Hindawi
Scientific Programming
Volume 2023, Article ID 9873894, 1 page
https://doi.org/10.1155/2023/9873894

Retraction

Retracted: Rapid Identification of Tobacco Mildew Based on Random Forest Algorithm

Scientific Programming

Received 18 July 2023; Accepted 18 July 2023; Published 19 July 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] Z. Jiang, W. Zhang, H. Huang et al., "Rapid Identification of Tobacco Mildew Based on Random Forest Algorithm," *Scientific Programming*, vol. 2022, Article ID 1818398, 10 pages, 2022.

Hindawi
Scientific Programming
Volume 2022, Article ID 1818398, 10 pages
https://doi.org/10.1155/2022/1818398

Research Article

Rapid Identification of Tobacco Mildew Based on Random Forest Algorithm

Zhimin Jiang,[1] Wenjun Zhang,[2] Haixia Huang,[3] Zhengguang Zhai,[2] Dairong chen,[4] Yongfeng Ai,[4] Bo Li,[1] and Xiaoxiang Chen [1]

[1]*China Tobacco Zhejiang Industry Co Ltd., Hangzhou 310008, Zhejiang, China*
[2]*Hunan Tobacco Corporation Changsha Company, Changsha 410007, Hunan, China*
[3]*Guangxi Zhuang Autonomous Region Tobacco Corporation Baise Company, Baise 533000, Guangxi, China*
[4]*Guizhou Tobacco Corporation Tongren Company, Tongren 0856, Guizhou, China*

Correspondence should be addressed to Xiaoxiang Chen; ql@bbc.edu.cn

Received 19 August 2022; Revised 27 August 2022; Accepted 1 September 2022; Published 24 September 2022

Academic Editor: Lianhui Li

In order to further improve the identification efficiency of tobacco mildew, a rapid identification model of tobacco mildew based on random forest algorithm was proposed in this study. In order to ensure the feasibility and pertinence of the model study, this study takes redried leaf tobacco as the research object, selects high-temperature and high-humidity environment as the experimental conditions, and obtains the sample data of the degree of tobacco mildew under different experimental conditions. At the same time, this paper constructs a rapid identification model of tobacco mildew with the help of random forest algorithm. Through the model experimental results, it is found that the accuracy of the model for the rapid identification of training samples can reach 93.82%, while the accuracy of independent testing is 94.84%. The experimental results fully reflect the availability and efficiency of the random forest algorithm model in the rapid identification of tobacco mildew.

1. Introduction

Tobacco leaf is a special leaf plant. In most cases, tobacco leaves need to be stored for a period of time after harvest, and tobacco leaf storage is directly related to the quality of tobacco leaves, because the tobacco leaves in the mature stage can only enter the cigarette production after being harvested through special processes such as baking, purchasing, transportation, redrying, and aging. However, it takes a long time for the tobacco leaves to be harvested and made into finished cigarettes. If the tobacco leaves are not properly managed in the storage process, it is easy to mildew and then form mold. Mold is a relatively wide range of fungal microorganisms. Its growth environment is relatively simple. As long as it meets the requirements of appropriate temperature and humidity, it will quickly reproduce, and then the tobacco leaves will become moldy, which will seriously affect the quality of tobacco leaves and even cause the deterioration of tobacco leaves. Based on this background, this paper starts with the process and key points of tobacco mildew and combines random forest algorithm to identify tobacco mildew more accurately and efficiently.

2. Literature Review

There are many related research works on the detection technology of the internal components of tobacco leaves after mildew, which can be roughly divided into the following types: using the near-infrared technology to detect specific components, such as using the near-infrared spectroscopy technology to establish and verify the quantitative prediction model of ergosterol, establish the GBA algorithm, screen the characteristic wavelengths of the basic spectral data, and establish the PLS-DA discrimination model [1]. The model is applied to the identification of tobacco leaf samples, and the accuracy is as high as 95.79%. The characteristics of grain respiration and microbial activity were explored by using carbon dioxide technology detection,

reserve pest detector, and colony counting method, so as to obtain the characteristics of carbon dioxide gas produced by grain respiration and microbial activity, respectively. Amino acid analyzer was used to determine the specific content of free amino acids in moldy tobacco leaves. At the same time, the changes of amino acid content were divided into two categories: the absolute content increased and the absolute content decreased. GC-MS technology was used to determine the changes of volatile components in tobacco before and after mildew. Some studies have shown that substances with large changes are obtained by comparing before and after mildew, and then the content changes of these substances are used to establish a model, so as to infer whether it is mildew [2].

At present, the most widely used control technology is chemical control, which uses a variety of reagents. The most common antifungal agents in food include benzoic acid, sodium benzoate, sorbic acid, and propionic acid, but the mechanism of action is too small, and harmful gases may be generated after high temperature. Therefore, only staying in the previous research on antimildew agents cannot meet the needs of modern social development and people's attention to health. For this reason, many researchers began to turn to other chemical control agents with less toxic and side effects and better effects [3]. Some scholars pointed out that dimethyl fumarate has the strongest inhibitory effect on mold in tobacco leaves. Some researchers selected four kinds of fungicides, 75% dakonine, 40% hexin, 98% kangzhuolin, and 100% dimethyl fumarate, for toxicological and antimildew tests. The results showed that, through bioassay, 40% nucleostar had the best antibacterial spectrum and antibacterial effect, and dimethyl fumarate was the second. Kangzhuoling is also a safe and efficient biological fungicide. Some scholars have used organic acid mold inhibitor KMC-LF2 to conduct mold proof test on corn and found that it has good mold proof effect [4]. In recent years, some researchers have also studied the control of microwave technology. Some scholars have studied the effects of pulse microwave on the mortality of rice weevil and parasitic *Aspergillus*, rice temperature, broken rice rate, and burst waist rate of rice and their sensory quality. The results showed that, with the increase of pulse microwave intensity, the mortality of rice weevil and parasitic *Aspergillus* increased, the hatching rate of insect eggs decreased significantly, the sensory quality remained basically unchanged, the rice temperature increased gradually, and the broken rice rate and waist burst rate also increased [5].

3. Introduction to Random Forest Algorithm

Random forest (RF) is an aggregation distribution algorithm that contains multiple decision trees and poll strategies. The main idea of RF is to select randomly (not all) vectors to grow a tree in a class, and the only difference between them when designing a tree is a small number of differences [6]. In other words, implementation variables and patterns are randomly divided. This random number is called random existence because it is used in division or regression analysis. The final decision tree is generated by voting through the potential random vector tree; that is, select the "class" with the most votes as the category of the corresponding sample [7].

3.1. Decision Tree. Deciduous trees are formed by root nodes, leaves, and petals. The algorithm can be seen from the 1970s and 1980s. The most commonly used algorithms are the ID3 logging algorithm, the C4.5 logging algorithm, and the push logging algorithm [8].

3.1.1. ID3 Decision Tree Algorithm. ID3 algorithm is based on entropy in information theory. Suppose that an event has k optional results, and the probability of each result is shown in the following formula:

$$P_i(i = I, ..., k). \tag{1}$$

After observing the result of this event, its information is described by entropy, and the definition is shown in the following formula:

$$I = -\left(P_1 \log_2 P_1 + P_2 \log_2 P_2 + ... + P_k \log_2 P_k\right) = -\sum_{i=1}^{k} P_i \log_2 P_i. \tag{2}$$

If the characteristic divides N samples into m parts and there are N_m samples in each part, the impurity reduction is expressed by the two following formulas:

$$\Delta I(N) = I(N) - \left(P_1 I(N_1)\right) + \left(P_2 I(N_2)\right) + ... + \left(P_m I(N_m)\right), \tag{3}$$

$$P_m = \frac{N_m}{N}. \tag{4}$$

First, calculate the entropy purity of the samples in the current leaf node, split the current node with different features, compare the information gain in the split node, that is, the uncertainty reduction (3), and take the feature with the maximum information gain as the best node feature. If the subsequent nodes only include one type of samples, the branches and leaves stop growing, and the final node is called a leaf node. If the subsequent nodes include different kinds of sample sets, continue to iterate the above steps until each branch reaches the leaf node [9].

3.1.2. C4.5 Decision Tree Algorithm. C4.5 algorithm uses information gain rate replacement formula to obtain information gain, as shown in the following formula:

$$\Delta I_R(N) = \frac{\Delta I(N)}{I(N)}. \tag{5}$$

Moreover, C4.5 algorithm can also solve the characteristic problems with continuous values. The basic principle is as follows: assuming that the numerical characteristic x has n values in the training sample, arrange these values in order from small to large, and obtain the following formula:

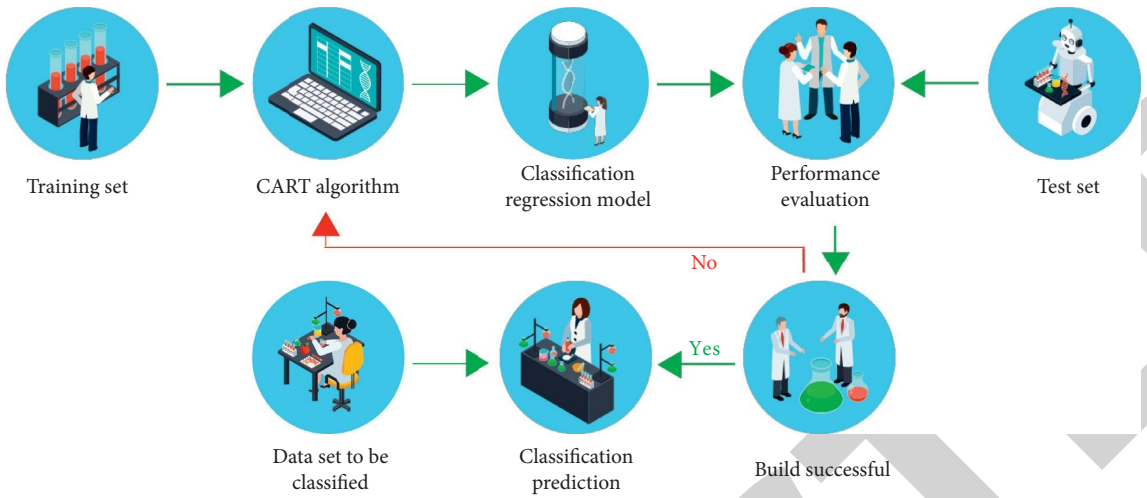

Figure 1: Flow chart of constructing classification regression tree.

Table 1: Advantages and disadvantages of decision tree.

Advantages
1. There is no need to introduce a priori assumption.
2. The decision tree is relatively simple to understand, its logical thinking is easier for people to understand, and it is easy to realize in practical application.
3. It has good stability for outliers and noise. The decision tree does not classify the data according to the specific value, and the outliers in the data have little impact on the whole result.
Disadvantages
1. In the process of top-down recursive construction, nodes store less and less information, and too little information will cause data fragmentation.
2. In the process of modeling, the unstable splitting of leaf nodes will cause overfitting.

$$v_i (i = I, \ldots, n). \tag{6}$$

Using dichotomy to divide the array, there are $n-1$ kinds of division methods. The information gain rate of each partition method is calculated. The continuous feature vector is changed into binary feature by selecting the method with the maximum gain rate, and then the decision tree is constructed together with other nonnumerical features. For the problem of feature discretization into multiple numerical values, the principle of the algorithm is the same; just increase the number of division methods.

3.1.3. Cart Decision Tree Algorithm. When constructing the classification regression tree model, we first randomly divide the sample set into training set and test set and then use cart algorithm to build the model in the training set, so as to obtain the classification regression model. Then we evaluate its performance through the test set. After successful construction, we apply this model to classify and predict the data of unknown categories. If the model construction fails, we return to the modeling process again and then conduct effective modeling and analysis through appropriate eigenvalues until the model is successfully constructed. Figure 1 shows the flow diagram of constructing classification regression tree [9].

Compared with neural network, support vector machine, and Bayesian algorithms, decision tree has its own advantages and disadvantages, as shown in Table 1.

3.2. Random Forest Model. Random forest is a model established by using a large number of decision trees (7).

$$\{h(x, \theta_k), k = 1, \ldots, K\}, \tag{7}$$

where $\{\theta_k\}$ represents a random vector subject to independent identically distributed, K is the number of decision trees, and each tree is the known variable x for optimal voting.

For the training sample set (8), N is the total number of samples, the object in X has an M-dimensional feature vector, and Y includes F different categories of information.

The classifier overemphasizes the classification of training samples, which makes the prediction of test samples worse. This phenomenon is called overfitting [10]. In the process of model construction, a variety of error analysis will be introduced. For random forest, generalization error is the key point to describe the overfitting problem. The generalization error is described by the edge function as follows:

$$mg(X, Y) = av_k I(h_k(X) = Y) - \max_{j \neq Y} av_k I(h_k(X) = j). \tag{8}$$

This function expresses the difference between the average value of correct votes obtained by correctly classifying random vector X into Y and the average value of votes obtained by other categories. The larger the function value is, the higher the classification accuracy is. Then the generalization error is defined as follows:

$$PE^* = P_{X,Y}(mg(X,Y) < 0). \tag{9}$$

According to the inference of the theorem of large numbers, the generalization error will eventually converge to the extreme value as the number of decision trees k increases, as shown in the following expression:

$$\lim_{k \longrightarrow \infty} PE^* = P_{X,Y}\left(P_\theta(h(X,\theta) = y) - \max_{j \neq Y} P_\theta(h(X,\theta) = j) < 0\right). \tag{10}$$

By solving the expected value of the edge function and the inference of Chebyshev inequality, an upper bound of the generalization error can be obtained as follows:

$$PE^* = \frac{\overline{\rho}(1-s)^2}{s^2}, \tag{11}$$

where $\overline{\rho}$ represents the mean value of the correlation factor ρ between trees and s represents the performance strength of the classifier. In order to better analyze the final performance of the classifier, the correlation factor and performance intensity are described by $c/s2$ ratio, and the smaller the ratio, the better the effect of the classifier, which is defined as follows:

$$\frac{c}{s2} = \frac{\overline{\rho}}{s^2}. \tag{12}$$

3.3. Random Forest Algorithm Regression Model. Suppose that the training set is taken from the distribution of random variables X and Y, which is similar to many predictors (see the following equation):

$$E_{XY}(Y - h(X))^2. \tag{13}$$

The above formula is the mean square random error, where $h(X)$ is the predicted value of the classifier. Since the predicted value is the average value of all trees, the form of mean square random error is as follows:

$$E_{XY}(Y - av_k h(X,\theta_k))^2. \tag{14}$$

When the total number of trees K is increasing, it is finally expressed as follows:

$$E_{XY}(Y - av_k h(X,\theta_k))^2 \longrightarrow E_{XY}(Y - E_\theta h(X,\theta_k))^2. \tag{15}$$

The regression function is expressed as follows:

$$Y = E_\theta h(X,\theta_k). \tag{16}$$

In practical application, it is often considered that the value of K is large enough and is replaced by the following:

$$Y = av_k h(X,\theta_k). \tag{17}$$

The average generalization error PE can be expressed in the form of a single tree as follows:

$$PE(\text{tree}) = E_\theta E_{XY}(Y - h(X))^2. \tag{18}$$

Table 2: Update of incremental random forest model.

Algorithm 3.3 incremental random forest←Update (x, y)
Steps:
1: For $t < 1$ to T do
2: For k←1 to P = Poisson (1) do
3: l = navigateToLeaf (x);
4: updateLeaf (l, (x, y));
5: If shouldSplit (l) then
6: Arg $\max_{\text{deD}} \Delta L$ (l,d);
7: createChild (l, d);
8: Endif
9: if P←0 then
10: $OOBE_t$←updateOOBE (t, (x, y));
11: End if
12: If s drawn from bern ($OOBE_t$)
13: rebuildTree (T);
14: Endif
15: End for
16: End for

Table 3: Node splitting conditions.

Algorithm 3.4 split function shouldSplit (l)
Steps:
1: If diffCluster (l) < 2 then
2: Return *false*;
3: If ΔL (l, d) > $\alpha \forall d \in D$ then
4: Return *false*;
5: Return *true*;

Through mathematical derivation and calculation, the average generalization error RE(forest) of random forest has an upper bound, and the form is as follows:

$$RE(\text{forest}) \leq \overline{\rho} PE(\text{tree}), \tag{19}$$

where $\overline{\rho}(0 < \overline{\rho} < 1)$ is the correlation factor. This formula shows that the generalization error of the whole random forest is $\overline{\rho}$ times that of a single decision tree, and rational design of the random forest model can better reduce the generalization error.

3.4. Incremental Random Forest. In the process of incremental random forest growth, we need to pay attention to two necessary factors: one is to sample the data set online, and the other is to reconstruct the splitting rules of nodes. Random forest generates a large number of test sets according to the characteristic randomness and then continuously calculates the quality measure of the test set to select the best splitting criterion. There are a large number of test sets at the nodes of each random tree when splitting. If the random forest randomly selects the set threshold instead of the threshold for a certain feature, it is called extreme random forest. The modeling time is long and the data cannot be updated in real time, but the accuracy of classification has been greatly improved. Tables 2 and 3 show the updating and node splitting conditions of the incremental random forest model.

This section selects 6 groups from UCI database 1 [9–11] as experimental data: Colon, Leukemia, Prostate, Lung,

Table 4: Comparison results of data classification.

UCI data	Training/testing	Category	Features	Random forest	Incremental random forest	OAB
Colon	6000/2150	2	62	0.107 ± 0.007	0.109 + 0.008	0.157 ± 0.010
Leukemia	7129/2500	2	72	0.128 ± 0.009	0.124 ± 0.008	0.188 ± 0.013
Prostate	6034/2300	2	102	0.109 ± 0.006	0.114 ± 0.006	0.215 ± 0.009
Lung	12533/4000	2	181	0.103 ± 0.008	0.108 ± 0.007	0.152 ± 0.011
Lymphoma	4026/1500	3	62	0.115 ± 0.010	0.118 ± 0.009	0.176 ± 0.013
SRBCT	2308/1000	4	63	0.113 ± 0.009	0.109 ± 0.006	0.189 ± 0.011

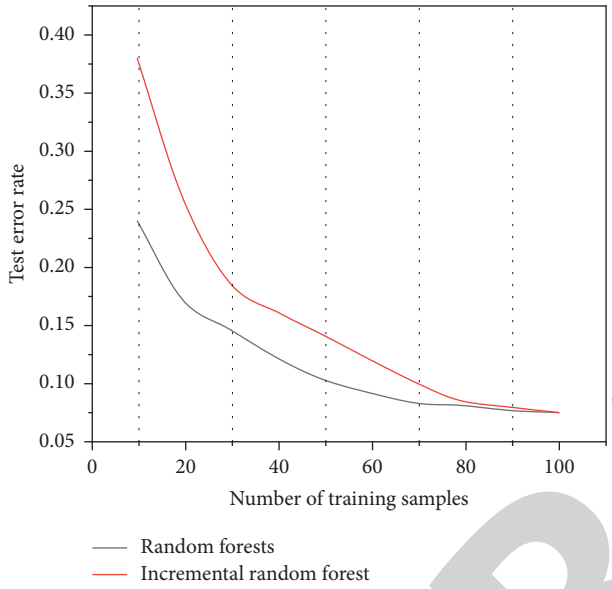

Figure 2: Effect of training sample number.

Lymphoma, and SRBCT. In the experiment, the total number of random trees is set to 100, and the number of randomly selected features is set to 10. In the experiment, the data are repeated 5 times and the average value is calculated. The standard deviation of classification error is shown in Table 4. The classification result of incremental random forest is very similar to that of original random forest. The prediction performance of the two methods is more accurate than that of OAB algorithm [11]. The main reason for the poor performance of OAB algorithm is that OAB algorithm can only solve the binary classification problem and cannot obtain the overall distribution of multifeature space in the data training process. Change the number of samples in the SRBCT data and compare the test error rates of the two methods. It can be concluded from Figure 2 that, with the gradual increase of the number of samples, the performance of the incremental random forest converges to the original algorithm.

4. Tobacco Mildew Image Recognition Based on Random Forest Algorithm and Neural Network

4.1. Simulation Experiment. In order to verify the effectiveness of the convolution neural network model for tobacco image classification, the tobacco image data set

Table 5: Experimental environment configuration.

Configuration	Content
CPU	Intel i7-8700
Computer memory	32 G
Computer operating system	Linux Ubuntu 19.10
GPU	NVIDIA GTX 1080 Ti
Deep learning framework	PyTorch

produced in this paper will be used to carry out experiments on the convolution neural network model and four classical convolution neural network models. During the whole experiment, each model was trained by multiple epochs to obtain the best parameters of the model, and the accuracy change curve of the training set and the test set was obtained, so as to visually see the effect and comprehensive performance of each model [12].

4.1.1. Experimental Environment and Data Set. The experiment in this chapter is carried out under the PyTorch framework. The specific software and hardware environment configuration of the computer used in the experiment is shown in Table 5. The main experiment environment is to use Ubuntu19.10 system + PyTorch deep learning framework + Python 3.6 and use the GPU with NVIDIA GeForce GTX 1080 Ti and 64g display memory to accelerate the experiment. The data set used in the experiment is a self-made tobacco leaf image data set. The tobacco leaf image data set mainly includes four categories of tobacco leaf image data: normal tobacco leaf, moldy tobacco leaf, green miscellaneous tobacco leaf, and variegated tobacco leaf. Each category has 3500 tobacco leaf images in the three categories of moldy tobacco leaf, green miscellaneous tobacco leaf, and variegated tobacco leaf. There are 4500 pieces of tobacco leaf data in the normal tobacco leaf category, and there are 15000 pieces of tobacco leaf image data in total. The tobacco leaf image data set is divided into 2 subsets in the ratio of 4 : 1, including 12000 training sets and 3000 test sets, so as to facilitate the use of experiments [13].

4.1.2. Experimental Design. In this paper, the tobacco leaf image data set is used as the experimental data, and the simulation experiments are carried out on four classical convolutional neural network models and the convolutional neural network model built in this paper. In the process of experiment, after many experiments, the model parameters and learning rate that can optimize the performance of the model are finally obtained. Then, they are applied to the

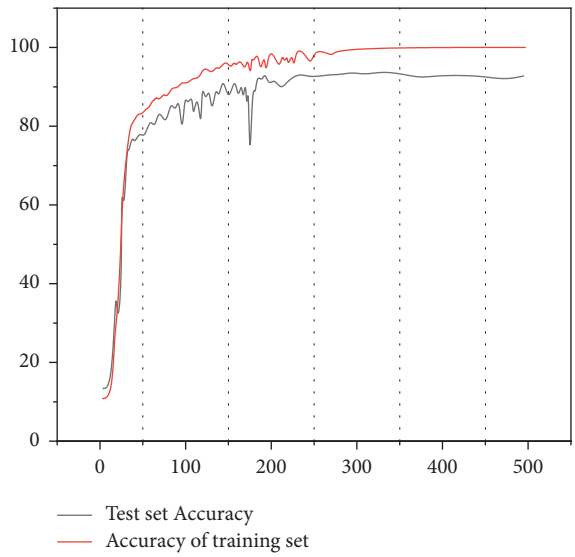

Figure 3: AlexNet model experiment results.

tobacco leaf image test set made in this paper for recognition [14]. During the experiment, the convolution neural network model has passed through 500 epochs, and the accuracy of the model test set is relatively stable. In addition, the visualization tool TensorboardX of PyTorch is used in the experiment. Its main function is to save and upload the data after model training to TensorboardX. TensorboardX is used to upload the coordinate data of the corresponding points of accuracy and generate csv files during model training. As the number of epochs increases, the accuracy change curve in the experiment is drawn. According to the change of the curve, we can more intuitively see the accuracy change and model effect of this test [15].

4.2. Analysis of Experimental Results

4.2.1. Results. Input the tobacco leaf image data set to AlexNet model for experiment. The network parameters are set as follows: the learning rate is set to 0.001, the dropout is set to 0.5, and the batch size is set to 64. This simulation experiment is based on 500 epochs and has been tested for many times. The accuracy change curve of AlexNet model simulation experiment is shown in Figure 3.

According to the experiment, the classification accuracy of AlexNet model test set is 95.5%. As can be seen from Figure 3, the convergence speed of the model is slow, and the accuracy of the model test set is relatively stable.

4.2.2. Result Analysis. We have made statistics on the test set loss value, training time of each epoch, and test set accuracy of the convolutional neural network model and four classical convolutional neural network models built in this paper [16]. The statistical results are shown in Table 6.

For the traditional image classification technology, this paper uses the HOG + SVM and HOG + KNN image classification methods. The main process is to first extract the HOG (histogram of oriented gradient) feature of the tobacco

Table 6: Comparison of experimental results of different convolutional neural network models.

Model	Loss	Time of each iteration (s)	Accuracy (%)
AlexNet	±0.002543	98	95.53
VGGNet16	±0.001822	120	96.23
GoogLeNet	±0.002015	132	96.56
ResNet18	±0.001555	115	97.36
This paper's model	±0.001424	85	98.08

Table 7: Comparison of experimental results of different algorithms.

Classification algorithm	Running time	Accuracy (%)
HOG + SVM	25 mins	71.36
HOG + KNN	13 mins	74.62
This paper's model	13 h	98.08

ptdata image and then use the support vector machine (SVM) algorithm and KNN algorithm for image classification [17]. The specific experimental process of traditional image classification method in this paper is as follows:

(1) HOG + SVM image classification method. Firstly, the moldy tobacco leaves, green miscellaneous tobacco leaves, and variegated tobacco leaves in the data set are taken as positive samples, and the normal tobacco leaves are taken as negative samples. Then the HOG features of positive and negative samples are collected to establish Feature Engineering, and then the support vector machine is used for image classification. As for support vector machine, LibSVM is used in this paper. It is a software library of support vector machine. It has the advantages of less input parameters and fast operation speed. It can easily classify or regress data.

(2) HOG + KNN image classification method. Firstly, take the moldy tobacco leaves, green miscellaneous tobacco leaves, and variegated tobacco leaves in the data set as positive samples and normal tobacco leaves as negative samples, and then collect the HOG characteristics of positive and negative samples to establish Feature Engineering [18]. For the KNN algorithm, this paper uses the KNeighborsClassifier function encapsulated by sklearn, where the K value is set to 3. The experimental results of the convolution neural network model built in this paper are compared with the experimental results of the traditional image classification methods. The comparison results are shown in Table 7.

The structure of the actual circulatory neural network formed when the image data generated in this form is used for image distribution. This text is much larger than the traditional image distribution method. In terms of runtime, the training time for curved neural network structures is much longer than that for traditional image distribution

methods, but traditional image distribution methods require more time to remove image features and establish functional engineering before splitting [19].

5. Fast Identification Experiment of Tobacco Mildew

5.1. Materials and Instruments. 116 kinds of redried leaf tobacco collected from 2015 to 2017 were selected from different places of origin, different parts, and different grades, so as to fully consider the different effects of different places of origin, different parts and grades, and different types and quantities of mold in tobacco leaves on the mildew of tobacco leaves. The samples were provided by a tobacco company. The instruments are as follows: TRH-1250 constant temperature and humidity box (for sample mildew test); MPA Fourier transform near infrared spectrometer; KBF constant temperature and humidity chamber (for microbial counting experiment); Double Biocao RNA/DNAultra clean workbench; Ba-2s flapping sterile homogenizer; MS204 balance; Milli-Q Integral 10 ultra pure water machine [20].

5.2. Preparation of Moldy Samples. Put the redried tobacco leaf sample under the environment of temperature $(22 \pm 2)^{\circ}$ C and humidity $(60 \pm 5)\%$ for 48 h. Put the balanced samples into the constant temperature and humidity box, adjust the temperature and humidity to 25°C and 85%, respectively, and carry out the tobacco mildew test. Take 21 days as the cycle, and take samples according to the following methods:

(1) The first sampling shall be carried out on day 0, that is, before putting the sample into a box with a constant temperature of 25°C and humidity of 85%.

(2) On the 3rd to 9th day, take the 2nd to 4th sampling, respectively, at the interval of 3 days.

(3) On the 11th to 21st day, take the 5th to 10th sampling, respectively, at 2-day intervals.

This process can completely collect the sample state that the redried tobacco leaves have never been mildewed to near mildewed and then to mildewed.

5.3. Mould Counting Test. According to YC/T 472-2013 microbiological examination of tobacco and tobacco products mold count, the mold count of redried tobacco samples with different degrees of mildew was detected. According to the mold count test results, the mold degree is divided into the following: nonmoldy samples (mold count $<2\times10^3$ CFU/g), adjacent moldy samples (2×10^3 CFU/g ≤ mold count $<10^4$ CFU/g), and moldy samples (mold count $\geq 10^4$ CFU/g).

5.4. Near-Infrared Spectrum Data Acquisition. MPA Fourier transform near-infrared spectrometer was used in the experiment. The spectrum acquisition range was 4000 ~12000 cm^{-1}, the resolution was 8 cm^{-1}, and the scanning times were 64. The tobacco samples with different degrees of mildew were loaded into the sample cup, and the near-infrared spectra of each sample were collected as the basic spectral information of each tobacco sample. Repeat the loading and determination for each sample twice, and then calculate the average result as the final spectrum [21].

Near-infrared spectroscopy is affected by a series of chemical and physical factors of samples. It is necessary to take mathematical pretreatment methods to reduce system noise, such as baseline change and light scattering. In this study, after comparing the first derivative (1-Der), second derivative (2-Der), multivariate scattering correction (MSC), standard normal variable (SNV) correction, and other preprocessing methods, discrete wavelet transform (DWT) is used to preprocess near-infrared spectral data [22].

The essence of wavelet transform is to decompose the signal into wavelet subspaces with different scales and frequencies. Choosing various mother wavelets according to the waveform or length makes wavelet transform more effective and flexible than other signal preprocessing methods in extracting the features of signals. Through wavelet transform, the signal is decomposed into low-frequency signal and high-frequency signal, approximation coefficient, and detail coefficient [23–29]. When wavelet transform is used to preprocess the near-infrared spectrum, there are two ways to establish the correlation model between independent variables and near-infrared signals: first is to reconstruct the spectrum with approximate coefficients and detail coefficients after denoising or data compression and to establish a model between the reconstructed spectrum and independent variables; second, the wavelet coefficients obtained by wavelet decomposition are directly used as variables to establish the model. Obviously, the latter method is much more convenient, time-saving, and widely used. In this study, wavelet coefficient was used as a variable to establish a prediction model for the mildew degree of redried tobacco leaves.

6. Results and Discussion

6.1. Classification of Moldy Tobacco Leaves. After the mildew experiment, 1160 tobacco samples with different degrees of mildew were obtained from 116 kinds of single flue-cured tobacco, which were sampled 10 times at different stages. Microbiological tests were carried out on 1160 samples according to YC/T 472-2013 microbiological examination of tobacco and tobacco products mold count. The research results of some scholars show that when the mold number reaches a certain amount (about 104 CFU/g), it will start to grow rapidly. In order to give early warning of tobacco mildew, it is necessary to prejudge the samples near mildew. Therefore, 104 CFU/g is taken as the critical point to judge the "near mildew" samples. According to the method of determining the degree of tobacco mildew, the tobacco mildew was finally divided into three categories: nonmildew, near mildew, and mildew. The three types of samples obtained from mildew test are shown in Table 8.

6.2. Near-Infrared Spectral Pretreatment. The original near-infrared spectra of 1160 redried tobacco samples with

Table 8: Classification of mildew degrees of tobacco leaf samples.

Mildewing degree	Sample quantity	Mold count (CFU/g)
Nonmildew	548	$<2\times10^3$
Near mildew	102	$[2\times103, 10^4)$
Mildew	510	$\geq10^4$

Table 9: Training results of random forest model based on different wavelet coefficients.

Wavelet coefficient	Variable length	Recognition rate/%	Recognition rate 1/%	Recognition rate 2/%	Recognition rate 3/%
cd1	1051	78.53	77.81	45.59	85.88
cd2	540	80.85	81.10	58.82	85.00
cd3	284	82.79	80.82	58.82	89.71
cd4	156	90.94	90.68	80.83	93.24
cd5	92	84.22	90.14	73.53	80.00
cd6	60	82.28	87.95	72.06	78.24
cd7	44	81.76	83.29	70.59	82.35
cd8	36	79.69	83.56	70.59	77.35
cd9	32	79.04	79.45	58.82	82.65
cd10	30	74.77	81.37	36.76	75.29
cd11	29	73.48	79.73	35.29	74.41
ca11	29	59.51	67.67	27.94	57.06
[cd4, cd5]	248	93.40	93.42	91.17	93.82
Primitive harmonic	2074	69.73	73.15	30.88	73.82

different degrees of mildew can be obtained from the original near-infrared spectra during the mildew process of redried tobacco. According to the principle of near-infrared spectroscopy, the near-infrared spectral absorption band of the sample is the frequency doubling and merging of hydrogen containing groups (O-H, N-H, C-H) in the mid infrared spectral region, as well as the superposition of differential absorption bands. When tobacco is mildewed, the organic substances such as C source and N source in the sample will change due to the catabolism of mold, and some chemical components related to the composition of mold cell wall such as ergosterol and chitin will be produced. Therefore, in theory, the near-infrared absorption bands of tobacco leaves with different degrees of mildew will change with the change of chemical composition in the sample. However, on the one hand, it is due to the serious overlap of near-infrared spectra; on the other hand, tobacco mildew is a complex process, and the changes of chemical components are also extremely complex. It is difficult to directly extract the information related to the degree of mildew from the near-infrared spectra of tobacco leaves and give a reasonable spectral analysis. It can be seen from the final results that the absorption bands related to mildew in the near-infrared spectra of tobacco leaf samples are difficult to be directly judged from the spectra.

In this study, discrete wavelet transform (DWT) was used to decompose the original near-infrared spectra of redried tobacco leaves in the process of mildew. When using DWT to process NIR spectra, there are two factors to be considered: the selection of mother wavelet and the determination of decomposition level. At present, there is no theory to follow for the selection of mother wavelets. The study investigated the influence of 15 mother wavelets, five Daubechies series wavelets (db2, db4, db6, db8, and db10), five Symlets series wavelets (sym2, sym4, sym6, sym8, and sym10) and five Coiflets series wavelets (coif1, coif2, coif3, coif4, and coif5), on the recognition accuracy of moldy tobacco leaves. The results of Daubechies series wavelets are basically the same, but, in general, the prediction results are better than those of the other two types of wavelets. Finally, db6 is determined as the mother wavelet. When determining the decomposition level, the dimension n of the input data should generally be considered, which generally does not exceed log2 (N). For the extraction of useful information, the decomposition level should be as large as possible. In this study, there are 2074 data points in the near-infrared spectrum, so this paper selects 11 as the decomposition level of wavelet transform. After each level of decomposition, a detail coefficient vector and an approximate coefficient vector are obtained. The approximation coefficient is further decomposed to obtain the detail coefficient and approximation coefficient until the 11th decomposition level. Finally, the vectors obtained by wavelet transform include the approximate coefficients (ca) of the spectral signal of each sample at the last decomposition level and the detail coefficients (cd) at all decomposition levels. A total of 12 groups of wavelet coefficients, ca11, cd11, cd10, ..., cd1, are obtained.

6.3. Establishment of Identification Model of Tobacco Mildew Degree. To compare the reception results of the different models, 1160 models were divided into training and test modes. Approximately 2/3 of the models are used as training, and 1/3 of the models are based on experiments. Finally, 773 models were selected for the training, of which 365 were standard, 68 were standard, and 340 were standard. The remaining 387 models were used as standardized

Table 10: Comparison of training results of cigarette model, cut tobacco model, cigarette end model, and comprehensive model.

Model	Total recognition rate/%	Recognition rate_1/%	Recognition rate_2/%	Recognition rate_3/%
Smoke model	93.40	93.42	91.17	93.82
Cut tobacco model	94.05	94.25	91.17	94.41
Smoke model	94.95	95.62	92.65	94.71
Integrated model	80.60	80.55	73.53	82.06

Table 11: Prediction results of test set based on [cd4, cd5] random forest model.

Parameter	Unmodified samples	Adjacent moldy samples	Moldy samples	Total
Number of test set samples	183	34	170	387
Accurately predict the number of samples	175	31	161	367
Correct prediction rate/%	95.63	91.18	94.71	94.84

experiments. This document categorizes mold-free designs, mold designs, and multimold designs as "1," "2," and "3," respectively. The accuracy of the training procedures and examination procedures is determined by the fees and assumptions.

It can be seen from Table 9 that the number of variables from cd1 to cd11 and ca11 and from 1051 to 29 decreases in turn. Among them, the random forest model constructed by the wavelet coefficient cd4 with a small number of variables (92) has the highest recognition rate: the total recognition rate is 90.04%, the recognition accuracy rate of nonmildewed samples is 90.68%, the recognition accuracy rate of near mildewed samples is 80.88%, and the recognition accuracy rate of mildewed samples is 93.24%, which are higher than the recognition ability of other wavelet coefficients.

As can be seen from Table 10, from the perspective of model accuracy, the prediction accuracy of the end of tobacco model > cut tobacco model > piece tobacco model > comprehensive model is similar to those of the end of tobacco model, cut tobacco model, and piece tobacco model.

In the measurement model, the flue gas model was used to calculate the sample size, and the calculated results are shown in Table 11. As shown in Table 11, 175 of the 183 models were correctly calculated, not moldy. The accuracy is 95.63%. Out of 34 samples close to the fungus, 31 samples were calculated correctly, with an accuracy of 91.18%; 161 out of 170 yeast samples were calculated correctly, with an accuracy of 94.71%. The accuracy of the test lumped forecast is 94.84%. The results showed that the model established by this method could effectively identify tobacco samples with different degrees of mildew.

7. Conclusion

In this study, the redried tobacco leaf was taken as the research object, the experimental conditions were high-temperature and high-humidity environment, and the sample data of tobacco mildew degree were obtained under different experimental conditions. In this paper, a rapid identification method of tobacco samples with different degrees of mildew was established by near-infrared spectroscopy, which provided a basis for early warning of tobacco mildew. The wavelet transform was used to process the spectral data, and [cd4, cd5] was selected as the spectral variable to establish the random forest recognition model of tobacco leaves with different degrees of mildew. The recognition rate and prediction rate of the model were 93.82% and 94.84%, respectively. The satisfactory recognition rates were achieved for the normal tobacco leaves, the adjacent moldy tobacco leaves, and the moldy tobacco leaves. It can be seen that near-infrared spectroscopy combined with wavelet transform and random forest algorithm can effectively identify tobacco samples with different degrees of mildew. This method can be considered feasible to quickly predict the degree of tobacco mildew.

Data Availability

The data set can be obtained from the corresponding author upon request.

Conflicts of Interest

The authors Zhimin Jiang, Bo Li, and Xiaoxiang Chen are affiliated to and funded by China Tobacco Zhejiang Industry Co., Ltd. The authors attest China Tobacco Zhejiang Industry Co., Ltd. has had no influence on design of this study or its outcomes.

The authors Wenjun Zhang and Zhengguang Zhai are affiliated to and funded by Hunan Tobacco Corporation Changsha Company. The authors attest Hunan Tobacco Corporation Changsha Company has had no influence on design of this study or its outcomes.

The authors Dairong Chen and Yongfeng Ai are affiliated to and funded by Guizhou Tobacco Corporation Tongren company. The authors attest Guizhou Tobacco Corporation Tongren company has had no influence on design of this study or its outcomes.

Acknowledgments

This research was funded by Project of China Tobacco Zhejiang Industry Co, Ltd. (ZJZY2021B009), Project of Hunan Tobacco Corporation Changsha Company (21-

23A04), and Project of Guizhou Tobacco Corporation Tongren company (202101).

References

[1] Y. A. Saadoon and R. H. Abdulamir, "Improved random forest algorithm performance for big data," *Journal of Physics: Conference Series*, vol. 1897, no. 1, Article ID 012071, 2021.

[2] J. Y. Kim, M. Lee, M. K. Lee et al., "Development of random forest algorithm based prediction model of alzheimer's disease using neurodegeneration pattern," *Psychiatry Investigation*, vol. 18, no. 1, pp. 69–79, 2021.

[3] S. Kim, G. K. Karahan, M. Sharma, and Y. Pachepsky, "The site-specific selection of the infiltration model based on the global dataset and random forest algorithm," *Vadose Zone Journal*, vol. 20, no. 3, 2021.

[4] S. l. Papineni, A. M. Reddy, S. Yarlagadda, S. Yarlagadda, and H. Akkineni, "An extensive analytical approach on human resources using random forest algorithm," *International Journal of Engineering Trends and Technology*, vol. 69, no. 5, pp. 119–127, 2021.

[5] Z. Tang, Z. Mei, W. Liu, and Y. Xia, "Identification of the key factors affecting Chinese carbon intensity and their historical trends using random forest algorithm," *Journal of Geographical Sciences*, vol. 30, no. 5, pp. 743–756, 2020.

[6] B. C. Kim, J. Kim, I. Lim, D. H. Kim, S. M. Lim, and S. K. Woo, "Machine learning model for lymph node metastasis prediction in breast cancer using random forest algorithm and mitochondrial metabolism hub genes," *Applied Sciences*, vol. 11, no. 7, p. 2897, 2021.

[7] W. Lin, W. Fan, H. Liu, Y. Xu, and J. Wu, "Classification of handheld laser scanning tree point cloud based on different knn algorithms and random forest algorithm," *Forests*, vol. 12, no. 3, p. 292, 2021.

[8] X. Dong, Z. Meng, Y. Wang, Y. Zhang, H. Sun, and Q. Wang, "Monitoring spatiotemporal changes of impervious surfaces in beijing city using random forest algorithm and textural features," *Remote Sensing*, vol. 13, no. 1, p. 153, 2021.

[9] L. Wan, K. Gong, G. Zhang, X. Yuan, and X. Deng, "An Efficient Rolling Bearing Fault Diagnosis Method Based on Spark and Improved Random forest Algorithm," *IEEE Access*, vol. 99, p. 1, 2021.

[10] J. Liang, "Problems and solutions of art professional service rural revitalization strategy based on random forest algorithm," *Wireless Communications and Mobile Computing*, vol. 2022, no. 1, pp. 1–11, Article ID 9752512, 2022.

[11] Y. Azhar, G. A. Mahesa, and M. C. Mustaqim, "Prediction of hotel bookings cancellation using hyperparameter optimization on random forest algorithm," *Jurnal Teknologi dan Sistem Komputer*, vol. 9, no. 1, pp. 15–21, 2021.

[12] T. Sui, J. Kan, T. Sun, and J. Liu, "Research on target classification method for dense matching point cloud based on improved random forest algorithm," *International Journal of Information and Communication Technology*, vol. 1, no. 1, p. 1, 2021.

[13] C. Zhang, C. Hu, S. Xie, and S. Cao, "Research on the application of decision tree and random forest algorithm in the main transformer fault evaluation," *Journal of Physics: Conference Series*, vol. 1732, no. 1, Article ID 012086, 2021.

[14] E. A. Urbanovich, D. A. Afonnikov, and S. V. Nikolaev, "Determination of the quantitative content of chlorophylls in leaves by reflection spectra using the random forest algorithm," *Vavilov Journal of Genetics and Breeding*, vol. 25, no. 1, pp. 64–70, 2021.

[15] M. A. Rasyidi, T. Bariyah, Y. I. Riskajaya, and A. D. Septyani, "Classification of handwritten Javanese script using random forest algorithm," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 3, pp. 1308–1315, 2021.

[16] X. D. Hoang and X. H. Vu, "An improved model for detecting dga botnets using random forest algorithm," *Information Security Journal: A Global Perspective*, vol. 31, no. 4, pp. 441–450, 2021.

[17] M. Onesime, Z. Yang, and Q. Dai, "Genomic island prediction via chi-square test and random forest algorithm," *Computational and Mathematical Methods in Medicine*, vol. 2021, no. 1, pp. 1–9, Article ID 9969751, 2021.

[18] G. Li, C. Wang, D. Zhang, and G. Yang, "An improved feature selection method based on random forest algorithm for wind turbine condition monitoring," *Sensors*, vol. 21, no. 16, p. 5654, 2021.

[19] M. Fan and A. Sharma, "Design and implementation of construction cost prediction model based on svm and lssvm in industries 4.0," *International Journal of Intelligent Computing and Cybernetics*, vol. 14, no. 2, pp. 145–157, 2021.

[20] J. Jayakumar, B. Nagaraj, S. Chacko, and P. Ajay, "Conceptual implementation of artificial intelligent based E-mobility controller in smart city environment," *Wireless Communications and Mobile Computing*, vol. 2021, pp. 1–8, Article ID 5325116, 2021.

[21] R. Huang, S. Zhang, W. Zhang, and X. Yang, "Progress of zinc oxide-based nanocomposites in the textile industry," *IET Collaborative Intelligent Manufacturing*, vol. 3, no. 3, pp. 281–289, 2021.

[22] Q. Zhang, "Relay vibration protection simulation experimental platform based on signal reconstruction of MATLAB software," *Nonlinear Engineering*, vol. 10, no. 1, pp. 461–468, 2021.

[23] L. Li, B. Lei, and C. Mao, "Digital twin in smart manufacturing," *Journal of Industrial Information Integration*, vol. 26, no. 9, Article ID 100289, 2022.

[24] L. Li, T. Qu, Y. Liu et al., "Sustainability assessment of intelligent manufacturing supported by digital twin," *IEEE Access*, vol. 8, pp. 174988–175008, 2020.

[25] L. Li and C. Mao, "Big data supported PSS evaluation decision in service-oriented manufacturing," *IEEE Access*, vol. 8, pp. 154663–154670, 2020.

[26] L. Li, C. Mao, H. Sun, Y. Yuan, and B. Lei, "Digital twin driven green performance evaluation methodology of intelligent manufacturing: hybrid model based on fuzzy rough-sets AHP, multistage weight synthesis, and PROMETHEE II," *Complexity*, vol. 2020, no. 6, pp. 1–24, Article ID 3853925, 2020.

[27] X. Zhao, X. Liu, J. Liu, J. Chen, S. Fu, and F. Zhong, "The effect of ionization energy and hydrogen weight fraction on the non-thermal plasma volatile organic compounds removal efficiency," *Journal of Physics D: Applied Physics*, vol. 52, no. 14, Article ID 145201, 2019.

[28] L. H. Li, J. C. Hang, Y. Gao, and C. Y. Mu, "Using an integrated group decision method based on SVM, TFN-RS-AHP, and TOPSIS-CD for cloud service supplier selection," *Mathematical Problems in Engineering*, vol. 2017, pp. 1–14, Article ID 3143502, 2017.

[29] L. H. Li, J. C. Hang, H. X. Sun, and L. Wang, "A conjunctive multiple-criteria decision-making approach for cloud service supplier selection of manufacturing enterprise," *Advances in Mechanical Engineering*, vol. 9, no. 3, 2017.