Discrimination of small sample tea varieties based on convolutional neural network and deep convolutional generative adversarial network enhanced near-infrared diffuse reflectance spectral dataset

Yulong Guo
Nanjing Forestry University
Zhengwei Huang
nupt_hzw@hotmail.com

Nanjing Forestry University
Yang Sheng
Nanjing Forestry University
Yan Teng
Nanjing Forestry University
Chunyang Li
Nanjing Forestry University
Chun Li
Nanjing Forestry University
Ling Jiang
Nanjing Forestry University

Research Article

Keywords: near infrared spectroscopy, data augmentation, deep convolutional generative adversarial network, tea variety identification

Posted Date: April 15th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4241593/v1

Additional Declarations: No competing interests reported.

Abstract

Near-infrared diffuse reflectance spectroscopy is widely recognized as a rapid, non-destructive, and environmentally friendly detection technology. However, in order to ensure the accuracy and stability of the detection model, a large number of sample data is necessary. This paper proposed the rapid and non-destructive detection of small sample tea variety recognition based on the near-infrared diffuse reflectance spectrum data extended by convolutional neural network (CNN) and deep convolutional generative adversarial network (DCGAN). The near-infrared diffuse reflectance spectra of 240 tea samples were obtained by Lambda 950 spectrometer using eight of the most popular tea varieties on the Chinese market. Firstly, the spectral data was enhanced using translation, linear superposition, noise addition, and DCGAN methods, and the quality of the generated spectra was evaluated using the support vector machine (SVM) and gradient boosting decision tree (GBDT) methods. Compared with other methods, the DCGAN has the highest accuracy of 91.75%. Secondly, the optimal number of iterations of DCGAN was confirmed to be 6000 by SVM and GBDT methods. To further augment the precision of identifying small samples of tea, two additional classification models of the Extreme Gradient Boosting (Xgboost) and CNN were applied to the DCGAN. Finally, the results demonstrated that the CNN achieved the highest identification accuracy of 98.68% compared with SVM (90.46%), GBDT (90.42%), and Xgboost (88.83%) with an additional 100 samples and 6000 iterations. Therefore, the combination of deep convolutional generative adversarial network enhanced near-infrared diffuse reflectance spectral dataset and the CNN successfully realizes the identification of small sample tea varieties. The experimental results strongly indicate that this method holds significant potential for practical implementation in the field of small sample tea varieties identification.

1 INTRODUCTION

Tea, originating from ancient China, is highly praised for its significant economic value and numerous health benefits, which is cherished by consumers worldwide[1]. It serves as a natural remedy for various human ailments, including enteritis, hepatitis, and heart diseases. Tea leaves are categorized based on their degree of fermentation, including unfermented (such as green tea, yellow tea, and black tea), semi-fermented (like oolong tea and white tea), and fully fermented (red tea). Each type of tea possesses distinct effects and properties. For instance, Bi Luo Chun is renowned for its anti-aging and skin whitening effects, while Longjing tea is believed to aid in preventing strokes and cardiovascular diseases. White tea is associated with stabilizing blood sugar levels, reducing glucose, blood pressure, and lipids. Green tea is known to inhibit brain aging [2], black tea can aid in weight reduction [3], and white tea possesses certain protective effects on the liver [4]. Currently, tea varietal identification relies predominantly on the integration of sensory evaluation and physicochemical methodologies [5]. Sensory analysis interprets organoleptic attributes including visual properties, infusion chromatics, aromatics, and gustation [6]. However, such human-centered techniques are intrinsically subjective, with precision potentially compromised by individual perceptual variabilities. Complementary physicochemical assessments involve analytical techniques like gas chromatography [7]and predominantly liquid chromatography [8] to substantiate intrinsic phytochemical profiles. While generating substantive compositional data, such methods possess limitations including significant economic overhead, protracted processing times, complex protocols, and requisite extensive analyst training. Recognizing the intrinsic shortcomings of sensory-centered and physicochemical identification paradigms, pursuit of rapid, non-invasive and economic means of varietal authentication seems prudent. Both authentication accuracy

and analytical throughput could potentially be optimized by harnessing adaptive algorithms or integrating emerging techniques.

Near-infrared spectroscopy offers myriad advantages befitting rapid, non-destructive tea authentication. As a high-throughput, environmentally-benign analytical technique requiring no sample preparation, NIR is well-suited for on-site assessment. These salient features support its versatility demonstrated across tea-related applications.Scholars have successfully leveraged NIR to distinguish tea cultivars [9], geographical origins [10], processed grades [11], and detect biotic stresses such as mildew infestation [12], through analyzing unique spectral profiles correlating with compositional and quality attributes. NIR discernment is founded on discerning subtle variances in phytochemical signatures among tea types, imparting taxonomic resolution. This makes NIR a powerful tool for rapid, non-destructive, and cost-effective tea analysis. Unlike direct measurement techniques used in chemical analysis, spectral analysis serves as an indirect method. It necessitates the integration with chemometric methods to construct a model that captures the correlation between the sample and its spectral data [13]. At present, the predominant chemometric methods include Support Vector Machines (SVM), Gradient Boosting Decision Tree (GBDT), Extreme Gradient Boosting (Xgboost), Random Forest (RF), among others. Due to the presence of spectral bandwidths, significant spectral overlap, and complex spectral information in the near-infrared spectra of different tea varieties, these methods face challenges in improving the accuracy of tea identification. In contrast, deep learning possesses a robust learning capacity. Specifically, CNN, characterized by local perception, weight sharing, and high operational efficiency, are widely utilized in fields such as computer vision, natural language processing [14], and speech recognition [15]. In recent years, several researchers have employed near-infrared spectroscopy in combination with CNN for the identification of various items such as macadamia nuts [16], tobacco leaves [17], peppers [18], and tea origins [19], all of which have yielded promising results.Traditional chemometrics methods and deep learning both require large amounts of spectral data to ensure model accuracy and stability. Acquiring spectral data is a process that can be time-consuming and laborious, involving steps such as sample preparation and testing. This becomes even more challenging when dealing with samples that are scarce or expensive, making it difficult to gather sufficient spectral data. Indeed, it has been concluded that data augmentation plays a role in regularization, preventing overfitting and enhancing model performance [20]. Data augmentation can effectively expand spectral datasets, typically employing methods such as panning, linear superposition, and noise addition. Generative Adversarial Networks (GAN) are a class of artificial intelligence algorithms used in unsupervised machine learning, introduced by Ian Goodfellow [21] and his team in 2014. They're widely used in the field of AI for tasks such as image synthesis, semantic image editing, style transfer, image super-resolution and classification. GANs are trained without any a priori assumptions and do not use maximum likelihood estimation. Instead, they use a unique training strategy. After forward propagation through the generative network, GANs can generate samples that approximate the true distribution of the training data. This means they can simulate or "create" new data that closely resembles the existing data [22]. In recent years, GAN have also been increasingly used to augment spectral data. Deng [23] achieved a method for the identification and classification of tomato diseases by combining computer imaging with Generative Adversarial Networks (GANs).The method of combining Raman spectroscopy with GAN for classifying marine pathogens enhances the accuracy of model classification [24]. Raman spectroscopy provides detailed information about the molecular composition of the pathogens, while GAN generate additional synthetic spectral data, augmenting the training dataset. This larger, more diverse dataset

helps the model to learn better and generalize more effectively, thus improving classification accuracy. The use of SVM in combination with GAN has proven effective in classifying foodborne pathogenic bacteria [25]. This innovative approach not only enhances the accuracy of the classification model but also makes the process more efficient and cost-effective, offering significant potential in food safety and related fields. The use of DCGAN to extract the internal features of Raman spectra of seven pharmaceuticals to generate new spectra [26]. By training the DCGAN on the Raman spectra of the pharmaceuticals, it can generate new, synthetic spectra that retain the essential features of the original data. These synthetic spectra are then compared with those obtained by traditional panning methods. The results show that the spectra generated by the DCGAN are of higher quality, indicating that the DCGAN has successfully learned and replicated the complex features of the pharmaceutical spectra. GAN have been effectively used for data augmentation in hyperspectral and Raman spectral analysis, enhancing the quality and diversity of the datasets. However, their application in NIR spectral data augmentation remains relatively unexplored. While much of the current research in the field of DCGAN is focused on improving classification accuracy, there is less emphasis on refining the DCGAN model itself or determining the optimal amount of data for generating samples. These factors are crucial as they directly impact the quality of the modeled data and, consequently, the accuracy of the classification model.

This paper introduces an innovative method of expanding the NIR spectral data of tea using DCGAN. It contrasts this approach with conventional data augmentation techniques and develops multiple models for tea variety classification. The study further investigates the usability of the generated spectra and performs a detailed analysis of the impact of the number of DCGAN iterations and the quantity of generated samples on prediction accuracy. This work could offer crucial insights into the best practices for using DCGAN for NIR spectral data augmentation, potentially improving the precision and efficiency of tea variety classification significantly.

2 Materials and methods

2.1 Sample collection and preparation

In this experiment, eight kinds of tea samples (Longjing, Yuhua, Biluochun, Jinjunmei, Tieguanyin, Xinyang Maojian, Huangshan Maofeng, and Liuan Gua Pian) were collected under certain conditions. Each kind of tea had 30 samples, so there were 240 samples in total. All tea samples were ground by a small-sized mill, and then these ground tea samples were filtered through a 100-mesh sieve. After this process, each tea sample powder (0.5 g) was pressed into a thin film for the following spectra collection.The information pertaining to the eight tea samples is displayed in Table 1.

Table 1
Information of eight kinds of tea samples

Breed	Category	Place of origin	Number of samples/PCS
Longjing	Green tea	Hangzhou, Zhejiang	30
Yuhua	Green tea	Nanjing, Jiangsu	30
Biluochun	Green tea	Suzhou, Jiangsu	30
Tieguanyin	Oolong tea	Anxi ,Fujian	30
Jinjunmei	Red tea	Wuyi Mountain, Fujian	30
Xinyang Maofeng	Green tea	Xinyang, Henan	30
Huangshan Maofeng	Green tea	Huangshan mountain, Anhui	30
Liuan Gua Pian	Green tea	Lu 'an, Anhui	30

2.2 Near infrared spectrum acquisition

The experiment utilized a PerkinElmerLambda950 UV-visible near-infrared spectrophotometer. The room temperature was maintained at approximately 25℃, and the relative humidity was kept between 45% and 50%. To ensure the stability of the instrument, it was preheated for half an hour before the test commenced. The sample was analyzed using a diffuse reflectance spectrum, which scans across a wavelength range of 800-2500nm at intervals of 1nm. To minimize measurement error, each sample was measured three times. The average of these three measurements was then taken to represent the sample's spectrum.

2.3 Traditional data enhancement methods

The traditional methods of spectral data enhancement include translation method, linear superposition method and adding noise method. The translation method involves shifting the spectral data within a small range without changing the spectrum's intensity. The linear superposition method overlays the original spectral data in a certain proportion. The adding noise method introduces random noise to the original spectrum.

2.4 Data enhancement method of DCGAN deep convolution generation adversarial network

GAN is a type of machine learning model that consists of two parts: a generator (G) and a discriminator (D). The G's role is to generate spectral data that closely resembles real data by learning the potential distribution of the real data. The D acts as a binary classifier. Its function is to determine, as accurately as possible, whether the spectral data generated by the generator is close to or deviates from the real data. The entire process necessitates that both the generator and discriminator enhance their respective abilities in data generation and discrimination. This is achieved through continuous learning and optimization, with the ultimate goal being to strike a balance between the two.

Figure 2 illustrates the fundamental structure of DCGAN. In the generator network of a DCGAN, a set of random noise data is defined as the input for the generator, $p_z(z)$ represents the probability distribution of

random noise data z, and $p_{data}\left(x\right)$ stands for the probability distribution of the real spectral data represented by x.The generator's task is to produce synthetic spectral data samples, denoted as $G\left(z\right)$.The probability distribution of these generated samples is represented as $p_G\left(z\right)$. The training goal of the G is to adjust its parameters (weights and biases) in such a way that $p_G\left(z\right)$ becomes as close as possible to$p_{data}\left(x\right)$. The discriminator network serves as a kind of "judge" that tries to distinguish between real and fake (generated) data.The input to the D can be either real spectral data x or generated spectral data $G\left(z\right)$. The output $D\left(G\left(z\right)\right)$ is a scalar value between 0 and 1, representing the discriminator's belief that the input data comes from the real data distribution $p_{data}\left(x\right)$. DCGAN, a fusion of CNN and GAN, capitalizes on the strengths of both, implementing its G and D components via CNN. This successful amalgamation addresses and effectively resolves the instability issues that plagued the training process in the original GAN network.

Figure 3 presents a designed 1D Convolutional Neural Network (1D-CNN) incorporated with a Deep Convolutional Generative Adversarial Network (DCGAN). The noise input to the generator G abides by a Gaussian distribution and has a dimension of 1x1x100. The said input is then subjected to four successive deconvolution operations and gets transformed into an output with a dimension of 1x1x1701. The input to the D in the DCGAN consists of 1x1x1701 dimensional tea spectral data (that is, normalized 1x1x1701 dimensional data) and the spectral data generated by the G. The output of the discriminator is a probability value indicating how well it can distinguish between real and generated data. Table 2 illustrates the specific parameters and the structure of the established DCGAN. Batch normalization is implemented in both the generator and the discriminator, which accelerates the training process and enhances the stability of model training by standardizing the inputs for each unit. In the DCGAN's convolutional network, the activation function is chosen as LeakyReLU with a slope value of 0.2. The network's learning rate is set at 0.01, and the optimizer for the convolutional layer is Adam.

Table 2
DCGAN structure and parameter settings

	No.	Controls	Input number	Output number	Kernel Size	Stride	Padding	Whether it is normalized	Activation function
G	1	Conv-trans1	nz	256	(1,28)	1	0	yes	relu
	2	Conv-trans2	256	128	(1,4)	2	1	yes	relu
	3	Conv-trans3	128	64	(1,4)	2	1	yes	relu
	4	Conv-trans4	64	1	(1,4)	2	1	no	tanh
D	5	Conv1	1	64	(1,4)	2	1	no	leakyrelu
	6	Conv2	64	128	(1,4)	2	1	yes	leakyrelu
	7	Conv3	128	256	(1,4)	2	1	yes	leakyrelu
	8	Conv4	256	1	(1,28)	1	0	no	sigmoid

2.5 Classification model

The learning strategy of the SVM algorithm, in the case of linear separability, is to maximize the margin between the closest points (support vectors) and the separating hyperplane. This strategy aims to find the classification hyperplane that has the maximum distance to the nearest training data points of any class, thereby ensuring the best possible generalization. The SVM algorithm transcends the limitations associated with conventional machine learning algorithms such as overfitting, the curse of dimensionality, and falling into local minima, boasting robust learning and superior generalization capabilities [27].

GBDT is an ensemble learning algorithm that combines multiple decision trees to improve the model's performance. Gradient Boosting is an algorithm under the umbrella of ensemble methods known as boosting. It iteratively improves the model by applying the principle of gradient descent [28]. Xgboost is indeed an efficient implementation of the GBDT algorithm. Notably, Xgboost enhances the standard GBDT method by introducing regularization terms in the loss function that effectively control the model's complexity, preventing overfitting and ultimately boosting its prediction capacity [29]. The node splitting approach of GBDT involves traversing all possible divisions of all features, and then selecting the optimal one for splitting. On the other hand, Xgboost employs a strategy known as the histogram and quantile sketch for approximate calculations. This method effectively reduces the computational load, making Xgboost more efficient without compromising the model's predictive power [30].

The CNN stands as one of the most classic models within the realm of deep learning. The fundamental structure of a CNN typically embodies three key parts: a convolutional layer, a pooling layer, and a fully connected layer. Convolution possesses the characteristic of "weight sharing", which serves to decrease the quantity of parameters, thereby preventing the model from overfitting. Pooling, on the other hand, is a nonlinear downsampling technique primarily aimed at reducing the dimensionality of the convolution layer's output eigenvalues, thus diminishing the computational scale [31]. Currently, pooling methods are primarily categorized into two types: maximum pooling and average pooling. In this particular study, the maximum pooling technique is employed to downsample the model.

3 Results and discussion

3.1 Near infrared spectrum of tea

In this study,there are several prominent absorption peaks within the ranges of 1400 to 1500 nm, 1900 to 2000 nm, and 2200 to 2400 nm. In conjunction with an analysis of the chemical components in tea, these absorption peaks correspond to the vibrational sequences of amino acids (R-NH), tea polyphenols (= C-H), and caffeine (-OH) compounds in the tea polyphenol group [32]. The NIR absorption peaks in the spectra of the eight tea varieties are essentially at the same position. However, the spectral data of some tea types exhibit overlapping and crossing characteristics. Therefore, employing a chemometric method is essential for accurately distinguishing between these tea varieties.

3.2 Spectral data enhancement results

Figure 4 illustrates a comparison of original spectral data from randomly selected Tieguanyin tea samples, both before and after the application of various data enhancement methods. Figure 4(a) presents the spectral data derived from the translation method, achieved by shifting the horizontal coordinate of the original spectrum randomly between 1 and 5 nanometers. Figure 4(b) illustrates the generation of spectral data using the linear superposition method, which involves summing the spectral data of two randomly selected samples and subsequently dividing the result proportionally. Figure 4(c) illustrates the spectral data derived using the noise addition method, a process that involves introducing Gaussian white noise ranging from 1 to 20dB. Meanwhile, Fig. 4(d) presents the spectral data produced by the DCGAN after undergoing 2000 iterations. The four methods mentioned above all generate data that closely resembles the original spectral data. However, it's challenging to identify the best data augmentation method through a simple visual comparison. Therefore, this study evaluates the quality of the generated spectra by integrating them with specific classification models.

Table 3
Classification results of SVM and GBDT models in different enhancement methods

Model	Real data	Fake data	Translation method	Linear Superposition Method	Noise method	DCGAN
	24	0	67.45%	67.45%	67.45%	67.45%
	24	20	69.72%	70.58%	72.37%	71.36%
SVM	24	40	81.26%	83.45%	84.52%	88.43%
	24	80	76.45%	81.36%	82.67%	85.28%
	24	160	75.42%	80.62%	80.45%	86.13%
	24	0	68.75%	68.75%	68.75%	68.75%
	24	20	70.47%	73.67%	72.65%	74.52%
GBDT	24	40	75.28%	80.37%	81.26%	82.67%
	24	80	78.85%	86.64%	83.54%	91.75%
	24	160	78.64%	85.73%	82.64%	88.45%

For each type of tea, 24 spectral data points were randomly chosen. The amount of generated sample data was then incrementally increased to collectively form the training set. The ratio between the training set and the test set was kept at 4:1. The classification accuracy was verified using 5-fold cross-validation. Table 3 portrays the classification outcomes attained by the SVM and GBDT models when applied to data enriched with four distinct methods. This table offers a visually discernible analysis of how these four data enhancement techniques affect the accuracy of the model's classification. When 40 data points were generated, the SVM model performed the best. The accuracy of the data set, when enhanced by the translation method, improved from 67.45–81.26%. The data set enhanced by the linear superposition method saw an increase in accuracy to 83.45%. The data set enhanced by the noise method saw its accuracy improve to 84.52%. Finally, the data set enhanced by the DCGAN method saw the most significant improvement, with its accuracy rising to 88.43%, which was better than the other three methods. Comparable outcomes were

observed with the GBDT model. When the dataset was expanded to 80 data points, the accuracy of the dataset enhanced by the DCGAN method reached 91.75%. In summary, when compared to the three data enhancement methods - the horizontal displacement method, the linear superposition method, and the noise addition method - the data generated by DCGAN proved to be the most accurate for identifying different types of tea. Therefore, DCGAN can be effectively used to expand the dataset of tea NIR spectral data.

3.3 Determination of DCGAN iterations

In pursuit of ascertaining the optimal iteration count for the DCGAN, a random set of original spectral data from Tieguanyin tea leaves was chosen. This dataset underwent iterative processes at varied magnitudes—500, 2000, 4000, 6000, 8000, and 10000 iterations respectively. Subsequently, the quality of the spectra was evaluated employing two models: SVM and GBDT. The training set for each category consisted of 24 real data instances and 40 generated data instances. The proportion between the training set and the test set was set at 4:1. The classification accuracy was verified using 5-fold cross-validation. Table 4 clearly illustrate that the number of DCGAN iterations used to generate spectral data significantly influences the accuracy of both SVM and GBDT models. Notably, when the DCGAN iteration count is set to 6000, the performance of both models surpasses that observed at other iteration counts. The sentence could be refined as follows: "This is due to the fact that, when the number of iterations is fewer than 6000, the generated spectrum incrementally approximates the real spectrum, thereby enhancing the model's accuracy. When the number of iterations exceeds 6000, the generated spectral data, constrained by the limited real spectral data, begins to diverge from the real spectrum, leading to a decrease in accuracy. As a result, the spectral data generated in this study employs a 6000 iteration method.

Table 4
Classification results of data generated by SVM and GBDT models at different epoch

Number of iterations	Real data	Fake data	SVM	GBDT
500	24	40	75.63%	73.56%
2000	24	40	88.43%	82.67%
4000	24	40	89.26%	83.84%
6000	24	40	90.46%	85.28%
8000	24	40	89.76%	84.75%
10000	24	40	88.86%	83.45%

3.4 Construction and training of CNN classification model

The accuracy of the model increases as the generated spectrum progressively approximates the real spectrum with an increase in the number of iterations, particularly when the iterations are less than 6000. However, when the iterations reach 6000, the accuracy of SVM and GBDT models in identifying tea varieties plateaus at around 90%. This falls short of practical requirements, prompting the adoption of a convolutional neural network model for more effective tea variety identification. A seven-layer convolutional neural network is constructed in this study, comprising an input layer, three convolutional layers, three pooling layers, two fully

connected layers, and an output layer. The specific parameters and structure are presented in Table 5 and Fig. 5. The convolutional kernel size is 1x3x1, and there is no zero-padding on the edges. The training and optimization of convolutional neural networks indeed hinge on a loss function. This function quantifies the discrepancy between the predicted and actual values. The error is then backpropagated from the final layer to the preceding layers of the network using a backpropagation algorithm. This process allows for the updating of the weights, thereby refining the network's performance over time. In the training process, the updated parameters are continually utilized, and this cycle repeats until the value of the loss function reaches its minimum. This signifies the achievement of the final training objective. The Adam optimizer is employed during the training process to determine the most effective direction for gradient descent, thereby facilitating faster convergence of the model. To counteract overfitting, a learning rate of 0.01 was established, a 20% dropout rate was applied to the fully connected layer, and cross-entropy was chosen as the loss function.

Table 5
CNN model parameter settings

Network layer	Model parameters
Input Layer	Near-infrared spectroscopy data of eight types of tea leaves
Convolution Layer Conv_1	Convolution kernel size 1x3x1, stride of 1, no zero padding at the edges, and 32 feature maps
Pool Layer MP_1	Max pooling, size of 1x2x32, number of feature maps 32, stride of 2
Convolution Layer Conv_2	Convolution kernel size 1x3x1, stride of 1, no zero padding at the edges, number of feature maps 64
Pool Layer MP_2	Max pooling, size of 1x2x64, number of feature maps 64, stride of 2
Convolution Layer Conv_3	Convolution kernel size 1x3x1, stride of 1, no zero padding at the edges, number of feature maps 128
Pool Layer MP_3	Max pooling, size of 1x2x128, number of feature maps 128, stride of 2
Fully Connected Layer FC_1	256 neurons
Fully Connected Layer FC_2	512 neurons
Output Layer	Probability values for distinguishing each variety

3.5 Effect of the number of generated samples on the identification rate of tea varieties

As evidenced in Table 3, the quantity of samples generated by various models differs when reaching peak discrimination accuracy. This necessitates an investigation into the optimal number of samples each model should generate. Based on the actual sample data collected, 1000 data points were generated from the

original spectrum of each variety through 6000 iterations of DCGAN. Prior to modeling, 24 spectral data points were randomly selected from each tea variety. The sample data volume was incrementally increased to form a combined training set. The ratio of the training set to the test set was consistently maintained at 4:1. The classification accuracy was then verified using 5-fold cross-validation.

Table 6
The prediction results of the training set with different number of samples

Real data	Fake data	SVM	GBDT	Xgboost	CNN
24	0	67.45%	68.75%	65.6%	73.36%
24	10	68.73%	68.27%	66.72%	75.67%
24	20	72.45%	73.62%	73.48%	83.56%
24	30	76.72%	78.26%	78.54%	88.93%
24	40	90.46%	85.28%	88.83%	93.45%
24	50	87.25%	88.43%	86.72%	97.62%
24	100	85.25%	90.42%	84.38%	98.68%
24	200	84.89%	87.72%	85.27%	96.72%
24	300	83.56%	85.63%	85.92%	95.45%
24	500	78.43%	80.44%	83.46%	95.21%
24	800	75.52%	76.52%	81.28%	95.67%

Table 6 presents the predictive outcomes of test sets with varying quantities of generated samples. For different classification models, dataset augmentation can enhance the model's accuracy. When the quantity of added samples for each tea variety is fewer than 40, the accuracy of both SVM and Xgboost models progressively improves with the increment in sample size. However, when the sample size exceeds 50, the accuracy of the SVM and Xgboost models begins to deteriorate. The GBDT and CNN models attained peak accuracy when the number of samples generated for each variety reached 100. GBDT achieved an accuracy of 90.42%, while the CNN model outperformed with a classification accuracy of 98.68%, surpassing the other three classification models. However, when the number of generated samples exceeded 100, there was a gradual decline in classification accuracy. In conclusion, as the number of samples generated by DCGAN increased, the accuracy of the test set correspondingly rose, peaking before starting to decline. This can be attributed to the fact that the diversity of samples did not increase with the rise in the number of generated samples. The generation of excessive duplicate samples, which did not significantly contribute to enhancing the model's accuracy, was a contributing factor to this decline. Simultaneously adding an excessive number of synthetic samples can escalate the conflict and interference within the data information, thereby diminishing the model's performance. Hence, integrating DCGAN with NIR spectral data to augment the sample data of eight tea varieties can markedly enhance the accuracy of the tea variety identification model. It is crucial to select an appropriate number of samples based on the specific classification models.

4 Conclusion

This study explores the impact of various data augmentation techniques on the identification outcomes of tea varieties. By juxtaposing DCGAN with traditional methods such as translation, linear superposition, and noise addition, it is discerned that DCGAN yields superior results compared to conventional data augmentation methods. The spectral data produced by DCGAN is primarily influenced by two parameters. The first parameter is the number of iterations. Research indicates that as the number of iterations increases, the model's classification accuracy follows a trend of initially increasing and then decreasing. The prediction accuracy peaks when the number of iterations reaches 6000, effectively preventing the data generated by excessive iterations from deviating from the actual data. The second parameter is the quantity of generated samples. An expanded dataset can significantly enhance the model's accuracy, with the CNN model demonstrating the highest classification accuracy, and the model's accuracy surges from 73.36–98.68%. As the number of samples generated by the training set increases, the test set's accuracy gradually rises, then begins to decline after reaching its apex. This is attributed to the fact that the diversity of samples did not proportionately increase with the number of generated samples, resulting in the generation of too many duplicate samples, which did not significantly contribute to improving the model's accuracy. Simultaneously, the addition of too many synthetic samples can increase conflict and interference in the data information, which can degrade the model's performance. In conclusion, the expansion of NIR spectral data utilizing the DCGAN model can significantly enhance the accuracy of tea identification. However, it's crucial to select the appropriate number of iterations and the volume of sample data tailored to different classification models. This approach can offer a fresh benchmark for fellow researchers, address the issue of limited dataset availability, and curtail the expenditure of human and material resources. Moreover, DCGAN can be synergistically utilized with a diverse array of spectral technologies, including Raman spectroscopy, laser-induced breakdown spectroscopy, and terahertz time-domain spectroscopy, for the detection of various substances. This collaboration could significantly advance the industrial application of spectral technology.

Declarations

Funding information

National Natural Science Foundation of China, Grant/Award Number: 62001235, 12273012;

Conflicts of interest

There are no conlicts of interest to declare.

Acknowledgements

We would like to thank the editors and reviewers for their valuable opinions and suggestions that improved this research. We also acknowledge the Advanced Analysis and Testing Center of Nanjing Forestry University for this work.

Author Contribution

G:Conceptualization, Methodology, Software, Investigation, Formal Analysis, Writing - Original Draft;S: Data Curation, Writing - Original Draft;T: Visualization, Investigation;L:Resources, Supervision;L:Software, ValidationJ: Visualization, Writing - Review & Editing,Funding Acquisition, ResourcesH:Conceptualization, Supervision, Writing - Review & Editing.All authors reviewed the manuscript

References

1. Xiao L, Ting Z, Yun W et al. (2019) Application of Near-infrared Spectroscopy in Tea: Research Progress. Chinese Agricultural Science Bulletin 35:80-84
2. Unno K, Nakamura Y (2021) Green Tea Suppresses Brain Aging. Molecules 26
3. Pan H, Gao Y, Tu Y (2016) Mechanisms of Body Weight Reduction by Black Tea Polyphenols. Molecules 21
4. Hamdy SM, El-Khayat Z, Farrag AR et al. (2020) Hepatoprotective effect of Raspberry ketone and white tea against acrylamide-induced toxicity in rats. Drug and Chemical Toxicology 45:722-730
5. Li C, Zong B, Guo H et al. (2020) Discrimination of white teas produced from fresh leaves with different maturity by near-infrared spectroscopy. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 227
6. Xiao-Li L, Yong H, Zheng-Jun Q (2007) Application PCA-ANN Method to Fast Discrimination of Tea Varieties Using Visible/Near Infrared Spectroscopy. Spectroscopy and Spectral Analysis:279-282
7. Shuyan W, Feng Z, Genghui R et al. (2021) Origin Difference Analysis of Aroma Components in Jasmine Tea Based on Electronic Nose and ATD-GC-MS. Science and Technology of Food Industry :234-239
8. Fu-Hua W (2018) Analysis and Determination of Free Amino Acids in Different Tea by HPLC. Food Research and Development 39:141-146
9. Pao L, Ru-Jia S, Shang-Ke L et al. (2019) Nondestructive Identification of Green Tea Based on Near InfraredSpectroscopy and Chemometric Methods. Spectroscopy and Spectral Analysis 39:2584-2589
10. Jin G, Xu Y, Cui C et al. (2022) Rapid identification of the geographic origin of Taiping Houkui green tea using near-infrared spectroscopy combined with a variable selection method. Journal of the Science of Food and Agriculture 102:6123-6130
11. Ding Y, Yan Y, Li J et al. (2022) Classification of Tea Quality Levels Using Near-Infrared Spectroscopy Based on CLPSO-SVM. Foods 11
12. Ren G, Sun Y, Li M et al. (2020) Cognitive spectroscopy for evaluating Chinese black tea grades (Camellia sinensis): near-infrared spectroscopy and evolutionary algorithms. Journal of the Science of Food and Agriculture 100:3950-3959
13. Pires FDC, Pereira RGFA, Baqueta MR et al. (2021) Near-infrared spectroscopy and multivariate calibration as an alternative to the Agtron to predict roasting degrees in coffee beans and ground coffees. Food Chemistry 365
14. Yu Z, Kai-Feng L, Quan-Xin Z et al. (2021) A Combined-Convolutional Neural Network for Chinese News Text Classification. Acta Electronica Sinica 49:1059-1067
15. Kui H, Pan J, Zong R et al. (2021) Heart sound classification based on log Mel-frequency spectral coefficients features and convolutional neural networks. Biomedical Signal Processing and Control 69

16. Jian D, Bing-Liang H, Yong-Zheng L et al. (2018) Study on Quality Identification of Macadamia nut Based on Convolutional Neural Networks and Spectral Features. Spectroscopy and Spectral Analysis 38:1514-1519
17. Meng-Yao L, Kai Y, Peng-Fei S et al. (2018) The Study of Classification Modeling Method for Near Infrared Spectroscopyof Tobacco Leaves Based on Convolution Neural Network. Spectroscopy and Spectral Analysis 38:3724-3728
18. Xi-Yu W, Shi-Ping Z, Hua H et al. (2018) Near Infrared Spectroscopy for Determination of the Geographical Originof Huajiao. Spectroscopy and Spectral Analysis 38:68-72
19. Qi C, Tian-Hong P, Yu-Qiang L et al. (2021) Geographical Origin Discrimination of Taiping Houkui Tea Using Convolutional Neural Network and Near-Infrared Spectroscopy. Spectroscopy and Spectral Analysis 41:2776-2781
20. Gao X, Deng F, Yue X (2020) Data augmentation in fault diagnosis based on the Wasserstein generative adversarial network with gradient penalty. Neurocomputing 396:487-494
21. Cai Z, Xiong Z, Xu H et al. (2021) Generative Adversarial Networks. ACM Computing Surveys 54:1-38
22. Besombes C, Pannekoucke O, Lapeyre C et al.
23. Deng H, Luo D, Chang Z et al. (2021) RAHC_GAN: A Data Augmentation Method for Tomato Leaf Disease Recognition. Symmetry 13
24. Wu M, Wang S, Pan S et al. (2021) Deep learning data augmentation for Raman spectroscopy cancer tissue classification. Scientific Reports 11
25. Du Y, Han D, Liu S et al. (2022) Raman spectroscopy-based adversarial network combined with SVM for detection of foodborne pathogenic bacteria. Talanta 237
26. Ling-Qiao L, Yan-Hui L, Lin-Lin Y et al. (2021) Data Augmentation of Raman Spectral and Its Application Research Basedon DCGAN. Spectroscopy and Spectral Analysis 41:400-407
27. Kanwal A, Mehmood T, Butt MM (2021) PLS and kernel SVM based hybrid classifier for discriminating FTIR spectrum data with limited sample size. Chemometrics and Intelligent Laboratory Systems 215
28. Rong G, Alu S, Li K et al. (2020) Rainfall Induced Landslide Susceptibility Mapping Based on Bayesian Optimized Random Forest and Gradient Boosting Decision Tree Models—A Case Study of Shuicheng County, China. Water 12
29. Liang W, Luo S, Zhao G et al. (2020) Predicting Hard Rock Pillar Stability Using GBDT, XGBoost, and LightGBM Algorithms. Mathematics 8
30. Sagi O, Rokach L (2021) Approximating XGBoost with an interpretable decision tree. Information Sciences 572:522-542
31. Wu Z, Jiang H, Liu S et al. (2021) A deep ensemble dense convolutional neural network for rolling bearing fault diagnosis. Measurement Science and Technology 32
32. Hazarika AK, Chanda S, Sabhapondit S et al. (2018) Quality assessment of fresh tea leaves by estimating total polyphenols using near infrared spectroscopy. Journal of Food Science and Technology 55:4867-4876

Figures

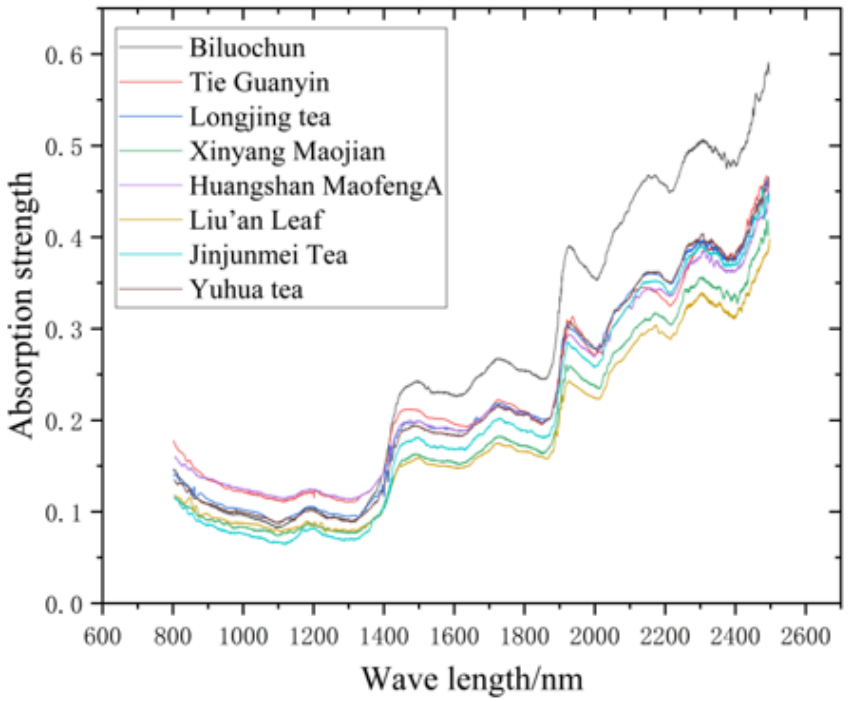

Figure 1

Near infrared spectra of 8 kinds of tea

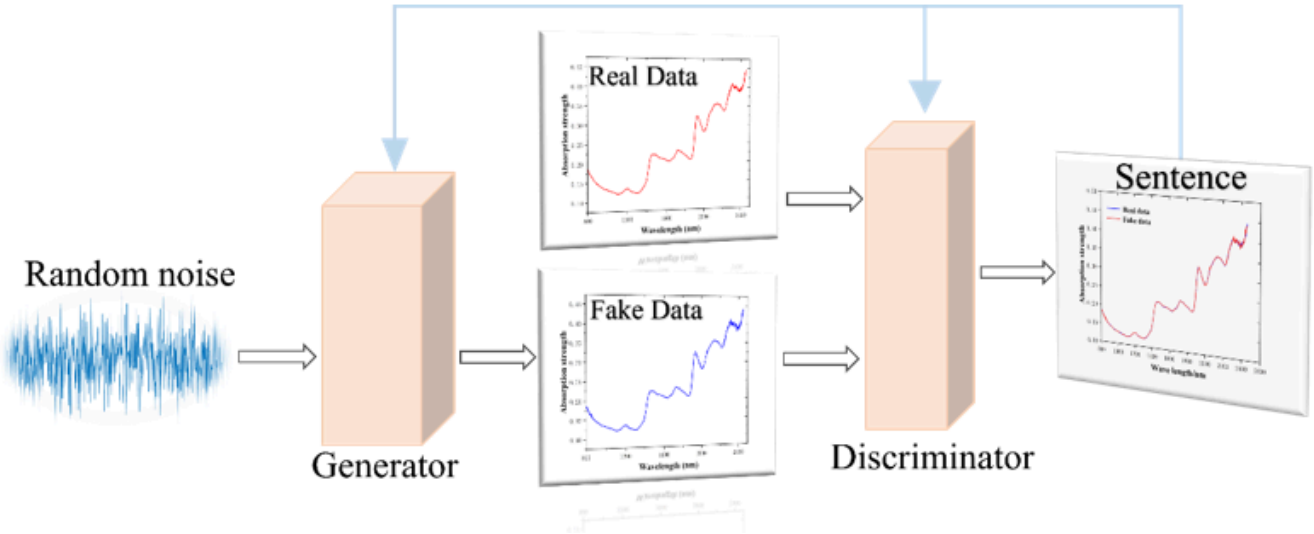

Figure 2

Basic structure of DCGAN

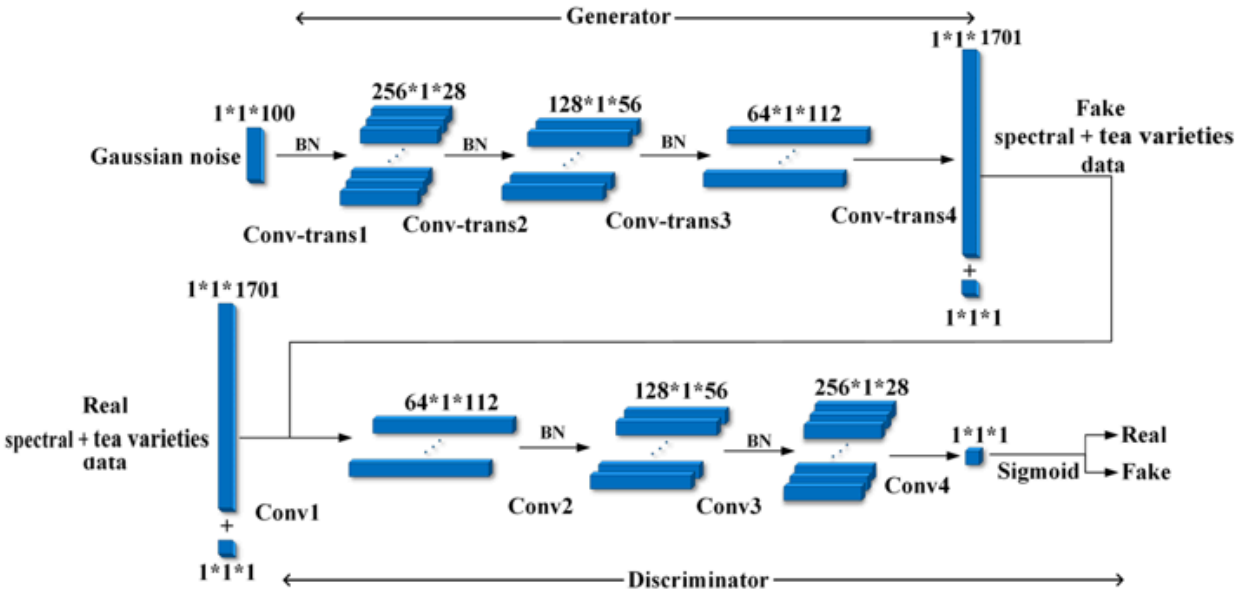

Figure 3

DCGAN structure and training process

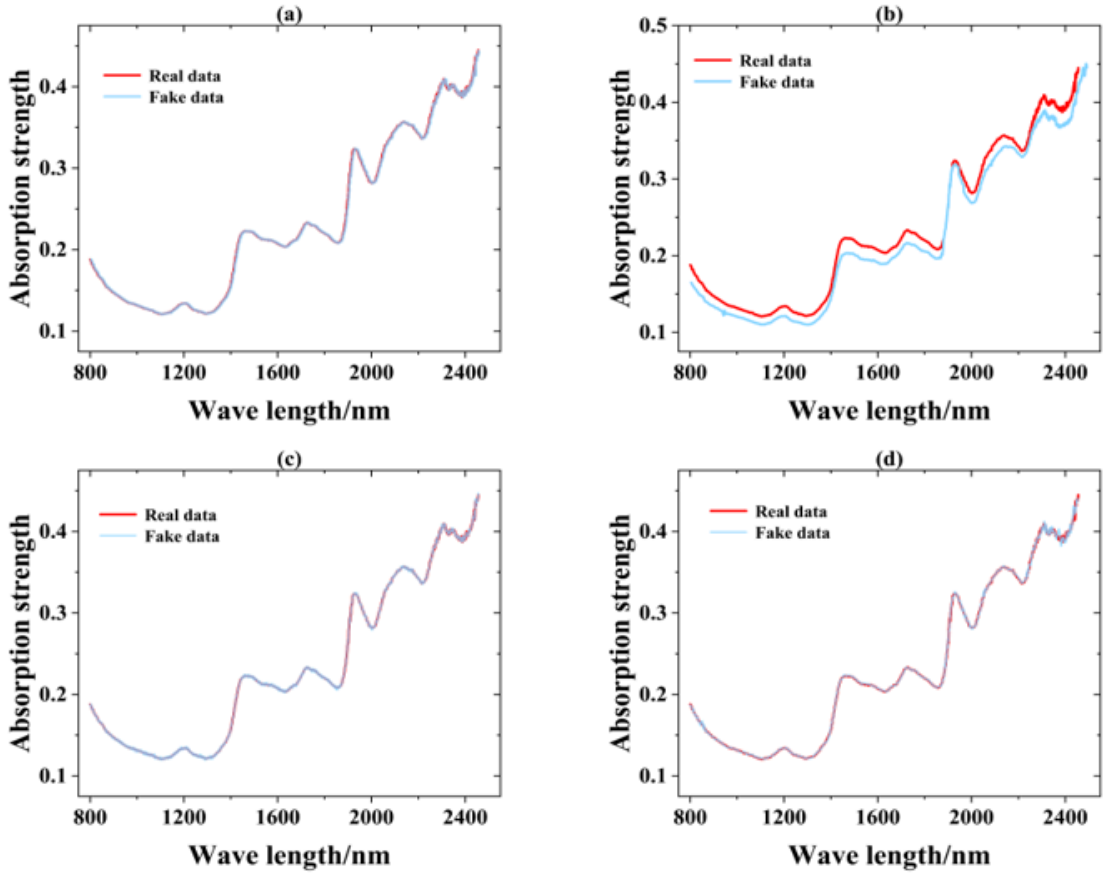

Figure 4

Spectral data enhancement results of different methods (a) Translation method; (b) Linear superposition method; (c) Noise addition method; (d) DCGAN method

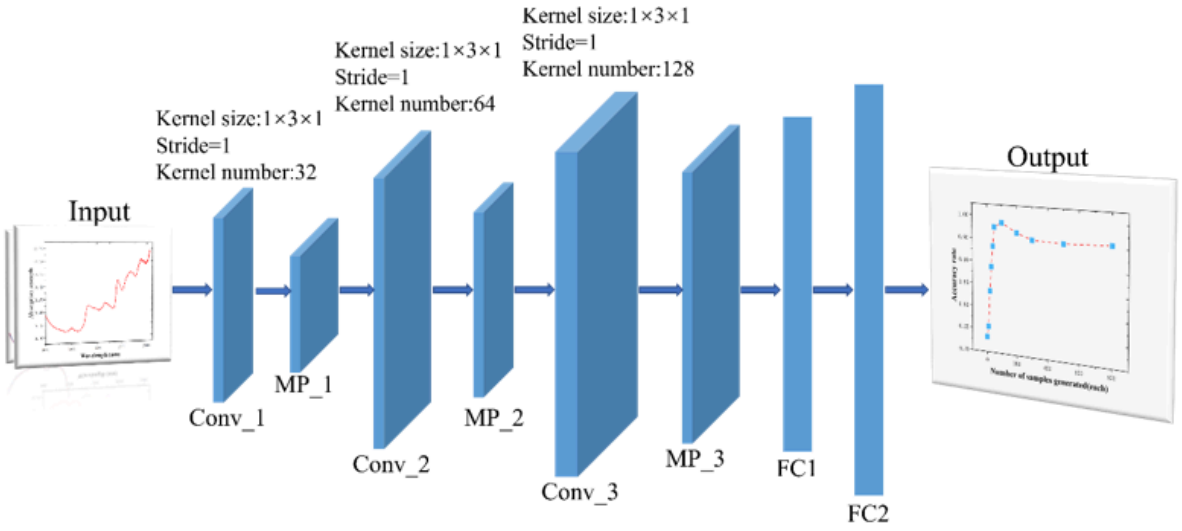

Figure 5

CNN Structure Diagram