D(U, l), D = \frac{1}{|U|}\sum_{\xi_i \in U} I(\xi_i; l), \tag{4}$$

$$\min R(U), R = \frac{1}{|U|^2}\sum_{\xi_i, \xi_j \in U} I\left(\xi_i; \xi_j\right), \tag{5}$$

where U is the feature subset, $|U|$ is the feature number, and l is category labels. $I(\xi_i; l)$ is the mutual information between feature i and l. $I(\xi_i; \xi_j)$ is the mutual information between feature i and feature j. D is the mean value between features and categories in feature subsets U, which reflects the correlation between features and category labels. R is the mutual information value between features, which reflects the degree of redundancy between features.

The criteria of mRMR are as follows:

$$\max \phi(D,R), \phi = D - R = \frac{1}{|U|}\sum_{\xi_i \in U} I(\xi_i; l) - \frac{1}{|U|^2}\sum_{\xi_i,\xi_j \in U} I\left(\xi_i; \xi_j\right). \tag{6}$$

mRMR = $\max \phi(D,R)$, and the ultimate goal is to find the set U with maximum correlation and minimum redundancy.

Support Vector Classification of Optimization by Gravitational Search Algorithm

The Gravity Search Algorithm (GSA) is a random heuristic search optimization algorithm (Rashedi et al., 2009). This algorithm is inspired by Newton's law of gravity and motion in physics. A particle is defined as a solution within the range of the solution set. There is an attraction between different solutions, which is affected by the distance between the mass of the solution and the solution. The value of the evaluation function can be used to describe the mass of the particle. In the range of solution set, one solution will get acceleration due to the attraction of other solutions, and the particle with a better evaluation function will provide more acceleration, and then get a better solution. The support vector classification (SVC) is a supervised machine learning algorithm based on the support vector machine for classification problems (Mustafa et al., 2016). The SVC is suitable for small sample learning and simple calculation. It can not only improve the generalization ability of the learning machine by seeking the minimum structural risk but also avoid the disaster of dimensionality in a sense (Manuel Serra et al., 2007). There are two important parameters in the SVC model, one is the *cost* and the other is the *gamma*. The *cost* is the penalty coefficient, which represents the tolerance to error, and the *gamma* is the kernel function parameter. GSA is used to search for the optimal *cost* and *gamma* parameters to improve the performance of the SVC classifier (MadhuSudana Rao et al., 2018).

The steps for GSA to optimize parameters of SVC are as follows:

Step 1. Determined the search space H, the population size K , and the number of iterations N. Initial the gravitational constant G^* , the attenuation coefficient α, the position $Z = (z_1, z_2, \cdots, z_i, \cdots z_n)$, and the speed $V = (v_1, v_2, \cdots, v_i, \cdots v_n)$ of the individual. Randomly generate n particles. The position of the particles corresponds to a set of *cost* and *gamma*.

Step 2. Determine the search range and iteration times of parameters *cost* and *gamma* in SVC.

Step 3. The data of training set is used as the input of the model, the training sample data is trained by SVC, and the classification error rate is used as the optimization objective function.

$$Error\_rate = 1 - ACC, \tag{7}$$

where ACC represents the ratio of samples correctly classified.

Step 4. Put the updated *cost* and *gamma* into the SVC model and calculate the fitness value of each individual in the fitness function $Fit(t)$ based on the minimum value of an objective function.

Step 5. Update the universal gravitation $G(t)$, mass $M(t)$, minimum fitness values $\min Fit_i(t)$, and $\max Fit_i(t)$ at time t.

Step 6. Calculate the gravity between individuals $F_{ij}(t)$ and individual acceleration $a_i(t)$ at time t.

$$F_{ij}(t) = G(t) \cdot \frac{M_i(t) \cdot M_j(t)}{d_{ij} + \tau} \cdot \left(z_i(t) - z_j(t)\right), \tag{8}$$

where $d_{ij}(t)$ represents the Euclidean distance between individuals i, j, and τ is a minimum positive constant.

$$M(t) = \frac{1}{\sum_{j=1}^{n} \frac{Fit_j(t) - worst(t)}{best(t) - worst(t)}} \cdot \frac{Fit_i(t) - worst(t)}{best(t) - worst(t)}, \tag{9}$$

where $best(t)$ represents the optimal fitness value. $worst(t)$ represents the worst fitness value.

$$a_i(t) = \frac{\sum_{j \in kbest} F_{ij}(t)}{M(t)}, \tag{10}$$

where *kbest* represents the number of best heavy mass elements, which can create the balance between exploration and exploitation processes. Some miscellaneous particles are filtered out to highlight the influence proportion of better individuals.

Step 7. Update the individual's position $z_i(t)$ and speed $v_i(t)$.

$$v_i(t+1) = Random \cdot v_i(t) + a_i(t), \tag{11}$$

$$z_i(t+1) = z_i(t) + v_i(t+1), \tag{12}$$

where *Random* is a random variable in the [0,1] interval.

Step 8. Judge whether the maximum number of iterations is reached, and stop when it is reached, otherwise go back to (2) to continue execution.

Step 9. Return the optimal solution of *cost* and *gamma* of SVC model.

Step 10. According to the optimal parameters *cost* and *gamma*, the optimal recognition model is obtained. The test set is used to evaluate the performance of GSA-SVC model.

Feature Construction

Machine learning model learns from training data, so it is very important to construct some features for the related task. Feature construction is achieved by studying the raw data samples combined with the experience of machine learning and professional knowledge in related fields. The existing features are combined or calculated with each other to manually create some new physical significant features, which are useful for model

TABLE 1 | Confusion matrix.

	Identification result	
Actual result	Abnormity	Normal
Abnormity	True positive (TP)	False negative (FN)
Normal	False positive (FP)	True negative (TN)

training and have certain engineering significance (Neshatian et al., 2012; Mahanipour and Nezamabadi-Pour, 2019). To better evaluate the performance and identification accuracy of the model, we constructed four new features based on the raw physical signs. These new features can be used as a measure of health, including the body mass index (BMI) (Opel et al., 2015), the pulse pressure (PP) (Gavish and Bursztyn, 2019), the mean arterial pressure (MAP) (Deverdun et al., 2016), and the rate-pressure product (RPP) (Keller-Ross et al., 2014).

The expression of BMI:

$$BMI = \frac{W}{H^2}, \tag{13}$$

where W represents the weight (kg) and H represents the height (m).

The expression of PP:

$$PP = SBP - DBP. \tag{14}$$

The expression of MAP:

$$MAP = \frac{1}{3}(SBP - DBP) + DBP. \tag{15}$$

The expression of RPP:

$$RPP = HR \cdot SBP, \tag{16}$$

where SBP represents systolic blood pressure, DBP represents diastolic blood pressure and HR represents heart rate.

Evaluation Indexes

The evaluation of model performance can guide the training process of classifiers and compare the performance of different classifiers. Accuracy is our most common evaluation index and is easy to understand. Generally speaking, the higher the accuracy, the better the performance of the model but the high accuracy does not necessarily indicate the algorithm is good. Especially when the uneven distribution of positive and negative samples leads to fewer data in some categories, it is not comprehensive to evaluate an algorithm model only by the index of accuracy. The confusion matrix is shown in **Table 1**. Through the confusion matrix, we can intuitively observe the distribution of each category. The confusion matrix includes four elements: TP, FP, FN, and TN. The first letter represents the match situation between the predictive result, the actual result of the sample, and the second letter represents the predictive result of the sample. The accuracy, precision, recall, specificity, geometric mean (G-mean), and Matthew's correlation coefficient (MCC) are used as quantitative evaluation indexes for the quality of algorithm or model parameters.

The expression of accuracy:

$$Accuracy = \frac{TN + TP}{TN + FN + FP + TP} \times 100\%. \tag{17}$$

The expression of precision:

$$Precision = \frac{TP}{FP + TP} \times 100\%. \tag{18}$$

The expression of recall:

$$Recall = \frac{TP}{FN + TP} \times 100\%. \tag{19}$$

The expression of specificity:

$$Specificity = \frac{TN}{FP + TN} \times 100\%. \tag{20}$$

G-mean is used to measure the evaluation index of the algorithm for a few types of samples and can evaluate the overall identification performance of two types of samples in an unbalanced data set. The closer the value is to one, the better the identification performance is. The closer the value is to zero, the worse the identification performance is. Its expression is

$$G - mean = \sqrt{\frac{TP}{FN + TP} \cdot \frac{TN}{FP + TN}}. \tag{21}$$

MCC is an evaluation index to measure the identification performance of the model. The closer the value is to one, the better the identification performance of the model is. The closer the value is to minus one, the worse the identification performance is. Its expression is

$$MCC = \frac{TN \cdot TP - FN \cdot FP}{\sqrt{(FP + TP)(FN + TP)(FP + TN)(FN + TN)}}. \tag{22}$$

RESULTS AND ANALYSIS

Physical Sign Screening of Elastic Net

The physical signs data of coal miners contains 25 variables after feature creation, but the physical health of coal miners only depends on a few unique and useful signs. Therefore, the screening of sign variables can not only reduce the identification cost but also effectively improve the accuracy, which is very important for the actual auxiliary diagnosis and evaluation process. The EN feature screening algorithm is used to select useful physical signs to avoid the adverse impact of redundant physical signs on auxiliary diagnosis results. The data set of coal miners' physical signs consists of 320 samples. These samples were randomly divided into the training set and the test set according to the ratio of 3:1. The training set contains 240 samples and the test set contains 80 samples.

The training set of data are fed into the EN algorithm. The output is the feature coefficient corresponding to different alpha. The features of the zero coefficient will be deleted. It can be seen from **Eq. 3** that the variable screening of EN is related to the two parameters of alpha and lambda in the optimization function. The value range of alpha is (0, 1), and the value of alpha is set to 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9

TABLE 2 | Results of physical sign screening.

Alpha	Deleted physical signs	Remaining number
0.1	x_{10}	24
0.2	x_2, x_{22}	23
0.3	x_2, x_{10}, x_{22}	22
0.4	x_2, x_5, x_{10}, x_{12}, x_{17}, x_{22}	19
0.5	x_2, x_5, x_{10}, x_{12}, x_{17}, x_{20}, x_{22}	18
0.6	x_2, x_5, x_{10}, x_{12}, x_{17}, x_{18}, x_{20}, x_{21}, x_{22}	16
0.7	x_2, x_4, x_5, x_7, x_{10}, x_{12}, x_{17}, x_{18}, x_{19}, x_{20}, x_{21}, x_{22}	13
0.8	x_2, x_4, x_5, x_7, x_{10}, x_{12}, x_{17}, x_{18}, x_{19}, x_{20}, x_{21}, x_{22}, x_{23}	12
0.9	x_2, x_4, x_5, x_7, x_{12}, x_{17}, x_{18}, x_{19}, x_{20}, x_{21}, x_{22}, x_{23}	13

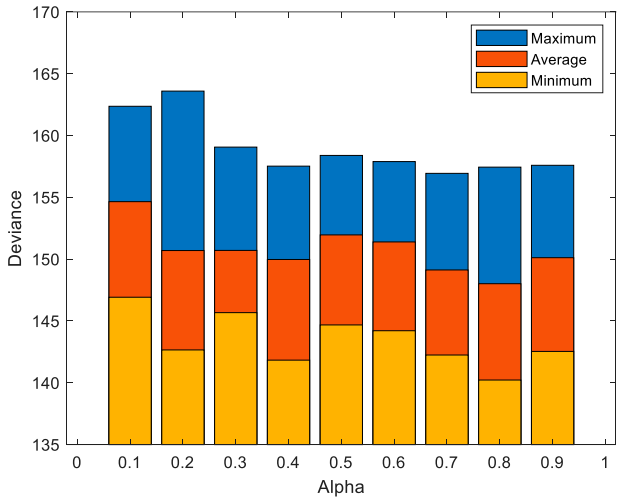

FIGURE 1 | Variation curves of different alpha values. The alpha values range from 0.1 to 0.9, and there are different degrees of differences between the maximum, minimum, and average values. When alpha = 0.8, the minimum and average value of deviance are the minimum values, which are 148.01 and 140.22, respectively.

at intervals of 0.1. Then, a 5-fold cross-validation is adopted based on the cross-validation of the deviation criterion. The results of physical signs screening are shown in **Table 2**. When alpha is 0.1, the number of deleted physical signs is the least, only the 10th physical sign is deleted, and the sign variable corresponding to the non-zero sparsity coefficient is the feature variable. With the continuous increase of alpha, more and more feature coefficients will become zero, and the number of deleted physical signs is gradually increasing. When the alpha is 0.8, the number of remaining physical signs is the least, a total of 12. When the alpha value is incremented from 0.3 to 0.4 and 0.6 to 0.7, the number of deleted features is larger than others. It can be roughly seen from the statistical information of sign parameters deleted by the different alpha that the LS and the measured values of FEV1/FVC appear the most times, a total of eight times. The HGT appeared seven times in total. The SBP, the BMI, and the predicted values of FVC appeared more times, more than five times in total. The deleted physical signs have little impact on the diagnostic results of coal miners' health assessment, which belong to redundant or useless information in the data.

To further judge the specific impact of different physical signs, each alpha value is substituted into EN to loop 10 times, and the value of alpha is finally determined by the maximum, minimum, and average value of deviation. The variation curves of different alpha values are shown in **Figure 1**. The alpha values range from 0.1 to 0.9, and there are different degrees of differences between the maximum, minimum, and average values. When alpha = 0.2, the maximum deviance reaches the maximum value of 163.59. When alpha = 0.1, the average value and minimum value of deviance reach the maximum, which are 154.65 and 146.92, respectively. When alpha = 0.8, the minimum and average value of deviance are the minimum values, which are 148.01 and 140.22, respectively. In conclusion, better screening results are obtained through the EN model. (AGE, TLS, DBP, MAP, RPP, WGT, ALT, TG, CHOL, GLU, FEV1%, and FEV1/FVC%) are the final result of selected features.

When the alpha of the cross-validation parameter is 0.8, the deviance variation curves of different regularization coefficients are shown in **Figure 2**. The circle at the green dotted line in the figure refers to the minimum deviance point. As can be seen from the figure, the value of minimum λ based on the deviance criterion is 0.0053.

Feature Selection of Max-Relevance and Min-Redundancy

The number of physical signs initially screened by EN is reduced by 13 from 25 which is 48%. This procedure greatly reduces redundant information. However, the mechanism of feature selection by EN is completed by deleting the variable with zero coefficient. There may be redundancy between the deleted 13 individual features. The correlation and redundancy between the physical signs are not fully considered, and the number of physical signs retained is still large. To achieve the accurate and efficient identification of coal miners with abnormal physical signs, the combination algorithm of filter and wrapper is employed for subsequent analysis. Firstly, the mRMR algorithm is used to sort the importance of features, then a specific classifier is used to select features, and the optimal feature subset is simplified according to the evaluation indexes. After the initial selection of EN, the 12 features are selected by the mRMR algorithm.

The importance score of different physical signs is shown in **Figure 3**. According to the value of importance score, the order from large to small is (DBP, ALT, CHOL, GLU, TG, RPP, WGT,

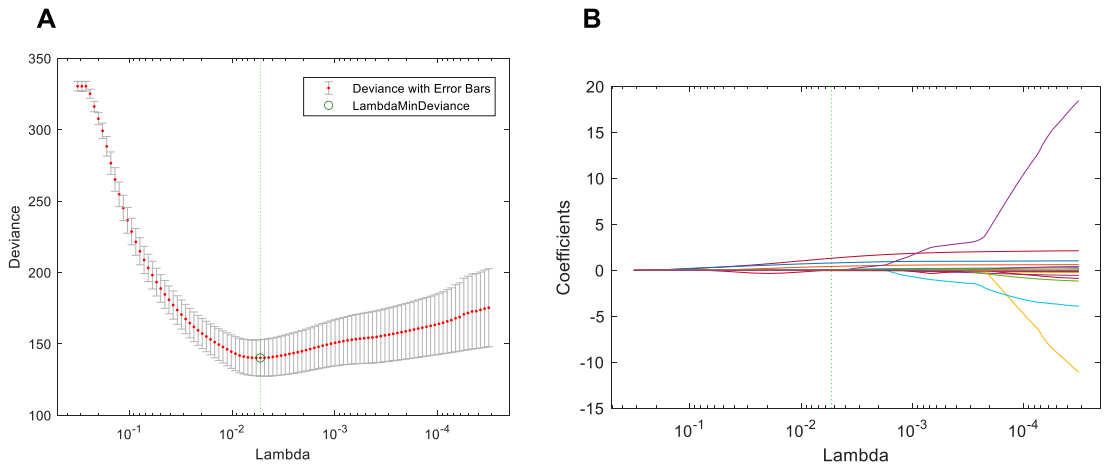

FIGURE 2 | Implementation results of the EN. **(A)** Deviance variation curves. The circle at the green dotted line in the figure refers to the minimum deviance point. The value of minimum λ based on the deviance criterion is 0.0053; **(B)** Variation track of coefficient. The position of the green dotted line corresponds to different coefficients.

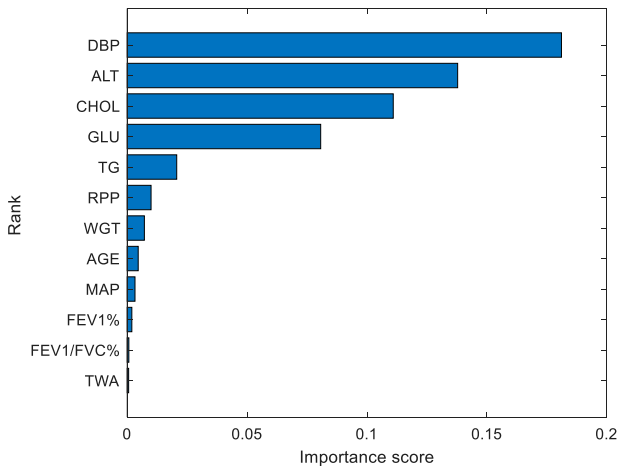

FIGURE 3 | Importance score of different physical signs. From top to bottom, the importance score of 12 physical signs are decreased continuously. The DBP has the highest importance, with a value of 0.18. The TWA is the least important, with a value of 0.0006.

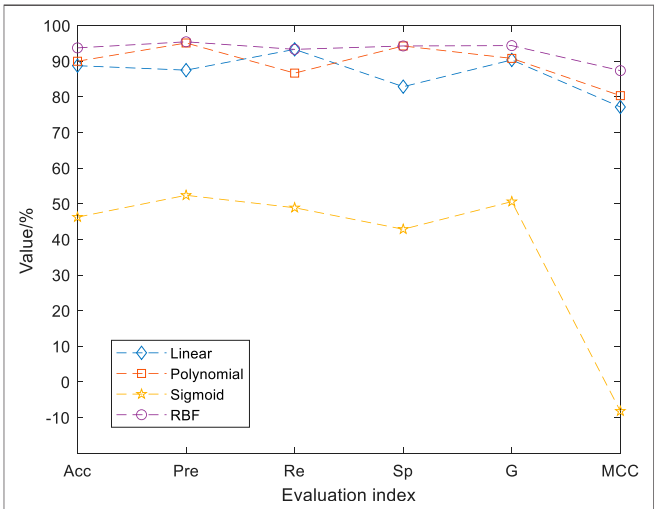

FIGURE 4 | Identification results of different kernel functions. The values of each evaluation index of RBF kernel function are high. The accuracy, precision, recall, specificity, G-mean, and MCC of the test set were 93.75%, 95.45%, 93.33%, 94.29%, 0.94, and 0.87. The values of each evaluation index of Sigmoid kernel function are low.

AGE, MAP, FEV1%, FEV1/FVC%, and TWA). From top to bottom, the importance score of DBP, ALT, CHOL, GLU, and TG decreased significantly. The decrease in importance score indicates the degree of reliability of feature selection. After that, the decline of the importance scores of the physical signs are not obvious, and they are stable.

Identification Model for Coal Miners With Abnormal Physical Signs

The initial population number of GSA-SVC is set to 20, the maximum number of iterations is set to 100, the search interval of penalty coefficient cost and kernel function parameter gamma are respectively set to (0,100), and the data are normalized to (0,1). Different kernel functions of SVC will affect the identification performance of the model. The SVC model of linear, polynomial, sigmoid, and radial basis function (RBF) kernel function are, respectively, established for the 12 physical signs data screened by EN. The training set of 12 physical signs data are fed into the GSA-SVC algorithm, and the identification model of coal miners with abnormal physical signs is established. The test set is used to evaluate the performance of GSA-SVC model. The output of the model is the category of the predictive coal miners, accuracy, precision, recall, specificity, G-mean, and MCC of the test set.

TABLE 3 | The identification results of the test set of different feature subsets.

Feature Subset	Number	Cost	Gamma	Acc/%	Pre/%	Re/%	Sp/%	G	MCC
$\{x_6\}$	1	93.4226	86.2252	72.5 (58/80)	73.47	80.00	62.86	0.77	0.44
$\{x_6, x_{13}\}$	2	4.3976	49.7381	82.5 (66/80)	91.89	75.56	91.43	0.83	0.67
$\{x_6, x_{13}, x_{15}\}$	3	15.0194	13.3854	87.5 (70/80)	94.87	82.22	94.29	0.88	0.76
$\{x_6, x_{13}, x_{15}, x_{16}\}$	4	17.9261	7.5884	87.5 (70/80)	94.87	82.22	94.29	0.88	0.76
$\{x_6, x_{13}, x_{15}, x_{16}, x_{14}\}$	5	67.081	14.4606	93.75 (75/80)	97.62	91.11	97.14	0.94	0.88
$\{x_6, x_{13}, x_{15}, x_{16}, x_{14}, x_9\}$	6	71.4777	12.4889	97.5 (78/80)	97.78	97.78	97.14	0.98	0.95
$\{x_6, x_{13}, x_{15}, x_{16}, x_{14}, x_9, x_{11}\}$	7	1.7972	42.3299	97.5 (78/80)	95.74	100	94.29	0.98	0.95
$\{x_6, x_{13}, x_{15}, x_{16}, x_{14}, x_9, x_{11}, x_1\}$	8	34.5154	3.3239	96.25 (77/80)	100	93.33	100	0.97	0.93
$\{x_6, x_{13}, x_{15}, x_{16}, x_{14}, x_9, x_{11}, x_1, x_8\}$	9	55.8913	2.5494	95 (76/80)	97.67	93.33	97.14	0.95	0.90
$\{x_6, x_{13}, x_{15}, x_{16}, x_{14}, x_9, x_{11}, x_1, x_8, x_{24}\}$	10	64.1792	1.7058	95 (76/80)	95.56	95.56	94.29	0.96	0.90
$\{x_6, x_{13}, x_{15}, x_{16}, x_{14}, x_9, x_{11}, x_1, x_8, x_{24}, x_{25}\}$	11	54.321	1.4341	93.75 (75/80)	95.45	93.33	94.29	0.94	0.87
$\{x_6, x_{13}, x_{15}, x_{16}, x_{14}, x_9, x_{11}, x_1, x_8, x_{24}, x_{25}, x_3\}$	12	63.1411	0.5812	93.75 (75/80)	95.45	93.33	94.29	0.94	0.87

The identification results are shown in **Figure 4**. According to the accuracy (Acc), precision (Pre), recall (Re), specificity (Sp), G-mean (G), and MCC, we find that the value of each evaluation index of sigmoid kernel function is very low, so the identification effect of the model is the worse, which cannot meet the needs of accurate identification. The values of each evaluation index of RBF kernel function are high, especially the values of accuracy, G-mean, and MCC are significantly greater than the corresponding values of other kernel functions. Therefore, the radial basis function (RBF) is selected as the kernel function of SVC.

Because the DBP is the feature with the highest importance score, it is the most important feature. Therefore, the subset composed of DBP is used as the first feature subset of the new feature ranking. Based on the importance ranking, new features with higher importance than the previous ones are continuously added and expanded into feature subsets with a different number of features. There are 12 different feature combinations until all 12 individual features are selected. The training set of different feature subset data are fed into the GSA-SVC algorithm, and the identification model of coal miners with abnormal physical signs is established. The test set is used to evaluate the performance of GSA-SVC model. The output of the model is the category of the predictive coal miners, accuracy, precision, recall, specificity, G-mean, and MCC of the test set. The final simplified feature subset is obtained according to the evaluation indexes of the test set.

The identification results of the test set of different feature subsets are shown in **Table 3**. When there is only one feature, the accuracy of the test set is the lowest, only 72.5%, and 22 samples are misclassified. From the first feature to the second feature, the accuracy, precision, and specificity of the test set increased rapidly, with an increase of 10, 18.42, and 28.57%, respectively, of which the increase of specificity was the most significant. When the number of features increases to six, each evaluation index of the test set reaches the maximum for the first time. The accuracy of the test set is 97.5%, and only two samples are misclassified, with the corresponding cost = 71.4777 and gamma = 12.4889. Compared with the model of six features, although the test set accuracy of the model with seven features is also 97.5%, the difference is mainly reflected in the three evaluation indexes of precision, recall, and specificity. However, the precision and specificity of the model with six features are better than that with seven features. The accuracy of the test set after the seven features decreases as a whole and finally tends to be stable. Based on the complexity of data and different evaluation indexes, the model that cannot only ensure the number of features as few as possible but also obtain the high performance is selected. (DBP, ALT, CHOL, GLU, TG, and RPP) as the result of mRMR algorithm selection. Blood pressure is the evaluating indicator of human cardiovascular function and maintains the oxygen and nutrition supply of body organs and tissues. The DBP value is above 90 mmHg, which represents hypertension. Abnormal blood pressure will damage the blood vessels of the human body and damage different organs such as heart, eyes, brain, and kidney. The injury of hepatocytes will release ALT into the blood and cause the increase in ALT content. Vigorous exercise, long-term and heavy drinking, eating habits, irregular life, and rest may all cause the increase in ALT content. CHOL contains all cholesterol in the human body. CHOL is produced and integrated by the liver. Abnormal CHOL will cause dyslipidemia, which will lead to angina pectoris, coronary heart disease, myocardial infarction, cerebral infarction, and other diseases. GLU is an indicator of blood glucose, abnormal GLU can lead to diabetes, human nervous system, and circulatory system diseases. TG is a blood lipid evaluation index. Too high content of TG will produce more fat, which not only increases the viscosity of blood but also causes blood vessel blockage and hypoxia in the brain and myocardium. RPP refers to the product of heart rate and systolic blood pressure. It is a sensitive reliability index of myocardial oxygen consumption. Abnormal RPP usually causes diseases such as stroke and heart failure. According to the screened physical signs, it can be seen that the abnormal physical signs of mine workers are mainly concentrated in the problem of cardiac function, liver function, blood lipid, and cardiac function.

The identification results of GSA-SVC of the raw 25 physical signs are shown in **Figure 5**. The label "−1" represents the diagnostic result indicates further clinical examination is required, and the label "1" represents the diagnostic result that no abnormality is found in the currently tested items. In **Figure 5B**, the blue diamond is the actual classification of

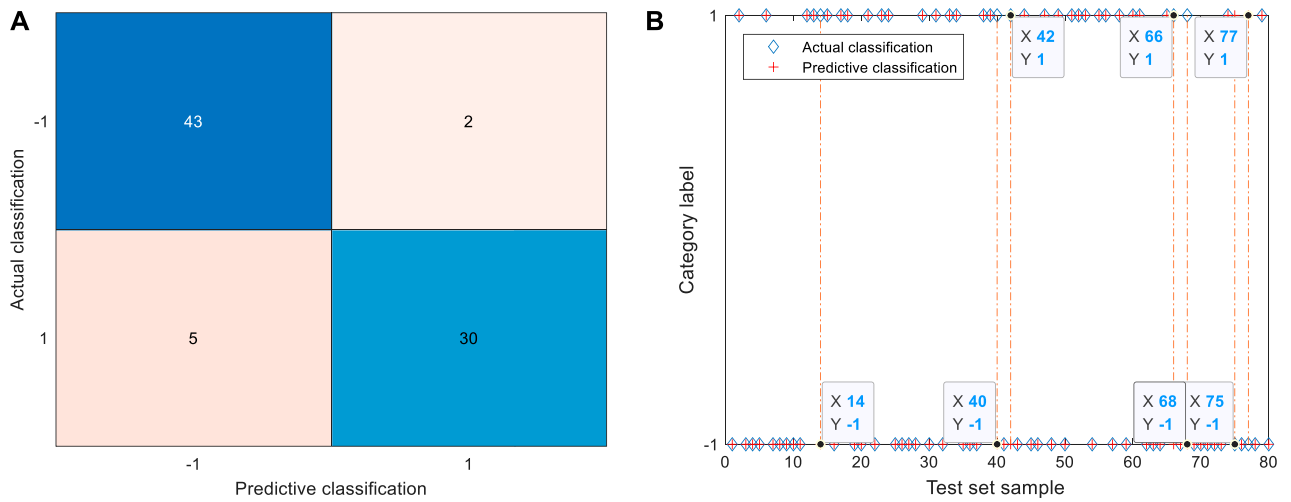

FIGURE 5 | Identification results of 25 physical signs by GSA-SVC. **(A)** Confusion matrix of test set. The TP is 43, TN is 30, FP is 5, and FN is 2; **(B)** Identification results of test set. The 14th, 40th, 42nd, 66th, and 68th samples were misclassified as abnormal coal miners in the test set. The 75th and 77th samples were misclassified as normal coal miners in the test set.

input samples, and the red plus sign is the result of the predictive classification of the model. If the plus sign can fill in the diamond, it means that samples are correctly identified. The orange dotted line corresponds to the misclassified sample. The color depth of different regions of the confusion matrix reflects the number of samples corresponding to the region. The darker the color, the larger the number of samples, otherwise the less the number of samples.

According to the confusion matrix and identification results of the test set, 43 coal miners with abnormal physical signs and 30 coal miners with normal physical signs were identified correctly. The five coal miners with normal physical signs in the test set were wrongly identified as coal miners with abnormal signs. These five coal miners belong to high-risk groups, and they are more likely to have physical abnormalities in the future, which should be paid attention to. The 14th, 40th, 42nd, 66th, and 68th samples were misclassified, corresponding to the 116th, 310th, 268th, 295th, and 33rd coal miners of the physical signs data set, respectively. Two coal miners with abnormal physical signs in the test set were incorrectly identified as normal coal miners. These two coal miners belong to the missed diagnosis population, and the coal miners with abnormal physical signs were not successfully identified. The 75th and 77th samples were misclassified, corresponding to the 176th and 119th coal miners in the physical signs data set, respectively. Among the identification results of raw 25 physical signs, the misclassification results occurred most frequently in the two departments of the excavation and coal preparation plant, and there was no misclassification in the electromechanical department.

The raw data are simplified to the greatest extent by EN and mRMR. The identification results of GSA-SVC of six physical signs are shown in **Figure 6**. According to the confusion matrix and identification results of the test set, 44 coal miners with abnormal physical signs and 34 coal miners with normal physical signs were identified correctly. One coal miner with normal physical signs was wrongly identified as a coal miner with abnormal physical signs, and the 66th sample in the test set was wrongly identified, corresponding to the 295th coal miner in the data set. One coal miner with abnormal physical signs in the test set is wrongly identified as a coal miner with normal physical signs, and the eighth sample is wrongly identified, corresponding to the 115th coal miner in the data set. The identification result of the model based on the features of raw data simplified by EN and mRMR is better than that of the raw data based on the model. The number of errors in identification has been reduced by five, which is mainly reflected by the number of real normal samples is four less than that of errors misidentified as abnormal samples.

Comparison With Other Classifiers

To verify the reliability of the method proposed for the identification of abnormal physical signs, we compare the algorithms proposed in different stages. EN and mRMR are separately employed to select the features of the raw physical signs data, and then the selected features are used as the input of the unoptimized SVC and GSA-SVC algorithms to establish the identification model of abnormal physical signs. In the libsvm-mat-3.1 toolkit, the default value of the penalty coefficient cost is set to one, and the default value of the kernel function parameter gamma is set to the reciprocal of the feature number. Then, the feature selection method of EN combined with mRMR is used to simplify the raw data and establish the identification model of abnormal physical signs. Finally, the performance of the model is evaluated by comparing the identification accuracy and the number of features of the test set with different processing methods.

The identification results of different processing methods are shown in **Table 4**. The number of features and evaluation indexes

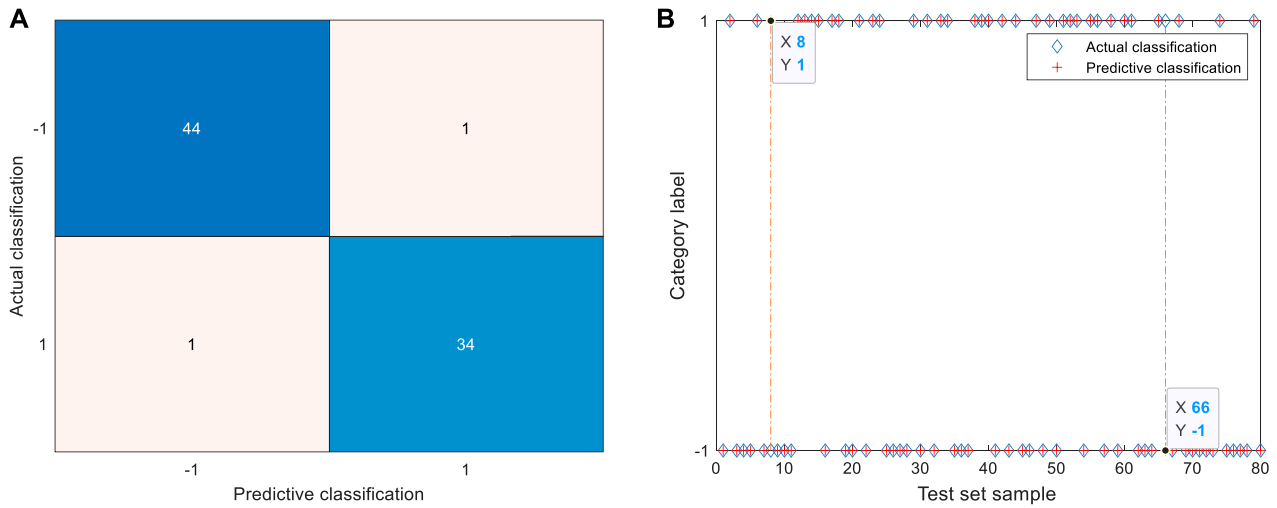

FIGURE 6 | Identification results of six physical signs by GSA-SVC. **(A)** Confusion matrix of test set. The TP is 44, TN is 34, FP is 1, and FN is 1; **(B)** Identification results of test set. The 66th samples were misclassified as abnormal coal miners in the test set. The eighth samples were misclassified as normal coal miners in the test set.

TABLE 4 | Identification results of different processing methods.

Method	Number	Cost	Gamma	Acc/%	Pre/%	Re/%	Sp/%	G	MCC
SVC	25	1	0.04	78.75 (63/80)	79.17	84.44	71.43	0.82	0.57
GSA-SVC	25	48.8545	0.5978	91.25 (73/80)	89.58	95.56	85.71	0.93	0.82
EN + SVC	12	1	0.0833	77.50 (62/80)	80	80	74.29	0.8	0.54
EN + GSA-SVC	12	63.1411	0.5812	93.75 (75/80)	95.45	93.33	94.29	0.94	0.87
mRMR + SVC	8	1	0.125	85.00 (68/80)	88.37	84.44	85.71	0.86	0.70
mRMR + GSA-SVC	12	25.8714	2.5898	96.25 (77/80)	95.65	97.78	94.29	0.97	0.92
EN + mRMR + SVC	7	1	0.1429	88.75 (71/80)	87.50	93.33	82.86	0.90	0.77
Proposed	6	71.4777	12.4889	97.50 (78/80)	97.78	97.78	97.14	0.98	0.95

of different methods is different in varying degrees. The accuracy and other evaluation indexes of the test set of the GSA-SVC model based on the raw 25 physical signs are significantly higher than those of the SVC model, which shows that the GSA is based on the swarm intelligence algorithm can greatly improve the identification accuracy of the SVC model. The accuracy of the test set of the SVC model with different methods is less than 90%, which shows that the unoptimized SVC model cannot achieve the accurate and efficient identification of abnormal physical signs. When the classifier is SVC, the test set obtained by using a single EN algorithm to process the raw data has the lowest accuracy and more features are deleted. The combination of EN and mRMR is used to process the raw data, the accuracy of the test set is high, and the number of deleted features is also small. The accuracy of the test set of the GSA-SVC model built by different methods is more than 90%. Compared with the EN algorithm, the model processed by mRMR has realized a better identification result. The value of the evaluation indexes of the test set obtained by the method proposed in this paper is larger than those of other processing methods in different stages, and the prediction effect is better. At the same time, the number of corresponding features is the least. Compared to the raw data and the data processed by single EN, the number of features is reduced by 76 and 50%, respectively.

We combine the processing method of EN and mRMR with SVC under different intelligent optimization algorithms to evaluate the performance of the model to verify its novelty and superiority of the proposed model. The compared intelligent optimization algorithms include the grid-search (GS), genetic algorithm (GA), and particle swarm optimization (PSO). The initial population number is set to 20, the maximum number of iterations is set to 100, the search interval of penalty coefficient cost and kernel function parameter gamma are respectively set to (0,100), and the data are normalized to (0,1) to ensure the unity of initial conditions.

The comparison results of different intelligent optimization algorithms are shown in **Table 5**. The SVC model with optimized parameters all achieves the best identification effect at six features. The test set accuracy of SVC optimized by GS, GA, and PSO is 96.25%, and other evaluation indexes can also reflect the same identification effect. The difference between the first three optimization strategies is mainly reflected in the evaluation index of the training set. The training set of GS and GSA has the highest accuracy. The model based on GSA can achieve a

TABLE 5 | Comparison results of different intelligent optimization algorithms.

Optimization algorithm	Number	Training set						Test set					
		Acc/ %	Pre/ %	Re%	Sp/ %	G	MCC	Acc/ %	Pre/ %	Re/ %	Sp/ %	G	MCC
GS- SVC	6	99.17	99.25	99.25	99.06	0.99	0.98	96.25	95.65	97.78	94.29	0.97	0.92
GA- SVC	6	97.08	95.68	99.25	94.34	0.97	0.94	96.25	95.65	97.78	94.29	0.97	0.92
PSO- SVC	6	97.92	97.08	99.25	96.23	0.98	0.96	96.25	95.65	97.78	94.29	0.97	0.92
GSA-SVC	6	99.17	99.25	99.25	99.06	0.99	0.98	97.50	97.78	97.78	97.14	0.98	0.95

TABLE 6 | Comparison results of different classifiers.

Classifier	Number	Training set						Test set					
		Acc/%	Pre/%	Re%	Sp/%	G	MCC	Acc/%	Pre/%	Re/%	Sp/%	G	MCC
DT	8	98.33	97.79	99.25	97.17	0.99	0.97	97.50	100.00	95.56	100.00	0.98	0.95
KNN	6	90.42	95.87	86.57	95.28	0.91	0.81	93.75	97.62	91.11	97.14	0.94	0.88
RF	6	100.00	100.00	100.00	100.00	1.00	1.00	96.25	95.65	97.78	94.29	0.97	0.92
NB	11	90.00	95.08	86.57	94.34	0.91	0.80	96.25	97.73	95.56	97.14	0.97	0.92
ELM	7	93.33	93.27	91.51	94.78	0.92	0.86	96.25	94.44	97.14	95.56	0.96	0.92
Proposed	6	99.17	99.25	99.25	99.06	0.99	0.98	97.50	97.78	97.78	97.14	0.98	0.95

better identification result by comprehensively comparing the evaluation indexes of the training set and the test set.

The decision tree (DT), k-nearest neighbors (KNN), random forest (RF), naive bayesian (NB), and extreme learning machine (ELM) are commonly used machine learning algorithms, and they belong to supervised learning. First of all, the training set of 25 feature data are fed into EN algorithms, and variables with zero coefficient are deleted. Then, the remaining variables are used as inputs to mRMR and different feature subsets are generated. Finally, the training set of different feature subset data are fed into the DT, KNN, RF, NB, and ELM algorithms. The identification model of coal miners with abnormal physical signs is established. The test set is used to evaluate the performance of the above models. The output of the model is the category of the predictive coal miners, accuracy, precision, recall, specificity, G-mean, and MCC of the training set and test set. The number of features is obtained according to evaluation indexes. Multiple evaluation indexes of the training set and test set of different classifiers are compared to evaluate the identification performance of different classifiers. The comparison results are shown in **Table 6**. The accuracy and other evaluation indexes of the training set and test set of the KNN are low. Although the accuracy of the test set of the RF, NB, and ELM is the same, which indicates that they have similar identification performance, the accuracy of the training set of the NB is the lowest. The identification effect of the test set of the DT is closest to the proposed classifier, but most of the evaluation indexes such as the accuracy of the training set are worse than the proposed classifier, and the number of reduced features is more. The KNN, RF, and the proposed classifier finally reduced the number of features to six. The number of features selected by NB is the most, which is about twice as many.

It needs to be verified that the new features generated by feature construction can improve the identification performance of the model, rather than adding some useless features to increase

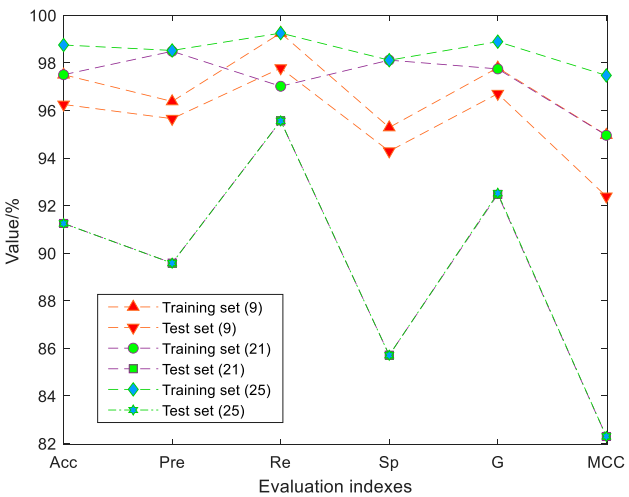

FIGURE 7 | Comparison of results after simplifying features. The values of each evaluation index of test set (9) are high. The accuracy, precision, recall, specificity, G-mean, and MCC of the test set were 96.25%, 95.65%, 97.78%, 94.29%, 0.97, and 0.92. The values of each evaluation index of test set (21) are low. The accuracy, precision, recall, specificity, G-mean, and MCC of the test set were 91.25%, 89.58%, 95.56%, 85.71%, 0.93, and 0.82.

the complexity of algorithm operation. We combined EN-mRMR with GSA-SVC to simplify the 21 physical signs data before feature construction. The simplified results are shown in **Figure 7**. When the number of physical signs is nine, the GSA-SVC model achieves the best identification results. The subset of nine features is (DBP, RPP, CHOL, GLU, TG, WGT, SBP, HR, AGE, FEV1%, FEV1/FVC%, and TWA). The accuracy, G-mean, and MCC of the training set of the identification model with nine physical signs and 21 physical signs data are the same,

TABLE 7 | Comparison results of different classifiers.

Classifier	Numbers	Training set						Test set					
		Acc/%	Pre/%	Re/%	Sp/%	G	MCC	Acc/%	Pr/%	Re/%	Sp/%	G	MCC
DT	7	99.58	100.00	99.25	100.00	1.00	0.99	95.00	100.00	91.11	100.00	0.95	0.90
KNN	7	91.25	90.65	94.03	87.74	0.92	0.82	95.00	95.56	95.56	94.29	0.96	0.90
RF	7	100.00	100.00	100.00	100.00	1.00	1.00	95.00	95.56	95.56	94.29	0.96	0.90
NB	11	89.58	96.58	84.33	96.23	0.90	0.80	97.50	100.00	95.56	100.00	0.98	0.95
ELM	8	93.75	93.33	92.45	94.78	0.93	0.87	95.00	94.29	94.29	95.56	0.94	0.90
SVC	8	83.33	83.57	87.31	78.30	0.85	0.66	87.50	85.71	93.33	80.00	0.89	0.75
Proposed	9	97.50	96.38	99.25	95.28	0.98	0.95	96.25	95.65	97.78	94.29	0.97	0.92

and the values are 97.5, 98, and 95%, respectively. There is no significant difference in the identification effect of the training set. The evaluation indexes of the test set are significantly better than that of 21 physical signs. Compared to the evaluation indexes of the model of the raw data with 21 physical signs, the EN-mRMR method can also improve the identification performance of the model. The performance indexes of the test set with 21 physical signs and the test set with 25 physical signs after feature construction are the same. The Identification performance of the model with the training set of 25 physical signs is better.

The data of 21 physical signs are also used as the input of the DT, KNN, RF, NB, and ELM and combined with EN-mRMR to simplify the feature number of the data. First of all, the training set of 21 feature data are fed into EN algorithms, and variables with zero coefficient are deleted. Then, the remaining variables are used as inputs to mRMR and different feature subsets are generated. Finally, the training set of different feature subset data are fed into the DT, KNN, RF, NB, and ELM algorithms. The identification model of coal miners with abnormal physical signs is established. The test set is used to evaluate the performance of the above models. The output of the model is the category of the predictive coal miners, accuracy, precision, recall, specificity, G-mean, and MCC of the training set and test set. The number of features is obtained according to evaluation indexes. Multiple evaluation indexes of the training set and test set are compared to evaluate the identification performance of different classifiers. The comparison results are shown in **Table 7**. The test set of the DT, KNN, RF, and ELM has a similar identification performance. The number of features reduced by DT, KNN, and RF is the same. Although the evaluation indexes of the test set are higher than those of other classifiers, the accuracy, recall, and MCC of the training set are very low, and the number of selected features is the largest. Overall, the EN-mRMR can still achieve better identification performance by simplifying the data without feature construction and combining the nine physical signs obtained with the GSA-SVC classifier.

Referring to **Tables 3**, **6**, **7**; **Figure 7**, it can be found that feature construction improves the identification performance of GSA-SVC and other classifiers at the same time. The EN-mRMR combined with different classifiers is used to simplify the data after feature construction, and the number of features after is less than the raw data before feature construction. In particular, the method of EN-mRMR combines with GSA-SVC for data with feature construction is better than the data without feature construction, and increases the evaluation indexes for the training set and test set.

DISCUSSION AND CONCLUSION

In this paper, we put forward a new strategy for identification of coal miners with abnormal physical signs based on EN-mRMR. Firstly, we constructed some features of related tasks through feature construction technics. Then the EN was used to delete the redundant physical parameters. Finally, we combined the mRMR with the GSA-SVC algorithm to establish the identification model of coal miners with abnormal physical signs. The features of the preliminarily selected data of physical signs were simplified, and the most important feature subset was obtained. To verify the reliability of the proposed strategy, we compared the identification performance of the proposed algorithms in different stages. We compared the identification performance of different classifiers to verify the novelty and superiority of the proposed model. We also compared the model after feature construction with the model before.

The analysis of the experimental process and the evaluation of the final results show that: 1) The single feature selection method of EN or mRMR can delete redundant physical signs, select useful features, reduce the complexity of data, and avoid the interference and indifferent impact of redundant data on the identification performance of the model. The number of physical signs selected by EN or mRMR is 48% of the number of all physical signs. 2) Compared to a single feature selection algorithm, the proposed strategy in this paper can reduce the number of features of the data to the greatest extent, and those fewer features are sufficient to reflect the key information of the raw data. The number of physical signs selected by EN-mRMR is only 24% of the number of all physical signs. 3) The GSA is employed to optimize the parameters of SVC. The established identification model not only has a higher identification accuracy but also avoids the over-fitting of the model to a certain extent. The accuracy of the training set and test set are 99.17 and 97.50%, respectively. 4) The data after feature construction are selected by EN-mRMR combined with different classifiers. The number of feature reduction and the identification performance of the model is improved to varying degrees. 5) This experiment verifies the feasibility of EN-mRMR combined with GSA-SVC for the

identification of coal miners with abnormal physical signs. The proposed strategy in this paper improves the modeling efficiency and model performance with fewer features and realizes the accurate identification of abnormal physical signs.

Despite the achievement of some research results, there are some limitations in this study. Due to the limitation of sample data and diagnosis results, the diagnosis results of normal and abnormal physical signs are identified based on a small sample size in this paper, which is primarily for early screening and early warning for the detection of coal miners' occupational and suspected occupational diseases. Therefore, we plan to collect more data samples of different kinds of diseases on the conditions permitted to identify and evaluate different types of occupational diseases.

DATA AVAILABILITY STATEMENT

The data in this study can be obtained from the corresponding author. Requests to access the datasets should be directed to Kai Bian, 422088134@qq.com.

AUTHOR CONTRIBUTIONS

MZ and KB conceived the study. KB and MZ developed the method. KB and FH built the hardware platform and software platform. KB implemented the algorithms. KB and WL validated and analyzed the results. MZ supervised the study. KB wrote the original draft. MZ reviewed and edited the manuscript. KB and FH saved the data. All authors read and approved the final manuscript and content of the work.

FUNDING

This work was supported by the Major Science and Technology Program of Anhui Province (Grant No. 201903a07020013), New Generation of Information Technology Innovation Project (Grant No. 2019ITA01010), National Key Research and Development Program of China (Grant No. 2018YFC0604503), Demonstration Project of Science Popularization Innovation and Scientific Research Education for College Students (Grant No. KYX202117), Energy Internet Joint Fund Project of Anhui Province (Grant No. 2008085UD06), Key Projects of Natural Science Research in Anhui Universities (Grant No. KJ2021A0470), and University-Level Key Projects of Anhui University of Science and Technology (Grant No. xjzd2020-06).

ACKNOWLEDGMENTS

The authors would like to highly thank the Hospital for the Prevention and Treatment of Occupational Disease and Huaihe Energy for the data.

REFERENCES

Ackermann, U. (2004). Regulation of Arterial Blood Pressure. *Surg. Oxf.* 22, 120a–120f. doi:10.1383/surg.22.5.120a.33383

Aghaeipoor, F., and Javidi, M. M. (2020). A Hybrid Fuzzy Feature Selection Algorithm for High-Dimensional Regression Problems: an mRMR-Based Framework. *Expert Syst. Appl.* 162, 113859. doi:10.1016/j.eswa.2020.113859

Blondeel, M., and Van de Graaf, T. (2018). Toward a Global Coal Mining Moratorium? a Comparative Analysis of Coal Mining Policies in the USA, China, India and Australia. *Clim. Change* 150, 89–101. doi:10.1007/s10584-017-2135-5

Bose, E., Maganti, S., Bowles, K. H., Brueshoff, B. L., and Monsen, K. A. (2019). Machine Learning Methods for Identifying Critical Data Elements in Nursing Documentation. *Nurs. Res.* 68, 65–72. doi:10.1097/NNR.0000000000000315

Bravo-Merodio, L., Williams, J. A., Gkoutos, G. V., and Acharjee, A. (2019). -Omics Biomarker Identification Pipeline for Translational Medicine. *J. Transl. Med.* 17, 155. doi:10.1186/s12967-019-1912-5

Cai, L., Huang, T., Su, J., Zhang, X., Chen, W., Zhang, F., et al. (2018). Implications of Newly Identified Brain eQTL Genes and Their Interactors in Schizophrenia. *Mol. Ther. - Nucleic Acids* 12, 433–442. doi:10.1016/j.omtn.2018.05.026

Chen, L., Pan, X., Zhang, Y.-H., Liu, M., Huang, T., and Cai, Y.-D. (2019). Classification of Widely and Rarely Expressed Genes with Recurrent Neural Network. *Comput. Struct. Biotechnol. J.* 17 (17), 49–60. doi:10.1016/j.csbj.2018.12.002

Deverdun, J., Akbaraly, T. N., Charroud, C., Abdennour, M., Brickman, A. M., Chemouny, S., et al. (2016). Mean Arterial Pressure Change Associated with Cerebral Blood Flow in Healthy Older Adults. *Neurobiol. Aging* 46, 49–57. doi:10.1016/j.neurobiolaging.2016.05.012

Fukushima, A., Sugimoto, M., Hiwa, S., and Hiroyasu, T. (2019). Elastic Net-Based Prediction of IFN-β Treatment Response of Patients with Multiple Sclerosis Using Time Series Microarray Gene Expression Profiles. *Sci. Rep.* 9, 1822. doi:10.1038/s41598-018-38441-2

Galarraga C., O. A., Vigneron, V., Dorizzi, B., Khouri, N., and Desailly, E. (2017). Predicting Postoperative Gait in Cerebral Palsy. *Gait Posture* 52, 45–51. doi:10.1016/j.gaitpost.2016.11.012

Gavish, B., and Bursztyn, M. (2019). Ambulatory Pulse Pressure Components. *J. Hypertens.* 37, 765–774. doi:10.1097/HJH.0000000000001920

Goswami, D., and Chakraborty, S. (2015). Parametric Optimization of Ultrasonic Machining Process Using Gravitational Search and Fireworks Algorithms. *Ain Shams Eng. J.* 6, 315–331. doi:10.1016/j.asej.2014.10.009

Grzywalski, T., Piecuch, M., Szajek, M., Bręborowicz, A., Hafke-Dys, H., Kociński, J., et al. (2019). Practical Implementation of Artificial Intelligence Algorithms in Pulmonary Auscultation Examination. *Eur. J. Pediatr.* 178, 883–890. doi:10.1007/s00431-019-03363-2

Guha, R., Ghosh, M., Chakrabarti, A., Sarkar, R., and Mirjalili, S. (2020). Introducing Clustering Based Population in Binary Gravitational Search Algorithm for Feature Selection. *Appl. Soft Comput.* 93, 106341. doi:10.1016/j.asoc.2020.106341

Han, X., Xiong, X., and Duan, F. (2015). A New Method for Image Segmentation Based on BP Neural Network and Gravitational Search Algorithm Enhanced by Cat Chaotic Mapping. *Appl. Intell.* 43, 855–873. doi:10.1007/s10489-015-0679-5

Hanchuan Peng, H., Fuhui Long, F., and Ding, C. (2005). Feature Selection Based on Mutual Information Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 1226–1238. doi:10.1109/TPAMI.2005.159

Hanoa, R., Baste, V., Kooij, A., Sommervold, L., and Moen, B. E. (2011). No Difference in Self Reported Health Among Coalminers in Two Different Shift Schedules at Spitsbergen, norway, a Two Years Follow-Up. *Ind. Health* 49, 652–657. doi:10.2486/indhealth.MS1280

Hoogendoorn, M., Szolovits, P., Moons, L. M. G., and Numans, M. E. (2016). Utilizing Uncoded Consultation Notes from Electronic Medical Records for

Predictive Modeling of Colorectal Cancer. *Artif. Intell. Med.* 69, 53–61. doi:10.1016/j.artmed.2016.03.003

Keller-Ross, M. L., Johnson, B. D., Joyner, M. J., and Olson, T. P. (2014). Influence of the Metaboreflex on Arterial Blood Pressure in Heart Failure Patients. *Am. Heart J.* 167, 521–528. doi:10.1016/j.ahj.2013.12.021

Kocsmár, É., Lotz, G., Kiss, A., Hoerner, M., Petrova, E., Freudenberg, N., et al. (2020). Prognostic Impact of Tumor Budding and EMT in Periampullary Adenocarcinoma: A Quantitative Approach. *J. Cancer* 11, 6474–6483. doi:10.7150/jca.46093

Koshimizu, H., Kojima, R., and Okuno, Y. (2020). Future Possibilities for Artificial Intelligence in the Practical Management of Hypertension. *Hypertens. Res.* 43, 1327–1337. doi:10.1038/s41440-020-0498-x

Lee, S., Lee, H., Choi, J. R., and Koh, S. B. (2020). Development and Validation of Prediction Model for Risk Reduction of Metabolic Syndrome by Body Weight Control: A Prospective Population-Based Study. *Sci. Rep.* 10, 10006. doi:10.1038/s41598-020-67238-5

Liu, Q., Meng, X., Li, X., and Luo, X. (2019). Risk Precontrol Continuum and Risk Gradient Control in Underground Coal Mining. *Process Saf. Environ. Prot.* 129, 210–219. doi:10.1016/j.psep.2019.06.031

Lu, Y., Zhang, Z., Yan, H., Rui, B., and Liu, J. (2020). Effects of Occupational Hazards on Job Stress and Mental Health of Factory Workers and Miners: A Propensity Score Analysis. *BioMed Res. Int.* 2020, 1–9. doi:10.1155/2020/1754897

Mackenzie Ross, S. (2016). Delayed Cognitive and Psychiatric Symptoms Following Methyl Iodide and Manganese Poisoning: Potential for Misdiagnosis. *Cortex* 74, 427–439. doi:10.1016/j.cortex.2015.06.031

MadhuSudana Rao, N., Kannan, K., Gao, X.-z., and Roy, D. S. (2018). Novel Classifiers for Intelligent Disease Diagnosis with Multi-Objective Parameter Evolution. *Comput. Electr. Eng.* 67, 483–496. doi:10.1016/j.compeleceng.2018.01.039

Mahanipour, A., and Nezamabadi-Pour, H. (2019). A Multiple Feature Construction Method Based on Gravitational Search Algorithm. *Expert Syst. Appl.* 127, 199–209. doi:10.1016/j.eswa.2019.03.015

Manuel Serra, J., Allen Baumes, L., Moliner, M., Serna, P., and Corma, A. (2007). Zeolite Synthesis Modelling with Support Vector Machines: a Combinatorial Approach. *Cchts* 10, 13–24. doi:10.2174/138620707779802779

Maxwell, A., Li, R., Yang, B., Weng, H., Ou, A., Hong, H., et al. (2017). Deep Learning Architectures for Multi-Label Classification of Intelligent Health Risk Prediction. *BMC Bioinforma.* 18, 523. doi:10.1186/s12859-017-1898-z

Mosa, M. A. (2019). Real-time Data Text Mining Based on Gravitational Search Algorithm. *Expert Syst. Appl.* 137, 117–129. doi:10.1016/j.eswa.2019.06.065

Mustafa, M. O., Varagnolo, D., Nikolakopoulos, G., and Gustafsson, T. (2016). Detecting Broken Rotor Bars in Induction Motors with Model-Based Support Vector Classifiers. *Control Eng. Pract.* 52, 15–23. doi:10.1016/j.conengprac.2016.03.019

Neshatian, K., Zhang, M., and Andreae, P. (2012). A Filter Approach to Multiple Feature Construction for Symbolic Learning Classifiers Using Genetic Programming. *IEEE Trans. Evol. Comput.* 16, 645–661. doi:10.1109/TEVC.2011.2166158

Opel, N., Redlich, R., Grotegerd, D., Dohm, K., Heindel, W., Kugel, H., et al. (2015). Obesity and Major Depression: Body-Mass Index (BMI) Is Associated with a Severe Course of Disease and Specific Neurostructural Alterations. *Psychoneuroendocrinology* 51, 219–226. doi:10.1016/j.psyneuen.2014.10.001

Özyurt, F. (2020). A Fused CNN Model for WBC Detection with MRMR Feature Selection and Extreme Learning Machine. *Soft Comput.* 24, 8163–8172. doi:10.1007/s00500-019-04383-8

Paul, A., Kumar, N., Kumar, P., and Singh, A. K. (2020). Application of CMRI-ISM RMR for Stability Analysis of Development Workings for Ballarpur Underground Coal Mine - an Empirical and Numerical Approach. *J. Geol. Soc. India* 96, 163–170. doi:10.1007/s12594-020-1524-y

Perret, J. L., Plush, B., Lachapelle, P., Hinks, T. S. C., Walter, C., Clarke, P., et al. (2017). Coal Mine Dust Lung Disease in the Modern Era. *Respirology* 22, 662–670. doi:10.1111/resp.13034

Pone, J. D. N., Hein, K. A. A., Stracher, G. B., Annegarn, H. J., Finkleman, R. B., Blake, D. R., et al. (2007). The Spontaneous Combustion of Coal and its By-Products in the Witbank and Sasolburg Coalfields of south africa. *Int. J. Coal Geol.* 72, 124–140. doi:10.1016/j.coal.2007.01.001

Pucciarelli, G., Greco, A., Paturzo, M., Jurgens, C. Y., Durante, A., Alvaro, R., et al. (2019). Psychometric Evaluation of the Heart Failure Somatic Perception Scale in a European Heart Failure Population. *Eur. J. Cardiovasc. Nurs.* 18, 484–491. doi:10.1177/1474515119846240

Rashedi, E., Nezamabadi-Pour, H., and Saryazdi, S. (2009). GSA: a Gravitational Search Algorithm. *Inf. Sci.* 179, 2232–2248. doi:10.1016/j.ins.2009.03.004

Takacs, B. C., Guffey, S. E., Wu, M., and Michael, K. (2015). Comparison of Noise Reduction Values for Fit Tests and Work in Coal Mines. *J. Acoust. Soc. Am.* 137, 2377. doi:10.1121/1.4920638

Teisseyre, P. (2017). CCnet: Joint Multi-Label Classification and Feature Selection Using Classifier Chains and Elastic Net Regularization. *Neurocomputing* 235, 98–111. doi:10.1016/j.neucom.2017.01.004

Volobaev, V. P., Sinitsky, M. Y., Larionov, A. V., Druzhinin, V. G., Gafarov, N. I., Minina, V. I., et al. (2016). Modifying Influence of Occupational Inflammatory Diseases on the Level of Chromosome Aberrations in Coal Miners. *Mutage* 31, 225–229. doi:10.1093/mutage/gev080

Watts, D., Moulden, H., Mamak, M., Upfold, C., Chaimowitz, G., and Kapczinski, F. (2021). Predicting Offenses Among Individuals with Psychiatric Disorders - A Machine Learning Approach. *J. Psychiatric Res.* 138, 146–154. doi:10.1016/j.jpsychires.2021.03.026

Wei, W., Hu, X.-w., Cheng, Q., Zhao, Y.-m., and Ge, Y.-q. (2020). Identification of Common and Severe COVID-19: the Value of CT Texture Analysis and Correlation with Clinical Characteristics. *Eur. Radiol.* 30, 6788–6796. doi:10.1007/s00330-020-07012-3

Wu, J.-H., Wei, W., Zhang, L., Wang, J., Damasevicius, R., Li, J., et al. (2019). Risk Assessment of Hypertension in Steel Workers Based on LVQ and Fisher-SVM Deep Excavation. *IEEE ACCESS* 7, 23109–23119. doi:10.1109/ACCESS.2019.2899625

Wu, Q., Han, L., Xu, M., Zhang, H., Ding, B., and Zhu, B. (2019). Effects of Occupational Exposure to Dust on Chest Radiograph, Pulmonary Function, Blood Pressure and Electrocardiogram Among Coal Miners in an Eastern Province, China. *BMC Public Health* 19, 1229. doi:10.1186/s12889-019-7568-5

Wu, Z., Shou, L., Wang, J., Huang, T., and Xu, X. (2020). The Methylation Pattern for Knee and Hip Osteoarthritis. *Front. Cell Dev. Biol.* 8, 602024. doi:10.3389/fcell.2020.602024

Xie, H., Liu, J., Gao, M., Liu, Y., Ma, T., Lu, Y., et al. (2020). Physical Symptoms and Mental Health Status in Deep Underground Miners. *Medicine* 99, e19294. doi:10.1097/MD.0000000000019294

Yang, M., Liu, C., Wang, X., Li, Y., Gao, H., Liu, X., et al. (2020). An Explainable Artificial Intelligence Predictor for Early Detection of Sepsis. *Crit. Care Med.* 48, e1091–e1096. doi:10.1097/CCM.0000000000004550

Zachurzok-Buczyńska, A., Klimek, K., Firek-Pedras, M., and Małecka-Tendera, E. (2011). Are Metabolic Syndrome and its Components in Obese Children Influenced by the Overweight Status or the Insulin Resistance? *Endokrynol. Pol.* 62, 102–108. doi:10.1038/nrendo.2011.18

Zhang, Y.-H., Guo, W., Zeng, T., Zhang, S., Chen, L., Gamarra, M., et al. (2021). Identification of Microbiota Biomarkers with Orthologous Gene Annotation for Type 2 Diabetes. *Front. Microbiol.* 12, 711244. doi:10.3389/fmicb.2021.711244

Zhu, A., Zhang, J., Zou, T., and Xiong, G. (2014). Associations of Blood Pressure, Glucose or Lipids with Stroke in Different Age or Gender. *Zhong Nan Da Xue Xue Bao Yi Xue Ban.* 39, 1271–1278. doi:10.11817/j.issn.1672-7347.2014.12.009

Zou, H., and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *J. R. Stat. Soc. B* 67, 301–320. doi:10.1111/j.1467-9868.2005.00503.x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.