

Article

Application of Semi-Supervised Learning Model to Coal Sample Classification

Dongming Wang ¹, Li Xu ^{2,*}, Wei Gao ², Hongwei Xia ², Ning Guo ¹ and Xiaohan Ren ^{1,*}

¹ Institute of Thermal Science and Technology, Shandong University, Jinan 250061, China; wangdm006@163.com (D.W.); 202334300@mail.sdu.edu.cn (N.G.)

² Anhui Special Equipment Inspection Institute, 45 Dalian Road, Hefei 230051, China; aseii_gaowei@163.com (W.G.); aseii_xiahongwei@163.com (H.X.)

* Correspondence: dps_1987@163.com (L.X.); renxh@sdu.edu.cn (X.R.)

Abstract: As an extremely important energy source, improving the efficiency and accuracy of coal classification is important for industrial production and pollution reduction. Laser-induced breakdown spectroscopy (LIBS) is a new technology for coal classification which has the ability to rapidly analyze coal compared with traditional coal analysis methods. In the practical application of LIBS, a large amount of labeling data is usually required, but it is quite difficult to obtain labeling data in industrial sites. In this paper, to address the problem of insufficient labeled data, a semi-supervised classification model (SGAN) based on adversarial neural network is proposed, which can utilize unlabeled data to improve the classification accuracy. The effects of labeled and unlabeled samples on the classification accuracy of the SGAN model are investigated, and the results show that the number of labeled and unlabeled samples are positively correlated, and the highest average classification accuracy that the model can achieve is 98.5%. In addition, the classification accuracies of SGAN and other models (e.g., CNN, RF) are also compared, and the results show that, with the same number of labeled samples in the three models, SGAN performs better after the number of unlabeled samples reaches a certain level, with an improvement of 0.7% and 2.5% compared to the CNN and RF models, respectively. This study provides new ideas for the application of semi-supervised learning in LIBS.

Keywords: semi-supervised learning; laser-induced breakdown spectrum; classification of coal samples; generating adversarial neural networks



Citation: Wang, D.; Xu, L.; Gao, W.; Xia, H.; Guo, N.; Ren, X. Application of Semi-Supervised Learning Model to Coal Sample Classification. *Appl. Sci.* **2024**, *14*, 1606. <https://doi.org/10.3390/app14041606>

Academic Editor: Rocco Furferi

Received: 10 January 2024

Revised: 4 February 2024

Accepted: 13 February 2024

Published: 17 February 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Coal is known as “black gold” and is one of the most important energy sources in the world, widely used in many industries such as power generation, steel smelting, chemical production, and construction. Especially in the field of electric power, coal occupies the position of one of the main sources of energy in the world. However, the current coal analysis technology mainly relies on the traditional assay analysis method, which has complicated processes, high labor costs, and long detection time, resulting in lagging coal quality data. Therefore, it is particularly urgent to develop rapid coal detection technology. Common rapid detection techniques include instantaneous gamma neutron activation analysis, X-ray fluorescence technology, inductively coupled plasma emission spectrometry, and so on. However, these techniques have a series of intractable problems, such as radiation hazards [1], limited analytical range [2], and large amount of argon gas consumption [3]. In this context, laser-induced breakdown spectroscopy (LIBS) technology emerged as an emerging rapid detection technology, which has significant advantages such as real-time detection and no need for sample pretreatment, etc. The LIBS system shows great potential for development in the field of coal analysis, and a series of analytical instruments have emerged on the market, with detection methods covering both online and offline modes. Based on this, Yao et al. [4] developed a fast coal analyzer based on laser-induced breakdown spectroscopy for coal quality analysis. The results showed that

the measurement accuracy of ash, volatile matter, fixed carbon, and calorific value met the requirements of neutron-activated coal online analyzer. Meanwhile, Yin et al. [5] also successfully developed a LIBS online inspection system, which realized automatic sampling and analysis, and could complete elemental analysis, coal quality analysis, and calorific value analysis. The relative error of elemental analysis was within 10%, while the relative error of ash analysis ranged from 2.29% to 13.47%. In addition to its application in the field of coal quality analysis, LIBS technology is also widely used in soil [6], metal [7], food [8], and other fields, showing its wide applicability in different fields.

Classification has been a topic of great interest in the field of LIBS, and people have been trying to find ways to improve the classification accuracy, and the commonly used classification algorithms for LIBS are Partial Least Squares Discriminant Analysis (PLS-DA), Random Forests (RF), Support Vector Machines (SVMs), KNN, and Neural Networks, and in addition to these classical machine learning algorithms, researchers have also proposed many special algorithms based on different principles.

Sirven et al. [9] applied the PLS-DA algorithm to the identification of rock types on the surface of Mars and found that the PLS-DA algorithm has good sensitivity in the task of classifying Martian rocks. Ma et al. [10] used the RF algorithm to classify the spectra of pulsed rocks and coal in order to separate the detritus from the coal, and improved the classification accuracy from 98.30% to 99.96%. Jin et al. [11] combined PCA with SVM to build a PCA-SVM model and optimized the model parameters using a grid search method, which ultimately achieved a classification accuracy of 98.52% on the test set. Zhang et al. [12] utilized a genetic algorithm (GA)-optimized support vector machine (SVM) to perform the classification, followed by regression prediction of each type separately, which significantly reduced the RMSECV of the test set. Cao et al. [13] used KNN algorithm to classify coal with an accuracy of 97.73%. Peng et al. [14] used K-Means to cluster coal, municipal sludge, and biomass samples, and then used SVM to further classify the biomass samples; the combined accuracy of the hybrid classification model reaches more than 98%, and it saves more time for the running time. Yang et al. [15] developed a PCA-ANN model to classify iron ore samples, and, after parameter optimization, the model achieved 99.19% classification accuracy on the test set. Zhang et al. [16] used wavelet neural network (WNN) to classify coal ash, and the results showed that WNN showed better classification performance. Cui et al. [17] used CNN, combined with multi-task regularization, proposed a transfer learning method which significantly reduced the RMSEP compared with the baseline method. Chen et al. [18] proposed a moisture-spectral intensity correction model, and established an ANN model to improve the model prediction capability. The plain Bayesian classification method is able to distinguish coal samples into different origins based on the probability distribution function. Zheng et al. [19] used this method to classify coal samples with a prediction accuracy of 96.7%. In addition to these classical machine learning algorithms, researchers have proposed many different methods, such as Piecewise Modeling [20] and Multiple-setting Spectra [21].

Based on the above discussion, it is clear that fully supervised learning has achieved good research results in LIBS. In the practical use of LIBS technology, we found that the acquisition of LIBS spectra is simple, but quantitative analysis for it requires a large amount of known data to build quantitative regression models, and the acquisition of labeled data is not easy in many domains, e.g., computer vision [22], Natural Language Processing (NLP) [23], medical image analysis [24], bioinformatics [25], the food [26], and other fields. Therefore, there is an urgent need to develop algorithms that can classify LIBS spectra in semi-supervised scenarios. Wang et al. [27] proposed a semi-supervised learning algorithm based on the KNN algorithm and the extended algorithm for the problem of classification of explosives, where a small amount of labeled data and a large amount of unlabeled data are used as the input variables, and after computation, all data will obtain their respective final labels. After testing, the accuracy of the model can reach 99.58%. Li et al. [28] proposed a semi-supervised LIBS quantitative analysis method based on least squares support vector machine (LS-SVM) co-training regression model, and

the prediction of Cr concentration in the high-alloy steel specimens showed that using the co-training technique and adding effective unlabeled specimens during the training process can effectively improve the performance of the regression model. Müller et al. [29] proposed a semi-supervised learning (SSL) classification model based on supervised linear discriminant analysis (LDA) and semi-supervised support vector machine (OC-SVM), which is able to classify known minerals and detect unknown substances in the sequence set. Yang et al. [30] proposed a semi-supervised Gene-SGAN algorithm, which dissects disease heterogeneity by jointly considering phenotypic and genetic data to confer genetic correlation between disease subtypes and associated endophenotypic features. Wang et al. [31] proposed a STFT-SACGAN-based bearing fault diagnosis method, which was validated and found to be a model with excellent fault recognition ability, and alleviated the problem of scarce labeled data.

In this paper, the semi-supervised learning model SGAN is applied to LIBS spectral classification to address the problem of the lack of labeled data that may occur in future LIBS applications. The advantage of the semi-supervised learning model is that only a small number of labeled samples, combined with a large number of unlabeled samples, can effectively improve the classification accuracy of the model. As a comparison, models such as Convolutional Neural Network (CNN) and Random Forest (RF) are also analyzed in this paper, and the comparison results show that when the number of labeled samples used for modeling is the same, an increase in the number of unlabeled samples results in an increase in the classification accuracy of the model. The results show that semi-supervised learning models are more advantageous than fully supervised learning models when there is a lack of labeled data.

2. Experimental Setup

2.1. LIBS Experimental Setup

As shown in Figure 1, the LIBS experimental system consists of a Q-switched Nd:YAG laser, a spectrometer, a laser attenuator, optics, and a computer. The laser is a Q-switched Nd:YAG laser manufactured by Beamtech Optronics with an output wavelength of 1064 nm, a pulse width of 7 ns, a spot diameter of 7 mm, an energy of 300 mJ, and a frequency of 5 Hz. The laser passes through a laser attenuator, which reduces the energy from 300 mJ to 100 mJ, and then passes through a UV-fused silica plano-convex lens with a focal length of 75.3 mm. In this way, the laser is focused on the surface of the sample, forming a coal plasma. The plasma signal was then transmitted to the spectrometer through two plano-convex lenses with a focal length of 40.1 mm each. Optimal experimental parameters were determined through parallel experimental tests, including a laser energy of 100 mJ, a Focus Depth of 2 mm (depth of focus buried in the sample), a delay time of 1.2 μ s, and a sample pressure of 40 MPa. Under these parameters, the average relative standard deviation (RSD) of the main peak of the sample spectrum was kept within 15%. The LIBS system diagram is shown in Figure 1.

2.2. Coal Samples

The 140 coal samples used in this study were from Hebei Province, China. The proximate analysis results were determined using an automatic muffle furnace. The proximate analyses were performed under dry conditions because the moisture in the coal samples is susceptible to environmental influences, while the results for ash and volatile matter did not change under dry conditions. Specific values are given in Appendix A. The sample was initially lump coal, which was processed in a pulverizer and then screened using a vibrating sieve machine to obtain coal particles with a diameter of less than 0.2 mm. Subsequently, the coal particles were placed in a powder compactor, and the pressure of the powder compactor was set at 40 MPa to compact the coal powder into a coal cake with a thickness of 1.3 mm and a diameter of 2 cm. Each briquette has 30 ablation points and each ablation point is excited 10 times. Therefore, 300 spectra were obtained for each briquette, and these 300 spectra were averaged to obtain one average spectrum representing the

current LIBS spectrum of the coal. According to the above experimental method, 4 samples were made for each coal sample, and a total of 560 average spectra were obtained after the experiment.

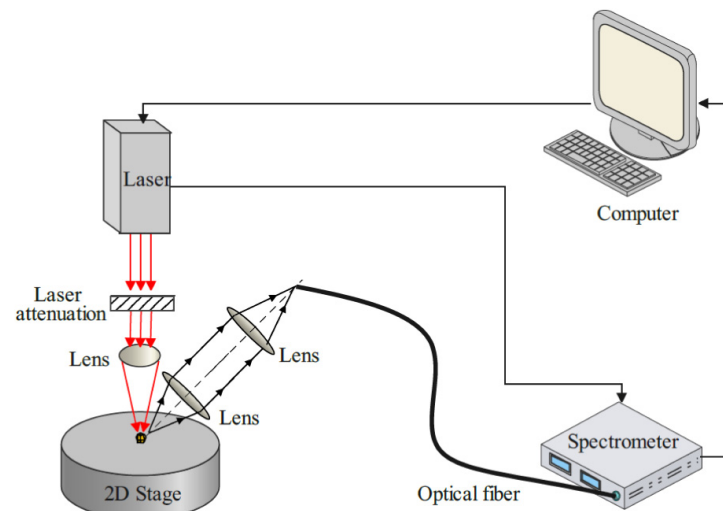


Figure 1. Diagram of the LIBS system.

3. Methods

In previous LIBS studies, fully supervised learning is the method that most researchers would use. However, considering that in practical applications, labeled data may not be simple to obtain, but unlabeled spectral data is easy to obtain. To cope with this situation, in this section, we introduce a semi-supervised learning technique based on GANs, which can combine labeled and unlabeled data to train classifiers.

3.1. Generative Adversarial Networks

Generative Adversarial Network (GAN) is an unsupervised generative model, proposed by Goodfellow et al. [32] in 2014, whose basic structure includes a Generator and a Discriminator. The Generator generates pseudo-data as realistic as possible based on the distribution of sample data, while the Discriminator is used to discriminate whether the input data are real data or pseudo-data generated by the Generator. Through the game between the generator and the discriminator, GAN can reach a Nash equilibrium, so that the generated data can fit the data distribution of the real samples. The network structure of GAN is shown in Figure 2.

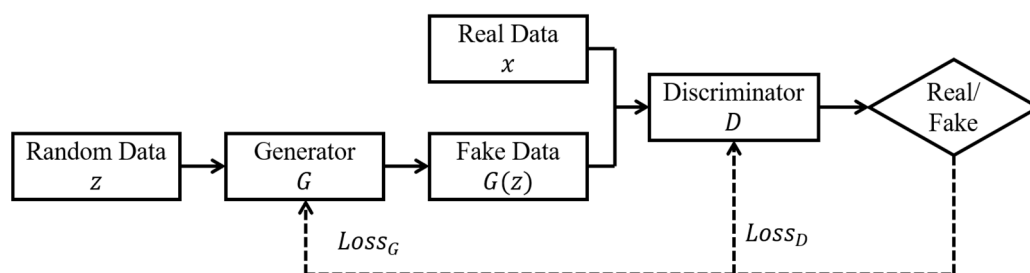


Figure 2. Network structure of GAN.

Typically, the generator G and discriminator D can be represented by a convolutional neural network or other functions. Generator G receives random noise as input for generating pseudo-data $G(z)$, while discriminator D discriminates between the input real data x and the pseudo-data $G(z)$ generated by the generator, and outputs the probability that it belongs to the real sample. The generator G and the discriminator D are trained by playing against each other through a loss function. The optimization process involves a very large very small game with an objective function of Equation (1):

$$\min_G \max_D V(D, G) = E_{x \sim P_{data}(x)} [\log D(x)] + E_{z \sim P_z(z)} [\log(1 - D(G(z)))] \quad (1)$$

where x represents the real data, P_{data} is the data distribution of x , z represents the random noise obeying the standard normal distribution, P_z only represents the data distribution of z , $G(z)$ represents the pseudo-data generated by the generator, and $D(x)$ represents the probability of the discriminator to judge that the input samples are from the real samples.

For the discriminator D , it is expected that the higher the accuracy of the discrimination, i.e., the closer $D(x)$ is to 1, the closer $D(G(z))$ is to 0, when the loss function $V(D, G)$ achieves a great value. For the generator G , it expects the generated data distribution to be closer to the real data distribution, i.e., it wants $D(G(z))$ to be closer to 1. In this case, the loss function $V(D, G)$ obtains a very small value.

When the loss function $V(D, G)$ obtains a very, very small value, the generative adversarial network reaches a Nash equilibrium, in which the generated data can fit the real data distribution. This game process promotes the discriminator to improve the accuracy, and at the same time prompts the generator to generate more realistic data, realizing the effective training of generative adversarial network.

3.2. Semi-Supervised Approach Based on GAN (SGAN)

Odena [33] proposed a semi-supervised classification model in a GAN-based framework. The model is based on a GAN network in which the image samples generated by the GAN generator G are added to the database images. For a K -category classification problem, assuming that the newly generated categories are labeled as $y = K + 1$, the dimension of the output SoftMax classifier will be extended from K to $K + 1$ accordingly. By using a semi-supervised training method that combines supervised loss and loss function of unsupervised GAN network, a large amount of unlabeled sample data will be learned to assist a small amount of labeled sample data, so as to improve the accuracy of semi-supervised classification. The SGAN network structure is shown in Figure 3.

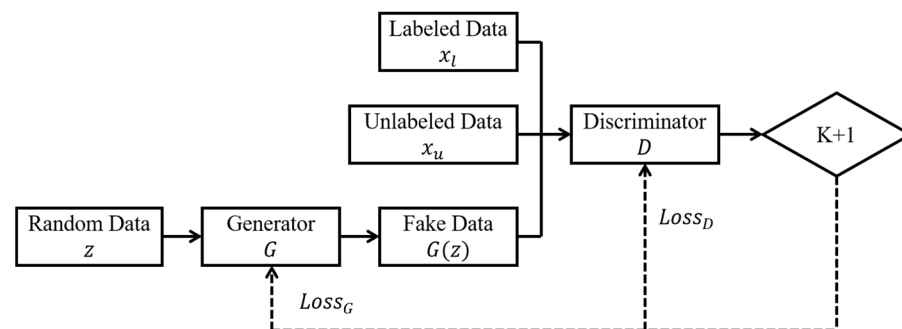


Figure 3. Network structure of SGAN.

The workflow of the model is as follows: the noise z that conforms to a specific distribution (e.g., Gaussian distribution, uniform distribution, etc.) is fed into the generator network G to generate the generated image sample $G(z)$ that conforms to the real data distribution as much as possible. Subsequently, the generated image samples $G(z)$ are fed into the discriminator network D together with the database samples, where the database sample data includes a small amount of labeled sample data and a large amount of unlabeled sample data, and the discriminator D consists of multiple convolutional layers and fully connected layers. Finally, a SoftMax output is used to characterize the normalized relative probabilities of the different classes.

At the beginning of training, both generator G and discriminator D have poor convergence. Through continuous alternate iteration training, generator G gradually fits the distribution of the database image samples and generates realistic image samples, while

the discriminator D performance of classifying and discriminating the categories of the input samples continues to improve.

4. LIBS Spectral Pretreatment

4.1. Baseline Removal

The baseline of the average spectrum is prone to drift, owing to the disturbance from the environment and equipment, which deteriorates the spectral accuracy and analysis results. Thus, baseline removal is indispensable for the improvement of the signal-to-noise ratio (SNR). It was indicated that the adaptive iterative reweighted penalty least squares algorithm (airPLS) [34] can better remove background noise. There was no user intervention and prior information, and it depended on the iterative weight change in the squared error between the fitted baseline and the original signal, which runs fast and flexibly. The effect of baseline removal is shown in Figure 4.

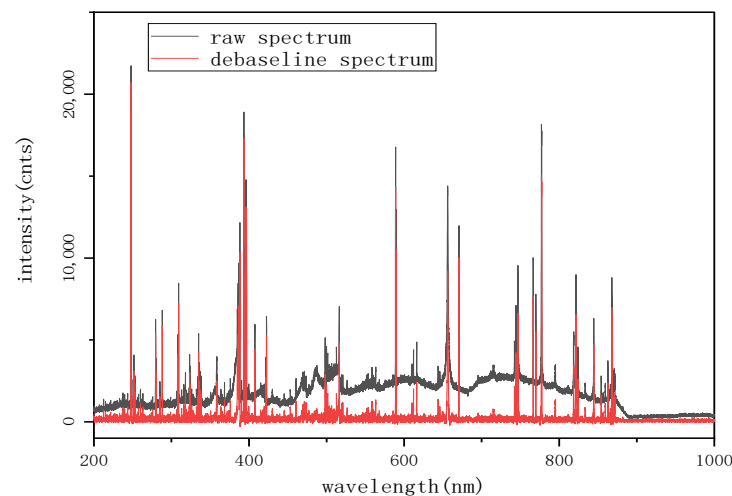


Figure 4. Effect of baseline removal on LIBS spectra.

4.2. Standardization

Spectral normalization is a data-processing technique designed to eliminate or reduce variations in spectral data to make them easier to compare and analyze. Through methods such as normalization, detrending or correction, spectral normalization ensures that spectral data obtained under different experimental conditions can be compared on the same scale. Normalization scales spectra to a standard range, detrending removes trends or drifts in the spectra, and calibration corrects errors caused by instrumental, environmental, and other factors. These processing tools help to improve the comparability of spectral data, making it more suitable for a variety of scientific research and application areas, such as quantitative analysis and modeling. Spectral standardization provides researchers with effective tools to interpret and compare spectral data more accurately and reliably. The equation is shown below:

$$x_{Z-scores} = \frac{x - \mu}{\sigma} \quad (2)$$

where $x_{Z-scores}$ is the spectral intensity after standardization; x is the spectral intensity; μ is the average spectral intensity of all spectra at the current wavelength; and σ is the standard deviation of the spectral intensity of all spectra at the current wavelength.

4.3. Evaluation Indicators

4.3.1. Clustering Model

The t-SNE method was used to cluster analyze the dataset to understand the distribution of the dataset. t-SNE (t-distributed Stochastic Neighbor Embedding) is a nonlinear dimensionality-reduction algorithm that captures the nonlinear relationships in the data better with K-Means than PCA. It performs well in visualizing high-dimensional data,

especially superior in preserving the local structure and intra-cluster similarity of the data. Meanwhile, t-SNE puts more emphasis on preserving local similarity relationships between data points in the embedding space. This makes it more superior in visualizing the clustering structure and so on. In terms of robustness, t-SNE tends to have better robustness because it is based on probability distribution in the optimization process, while the optimization process of PCA is based on variance, and the optimization process of K-Means is based on Euclidean distance. Figure 5 shows the clustering effect of t-SNE.

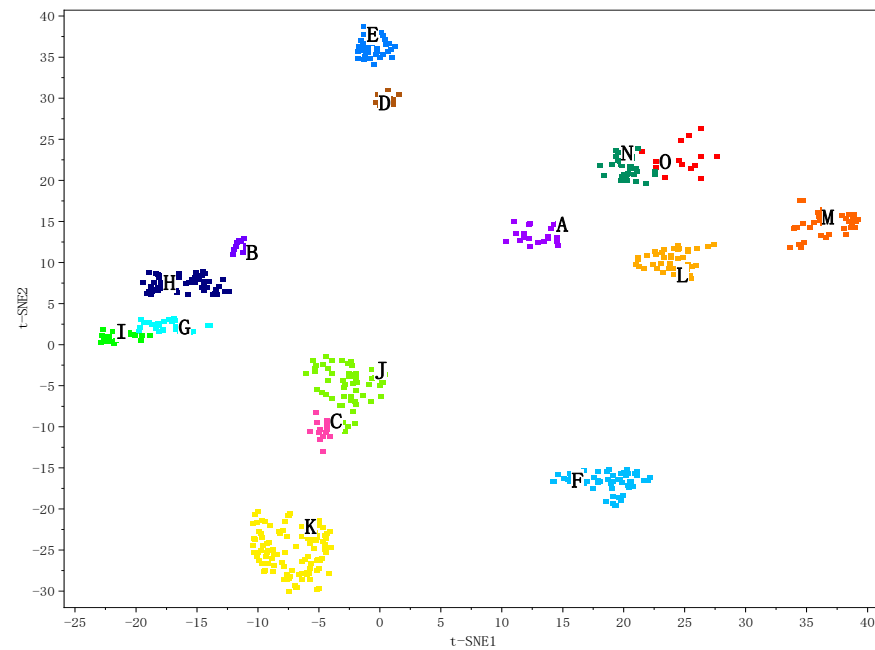


Figure 5. t-SNE calculation results. (A to O represent different types of coal samples).

The t-SNE is implemented as follows:

For each data point i , define this probability distribution as follows:

$$P(j|i) = \frac{S(x_i, x_j)}{\sum_{k \neq i} S(x_i, x_k)}, j \neq i, i = 1, 2, \dots, n \quad (3)$$

where $S(x_i, x_j)$ is the similarity between data i and data j . The closer the distance, the more similar. Suppose there are n data in total, then for these n data, we will define n probability distributions.

Similarly, suppose the data after dimensionality reduction are $z_1, z_2, \dots, z_n$, for this batch of data one can define n probability distributions as:

$$Q(j|i) = \frac{S'(z_i, z_j)}{\sum_{k \neq i} S'(z_i, z_k)}, j \neq i, i = 1, 2, \dots, n \quad (4)$$

It is important to note that different measures can be used for the similarity between high and low dimensions. The denominator is introduced here, on the one hand, to turn it into a probability distribution (the sum of probabilities is 1); on the other hand, it is because we ultimately want to compare whether they have the same distance structure before and after the dimensionality reduction, and the numerator's measure is different, so the denominator must be introduced to normalize it.

4.3.2. Confusion Matrix

The confusion matrix is a table used to evaluate the performance of a classification model, as shown in Figure 6, which demonstrates the relationship between the model's predictions on the test set and the true labels. Confusion matrices are mainly used for binary classification problems, but can be extended to multi-category classification. In the

confusion matrix, the rows represent the actual categories, and the columns A represent the categories predicted by the model. Equation (3) is the formula for the accuracy rate:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

where *TP* is the true positive result; *FP* is the false positive result; *TN* is the true negative result; and *FN* is the false negative result.

True label	A	True positive (TP)	False negative (FN)
	B	False positive (FP)	True negative (TN)
		A	B
		Predicted label	

Figure 6. Confusion matrix diagram of binary classification.

Based on these metrics, a series of classification performance metrics, such as accuracy, precision, recall, and F1 score, can be computed to more comprehensively evaluate the model's performance on different categories. Confusion matrices provide detailed insights into the performance of classification models, helping to analyze model performance across categories and identify potential problems.

5. Results and Discussion

In this section, we will perform SGAN, CNN, and RF algorithms on the spectral dataset. From the total dataset, 200 samples are selected as labeled data, 260 samples are selected as unlabeled data, and 100 data are selected as test data, which are independent of each other. For 200 labeled data, we performed 5-fold cross-validation to determine the best parameters for the model. In this experiment, we use a single CNN classifier with a similar structure to the one used in the SGAN model. The difference between the two lies in the training strategy: the single CNN classifier is trained using full supervision, while the SGAN model is trained using semi-supervision. First, the effect of the number of labeled samples and the number of unlabeled samples on the SGAN model was tested. Finally, the effects of the number of labeled samples on the CNN and RF models were also tested and analyzed in comparison with the classification effect of the SGAN model. Detailed data for this section are shown in Appendix B.

5.1. Convolutional Neural Network

Convolutional neural network (CNN) is a kind of feedforward neural network with convolutional computation and deep structure, which is one of the representative algorithms of deep learning. Convolutional neural networks have the ability of representation learning, and can classify input information according to its hierarchical structure. The convolutional neural network mimics the biological visual perception mechanism, and can carry out supervised and unsupervised learning. The parameter sharing of convolutional kernel in the hidden layer and the sparsity of inter-layer connections enable the convolutional neural network to carry out lattice features with a small amount of computation. Through previous studies, the convolutional neural network has made certain achievements in the field of LIBS.

In order to test the classification efficiency of the CNN model and compare it with the SGAN model, four sets of experiments were conducted. The number of labeled spectral data was chosen to be 100, 135, 165, and 200, respectively. As can be seen in Figure 7, the average classification accuracy of the CNN model increases from 88% to 97.8% in the prediction set.

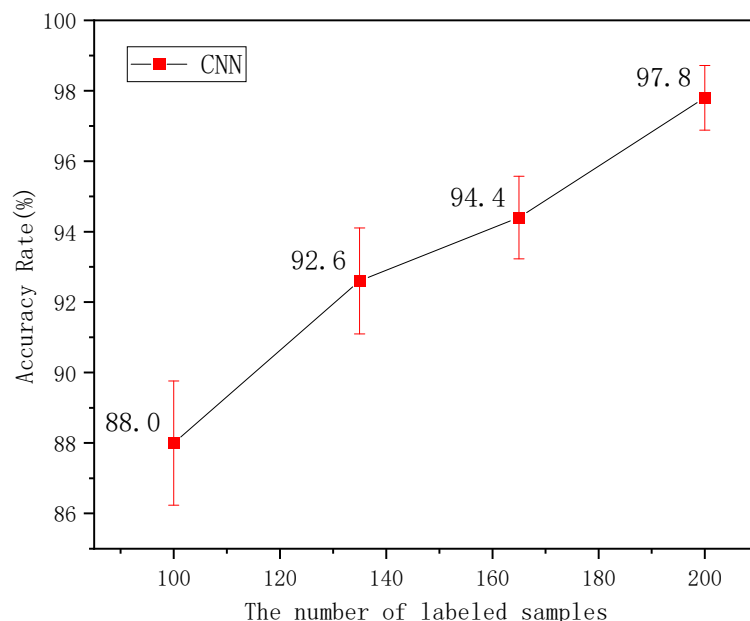


Figure 7. Effect of the number of labeled samples on the accuracy of CNN models.

In the confusion matrix diagram, blue represents correctly classified samples and red represents incorrectly classified samples, which is the same rule in this paper. From Figure 8, it can be seen that the model misclassifies 1 sample in category J as category C and 1 sample in category N as category O. The model also misclassified 1 sample in class N as class O in the prediction set. After spectral comparison of the industrial analysis components of coal, it can be found that there is a certain intersection of the above misclassified categories, with a high degree of spectral similarity, and at the same time, it is verified that their industrial analysis data C are also similar.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
A	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	12	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0
J	0	0	1	0	0	0	0	0	0	14	0	0	0	0	0
K	0	0	0	0	0	0	0	0	0	0	14	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0	0	5	1
O	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2

Figure 8. Confusion matrix plot for CNN models.

5.2. Random Forest

Random Forest (RF) is an integrated algorithm of Bagging type, which is an algorithm that obtains results by voting by combining multiple decision tree-based classifiers, including Random Forest classification (RFC) and Random Forest regression (RFR) models. The model randomly selects some features at random and generates nodes of the tree with

variance. When nodes can no longer split, a decision tree is formed and a forest with multiple decision trees is repeatedly built. Finally, classification results are obtained by voting results of all decision trees.

Before training the model, the model parameters were optimized, and 1000 decision trees were selected for modeling. A total of 4 experiments were conducted and 100, 135, 165, and 200 labeled spectra were selected as the training dataset. The results are shown in Figure 9, in which the best classification accuracy of the RF model is 96% in this experiment.

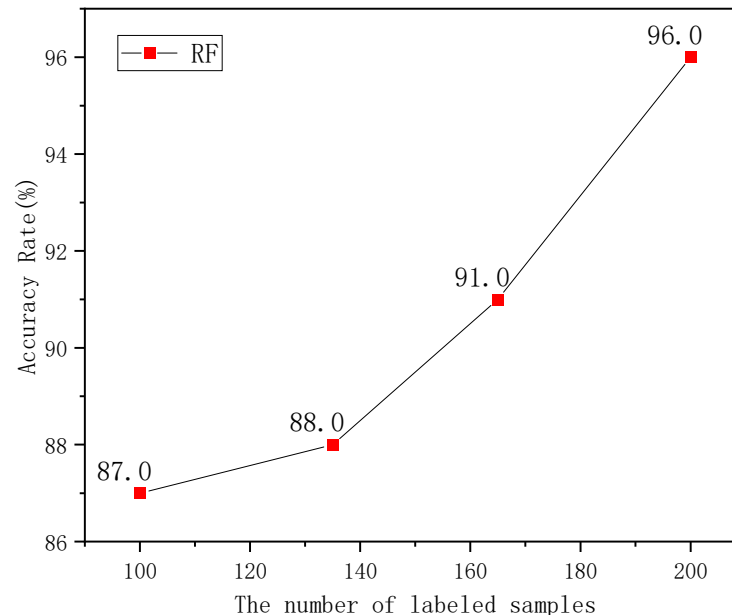


Figure 9. Effect of the number of labeled samples on the accuracy of RF models.

As can be seen in Figure 10, the model misclassified one sample in class G as class I, two samples in class J as class C, and one sample in class N as class O. The model also misclassified two samples in class J as class C, and one sample in class N as class O. After comparing the spectra of industrial analyzed constituent coals, it can be found that there is some intersection of the above misclassified categories, but compared with CNN, Random Forest underperforms in LIBS spectral classification.

5.3. Semi-Supervised GAN

The purpose of our work is to improve the performance of LIBS spectral classification model using SGAN method and to overcome the lack of labeled spectral data in the training dataset using semi-supervised approach. First, to examine the contribution of unlabeled data in model training, we selected 200 labeled samples and 100 test samples from the full dataset and conducted the following 5 experiments.

The unlabeled dataset in the first experiment contains 50 samples, the unlabeled dataset in the second experiment contains 100 samples, the unlabeled dataset in the third experiment contains 150 samples, the unlabeled dataset in the fourth experiment contains 200 samples, and the unlabeled dataset in the fifth experiment contains 260 samples. We evaluated the classification accuracy of the test set, and the results are as follows. As can be seen from Figure 11, the classification accuracy improves significantly as the number of unlabeled data increases, from 93.4% when there are only 50 unlabeled samples to 98.5% when there are 260 unlabeled samples. At the same time, it is not difficult to find that the improvement of the accuracy rate is gradually slowing down.

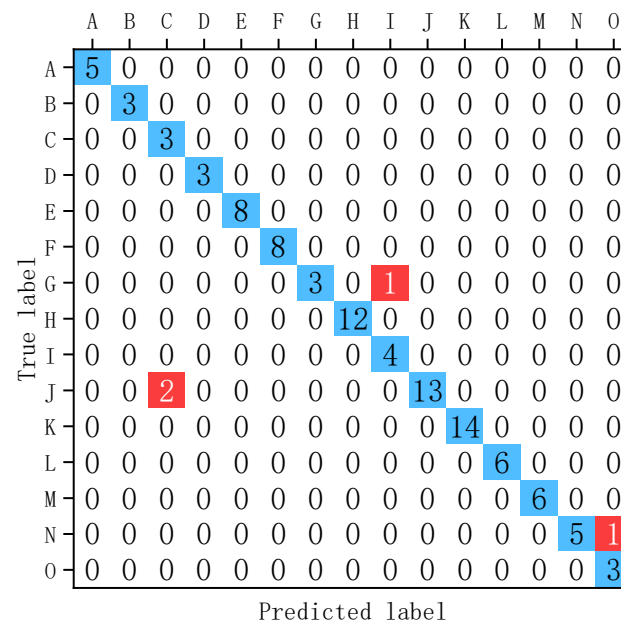


Figure 10. Confusion matrix plot for RF models.

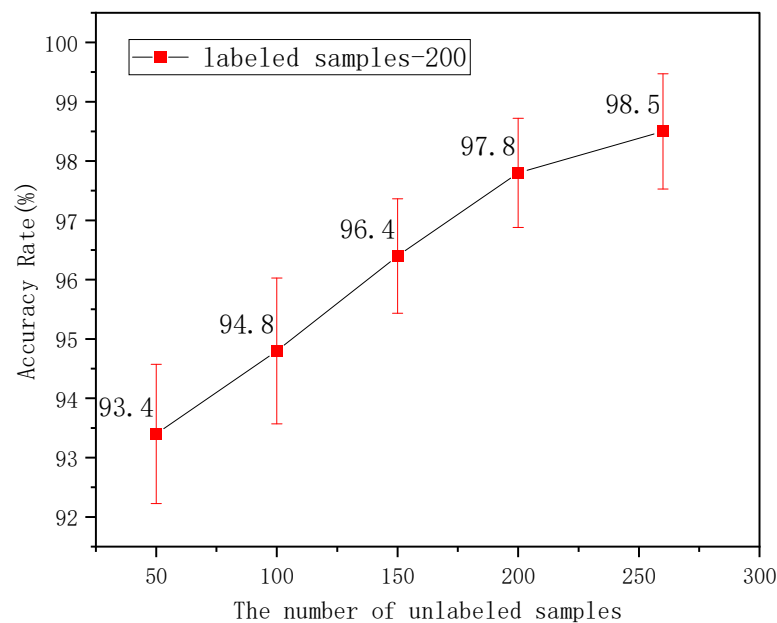


Figure 11. Effect of the number of unlabeled samples on the accuracy of SGAN models.

The most important factor in selecting the optimal number of unlabeled samples is the classification accuracy; in our expectation, there should be a turning point in the curve of classification accuracy versus the number of unlabeled samples, before which increasing the number of samples is necessary, and after which increasing the number of samples will not contribute much to the classification accuracy, but instead will use more computational resources.

After verifying the effect of the number of unlabeled samples on model training, we investigated the effect of the number of labeled samples on model training in order to better evaluate the SGAN algorithm and more easily compare it with other models. In total, we conducted four experiments. The first labeled dataset contained 100 samples, the second labeled dataset contained 135 samples, the third labeled dataset contained 165 samples, and the fourth labeled dataset contained 200 samples. As shown in Figure 12, it can be seen that the number of labeled data has a great impact on the accuracy. When

the number of unlabeled samples is 260, the classification accuracy increases from 90.6% when using 100 labeled samples to 98.5% when using 200 labeled samples, achieving a high classification accuracy.

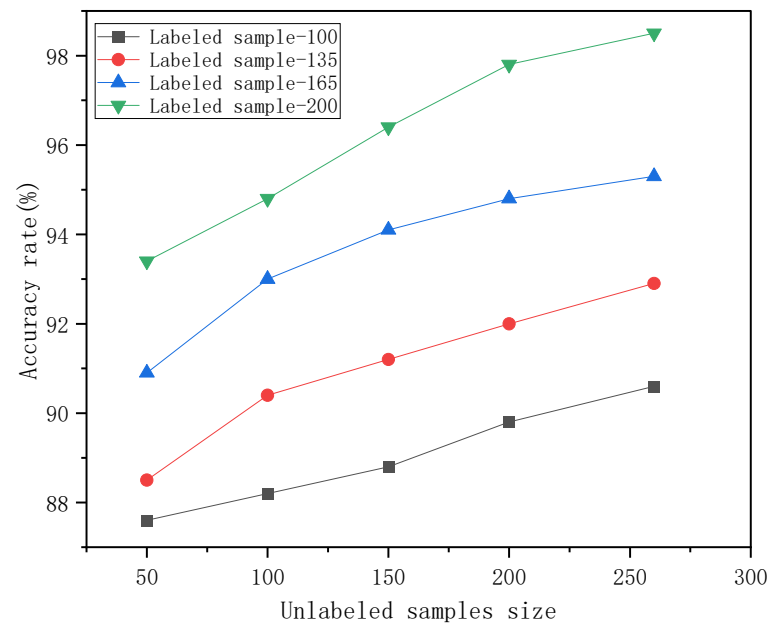


Figure 12. Effect of the number of unlabeled and labeled samples on the accuracy of SGAN models.

As can be seen in Figure 13, the model misclassifies 1 samples of class G into class H, 1 sample of class N into class M. Through spectral comparison and the t-SNE analysis mentioned above, it can be found that there is a certain intersection in the above misclassified categories, and the spectral similarity is high. Meanwhile, it has been verified that the industrial analysis data are also similar, and the impact on rough quantification is small.

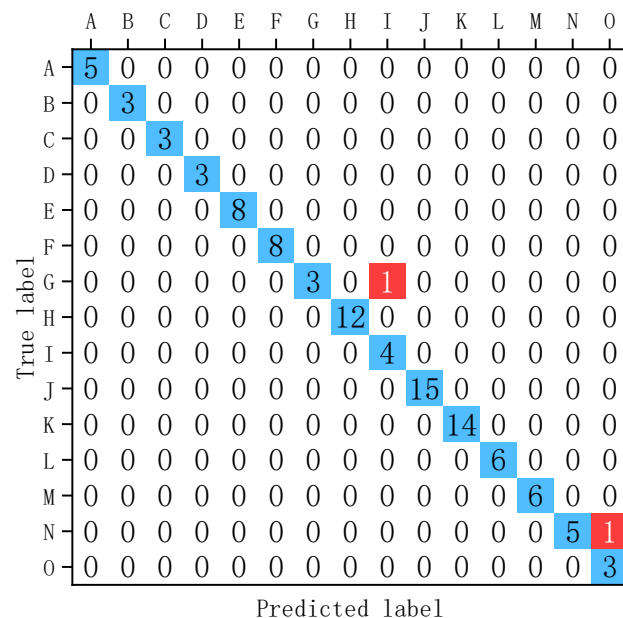


Figure 13. Confusion matrix plot for SGAN models.

5.4. Comparison

Comparing the CNN model when the labeled dataset is 200 with the SGAN model when the labeled dataset is 200, the results are shown in Figure 14, which shows that when the number of unlabeled data reach 260, the classification accuracy of the SGAN

model is obviously better than that of the CNN model; when the number of unlabeled data reach 200, the classification accuracies of the SGAN model and the CNN model are basically the same; when the number of unlabeled data are lower than 150, the classification accuracy of the CNN model is obviously better than that of the SGAN model. When the number of unlabeled data reach 200, the classification accuracy of SGAN model and CNN model is basically the same. When the number of unlabeled data are lower than 150, the classification accuracy of CNN model is significantly better than SGAN model. This also shows that semi-supervised learning is more flexible and can achieve better results when it is difficult to obtain labeled data.

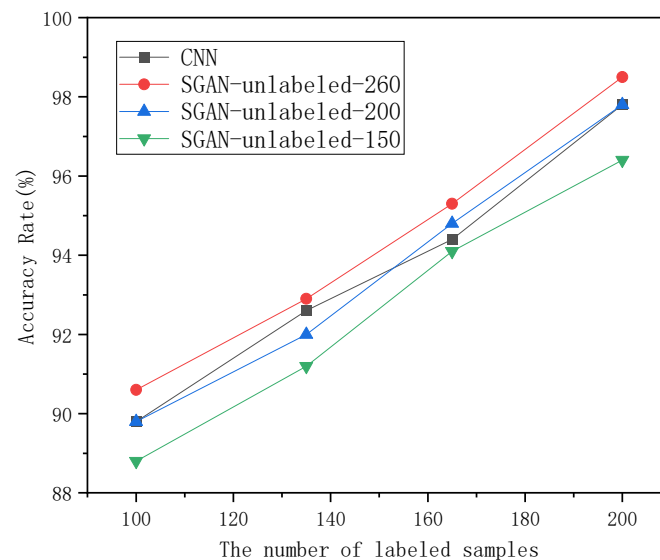


Figure 14. Comparison of classification accuracy between SGAN and CNN.

The RF model when the labeled dataset is 200 is compared with the SGAN model when the labeled dataset is 200, and the results are shown in Figure 15, which reveals that the classification accuracy of the SGAN model has a great advantage over the RF model when the number of unlabeled data reach 260, and the difference in classification accuracy between the SGAN model and the RF model is not great when the number of unlabeled data reach 100. However, the overall classification effect of the RF model is poor.

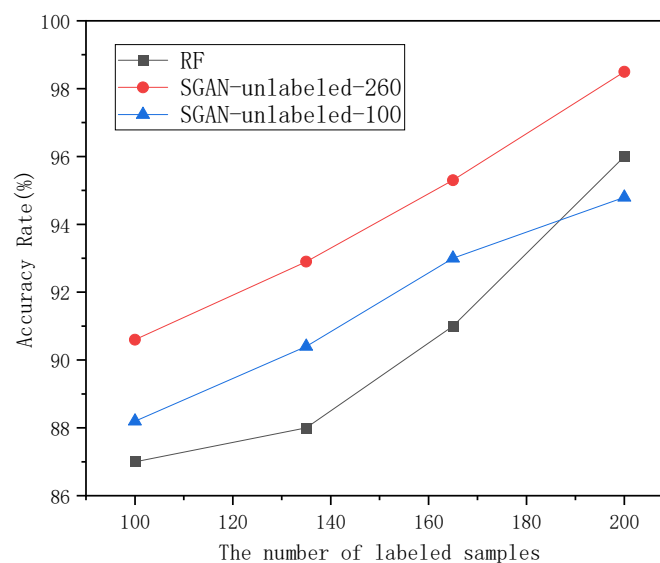


Figure 15. Comparison of classification accuracy between SGAN and RF.

6. Conclusions

In order to solve the problem of insufficient labeled data that may occur during the application of LIBS in power plants, a semi-supervised spectral classification model based on SGAN is proposed to extend the labeled dataset through the use of unlabeled data, so as to effectively improve the accuracy of semi-supervised classification. The experimental results show that under the condition of a small number of labeled samples, increasing the number of unlabeled samples can effectively improve the classification accuracy of coal samples and reduce the dependence on labeled spectral data to a certain extent. In addition, comparisons are made with two commonly used classification models (CNN, RF), and the results show that SGAN has obvious advantages, which illustrates that semi-supervised learning has great potential for development in the field of LIBS.

In our opinion, it is very important to develop semi-supervised learning models suitable for LIBS, as no matter what the domain is, one may need to deal with a large amount of data, and the biggest advantage of semi-supervised learning is its extremely efficient data utilization. Any data can be provided to the semi-supervised model after simple processing without manual labeling, which will greatly increase efficiency. Semi-supervised learning also faces challenges; for example, the samples analyzed by LIBS may be heterogeneous, and unlabeled data may contain various types of samples, for which sufficient prior knowledge needs to be provided for a wider applicability if the model is expected to have better results. Take the SGAN model used in this paper as an example; SGAN includes generator and discriminator, and the complex deep network structure of these two parts will reduce the interpretability of the model, which requires us to do a good job of cross-validation and optimization of the model structure parameters in each test. The dataset also has an impact on the model interpretability and robustness; for example, there is a category imbalance in the labeled dataset, i.e., the number of samples of certain categories is more or less, the model may produce inconsistent interpretations between different categories, the impact of category imbalance can be minimized by stratified sampling, etc. Semi-supervised learning requires more assumptions to be met, and is more demanding to implement than fully supervised learning. Any algorithm has both advantages and disadvantages, and all we have to do is to develop models that are suitable for the current application scenario.

Author Contributions: Conceptualization, D.W. and L.X.; methodology, D.W. and N.G.; validation, W.G. and H.X.; investigation, N.G. and W.G.; resources, W.G. and H.X.; writing—original draft preparation, D.W. and N.G.; writing—review and editing, X.R. and L.X.; supervision, X.R. and L.X.; funding acquisition, X.R. and H.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by [Technology Plan of State Administration for Market Regulation] grant number [2022MK060] and [Anhui Province Quality Infrastructure Standardization Special Project] grant number [2023MKS19].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Specific Values for All Coal Samples

Number	Label *	Ash d ** (%)	Volatiles d (%)	Fixed Carbon d (%)
1	A	16.63	31.5	51.87
2	A	16.47	31.59	51.94
3	A	16.2	31.91	51.89
4	A	15.42	32.08	52.5
5	A	15.42	32.08	52.5

Number	Label *	Ash d ** (%)	Volatiles d (%)	Fixed Carbon d (%)
6	B	21.38	28.11	50.51
7	B	21.55	27.88	50.57
8	B	21.98	27.47	50.55
9	C	25.26	29.47	45.27
10	C	27.1	28.74	44.16
11	C	30.24	27.66	42.1
12	C	28.48	28.06	43.46
13	D	18.23	12.41	69.36
14	D	16.12	11.69	72.19
15	D	16.93	12.58	70.49
16	D	15.85	12.26	71.89
17	E	11.32	22.13	66.55
18	E	10.75	21.58	67.67
19	E	11.38	21.97	66.65
20	E	11.72	21.58	66.7
21	E	10.85	21.57	67.58
22	E	10.57	21.98	67.45
23	E	11.04	21.83	67.13
24	E	10.91	21.94	67.15
25	E	10.82	21.48	67.7
26	F	19.55	30.06	50.39
27	F	17.39	30.51	52.1
28	F	20.7	30.64	48.66
29	F	17.82	31.05	51.13
30	F	24.66	29.01	46.33
31	F	19.5	30.71	49.79
32	F	21.47	29.58	48.95
33	F	19.21	30.3	50.49
34	F	23.16	29.53	47.31
35	F	20.26	30.29	49.45
36	F	21.83	29.61	48.56
37	F	23.53	29.72	46.75
38	F	14.44	30.92	54.64
39	F	19.29	29.88	50.83
40	G	24.38	29.38	46.24
41	G	22.09	30.48	47.43
42	G	26.16	29.29	44.55
43	G	22.89	29.43	47.68
44	G	22.39	30.2	47.41
45	G	21.58	30.23	48.19
46	G	22.93	30.02	47.05
47	G	22.96	29.7	47.34
48	H	18.97	32.43	48.6
49	H	17.86	33.21	48.93
50	H	19.25	32.5	48.25
51	H	19.66	32.4	47.94
52	H	18.96	32.64	48.4
53	H	19.75	32.28	47.97
54	H	19.31	32.59	48.1
55	H	19.73	32.66	47.61
56	H	19.08	32.19	48.73
57	H	19.99	32.06	47.95
58	H	21.07	30.84	48.09
59	H	20.09	32.29	47.62
60	H	19.35	32.29	48.36
61	I	22.96	30.24	46.8
62	I	23.03	30.47	46.5

Number	Label *	Ash d ** (%)	Volatiles d (%)	Fixed Carbon d (%)
63	I	23.37	30.07	46.56
64	I	22.42	30.49	47.09
65	I	23.69	30.23	46.08
66	J	25.04	29.86	45.1
67	J	25.03	30.41	44.56
68	J	25.16	30.74	44.1
69	J	25.56	29.5	44.94
70	J	24.08	30.26	45.66
71	J	25.8	29.98	44.22
72	J	28.22	29.9	41.88
73	J	27.02	29.57	43.41
74	J	27.94	28.11	43.95
75	J	24.98	29.93	45.09
76	J	25.46	29.67	44.87
77	J	26.04	29.44	44.52
78	J	24.64	30.35	45.01
79	J	23.81	29.9	46.29
80	J	25.98	29.9	44.12
81	K	32.56	28.90	38.54
82	K	30.03	29.06	40.91
83	K	26.56	29.02	44.42
84	K	29.79	27.73	42.48
85	K	28.72	28.37	42.91
86	K	30.31	28.15	41.54
87	K	32.07	27.52	40.41
88	K	32.07	27.55	40.38
89	K	31.52	27.12	41.36
90	K	30.85	28.26	40.89
91	K	28.43	29.24	42.33
92	K	30.96	27.63	41.41
93	K	28.36	29.12	42.52
94	K	29.70	28.24	42.06
95	K	30.23	28.49	41.28
96	K	33.64	26.67	39.69
97	K	27.53	29.59	42.88
98	K	31.53	28.11	40.36
99	K	33.94	27.42	38.64
100	K	33.88	26.62	39.50
101	K	30.53	28.72	40.75
102	K	32.18	29.26	38.56
103	K	31.63	29.44	38.93
104	K	30.61	28.95	40.44
105	K	31.53	28.30	40.17
106	K	33.59	26.79	39.62
107	K	30.16	28.24	41.60
108	K	34.83	27.10	38.07
109	L	11.29	34.71	54.00
110	L	11.96	34.64	53.40
111	L	11.34	35.06	53.60
112	L	11.85	33.89	54.26
113	L	12.22	34.39	53.39
114	L	11.90	34.62	53.48
115	L	10.45	34.73	54.82
116	L	13.10	33.89	53.01
117	L	12.54	34.01	53.45
118	L	12.56	33.80	53.64
119	M	18.71	32.12	49.17

Number	Label *	Ash d ** (%)	Volatiles d (%)	Fixed Carbon d (%)
120	M	17.66	32.28	50.06
121	M	15.92	32.90	51.18
122	M	17.89	32.38	49.73
123	M	17.76	31.82	50.42
124	M	20.63	31.35	48.02
125	M	19.66	31.44	48.90
126	M	17.37	32.56	50.07
127	M	19.87	31.49	48.64
128	M	17.78	31.51	50.71
129	N	13.96	33.57	52.47
130	N	15.03	33.50	51.47
131	N	15.72	33.25	51.03
132	N	14.31	33.75	51.94
133	N	15.42	33.18	51.40
134	N	13.73	34.18	52.09
135	N	13.86	33.88	52.26
136	N	16.34	33.12	50.54
137	O	25.16	30.26	44.58
138	O	24.80	30.37	44.83
139	O	29.09	29.14	41.77
140	O	24.40	30.41	45.19

* The label was determined by the origin of coal. ** Dry basis.

Appendix B. The Specific Values of All Sample Points

(a) All Data Points of the CNN Model					
The number of labeled samples		100	135	165	200
Accuracy rate (%)		88 ± 1.764	92.6 ± 1.506	94.4 ± 1.174	97.8 ± 0.919
(b) All data points of the RF model					
The number of labeled samples		100	135	165	200
Accuracy rate (%)		87.0	88.0	91.0	96.0
(c) All data points of the SGAN model					
Accuracy rate (%)	The number of unlabeled samples	The number of labeled samples			
		100	135	165	200
		87.6 ± 1.174	88.5 ± 1.269	90.9 ± 1.370	93.4 ± 1.174
		88.2 ± 1.229	90.4 ± 1.075	93.0 ± 1.333	94.8 ± 1.230
		88.8 ± 1.135	91.2 ± 0.919	94.1 ± 1.449	96.4 ± 0.966
		89.8 ± 1.398	92.0 ± 1.155	94.8 ± 1.135	97.8 ± 0.919
	260	90.6 ± 1.265	92.9 ± 1.287	95.3 ± 0.949	98.5 ± 0.972

References

1. Charbucinski, J.; Nichols, W. Application of spectrometric nuclear borehole logging for reserves estimation and mine planning at Callide coalfields open-cut mine. *Appl. Energy* **2003**, *74*, 313–322. [\[CrossRef\]](#)
2. Parus, J.; Kierzek, J.; Małżewska-Bućko, B. Determination of the carbon content in coal and ash by XRF. *X-ray Spectrom. Int. J.* **2000**, *29*, 192–195. [\[CrossRef\]](#)
3. Ctvrtnickova, T.; Mateo, M.P.; Yanez, A.; Nicolas, G. Application of LIBS and TMA for the determination of combustion predictive indices of coals and coal blends. *Appl. Surf. Sci.* **2011**, *257*, 5447–5451. [\[CrossRef\]](#)
4. Yao, S.; Mo, J.; Zhao, J.; Li, Y.; Zhang, X.; Lu, W.; Lu, Z. Development of a rapid coal analyzer using laser-induced breakdown spectroscopy (LIBS). *Appl. Spectrosc.* **2018**, *72*, 1225–1233. [\[CrossRef\]](#)
5. Zhang, L.; Ma, W.; Dong, L.; Yan, X.; Hu, Z.; Li, Z.; Zhang, Y.; Le, W.; Yin, W.; Jia, S. Development of an apparatus for on-line analysis of unburned carbon in fly ash using laser-induced breakdown spectroscopy (LIBS). *Appl. Spectrosc.* **2011**, *65*, 790–796. [\[CrossRef\]](#)

6. Bousquet, B.; Sirven, J.B.; Canioni, L. Towards quantitative laser-induced breakdown spectroscopy analysis of soil samples. *Spectrochim. Acta Part B At. Spectrosc.* **2007**, *62*, 1582–1589. [\[CrossRef\]](#)
7. Sabsabi, M.; Cielo, P. Quantitative analysis of aluminum alloys by laser-induced breakdown spectroscopy and plasma characterization. *Appl. Spectrosc.* **1995**, *49*, 499–507. [\[CrossRef\]](#)
8. Moncayo, S.; Manzoor, S.; Rosales, J.; Anzano, J.; Caceres, J. Qualitative and quantitative analysis of milk for the detection of adulteration by Laser Induced Breakdown Spectroscopy (LIBS). *Food Chem.* **2017**, *232*, 322–328. [\[CrossRef\]](#)
9. Sirven, J.-B.; Sallé, B.; Mauchien, P.; Lacour, J.-L.; Maurice, S.; Manhès, G. Feasibility study of rock identification at the surface of Mars by remote laser-induced breakdown spectroscopy and three chemometric methods. *J. Anal. At. Spectrom.* **2007**, *22*, 1471–1480. [\[CrossRef\]](#)
10. Ma, W.; Yu, Z.; Lu, Z.; Ma, Q.; Yao, S. A Step-By-Step Classification Method of Coal And Miscellaneous Materials By Laser-Induced Breakdown Spectroscopy. *At. Spectrosc.* **2023**, *44*, 160–168.
11. Jin, H.; Hao, X.; Yang, Y. Laser-induced breakdown spectroscopy combined with principal component analysis-based support vector machine for rapid classification of coal from different mining areas. *Optik* **2023**, *286*, 170990. [\[CrossRef\]](#)
12. Zhang, W.; Zhuo, Z.; Lu, P.; Tang, J.; Tang, H.; Lu, J.; Xing, T.; Wang, Y. LIBS analysis of the ash content, volatile matter, and calorific value in coal by partial least squares regression based on ash classification. *J. Anal. At. Spectrom.* **2020**, *35*, 1621–1631. [\[CrossRef\]](#)
13. Cao, Z.; Cheng, J.; Han, X.; Li, L.; Wang, J.; Fan, Q.; Lin, Q. Rapid classification of coal by laser-induced breakdown spectroscopy (LIBS) with K-nearest neighbor (KNN) chemometrics. *Instrum. Sci. Technol.* **2023**, *51*, 59–67. [\[CrossRef\]](#)
14. Peng, H.; Chen, G.; Chen, X.; Lu, Z.; Yao, S. Hybrid classification of coal and biomass by laser-induced breakdown spectroscopy combined with K-means and SVM. *Plasma Sci. Technol.* **2018**, *21*, 034008. [\[CrossRef\]](#)
15. Yang, Y.; Li, C.; Liu, S.; Min, H.; Yan, C.; Yang, M.; Yu, J. Classification and identification of brands of iron ores using laser-induced breakdown spectroscopy combined with principal component analysis and artificial neural networks. *Anal. Methods* **2020**, *12*, 1316–1323. [\[CrossRef\]](#)
16. Zhang, T.; Yan, C.; Qi, J.; Tang, H.; Li, H. Classification and discrimination of coal ash by laser-induced breakdown spectroscopy (LIBS) coupled with advanced chemometric methods. *J. Anal. At. Spectrom.* **2017**, *32*, 1960–1965. [\[CrossRef\]](#)
17. Cui, J.; Song, W.; Hou, Z.; Gu, W.; Wang, Z. A transferred multitask regularization convolutional neural network (TrMR-CNN) for laser-induced breakdown spectroscopy quantitative analysis. *J. Anal. At. Spectrom.* **2022**, *37*, 2059–2068. [\[CrossRef\]](#)
18. Chen, J.; Li, Q.; Liu, K.; Li, X.; Lu, B.; Li, G. Correction of moisture interference in laser-induced breakdown spectroscopy detection of coal by combining neural networks and random spectral attenuation. *J. Anal. At. Spectrom.* **2022**, *37*, 1658–1664. [\[CrossRef\]](#)
19. Zheng, Y.; Lu, Q.; Chen, A.; Liu, Y.; Ren, X. Rapid Classification and Quantification of Coal by Using Laser-Induced Breakdown Spectroscopy and Machine Learning. *Appl. Sci.* **2023**, *13*, 8158. [\[CrossRef\]](#)
20. Bai, Y.; Li, J.; Zhang, W.; Zhang, L.; Hou, J.; Zhao, Y.; Chen, F.; Wang, S.; Wang, G.; Ma, X.; et al. Accuracy enhancement of LIBS-XRF coal quality analysis through spectral intensity correction and piecewise modeling. *Front. Phys.* **2022**, *9*, 820. [\[CrossRef\]](#)
21. Song, Y.; Song, W.; Yu, X.; Afgan, M.S.; Liu, J.; Gu, W.; Hou, Z.; Wang, Z.; Li, Z.; Yan, G.; et al. Improvement of sample discrimination using laser-induced breakdown spectroscopy with multiple-setting spectra. *Anal. Chim. Acta* **2021**, *1184*, 339053. [\[CrossRef\]](#)
22. Mahadevkar, S.V.; Khemani, B.; Patil, S.; Kotecha, K.; Vora, D.R.; Abraham, A.; Gabralla, L.A. A review on machine learning styles in computer vision—Techniques and future directions. *IEEE Access* **2022**, *10*, 107293–107329. [\[CrossRef\]](#)
23. Feng, Y.; Chen, J.; Zhang, T.; He, S.; Xu, E.; Zhou, Z. Semi-supervised meta-learning networks with squeeze-and-excitation attention for few-shot fault diagnosis. *ISA Trans.* **2022**, *120*, 383–401. [\[CrossRef\]](#)
24. Chen, X.; Wang, X.; Zhang, K.; Fung, K.-M.; Thai, T.C.; Moore, K.; Mannel, R.S.; Liu, H.; Zheng, B.; Qiu, Y. Recent advances and clinical applications of deep learning in medical image analysis. *Med. Image Anal.* **2022**, *79*, 102444. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Zhong, L.; Ming, Z.; Xie, G.; Fan, C.; Piao, X. Recent advances on the semi-supervised learning for long non-coding RNA-protein interactions prediction: A review. *Protein Pept. Lett.* **2020**, *27*, 385–391. [\[CrossRef\]](#)
26. Zhang, Y.; Deng, L.; Zhu, H.; Wang, W.; Ren, Z.; Zhou, Q.; Lu, S.; Sun, S.; Zhu, Z.; Gorris, J.M.; et al. Deep Learning in Food Category Recognition. *Inf. Fusion* **2023**, *98*, 101859. [\[CrossRef\]](#)
27. Wang, Q.; Teng, G.; Li, C.; Zhao, Y.; Peng, Z. Identification and classification of explosives using semi-supervised learning and laser-induced breakdown spectroscopy. *J. Hazard. Mater.* **2019**, *369*, 423–429. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Li, X.; Lu, H.; Yang, J.; Fu, C. Semi-supervised LIBS quantitative analysis method based on co-training regression model with selection of effective unlabeled samples. *Plasma Sci. Technol.* **2018**, *21*, 034015. [\[CrossRef\]](#)
29. Müller, S.; Meima, J.A. Mineral classification of lithium-bearing pegmatites based on laser-induced breakdown spectroscopy: Application of semi-supervised learning to detect known minerals and unknown material. *Spectrochim. Acta Part B At. Spectrosc.* **2022**, *189*, 106370. [\[CrossRef\]](#)
30. Yang, Z.; Wen, J.; Abdulkadir, A.; Cui, Y.; Erus, G.; Mamourian, E.; Melhem, R.; Srinivasan, D.; Govindarajan, S.T.; Chen, J. Gene-SGAN: Discovering disease subtypes with imaging and genetic signatures via multi-view weakly-supervised deep clustering. *Nat. Commun.* **2024**, *15*, 354. [\[CrossRef\]](#)
31. Wang, H.; Zhu, H.; Li, H. Multi-Mode Data Generation and Fault Diagnosis of Bearings Based on STFT-SACGAN. *Electronics* **2023**, *12*, 1910. [\[CrossRef\]](#)

32. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems 27, Montreal, QC, Canada, 8–13 December 2014.
33. Odena, A. Semi-Supervised Learning with Generative Adversarial Networks. *arXiv* **2016**, arXiv:1606.01583.
34. Zhang, Z.M.; Chen, S.; Liang, Y.Z. Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst* **2010**, *135*, 1138–1146. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.