Back Propagation Neural Network-Enhanced Generative Model for Drying Process Control

Yonggang Liu*, Hongliang Zhang, Xiang Wei, Mei Li
Hebei Baisha Tobacco Co., Ltd. Baoding Cigarette Factory, Baoding 071000, China
E-mail: 15631295011@163.com
*Corresponding Author

Keywords: drying control, back propagation neural network, data generation model, process parameters, insufficient sample space

Received: October 25, 2024

To improve the control precision and stability of the drying process, this work investigates a drying process control model based on a Back Propagation Neural Network (BPNN). The model includes a generative adversarial network framework, and combines the discriminator and generator structure of BPNN, optimizing their respective loss functions. It aims to solve the problem of insufficient sample space in the process parameters of the drying machine. The discriminator receives 7 key process parameters and uses a three-layer fully connected network for processing. The generator generates fake samples based on 7 process parameters. The model's performance is validated through experiments, including fit analysis of the generated results and the model's reliability analysis. The results show that the composite R² value of the data generation model reaches 0.93915 in the fit analysis. This consistency validates the model's ability to accurately fit the global data distribution, reflecting its generalization ability. The mean square error and mean absolute error are 0.275 and 0.185 respectively, which are better than other models, further verifying the model's performance. Additionally, significance analysis reveals that the H values of the process parameters in the datasets generated by the data generation model and the original datasets are all 0, with p-values greater than 0.05. This indicates no significant statistical difference between the two, confirming the reliability of the data generation model in filling the insufficient sample space. It suggests that the model can effectively enhance the completeness of the dataset without affecting the data distribution characteristics. The findings of this work provide theoretical and practical guidance for optimizing control in the drying process, contributing to improved control precision and stability in industrial drying operations.

Povzetek: Model za nadzor sušenja temelji na nevronski mreži s povratnim razširjanjem (BPNN) in vključuje generativni nasprotni okvir in združuje strukture diskriminatorja in generatorja BPNN za reševanje problema nezadostnega vzorčnega prostora. Rezultati kažejo, da model učinkovito povečuje nabor podatkov, ne da bi pri tem vplival na značilnosti porazdelitve podatkov.

1 Introduction

The drying process is one of the key steps in tobacco processing, and its control quality directly impacts the flavor and quality of the final product. Parameters such as temperature and humidity during the drying process must be precisely controlled to ensure that the tobacco leaves reach the desired moisture content and optimal physical properties [1,2]. However, due to the nonlinear and time-varying characteristics of the drying process, coupled with the complex environment in industrial production, traditional control methods, such as Proportional-Integral-Derivative (PID) control based on experience, often fail to achieve the desired results [3-5]. This makes achieving precise control of the drying process a pressing issue within the industry.

With the development of artificial intelligence technology, particularly the widespread application of neural networks in pattern recognition and process control, researchers have begun to explore their use in drying control [6]. The Back Propagation Neural Network (BPNN), a classic feedforward neural network, has been widely used in modeling and controlling complex industrial processes due to its powerful nonlinear mapping capabilities and adaptability [7,8]. However, the performance of neural network models depends on a large amount of high-quality training data. In actual production, due to the high cost of data collection or frequent changes in production batches, it is often difficult to obtain sufficient sample data. This limits the further application of neural network models in drying control [9,10]. Therefore, solving the problem of insufficient data samples has become the key to improving the neural networks' application effect.

Generative models like the Generative Adversarial Network (GAN) have gained widespread attention in the data augmentation field to overcome the issue of insufficient samples. Although these models have achieved remarkable results in areas such as image processing, their application in industrial process control is still in the exploratory stage. Specifically, in drying control, generating high-quality synthetic process parameter data remains a challenging research topic.

Therefore, this work constructs a BPNN-based data generation model to address the insufficient sample issues in the drying process control. This model combines the BPNN's generator and discriminator structures to expand the training sample space by generating high-quality synthetic data, thereby improving the drying control model's performance.

The main contributions of this work are as follows:

(1) A data generation model for insufficient sample space of process parameters based on BPNN is proposed. By generating high-quality synthetic data, the model expands the training sample space, effectively solving the problem of data scarcity in drying process control.

(2) A generator and discriminator model based on BPNN is designed and optimized, significantly improving the accuracy and reliability of the generated data through detailed analysis of their network structures and optimization strategies.

(3) The effectiveness of the proposed model is validated through experiments, demonstrating its remarkable advantages in handling the under-sampling problem in the drying process. This provides a novel solution for the intelligent control of industrial processes.

2 Related work

In the field of drying control, the stability and quality of the drying process directly influence the flavor and quality of tobacco products. In recent years, more research has focused on exploring the application of intelligent control methods in the drying process. Mu et al. proposed an innovative humidity prediction method. This model transformed raw data into image data and used a multilayer convolutional neural network (CNN) for real-time prediction, successfully addressing the delay issues in traditional detection methods. Experimental results demonstrated that this method effectively predicted humidity and optimized the production process using real production data [11]. Furthermore, Odabas et al. introduced a correlated color temperature-based method to determine the optimal drying time for tobacco leaves. They found that the 17th day of drying yielded the best quality, with a correlated color temperature of 3000 K and a quality grade of 100% [12]. This method effectively guided the tobacco leaf drying process to ensure product quality. Similarly, Li et al. examined the relationship between thermal energy consumption and process parameters in tobacco drying using statistical analysis and machine learning. Their findings showed that reducing the main steam temperature significantly lowered energy consumption while optimizing other equipment parameters further improved energy efficiency [13]. This research provided valuable insights for green upgrading and energy conservation in tobacco manufacturing.

As research progresses and data scarcity in process control persists, generating high-quality synthetic data to improve model performance in under-sampled environments has become a hot topic. Zhou and Chiam proposed a new synthetic data generation strategy for knowledge distillation in regression tasks without original training data. By optimizing the bounded difference between student and teacher models, this strategy significantly improved the student model's performance, as confirmed by experimental results [14]. Barbierato et al. used probabilistic networks and structural equation modeling to address data bias, proposing a new synthetic dataset generation method. Validated on simple and loan approval datasets, this method effectively regulated bias and fairness with fewer parameters [15]. This method provided a powerful tool for developing and validating fair decision algorithms. Moreover, Zhang and Mikelsons introduced a novel sensitivity-guided iterative parameter identification and data generation algorithm, replacing traditional manual intervention. Using BayesFlow for parameter identification and a spatial variational autoencoder for data generation, the method remarkably accelerated the model calibration process. Experiments on a vehicle cabin thermal model showed that this method could greatly improve calibration accuracy and reduce calibration time from over a week to just one day [16]. This method effectively improved model calibration accuracy. The above studies have been summarized and compared with the method proposed in this work, resulting in Table 1:

Table 1: Comparison of existing methods with the proposed method

Research methods	**Data generation model**	**Main advantages**	**Limitations**	**The advantages of the proposed method**
Mu et al. [11]	Image data transformation + CNN	It improved humidity prediction accuracy and optimized the production process.	It was only applicable for humidity prediction and difficult to extend to other drying parameters.	The generation model proposed here can handle more drying process parameters and has wider applicability.
Odabas et al. [12]	Drying time prediction based on color temperature.	It provided effective recommendations for tobacco leaf drying time.	It was only suitable for predicting drying time,	The proposed method could adapt to more comple

			failed coping with the complex and variable drying process.	x drying processes by generating diverse process parameter data.
Li et al. [13]	Statistical analysis and machine learning	It offered effective guidance for energy conservation.	It mainly focused on the relationship between thermal energy consumption and parameters, neglecting dynamic control.	The proposed method could generate more dimensional process data, covering more variables to optimize drying control.
Zhou and Chiam [14]	Data generation strategy for knowledge distillation	It enhanced the learning effect of the student model.	It was primarily used for regression tasks and lacked customization for process control	This study applied synthetic data generation to the drying process, which exhibited higher industry adaptability.
Barbierato et al. [15]	Probabilistic networks and structural equation modeling	It controlled bias and fairness.	It was primarily used for regression tasks and lacked customization for process control.	This study applied synthetic data generation to the drying process, which exhibited higher industry adaptability.

In Table 1, existing methods generally exhibit several shortcomings. First, many methods are primarily designed for simplified datasets and single-task scenarios, lacking adaptability to complex industrial processes. Second, current methods often fail to ensure the interpretability of generated data and its high consistency with actual process parameters, resulting in significant discrepancies between generated and real data. Third, most synthetic data generation methods do not adequately consider the dynamic changes and multivariate characteristics of the drying process, leading to poor scalability in practical applications. Therefore, this work proposes a data generation model based on BPNN to handle the insufficient sample space of drying machine process parameters. By generating high-quality synthetic data, the model aims to expand the training sample space and improve the application performance of BPNN in the drying process. Compared with existing methods, the proposed model emphasizes practical applicability while focusing on the interpretability of the generated data and its close alignment with actual process parameters. Furthermore, the proposed model effectively handles complex, multivariate drying process data, thus overcoming the limitations of current methods in multidimensional data generation and process control.

3 The BPNN-based data generation model for drying machine process parameters with insufficient sample space

Accurate process parameter data is crucial for precise drying process control [17]. However, in practice, insufficient sample data often presents challenges for model development. To address this issue and enhance the application of BPNN in drying control, this work proposes a BPNN-based data generation model. The model aims to improve control accuracy in the drying process by generating high-quality synthetic data to expand the training sample space. Unlike traditional methods, this work focuses on solving data scarcity issues through advanced data generation techniques, providing a novel solution for intelligent drying process control. This section outlines the proposed data generation model's overall structure and key components.

3.1 Overall structure of the data generation model

GAN is a powerful tool for data generation through adversarial training. Here, the GAN's basic framework is used to construct the data generation model, producing high-quality drying machine process parameter data. The model's overall workflow is displayed in Figure 1 [18,19].

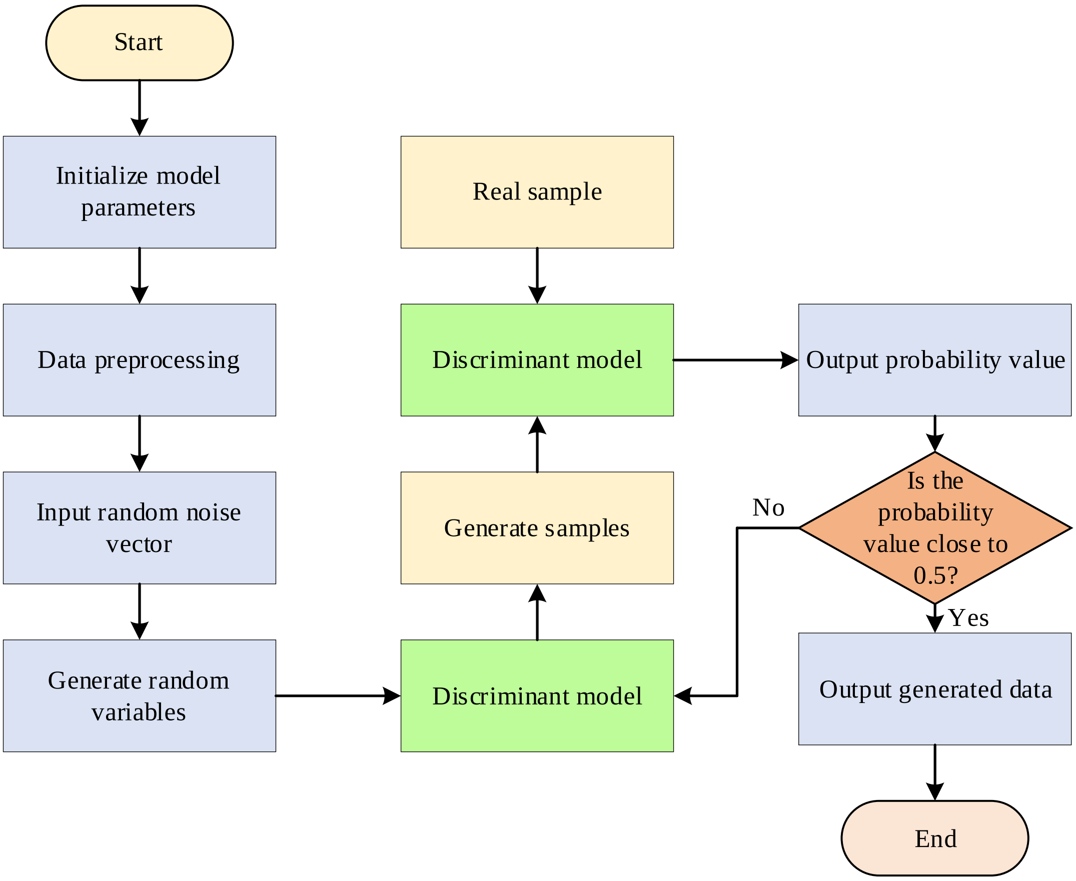


Figure 1: Workflow of the data generation model for drying machine process parameter

The workflow can be broken down into the following key steps:

(1) Data Initialization: In the initial phase of the model, a subset of samples is first extracted from the actual drying process dataset for subsequent training of the generator and discriminator. Data initialization involves randomly initializing the model parameters to start the training process. Additionally, a preprocessing step is applied to enhance training efficiency and data consistency, which often includes normalization. Normalization adjusts the values of each input data feature to a uniform range, typically [0, 1] or [-1, 1], which accelerates the convergence of model training and improves training performance.

(2) Generating Random Variables: The primary goal of the generator is to learn how to map random noise vectors into data samples that possess the characteristics of real samples. During this process, the generator receives a random noise vector as input, typically sampled from a Gaussian or uniform distribution. It then processes these random noise inputs through a neural network to generate new data samples. The objective is for these generated samples to exhibit distributional characteristics similar to those of the real data. The training goal of the generator is to progressively optimize its parameters to produce samples increasingly close to real data.

(3) Sample Mixing: To train the discriminator, the generated samples are mixed with real samples to form a training dataset. The discriminator's task is to distinguish which samples are fake, generated by the generator, and which come from real data. At this stage, the generator and the discriminator are adversarial. The generator aims to produce samples that the discriminator classifies as real, while the discriminator strives to identify the fake samples accurately. The mixed dataset includes real and generated fake samples for the discriminator's training.

(4) Discriminator Output Probability: The discriminator processes the input samples through a neural network and outputs a scalar probability value, representing the probability that a given sample is real. This probability ranges from 0 to 1. Ideally, the generator aims for its generated samples to yield probability values close to 1, meaning the discriminator misclassifies them as real samples. Conversely, for real data, the discriminator strives to output probabilities close to 0. The generator's objective is to make the discriminator classify its generated samples as real, resulting in probability values near 1.

(5) Training Objective Adjustment and Optimization: Based on the discriminator's output probabilities, the training objectives of both the generator and the discriminator are continuously adjusted. During each training iteration, if the discriminator's judgment on the generated samples approaches 50% (i.e., output probability near 0.5), indicating that the generated and real samples are nearly indistinguishable. In this case, the generator performs well and can continue optimization with its current parameters. However, if the

discriminator's output probabilities deviate significantly from 0.5, it implies poor sample quality from the generator. The discriminator's loss is then backpropagated to the generator, prompting parameter adjustments to enhance its performance. Through this adversarial training process, the generator and discriminator improve iteratively, with the generator progressively learning to produce more realistic data samples.

The above describes the detailed workflow of the data generation model for the drying machine process parameters illustrated in Figure 1. Each step aims to optimize the generator and discriminator through adversarial training, ultimately enabling the generated samples to closely resemble real data. Based on the above data processing in the generation model, the modeling process of the data generation model is obtained, as suggested in Figure 2.

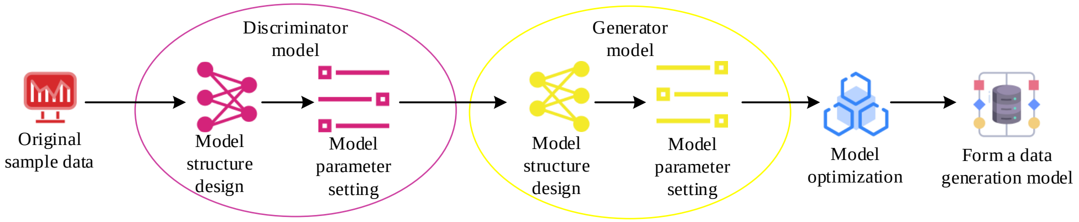


Figure 2: Modeling process of the data generation model for drying machine process parameters

3.2 BPNN-based discriminator and generator model network structure

In the data generation model, the design of the discriminator and generator is crucial for the quality and accuracy of the generated data. The discriminator evaluates whether the generated data matches the real data distribution, while the generator aims to produce fake data that aligns with the real distribution.

The main function of the discriminator is to perform binary classification between the fake samples generated by the generator and the real samples. Its input consists of a mixed dataset of fake samples generated by the generator and real samples of drying machine process parameters, which exhibit nonlinear relationships among the data groups. This dataset includes seven key parameters influencing the moisture content of the tobacco leaves during the drying process. They are the vent valve opening, drum wall temperature, moisture content of dried leaves, hot air temperature, steam valve opening of the drum wall, steam valve opening of the circulating air, and moisture content of cooled leaves [20-22]. The output value is a probability scalar indicating the likelihood that the generated fake sample data are classified as real. Due to the complex nonlinear relationships between these samples, the discriminator needs to have certain nonlinear mapping capabilities. Given the dataset's relatively low dimensionality and limited features, a three-layer fully connected BPNN is used as the discriminator. This network structure effectively handles simple binary classification tasks, performing both linear and nonlinear features, and ensuring the generated data meets accuracy requirements.

The generator's goal is to produce fake data that are consistent with the real data distribution. To achieve this, the generator's network structure must formalize the characteristics of the fake dataset it generates and possess strong nonlinear learning capabilities. The input and output data of the generator are analyzed. Its input consists of randomly generated noise data with the same number as the real dataset, and its output is the data processed by the generator network to be closer to the real values. Given that the input data is one-dimensional and relatively simple, a three-layer fully connected BPNN is employed for the generator.

The drying machine process parameter data have highly nonlinear characteristics. The BPNN's advantage lies in its strong nonlinear mapping capability. This allows it to automatically adjust structural parameters during training to establish complex input-output relationships without requiring an understanding of their intrinsic connections [23,24]. This makes it particularly well-suited for handling the nonlinear characteristics and intricate data relationships in drying machine process parameters. The three-layer network structure provides sufficient nonlinear representation capacity while avoiding overcomplicating the model. The design of a three-layer fully connected BPNN strikes a reasonable balance between computational efficiency and model complexity. Compared with deeper networks, the three-layer network has fewer parameters, reducing the risk of overfitting during training. This is particularly advantageous in the context of drying control, where the number of data samples is limited, as overly deep network structures may compromise the model's generalization ability. Additionally, the three-layer structure is relatively simple, efficiently performing linear and nonlinear mappings. Therefore, a three-layer fully connected BPNN is considered an effective choice for this work's tasks. Figure 3 presents the network structure of the discriminator and generator models based on BPNN.

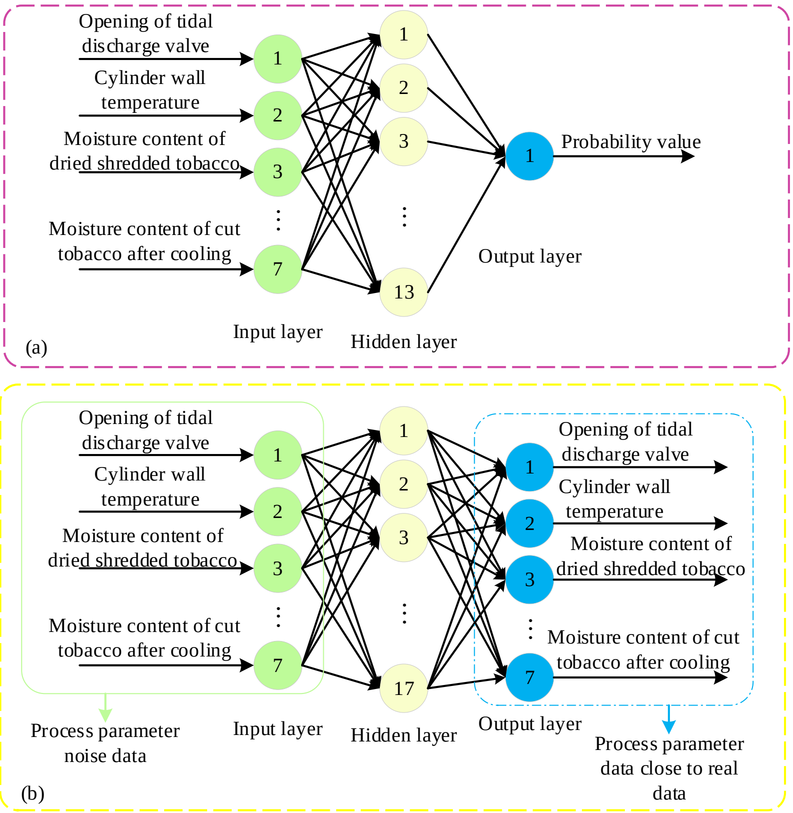
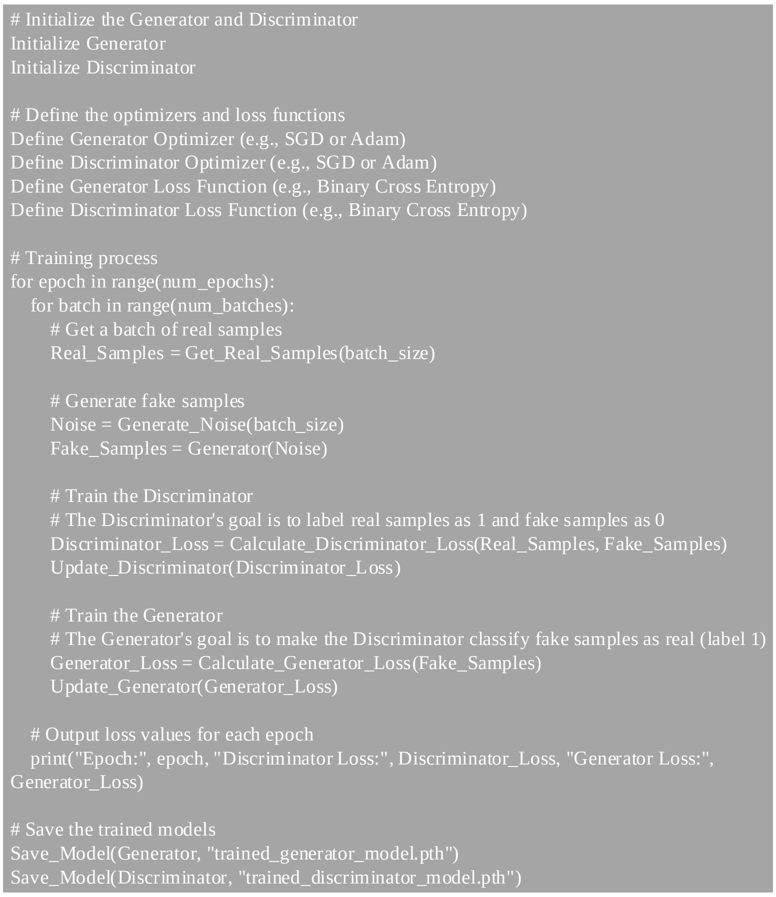

Opening of tidal discharge valve
Cylinder wall temperature
Moisture content of dried shredded tobacco
Moisture content of cut tobacco after cooling
1
2
3
7
1
2
3
13
1
Probability value
Output layer
Input layer
Hidden layer
(a)
Opening of tidal discharge valve
Cylinder wall temperature
Moisture content of dried shredded tobacco
Moisture content of cut tobacco after cooling
1
2
3
7
1
2
3
17
1
2
3
7
Opening of tidal discharge valve
Cylinder wall temperature
Moisture content of dried shredded tobacco
Moisture content of cut tobacco after cooling
Process parameter noise data
Input layer
Hidden layer
Output layer
Process parameter data close to real data
(b)

Figure 3: Diagram of the Discriminator and Generator Model Network Structure (a: Discriminator model structure; b: Generator model structure)

The pseudocode for the entire model is presented in Figure 4.



```
# Initialize the Generator and Discriminator
Initialize Generator
Initialize Discriminator

# Define the optimizers and loss functions
Define Generator Optimizer (e.g., SGD or Adam)
Define Discriminator Optimizer (e.g., SGD or Adam)
Define Generator Loss Function (e.g., Binary Cross Entropy)
Define Discriminator Loss Function (e.g., Binary Cross Entropy)

# Training process
for epoch in range(num_epochs):
    for batch in range(num_batches):
        # Get a batch of real samples
        Real_Samples = Get_Real_Samples(batch_size)

        # Generate fake samples
        Noise = Generate_Noise(batch_size)
        Fake_Samples = Generator(Noise)

        # Train the Discriminator
        # The Discriminator's goal is to label real samples as 1 and fake samples as 0
        Discriminator_Loss = Calculate_Discriminator_Loss(Real_Samples, Fake_Samples)
        Update_Discriminator(Discriminator_Loss)

        # Train the Generator
        # The Generator's goal is to make the Discriminator classify fake samples as real (label 1)
        Generator_Loss = Calculate_Generator_Loss(Fake_Samples)
        Update_Generator(Generator_Loss)

    # Output loss values for each epoch
    print("Epoch:", epoch, "Discriminator Loss:", Discriminator_Loss, "Generator Loss:",
Generator_Loss)

# Save the trained models
Save_Model(Generator, "trained_generator_model.pth")
Save_Model(Discriminator, "trained_discriminator_model.pth")
```

Figure 4: Pseudocode for model training

The carefully designed network structures of the discriminator and generator enable efficient data generation and classification, forming a robust foundation for modeling drying process parameters.

3.3 Optimization of the data generation model

Due to the intra-group data correlations and nonlinear characteristics between groups in the drying machine process parameter data with insufficient sample space, the original GAN model's loss function is not suitable for the subject of this work. Thus, optimizing the loss functions of both the discriminator and generator is crucial.

During the data generation model's operation, the discriminator and generator models are trained alternately. After each generator iteration, the discriminator model is optimized to achieve overall model improvement. The optimization goal of the model is to maximize the ability of the discriminator, which in turn drives the generator to produce more realistic data.

In the process of the data generation model, the discriminator and generator models are trained alternately. After each iteration of the generative model, the discriminator model is optimized to achieve the overall optimization of the model. The optimization process of the model can be expressed by equation (1):

$$\min_S \max_P H(P,S) = \mathbb{E}_{x-z_0}[\log P(x)] + \mathbb{E}_{c-z_c}[\log(1 - P(S(c)))] \qquad (1)$$

x represents the real samples; c is the samples produced by the generator; z_0 refers to the distribution of real samples; z_c stands for the distribution of samples generated by the generator; $P(x)$ and $P(S(c))$ indicate the probability that the discriminator model classifies the real and generated samples as true; $\mathbb{E}$ denotes the expectation symbol; $\mathbb{E}_{x-z_0}$ means the expectation for real samples x; $\mathbb{E}_{c-z_c}$ represents the expectation for generated samples c.

The loss function for the discriminator enhances its ability to distinguish between real and generated samples by minimizing the probability of generated samples being classified as real.

In contrast, the generator optimizes the quality of its output by maximizing the discriminator's probability of misclassifying fake samples. The generator's goal is to produce fake samples that achieve the highest possible score from the discriminator, effectively deceiving it into classifying the fake samples as real. Thus, the generator's loss function is formulated as follows:

$$F_S = -\frac{1}{m}\sum_{i=1}^{m} \log(P(S(c^i))) \qquad (2)$$

m represents the number of samples; c^i represents the ith generated sample; $P(S(c^i))$ denotes the probability that the discriminator predicts the generated sample c^i as a real sample. The generator aims to minimize this loss function, ensuring its fake samples are increasingly likely to be judged as real by the discriminator. The discriminator's primary task is to distinguish real data from generated data, and its loss function is designed to accurately evaluate the similarity between generated and real samples. In the original GAN model, the loss function is typically based on cross-entropy loss, calculating the binary classification probability between generated and real samples. The loss function for the discriminator reads:

$$F_P = -\frac{1}{m}\sum_{i=1}^{m} \log P(x^i) - \frac{1}{m}\sum_{i=1}^{m} \log(1 - P(S(c^i)))(3)$$

For the discriminator in the data generation model, an input x without a label is obtained. Since there are no labels and it is unknown whether the input is a real sample or a generated fake sample, the contribution $Q(x)$ to the loss function of the data generation model is given by:

$$Q(x) = -z_0(x)\log P(x) - z_c(x)\log[\,1 - P(x)] \qquad (4)$$

$z_0(x)$ and $z_c(x)$ represent whether sample x belongs to the real data distribution z_0 or the generated data distribution z_c, respectively; $P(x)$ denotes the probability assigned by the discriminator to sample x being real. By solving the derivative of $Q(x)$ with respect to $P(x)$, the optimal decision condition for the discriminator is obtained:

$$-\frac{z_0(x)}{P(x)} + \frac{z_c(x)}{1-P(x)} = 0 \qquad (5)$$

Further simplification, the optimal discriminator decision function is obtained:

$$P^*(x) = \frac{z_0(x)}{z_0(x)+z_c(x)} \qquad (6)$$

Analyzing the optimal discriminator, if $z_0(x) = 0$ and $z_c(x) \neq 0$, the optimal discriminator sets the probability to 0, indicating that the discriminator incorrectly classifies real samples as fake. If $z_0(x) = z_c(x)$, the discriminator considers the sample to have a 50% chance of being either a real or fake sample, resulting in an output of 0.5.

To further improve model performance and address the vanishing gradient problem during generator training, Jensen-Shannon (JS) divergence is introduced to measure the similarity between the real distribution N and the generated distribution W. JS divergence effectively quantifies the difference between two probability distributions and is defined as:

$$\begin{cases} KL(z_1\|z_2) = E_{x-z_1}\log\frac{z_1}{z_2} \\ JS(z_1\|z_2) = \frac{1}{2}KL\left(z_1\,\middle\|\,\frac{z_1+z_2}{2}\right) + \frac{1}{2}KL\left(z_2\,\middle\|\,\frac{z_1+z_2}{2}\right) \end{cases} \qquad (7)$$

KL represents the Kullback-Leibler Divergence. By applying JS divergence, the optimized form of the generator's loss function is derived:

$$\min_S \max_P H(P,S) = 2JS(z_0\|z_c) - 2\log 2 \qquad (8)$$

This improvement not only mitigates the vanishing gradient problem for the generator but also enhances the similarity between the distributions of the generated and real samples.

To address the vanishing gradient issue, the loss function is further simplified by removing the logarithmic operation. The optimized loss functions for the discriminator and generator are as follows:

$$\begin{cases} F_P = -\frac{1}{m}\sum_{i=1}^{m} P(x^i) - \frac{1}{m}\sum_{i=1}^{m}[1 - P(S(c^i))] \\ F_S = -\frac{1}{m}\sum_{i=1}^{m} P(S(c^i)) \end{cases} \qquad (9)$$

These optimization measures are aimed at enhancing the training effectiveness of the generator and discriminator, overcoming the gradient vanishing problem in GAN, and ultimately generating higher-quality drying machine process parameter data.

4 Model performance validation

4.1 Experimental setup

The model is trained and validated using the Matlab platform in this experiment.

The dataset used in the experiment includes both real drying machine process parameter data and generated noise data. The real dataset comprises multiple samples of drying machine process parameters collected from actual operations, reflecting authentic conditions across various drying scenarios. Collecting these data involves gathering process parameters from a real production environment, followed by detailed annotation and processing. The generated noise data are produced by the generator network, to mimic the characteristics of the real data distribution. By inputting random noise into the generator, the network learns to generate samples similar to the real data, continuously optimizing the quality of the generated data. These noise data are used during training to challenge the real samples, driving the generator to produce data closely resembling the real samples.

First, the real training samples are imported into the Matlab environment, and all of these real samples are labeled as "1" to represent positive samples. Simultaneously, the noise samples generated by the generator are labeled as "0" to denote negative samples. This labeling setup ensures a clear distinction between real and generated data, effectively training the discriminator and generator models. Table 2 provides the parameter settings for the data generation model used in the experiment:

Table 2: Experimental parameter setting

Model	Parameter Name	Value
Discriminator	The Number of Input Layer Neurons	7
	The Number of Hidden Layer Neurons	13
	The Number of Output Layer Neurons	1
	Batch Size	62
	Initial Learning Rate	0.00001
	Optimizer	SGD Optimizer
Generator	The Number of Input Layer Neurons	7
	The Number of Hidden Layer Neurons	14
	The Number of Output Layer Neurons	7
	Batch Size	62
	Initial Learning Rate	0.00001
	Activation Function	Sigmoid
Overall Model	Iterations	30000

The input to the discriminator consists of seven key parameters from the drying machine process that primarily affect the moisture content of the tobacco leaves. Hence, the number of neurons in the input layer is 7, while the output is a probability value, so the output layer contains one neuron. The number of hidden layer neurons is determined based on the input and output layer neurons using equation (10):

$$Z = \sqrt{i + j} + a \tag{10}$$

i and j represent the number of input and output layer neurons; a refers to a constant in the range [0,10], set to a=10. Using this equation, the number of hidden layer neurons for the discriminator is calculated to be 13. Given that the dataset contains 309 highly nonlinear samples, conventional full-dataset and online learning approaches are not suitable. To improve training accuracy while maintaining reasonable training time, the batch size is appropriately reduced based on the full dataset. In this experiment, the 309 real samples are divided into five groups, setting the batch size to 62. This configuration retains sufficient data characteristics to effectively guide model optimization while avoiding overly large data inputs that could slow training or overburden computational resources, ensuring the efficiency of the training process. Due to the activation function of the hidden layer being Sigmoid, the initial setting of the learning rate requires a small value to meet the changing requirements of the learning rate. Therefore, the initial learning rate is set to 0.00001.

The number of neurons in the input layer of the generator is determined by the number of input parameters. The input consists of the process parameters with missing data and the process parameters that are related to them, totaling seven process parameters. Therefore, the number of neurons in the input layer is 7. The number of neurons in the output layer corresponds to the number of generated samples. To ensure the accuracy of the generator’s learning, the output layer should have the same number of neurons as the input layer, hence the output layer also has 7 neurons. Similarly, using the equation, the number of neurons in the generator's hidden layer is calculated to be 14. The batch size of the generator is determined by the number of groups of input data. After excluding the samples with missing data, 310 valid samples are remaining. These valid samples are divided into 5 groups, with 62 samples per group, and the data are fed into the generator network. Consequently, the batch size for the generator is set to 62. The data preprocessing steps in the experiment include normalization to ensure that the values of each parameter are within a similar range. This helps to enhance the model's training effectiveness and convergence speed. All experimental data are standardized and split into training and testing sets in a 3:1 ratio, ensuring the reliability and reproducibility of the model's training and validation results.

In this work, the model may encounter overfitting issues. To prevent this, early stopping is employed to enhance the model's robustness and generalization ability. During training, the error on the validation set is monitored, and if the error no longer decreases over several consecutive training rounds, the training is stopped early to prevent overfitting on the training set.

4.2 Analysis of the fit of the data generation model's output

Figure 5 shows the fit of the training set.

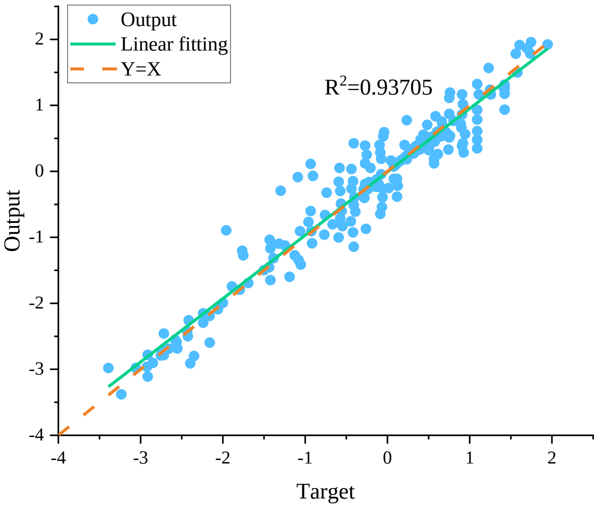


Figure 5: The fit of the data generation model on the training set

Figure 5 reveals that the model exhibits an optimal fit. Specifically, the R² value reaches 0.93705, indicating that the generative model effectively captures the distribution characteristics of the sample data within the training set. An R² value approaching 1 suggests a high correlation between the generated and actual sample data within the training set, demonstrating that the model has accurately learned and replicated the features of the real data.

Figure 6 displays the fit of the data generation model on the testing set.

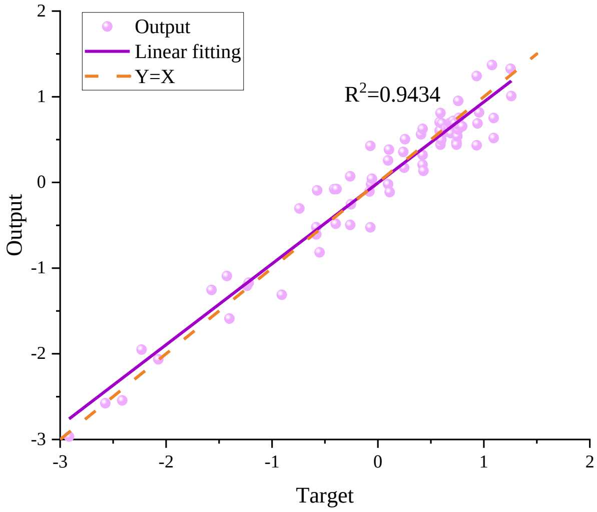


Figure 6: The fit of the data generation model on the testing set

The fit curve of the data generation model on the testing set, as shown in Figure 6, demonstrates that the model also performs exceptionally well on the testing data, with an R² value reaching 0.9434. This result indicates that the generative model can accurately replicate the distribution characteristics of real data, even when applied to previously unseen data, reflecting the model's strong generalization ability.

Analyzing the model's fit across the training and testing sets provides an overall assessment of its performance on the entire dataset. Figure 7 depicts the results.

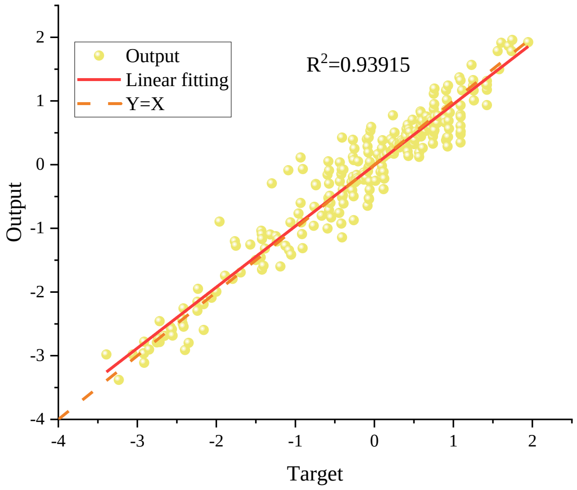


Figure 7: The overall fit of the data generation model

Figure 7 shows the overall fit of the data generation model. The results demonstrate the model's robustness and generalizability across the entire dataset. By analyzing the training and testing sets, the model maintains a high fit accuracy across a broader sample space, achieving an R^2 value of 0.93915. This indicates that the model effectively captures patterns in the training data and extends these patterns when encountering unseen test data. As a result, the quality and authenticity of the generated samples are preserved. This consistency confirms the model's robust ability to fit the global data distribution, highlighting its generalization ability.

4.3 Reliability analysis of the data generation model

After completing the training, the data generation model is used to interpolate samples into the under-sampled space of the drying machine process parameters. To verify the reliability of the data generation model, a significance analysis is conducted comparing the interpolated new data with the original dataset. This aims to ensure that the interpolated samples are consistent with the original samples in data distribution. Specifically, the significance of seven process parameters is tested at a 0.05 significance level to determine whether the interpolated data can preserve the distribution characteristics of the original dataset.

In the significance analysis, the null hypothesis is that the new and original datasets share the same data distribution at a significance level of 0.05. The H values and p-values for each process parameter are obtained by calculating the significance test parameters, as denoted in Figure 8.

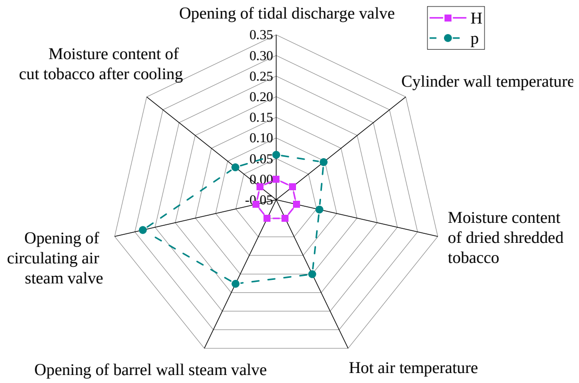


Figure 8: Significance test of the data generation model

In Figure 8, the H values for all process parameters are 0, and the p-values are greater than 0.05. There is insufficient evidence to reject the null hypothesis at the specified significance level. Therefore, it can be inferred that there is no statistically significant difference between the new and original datasets. Specifically, the significance test results for each process parameter demonstrate that the interpolated data statistically aligns with the original data's distribution. This outcome verifies the data generation model's reliability, showing that the generated interpolated values successfully learn and preserve the data distribution characteristics of the original sample space. Consequently, the data generation model is highly reliable in filling under-sampled spaces, effectively enhancing the dataset's completeness without altering its distribution characteristics.

An ablation experiment is conducted to evaluate the effectiveness of the loss function optimization. Model-1 represents the original model; Model-2 represents the model without optimization of the discriminator's loss function; Model-3 represents the model without optimization of the generator's loss function; Model-4 represents the model with neither the discriminator nor the generator optimized. The performance is evaluated using the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) as metrics, and the results are indicated in Figure 9.

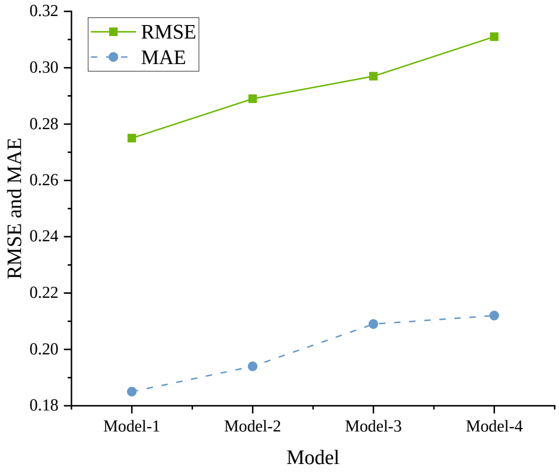


Figure 9: Results of the ablation experiment

As observed from Figure 9, model optimization significantly improves the quality of generated data. Model-1 performs the best in terms of RMSE and MAE, with values of 0.275 and 0.185, respectively. The removal of optimization for the discriminator and generator loss functions results in a decrease in performance. In Model-4, where neither the discriminator nor the generator's loss function is optimized, the performance is the worst, with RMSE and MAE values of 0.311 and 0.212, respectively. Overall, optimizing the loss functions leads to a significant improvement in model performance, particularly when both the discriminator and the generator are optimized simultaneously, which maximizes the precision of data generation.

To further validate the effectiveness of the model, this work compares the proposed model with several traditional modeling methods for drying machine process parameters. These methods include Support Vector Machine (SVM), Random Forest (RF), and Multi-Layer Perceptron (MLP), among other classical machine learning models. The results are listed in Table 3.

Table 3: Comparison of different models

Model	RMSE	MAE
The proposed model	0.275	0.185
SVM	0.307	0.215
RF	0.292	0.204
MLP	0.285	0.198

Table 3 shows that the proposed model outperforms the SVM, RF, and MLP models in terms of both RMSE and MAE. This indicates that the optimized BPNN model can better capture the complex nonlinear relationships in the drying process. It generates samples that are closer to real data, particularly demonstrating superior performance on the test set with strong generalization ability.

In addition, to further evaluate the model's robustness and the repeatability of the results, cross-validation is conducted. By employing a 5-fold cross-validation method, the model undergoes multiple rounds of training and validation to reduce the possibility of overfitting and ensure the stability of the results. The cross-validation results reveal that the model's performance is consistent across all folds, with a standard deviation of the R^2 value of only 0.0025, further demonstrating the model's robustness and reliability under different data splits.

4.4 Discussion

In this work, the proposed BPNN-based data generation model for drying machine process parameters with insufficient sample space achieves promising results in the experiments. Specifically, the model's R^2 value on the training set is 0.9322, and on the test set, it is 0.9434. The combined R^2 value for the training and test sets reaches 0.93915, indicating that the model has good generalization ability within the sample space. To compare the results of this work, the experimental outcomes are contrasted with existing data generation methods in the literature. Zhou and Chiam [14] proposed a synthetic data generation strategy for the MNIST regression task, with an MAE of

1.179 and an RMSE of 1.978, significantly outperforming traditional random sampling and generator methods. Although this method demonstrates significant performance in simpler tasks, its limitation lies in its primary applicability to simplified datasets and tasks. It has not been sufficiently validated for complex process control and high-dimensional data generation in multi-variable scenarios. In contrast, the proposed method optimizes the generation of multi-dimensional, multi-variable parameters in the drying process. The model achieves high R^2 values on the training and test sets, indicating strong adaptability to real process data. Barbierato et al. [15] proposed a synthetic dataset generation method for a simple dataset with six features, evaluating the impact of mutual information between the features. This method focuses on single-dimensional data generation and adjusting feature importance, making it suitable for relatively simple decision-making problems. Compared to this, the proposed model can generate more complex multi-dimensional data while considering the interdependencies among process parameters in the drying process, offering higher industrial application value. By validating the model's fitting accuracy across a broader sample space, the proposed model demonstrates better scalability and reliability than the method of Barbierato et al. Zhang and Mikelsons [16] presented a data generation algorithm. This algorithm performed well in calibration models, achieving a mean relative error rate of 1.62% for all parameter estimates and an MAE of 0.108°C for the calibration model's output. This method effectively improves calibration accuracy and reduces calibration time. However, its core focus is on model calibration, primarily applied to a few variables such as temperature, limiting its use in complex multi-variable drying control problems. In contrast, the proposed generation model demonstrates high fitting accuracy and handles multiple process parameters in the drying process. Moreover, significance analysis is conducted on the generated data, verifying its reliability and statistical consistency with the original data. This ensures the effectiveness and practicality of the generated data for more complex control tasks.

Through comparison, it is evident that existing methods have achieved good results in simpler tasks or specific control scenarios. However, they often face challenges such as inadequate adaptability, poor interpretability, or a disconnect between generated data and actual process data when addressing multi-dimensional, complex process data. The BPNN-based data generation model for drying machine process parameters with insufficient sample space proposed in this work effectively expands the training sample space. Meanwhile, it improves the generation quality of multi-variable parameters and maintains high fitting accuracy across a broader sample space. Through comprehensive analysis of the training and test sets, the model demonstrates strong generalization ability and robustness.

However, despite the strong performance of the proposed method in several aspects, there are still some limitations. First, although the proposed model performs well with existing process parameter data, its generation capabilities may be limited when facing entirely new or unseen process conditions. Particularly, in extreme or special process scenarios, the generated data may not fully reflect the details of the actual process. Therefore, future work could focus on enhancing the adaptability of the generation model to handle more variable and complex process parameter spaces. Second, despite improving the quality of generated data through optimization of the generator network structure, the interpretability of the data generation process remains a challenge. In practical applications, ensuring strong interpretability between the generated data and the process parameters in actual production remains a crucial area for future research.

Overall, the proposed model has demonstrated good performance in enhancing the accuracy and stability of the drying process control. Compared to existing methods, the proposed model better accommodates the multidimensional and multivariable data generation requirements of the drying process, showing high fitting accuracy on both the training and testing datasets. Nevertheless, the generative model has certain limitations when dealing with extreme process conditions or special data requirements. Future research should focus on further optimizing the adaptability and interpretability of the data generation algorithm to better support drying control in actual production settings.

In practical applications, although the proposed method has yielded favorable results in generating drying process parameter data, significant differences in the process characteristics of different industrial processes may exist. For example, some industrial processes may involve more complex nonlinear relationships or multivariable interactions, which could increase the difficulty of model training. In these cases, further adjustments to the network structure of the generator and discriminator may be necessary to accommodate specific data distributions and process requirements. Additionally, data distributions under different operating conditions may vary. Factors such as changes in the production environment, equipment aging, and raw material differences can all affect the data generation and prediction performance. The model's stability and robustness may be tested under these variations, making it critical to adjust model parameters in response to new operating conditions to maintain the quality of generated samples. In this regard, adaptive training strategies or online learning methods may need to be introduced to allow the generative model to adapt to environmental changes in real-time. Moreover, the quality and quantity of the data are also significant factors affecting the model's generalization ability. The data generation model used in this work assumes that the input data is of high quality and sufficiently diverse. However, in some industrial scenarios, data may be missing, noisy, or incomplete, which can affect model training and prediction outcomes. Therefore, additional denoising and data supplementation strategies may need to be incorporated during the data preprocessing stage to ensure data quality and improve the model's generalization ability. In conclusion, although the proposed model performs well in drying control, it still requires appropriate adjustments and optimization when applied to other industrial processes or under different operating conditions.

5 Conclusion

This work proposes a BPNN-based data generation model to address the insufficient sample space for drying machine process parameters. It demonstrates the model's effectiveness in expanding the training sample space and improving drying control through experiments. The following conclusions are drawn. (1) The constructed data generation model achieves a fitting R^2 of 0.93705 on the training set and an R^2 of 0.9434 on the test set. This indicates that the model performs well in the training set and maintains high predictive accuracy in the test set. (2) By integrating data from the training and test sets, the model maintains a high fitting accuracy across a broader sample space, with an R^2 value of 0.93915. This reveals that the model effectively captures patterns in the training data and extends these patterns to unseen test data, ensuring the quality and authenticity of the generated samples. (3) Significance analysis shows that the H values for all process parameters are 0, and the p-values are greater than 0.05. It illustrates that generated dataset shows no significant statistical difference from the original, validating the model's reliability.

The proposed BPNN-based data generation model has achieved relatively ideal experimental results in generating and controlling drying machine process parameter data. However, its adaptability under diverse production conditions still requires further research and validation. In actual production environments, process conditions often experience significant fluctuations, such as variations in raw materials, equipment aging, and environmental temperature changes, which may impact the model's performance. Therefore, future research could consider incorporating a wider range of production conditions for validation, especially under high noise, dynamic changes, or extreme operating conditions, to comprehensively assess the model's stability, adaptability, and generalization ability. For instance, testing the model's performance under different seasons, production batches, or equipment scenarios could further validate its effectiveness in complex and dynamic production environments.

Furthermore, the proposed method primarily relies on a fixed process parameter dataset, whereas different industrial processes, equipment, and production lines may exhibit substantial differences in parameter settings, production environments, and data characteristics. Hence, the model's cross-domain adaptability can be an important focus of future research. To address this, it should explore methods such as multi-task learning and transfer learning to enhance the model's self-adaptive ability under varying process conditions and production environments, thereby handling the heterogeneity of different industrial processes. Additionally, the proposed model has not been thoroughly validated in scenarios where data quality is lacking or missing. In practical applications, data may be incomplete, missing, or subject to high noise interference, which can significantly affect the model's training outcomes. Future research could explore data preprocessing, missing data imputation, and adaptive algorithms to improve the model's robustness under conditions of unstable data quality.

Additionally, although the proposed model performs well under conditions of high-quality data, its robustness still needs further improvement in environments with poor data quality, high missing values, or significant noise. Therefore, future research could focus on exploring ways to enhance the model's performance under incomplete data conditions by introducing data preprocessing techniques, missing data imputation, or adaptive algorithms. These approaches can help the model address more complex real-world application scenarios, improving its stability and accuracy.

Regarding the selection of model architecture, this work employs the traditional BPNN architecture, which performs well in handling nonlinear data. However, for more complex and dynamic data, other architectures, such as deep CNNs or long short-term memory networks, may perform better in capturing temporal relationships and local features. Therefore, future research should consider a comparison and selection of various neural network architectures, examining the strengths and weaknesses of different architectures under various process conditions. Meanwhile, it should discuss the specific constraints faced in architecture selection, such as computational resources, training time, and model complexity.

In conclusion, future research could further deepen in the following areas. First, the model's adaptability testing is expanded to cover a wider range of production conditions, particularly considering the impact of complex environments such as high noise and dynamic changes on model performance. Second, cross-domain data generation and optimization issues under different process conditions and production environments are considered to enhance the model's adaptability to heterogeneous data. Finally, the challenges of unstable data quality are addressed by further exploring optimizing data preprocessing, and advanced missing data imputation techniques are adopted to strengthen the model's robustness and stability. With these improvements, future work could further enhance the model's adaptability and robustness, providing more reliable and universal technical support for data generation and process control in intelligent systems.

References

[1] Luo, D., Li, Y., Tang, S., Liu, A., & Zhang, L. (2022). The Tobacco Leaf Redrying Process Parameter Optimization Based on IPSO Hybrid Adaptive Penalty Function. *Processes*, 10(12), pp. 2747. https://doi.org/10.3390/pr10122747

[2] Losso, K., Cardini, J., Huber, S., Kappacher, C., Jakschitz, T., Rainer, M., & Bonn, G. K. (2022). Rapid differentiation and quality control of tobacco products using direct analysis in real time mass spectrometry and liquid chromatography mass spectrometry. *Talanta*, 238, pp. 123057. https://doi.org/10.1016/j.talanta.2021.123057

[3] Maiyo, A. K., Kibet, J. K., & Kengara, F. O. (2023). A review of the characteristic properties of selected tobacco chemicals and their associated etiological risks. *Reviews on Environmental Health*, 38(3), pp. 479-491. https://doi.org/10.1515/reveh-2022-0013

[4] Lingayat, A., Zachariah, R., & Modi, A. (2022). Current status and prospect of integrating solar air heating systems for drying in various sectors and industries. *Sustainable Energy Technologies and Assessments*, 52, pp. 102274. https://doi.org/10.1016/j.seta.2022.102274

[5] Hecht, S. S., & Hatsukami, D. K. (2022). Smokeless tobacco and cigarette smoking: chemical mechanisms and cancer prevention. *Nature Reviews Cancer*, 22(3), pp. 143-155. https://doi.org/10.1038/s41568-021-00423-4

[6] Zare, A., & Payvandy, P. (2023). The prediction of optimal conditions for the surface grafting of β-cyclodextrin onto silk fabrics by an artificial neural network (ANN). *Pigment & Resin Technology*, 52(2), pp. 183-191. https://doi.org/10.1108/PRT-08-2021-0090

[7] Wu, J., & Liu, Y. (2024). Optimization Design of Hull Compartment Structure Based on 3D Modeling. *Informatica*, 48(19), pp. 159-178. https://doi.org/10.31449/inf.v48i19.6432

[8] Evandari, K., Soeleman, M. A., & Pramunendar, R. A. (2023). BPNN Optimization With Genetic Algorithm For Classification of Tobacco Leaves With GLCM Extraction Features. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 7(2), pp. 293-301. https://doi.org/10.29207/resti.v7i2.4743

[9] Alighaleh, P., Khosravi, H., Rohani, A., Saeidirad, M. H., & Einafshar, S. (2022). The detection of saffron adulterants using a deep neural network approach based on RGB images taken under uncontrolled conditions. *Expert Systems with Applications*, 198, pp. 116890. https://doi.org/10.1016/j.eswa.2022.116890

[10] Kanjanawanishkul, K. (2022). An image-based eri silkworm pupa grading method using shape, color, and size. *International Journal of Automation and Smart Technology*, 12(1), pp. 2331-2331. https://doi.org/10.5875/ausmt.v12i1.2331

[11] Mu, L., Bi, S., Yu, S., Liu, X., & Ding, X. (2022). An intelligent moisture prediction method for tobacco drying process using a multi-hierarchical convolutional neural network. *Drying Technology*, 40(9), pp. 1791-1803. https://doi.org/10.1080/07373937.2021.1876722

[12] Odabas, M. S., Şenyer, N., & Kurt, D. (2023). Determination of quality grade of tobacco leaf by image processing on correlated color temperature. *Concurrency and Computation: Practice and Experience*, 35(2), pp. e7506. https://doi.org/10.1002/cpe.7506

[13] Li, Z., Feng, Z., Zhang, Z., Sun, S., Chen, J., Gao, Y., ... & Wu, Y. (2024). Analysis of energy consumption of tobacco drying process based on industrial big data. *Drying Technology*, 42(2), pp. 307-317. https://doi.org/10.1080/07373937.2023.2288667

[14] Zhou, T., & Chiam, K. H. (2023). Synthetic data generation method for data-free knowledge distillation in regression neural networks. *Expert Systems with Applications*, 227, pp. 120327. https://doi.org/10.1016/j.eswa.2023.120327

[15] Barbierato, E., Vedova, M. L. D., Tessera, D., Toti, D., & Vanoli, N. (2022). A methodology for controlling bias and fairness in synthetic data generation. *Applied Sciences*, 12(9), pp. 4619. https://doi.org/10.3390/app12094619

[16] Zhang, Y., & Mikelsons, L. (2023). Sensitivity-guided iterative parameter identification and data generation with BayesFlow and PELS-VAE for model calibration. *Advanced Modeling and Simulation in Engineering Sciences*, 10(1), pp. 9. https://doi.org/10.1186/s40323-023-00246-y

[17] Losso, K., Cardini, J., Huber, S., Kappacher, C., Jakschitz, T., Rainer, M., & Bonn, G. K. (2022). Rapid differentiation and quality control of tobacco products using direct analysis in real time mass spectrometry and liquid chromatography mass spectrometry. *Talanta*, 238, pp. 123057. https://doi.org/10.1016/j.talanta.2021.123057

[18] Li, M. (2024). Application of GAN-Based Data Encryption Technology in Computer Communication System. *Informatica*, 48(15), pp. 17-34. https://doi.org/10.31449/inf.v48i15.6390

[19] Chen, P., Liu, H., Xin, R., Carval, T., Zhao, J., Xia, Y., & Zhao, Z. (2022). Effectively detecting operational anomalies in large-scale IoT data infrastructures by using a GAN-based predictive model. *The Computer Journal*, 65(11), pp. 2909-2925. https://doi.org/10.1093/comjnl/bxac085

[20] Kiš, D., Budimir, A., Svitlica, B., Kalambura, S., & Kujundžić, S. (2024). Analysis of Tobacco Drying with Different Energy Sources. *Tehnički vjesnik*, 31(3), pp. 701-705. https://doi.org/10.17559/TV-20230428000586

[21] Zong, J., He, X., Lin, Z., Hu, M., Xu, A., Chen, Y., ... & Zou, C. (2022). Effect of two drying methods on chemical transformations in flue-cured tobacco. *Drying Technology*, 40(1), pp. 188-196. https://doi.org/10.1080/07373937.2020.1779287

[22] Khudyakov, D. A., Shorstkii, I. A., Ulyanenko, E. E., & Gnuchykh, E. V. (2022). Influences of cold atmospheric plasma pretreatment on drying kinetics, structural, fractional and chemical characteristics of tobacco leaves. *Drying Technology*, 40(15), pp. 3285-3291. https://doi.org/10.1080/07373937.2021.2021230

[23] Ouyang, L. (2024). Financial Risk Control of Listed Enterprises Based on Risk Warning Model. *Informatica*, 48(11), pp. 125-132. https://doi.org/10.31449/inf.v48i11.6026

[24] Lai, H. (2024). Predicting the Growth Value of Technology Enterprises with an Optimized Back-propagation Neural Network. *Informatica*, 48(16), pp. 105-112. https://doi.org/10.31449/inf.v48i16.6437