Chinese forest fire occurrence prediction based on machine learning methods

yudong Li
Beijing Forestry University
Zhongke Feng (zhongkefeng@bjfu.edu.cn)
Beijing Forestry University https://orcid.org/0000-0003-1602-5045
Ziyu Zhao
Beijing Forestry University
Wenyuan Ma
Beijing Forestry University
Shilin Chen
Beijing Forestry University
Hanyue Zhang
Beijing Forestry University

Research

Keywords: forest fire occurrence in China, feature selection, forest fire driving factors, machine learning, prediction model, forest fire prevention and control

Posted Date: January 12th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-62305/v2

[Title Page]

Chinese forest fire occurrence prediction based on machine learning methods

Yudong Li,[1] Zhongke Feng,[1] Ziyu Zhao,[1]Wenyuan Ma,[1]Shilin Chen,[1]Hanyue Zhang,[1]

1 Precision Forestry Key Laboratory of Beijing, Beijing Forestry University, Beijing 100083, China.

E-mail address ：

Yudong Li： lyd85842@163.com; **Zhongke Feng：** zhongkefeng@bjfu.edu.cn

Ziyu Zhao： zhaozy0315@126.com; **Wenyuan Ma:imawenyuan@126.com;**

Shilin Chen: chenshilin@bjfu.edu.cn; **Hanyue Zhang：hanyue.zhang@foxmail.com**

Correspondence should be addressed to Zhongke Feng: zhongkefeng@bjfu.edu.cn

Permanent address: Beijing Forestry University

[Manuscript]

Chinese forest fire occurrence prediction based on machine learning methods

Yudong Li,[1] Zhongke Feng,[1] Ziyu Zhao,[1] Wenyuan Ma,[1]Shilin Chen,[1]Hanyue Zhang,[1]

1 Precision Forestry Key Laboratory of Beijing, Beijing Forestry University, Beijing 100083, China.

Correspondence should be addressed to Zhongke Feng: zhongkefeng@bjfu.edu.cn

Abstract

Forest fires can cause serious harm. Scientifically predicting forest fires is an important basis for preventing them. Currently, there is little research on the prediction of long time-series forest fires in

China. Choosing a suitable forest fire prediction model and predicting the probability of Chinese forest fire occurrence are of great importance to China's forest fire prevention and control work. Based on fire hotspot, meteorological, terrain, vegetation, infrastructure, and socioeconomic data collected from 2003 to 2016, we used a random forest model as a feature-selection method to identify 13 major drivers of forest fires in China. The forest fire prediction models developed in this study are based on four machine-learning algorithms: an artificial neural network, a radial basis function network, a support-vector machine, and a random forest. The models were evaluated using the five performance indicators of accuracy, precision, recall, f1 value, and area under the curve. We used the optimal model to obtain the probability of forest fire occurrence in various provinces in China and created a spatial distribution map of the areas with high incidences of forest fires. The results showed that the prediction accuracy of the four forest fire prediction models was between 75.8% and 89.2%, and the area under the curve value was between 0.840 and 0.960. The random forest model had the highest accuracy (89.2%) and area under the curve value (0.96); thus, it was used as the optimal model to predict the probability of forest fire occurrence in China. The prediction results indicate that the areas with high incidences of forest fires are mainly concentrated in north-eastern China (Heilongjiang Province and northern Inner Mongolia Autonomous Region) and south-eastern China (including Fujian Province and Jiangxi Province). In areas at high risk of forest fire, management departments can improve forest fire prevention and control by establishing watch towers and using other monitoring equipment. This study helps in understanding the main drivers of forest fires in China, provides a reference for the selection of high-precision forest fire prediction models, and provides a scientific basis for China's forest fire prevention and control work.

Keywords: forest fire occurrence in China; feature selection; forest fire driving factors; machine learning; prediction model; forest fire prevention and control

1. Introduction

Forest fires are one of the most dangerous natural disasters. They have received worldwide attention due to their rapid spread, their low controllability, and the hazards they pose[1][2]. Forest fires have varying degrees of impact on human health and safety, the ecological environment and resources, and society and the economy[3]. Forest fire prevention has therefore become a key research topic in the fields of forestry and ecology[4][5][6].

The most effective way to control forest fires is to detect them quickly. Detection is usually divided into three categories: satellite monitoring, smoke detection, and local perception (such as data analysis). Satellite monitoring is expensive, is affected by delays, and is not fully applicable to all locations[7]. Smoke detection also requires expensive equipment and maintenance work. In contrast, data for forest fire analysis are easy to obtain, and data analysis is less expensive than the first two methods[8].

In recent years, researchers have mostly used data analysis methods to obtain the degree of risk of forest fires in the study area[9][10][11]. Many researchers have used geospatial technologies such as GIS and remote sensing (RS) and have used data analysis models to analyse and evaluate an area's sensitivity to forest fire[12][13][14]. Forest fire susceptibility is the probability of wildfire occurrence based on certain thresholds[15]. When the probability of wildfire is high, the potential forest fire risk is high. By establishing a forest fire prediction model, we can predict the probability of the occurrence of a forest fire and then strictly manage the area where the fire is likely to occur. This approach can directly reduce the occurrence of forest fires as well as the associated potential casualties and economic losses[16][17]. This method is therefore of great significance in forest fire prediction and prevention[18]

Much research has been conducted on forest fire prediction models. Logistic regression models are the

most commonly used models, and they have the advantage of solving the classification problem[19][20]. In recent years, geographically weighted regression models have also been used[21][22]. This method can provide a reasonable explanation for spatial heterogeneity, but the regression analyses can be performed only on continuous variables; the method lacks analysis of categorical variables. Similarly, many researchers have used generalized linear regression models for forest fire prediction. Liao et al. (2008) used the zero-inflated Poisson model to predict the frequency of forest fires in Japan in 2000 [23]. Mandallaz et al. (1997) used the Poisson model to predict forest fires in France, Italy, etc. [24]. Guo et al. (2010) used ordinary least squares regression, zero-inflated negative binomial regression and the zero-inflated negative binomial model to predict the number of forest fires in the Greater Xing'an Mountains area of Heilongjiang Province, China[25].

The development of artificial intelligence has led researchers to focus on building a forest fire prediction model using machine learning algorithms[26][27][28][29][30][31][32][33][34][35]. Researchers have used machine learning models to predict wildfires with greater accuracy and faster automation. For example, Camp et al.(1997)used decision trees (DTs) to identify historic forest fire shelters[36]; Pourghasemi et al. (2016)used the Mamdani fuzzy logic model to evaluate forest fire prediction ability[37]; and Li et al. (2020)used the long-term memory (LSTM) neural network model to study the burning area of wildfires[38]. For forest fire prediction, researchers have used more machine learning methods, such as artificial neural network models and support vector machine models.

Artificial neural networks (ANNs) consist of neurons with adjustable connection weights. Compared with traditional multiple linear regression models or parametric regression models, neural networks have better self-organization and self-learning capabilities, and they have been widely used in forest fire

prediction[39][40][41]. For example, Maeda et al. (2009) used ANNs and multitemporal images from MODIS/Terra-Aqua sensors to detect areas at high risk of forest fires in the Amazon region of Brazil [42]. The results showed that the error was small, and the predictions were accurate. Sakr et al. (2011) predicted the occurrence of forest fires in developing countries through two meteorological factors using artificial neural networks [43]. A radial basis function (RBF) neural network is a three-layer neural network, and it is a special case of a back-propagation neural network. At present, little research has used RBF neural networks for forest fire prediction. Samaher (2018) used an RBF neural network to predict the forest fire risk in natural parks in Portugal [27]. Support-vector machines (SVMs) are most suitable for the binary classification of data in the form of supervised learning. SVMs apply the principle of structural risk minimization and have good learning ability. In recent years, researchers have begun to use SVMs to predict forest fires[8][44][45][46]. Samaher (2018) used five different soft computing (SC) technologies, including an SVM algorithm, to predict areas at risk of forest fires[27]. He determined that the SVM algorithm provides more accurate predictions than the other four algorithms. Cortez et al. (2007) used five different data mining (DM) algorithms to predict the area at risk of forest fires in the north-eastern region of Portugal[8]. Their results showed that the prediction effect of the model was good. Based on Cortez's research, Xu et al. (2012) used the semidefinite programming model to select the optimal kernel function of the SVM to establish an SVM model for forest fire prediction [7]. The mean square error was small, and the model effect was good. The random forest (RF) algorithm is a well-known integrated learning algorithm that can provide higher accuracy than other algorithms. At present, the use of RFs to predict forest fires is relatively established[47][48][49]. Liang et al. (2016) used an RF model to predict the occurrence of forest fires in Fujian Province, China, with an accuracy rate of 85% [50]. Pourtaghi et al. (2016) used an RF algorithm to study the sensitivity of forest fires in Golestan

Province, Iran, and their results showed that the model achieved the desired accuracy[51].

Most of the current research focuses on certain areas, and there are few studies on the prediction and analysis of long time-series in China. Most studies have concentrated on the temporal and spatial changes and influencing factors of forest fires in specific years[52][53][54][55][56]. The results of previous research are therefore localized and limited, and there is a lack of research investigating the most suitable and high-precision forest fire prediction model on the national scale.

In this study, we selected a variety of forest fire driving factors to build four prediction models based on machine learning algorithms. The models were evaluated using data on Chinese forest fires from 2003 to 2016. The study has three objectives: (1) identify the main forest fire driving factors and their impacts in China; (2) select the most suitable model for forest fire prediction in China by creating four models and comparing and analysing the fitting results; and (3) use the model that offers the most accurate predictions to create a forest fire probability map for China and put forward recommendations for forest fire prevention.

2. Materials and Methods

2.1 Study Area and Data Resources

Located in East Asia on the west coast of the Pacific Ocean, China's territory is vast, with a total land area of approximately 9.6 million square kilometres. The topography is high in the west, with vast mountains and plateaus, and low in the east. The distance between the east and the west of the country is approximately 5,000 kilometres; the coastline of the mainland is more than 18,000 kilometres in length; and the temperature and precipitation are diverse, forming a variety of climates. The distribution of forest

resources in China is uneven and is mainly distributed in the north-eastern, southern, and south-western regions. The forested area is 220 million hectares, and the forest coverage rate is 22.96%.

The research data were divided into six parts: fire ignition data, meteorological data, terrain data, vegetation data, infrastructure data, and socioeconomic data[70]. The fire point data were derived from NASA's *Global Fire Atlas with Characteristics of Individual Fires, 2003–2016* (https://daac.ornl.gov/). The *Global Fire Atlas* is a global dataset that tracks the daily dynamics of single fires. For each individual fire, the dataset provides information about the fire's timing and location, scale, perimeter, duration, speed, and direction of spread. These individual fire characteristics are based on the *Global Fire Atlas* algorithms and estimated combustion day information from a 500-metre resolution product of the 6 MCD64A1 combustion zone product of the Medium Resolution Imaging Spectroradiometer (MODIS) collection.

This study used fire point data for forest land in China from 2003 to 2016. The final number of fire point data points was 32,746 (excluding Taiwan). The meteorological data were derived from the 14-day daily value dataset of the China Meteorological Data Network (http://data.cma.cn/dataService/). The dataset includes eight elements, such as the barometric pressure, temperature, relative humidity, and precipitation at the station. Digital elevation model (DEM) data were obtained through the Geospatial Data Cloud website (http://www.gscloud.cn/). Vegetation data were represented by the normalized difference vegetation index (NDVI), and the spatial distribution dataset of China's Quarterly Vegetation Index came from the Resource and Environment Data Cloud Platform (http://www.resdc.cn/). The basic geographic data were taken from the "National Basic Geographic Database of 1:1 Million" on the website of the National Geographic Information Resources Directory System. The data include the locations of railways,

highways, water systems, and residential areas. The socioeconomic data include population density and GDP per capita, and the grid data of the spatial distribution of population and GDP were obtained from the Resource and Environment Data Cloud Platform. Figure 1 shows the map of the study area.

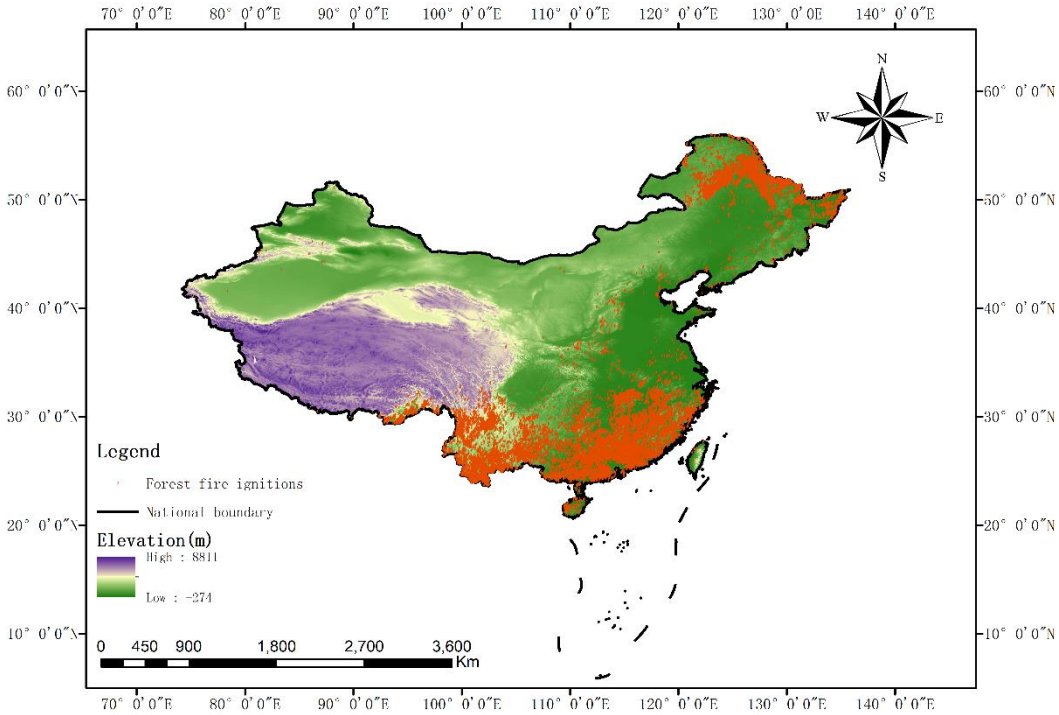


Fig. 1 Map of the study area

2.2 Data Pre-processing

2.2.1 Variable Handling

The dependent variable is a binary variable (i.e., whether or not a forest fire occurs); thus, we used ArcGIS 10.4 to create a certain percentage of random points (non-fire points) and assigned a value of 1 to fire points and a value of 0 to non-fire points[57]. To ensure that the data were not overdispersed, random points were selected according to experience in a ratio of 1:1[58], and in principle, randomness in space and time should be followed[59]. We used ArcGIS 10.4 software to create random points and

160 then used the 2015 national land-use data as a basis to exclude random points that were located in bodies
161 of water or urban land. We obtained a total of 65,492 fire points and random points.

162 For the meteorological data, we first used ArcGIS 10.4 to match the sample points with the nearest
163 meteorological station using the Thiessen polygon method. We then extracted the corresponding sample
164 point weather data and used an SQL server database to match the daily weather data. For the terrain data,
165 we used the spatial analysis tool in ArcGIS 10.4 to extract the slope and aspect of the obtained DEM data.
166 Seasonal climatic differences have an impact on vegetation status; thus, we divided the year into spring
167 (March, April, May), summer (June, July, August), autumn (September, October, November), and winter
168 (December, January, February)[60]. We used the extraction and analysis tools of ArcGIS to extract the
169 NDVI data for the sample points on an annual basis and a quarterly basis.

170 Similarly, from the infrastructure data and socioeconomic data, we extracted the information
171 corresponding to the sample points. We set the aspect and special festivals as categorical variables and
172 the others as continuous variables. Table 1 shows the classification of aspect. During certain traditional
173 festivals in China, people burn paper to commemorate their loved ones, increasing the probability of a
174 forest fire. We classified (value 1) the following dates of these events as special festivals: Chinese New
175 Year's Eve, the first day of the first lunar month, the second day of the first lunar month, the fifteenth
176 day of the first lunar month, and Qingming Festival and Zhongyuan Festival (July 15th of the lunar
177 calendar). Non-special festivals were set to 0.

178 **Table 1: Descriptions of aspect classifications**

Aspect	Azimuth (degree)	Classification

Gentle Slope	-1	0
Shady Slope	0~67.5, 337.5~360	1
Semi-shady Slope	67.5~112.5, 292.5~337.5	2
Sunny Slope	157.5~247.5	3
Semi-sunny Slope	112.5~157.5, 247.5~292.5	4

After processing, we obtained 20 independent variables and their possible values (see Table 2). Finally, we performed data cleaning on the sample points and the various types of data extracted to remove abnormal samples from the original dataset (including some samples with missing data and samples with observations that were significantly outside the normal range).

Table 2: Descriptions of independent variables

Category	Independent Variable	Symbol	Variable Type
Location	Longitude (°)	*Lon*	Continuous Variable
	Latitude (°)	*Lat*	Continuous Variable
Terrain	Altitude (m)	*Alt*	Continuous Variable
	Slope (°)	*Slo*	Continuous Variable
	Aspect	*Asp*	Categorical Variable
Meteorology	Average Surface Temperature (°C)	*Avst*	Continuous Variable
	Daily Maximum Surface temperature (°C)	*Mast*	Continuous Variable
	Cumulative Precipitation at 20–20 (mm)	*Pre*	Continuous Variable
	Average Relative Humidity (%)	*Arh*	Continuous Variable
	Hours of Sunshine (h)	*Suh*	Continuous Variable

	Average Temperature (°C)	*Ate*	Continuous Variable
	Daily Maximum Temperature (°C)	*Mate*	Continuous Variable
	Average Wind Speed (m/s)	*Aws*	Continuous Variable
	Maximum Wind Speed (m/s)	*Mws*	Continuous Variable
Infrastructure	Distance from Fire Point to Highway (m)	*Hig*	Continuous Variable
	Closest Distance from Fire Point to Residential Area (m)	*Set*	Continuous Variable
Social humanity	Population	*Pop*	Continuous Variable
	GDP	*GDP*	Continuous Variable
	Special Festival	*Sfe*	Categorical Variable
Vegetation	NDVI	*NDVI*	Continuous Variable

2.2.2 Data Normalization

Given the different dimensions and magnitudes of the factors above, the data were normalized to eliminate the variation in dimensions, avoid large differences in the magnitudes of the input and output data, and balance the contributions of various factors. All data were converted to values between 0 and 1. Table 3 shows the normalized formulas and specific interpretations of the independent variables.

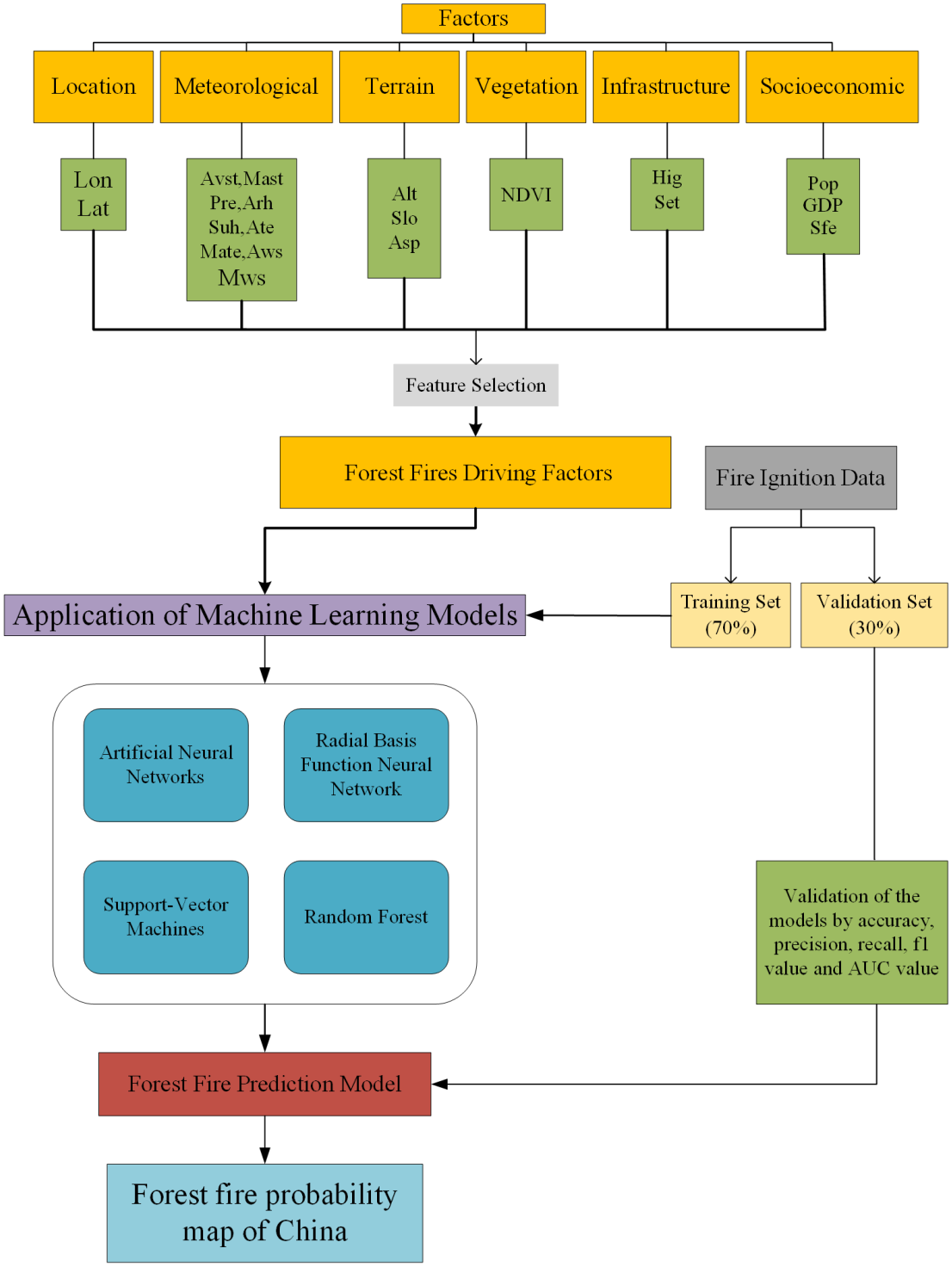
Table 3: Normalized formulas and explanations

No.	Formula	Explanation	Variables using this formula
(1)	$x_i^* = \frac{x_i - x_{min}}{x_{max} - x_{min}}$	x_i and x_i^* are the values before and after data normalization, respectively; x_{max} and x_{min} are the maximum and minimum values of the full sample data, respectively.	*Lon, Lat, Alt, Avst, Mast, Pre, Suh, Ate, Mate, Aws, Mws, Hig, Set, Pop, GDP*

(2)	$x_\alpha = \sin\alpha$	α is the slope value.	*Slo*
(3)	$x_\gamma = \frac{\gamma}{100}$	γ is the humidity value.	*Arh*

2.3 Research Method

This study provides a methodological framework for predicting the occurrence of forest fires in China, as shown in Figure 2. First, all the forest fire correlation factors are selected by feature selection to obtain the forest fire driving factors that have great influence on fires. These factors are then used as input data of the forest fire prediction model, and machine learning models (ANNs, RBF neural networks, SVMs and RFs) are applied to obtain corresponding results. Finally, the model accuracy is obtained through evaluation indexes such as the AUC value. The noun abbreviations in the following articles are shown in Table 4.



198

199 **Fig. 2 Flowchart of the Chinese forest fire occurrence prediction**

200 **Table 4: Abbreviations list**

Full name	Abbreviation
Artificial Neural Network	ANN
Radial Basis Function	RBF

Radial Basis Function Neural Network	RBFNN
Support-Vector Machine	SVM
Random Forest	RF
Out-of-Bag	OOB
Area Under the Curve	AUC
True Positive	TP
True Negative	TN
False Positive	FP
False Negative	FN
Receiver Operating Characteristic	ROC
Support Vector Classification	SVC

2.3.1 Artificial Neural Networks

ANNs have become widely used in feedforward networks due to their clear structure, fast operation, easy implementation, and abilities for self-learning and adaption to the environment[61][62]. ANNs consist of three parts: an input layer, an output layer, and a hidden layer. The hidden layer may be a topological structure of one or more layers, as shown in Figure 3. The input layer does not perform any calculations; rather, it is used to receive data, that is, to transfer data to the adjacent hidden layer with different weights. The hidden layer processes the data through a nonlinear activation function and then passes it to the output layer. The final result is obtained from the output layer. The mathematical principle is as follows:

$$\begin{cases} \mathrm{h}^{(1)} = \varphi^{(1)}(\sum_{i=1}^{n} x_i \cdot {\omega_j}^{(1)} + b^{(1)} \\ y = \varphi^{(2)}(\sum_{j=1}^{n} {h_i}^{(1)} \cdot {\omega_j}^{(2)} + b^{(2)} \end{cases} \tag{4}$$

In the formula, the input layer is $x \in R^m$, the hidden layer output is $\mathrm{h} \in R^n$, the output layer is $y \in R^K$, the input layer to the hidden layer weight connection matrix is $\omega^{(1)} \in R^{m \times n}$, the weight connection bias from the input layer to the hidden layer is $b^{(1)} \in R^n$, and the weight connection matrix and the bias from the hidden layer to the output layer are $\omega^{(2)} \in R^{n \times K}$ and $b^{(2)} \in R^{n \times K}$, respectively.

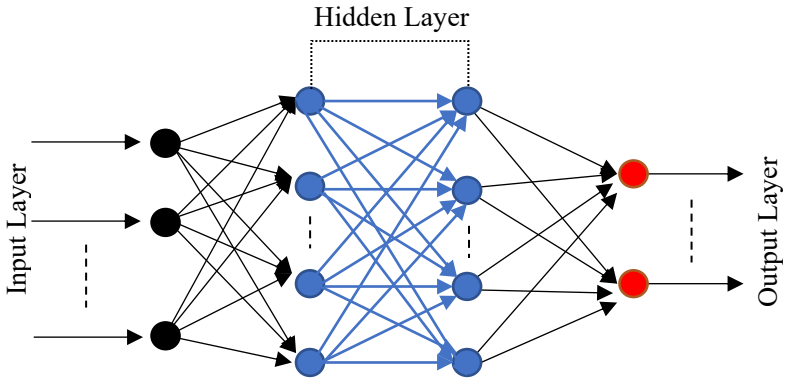


Fig. 3 Diagram of the structure of an ANN

2.3.2 Radial Basis Function Neural Network

The RBF neural network structure is a feedforward structure with an input layer, a single hidden layer, and an output layer. Its advantages are concise training and fast learning convergence speed, which can approximate any nonlinear function. This method has been widely used in time-series forecasting, nonlinear control systems, and the graphics-processing field. The basic idea of an RBF neural network is as follows. The RBF is used as the "base" of the hidden unit to form the hidden layer space. The hidden layer transforms the input vector and transforms the low-dimensional pattern input data into the high-dimensional space. The result is that the data are linearly separable in the high-dimensional space. The output of the RBF neural network is as follows:

$$y_i = \sum_{i=1}^{h} \omega_{ij} \exp\left(-\frac{1}{2\sigma^2}\left\|x_p - c_i\right\|^2\right) \quad j = 1,2,\cdots,n \tag{5}$$

where $x_p = (x_1{}^p, x_2{}^p, \cdots, x_m{}^p)^T$ is the p^{th} input sample (p = 1,2,3,...,P), P is the total number of samples, c_i is the centre of the hidden layer node of the network, ω_{ij} is the connection weight from the hidden layer to the output layer, i = 1,2,3,...,h is the number of hidden layer nodes, and y_i is the actual output of the j^{th} output node of the network corresponding to the input sample[63].

2.3.3 Support-Vector Machines

SVMs are mainly used for pattern classification and nonlinear regression. They are general learning algorithms based on the principle of structural risk minimization. The core idea of an SVM is to establish a classification hyperplane as a decision surface to maximize the isolation edge between the positive and negative examples, thereby providing a high generalization performance[64]. SVMs can improve the ability to transform data from high-dimensional spaces by flexibly using kernel functions when dealing with various nonlinear problems. Taking a two-class SVM as an example, given a training set $T = \{(x_1, y_1), \cdots (x_l, y_l)\} \in (X \times Y)^l$, where $x_i \in X = R^n, y_i \in \{1, -1\} (i = 1,2, \cdots l)$, x_i is the feature vector. The penalty parameter *C* and the kernel function $K(x, x')$ are first selected, and the optimization problem is then constructed and solved as follows[63]:

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{j} \sum_{j=1}^{l} y_i y_j a_i a_j K(x, x') - \sum_{j=1}^{l} \alpha_j \tag{6}$$

$$s.t. \ \sum_{i=1}^{l} y_i \alpha_i = 0, \ 0 \le \alpha_i \le C, i = 1, \cdots, l \tag{7}$$

The optimal solution is then obtained: $\alpha^* = (\alpha_1{}^*, \cdots, \alpha_l{}^*)^T$. A positive component of α^*:$0 \le \alpha_j{}^* \le C$ is then selected, and the threshold is calculated as follows:

$$b^* = y_j - \sum_{i=1}^{l} y_i \alpha_i{}^{'} K(x_i - x_j) \tag{8}$$

Finally, the decision function is constructed:

$$f(x) = sgn(\sum_{i=1}^{l} \alpha_i{}^* y_i K(x, x_i) + b^* \tag{9}$$

2.3.4 Feature Selection and Random Forest

Feature selection refers to the selection of subsets from the original feature set to optimize a certain evaluation criterion so that the model established with the optimal feature subset can achieve a prediction accuracy similar to or better than that of the model established without feature selection[65][66]. RFs have been demonstrated to have a high prediction accuracy and high tolerance to outliers and 'noise'[67]. This method can be used to evaluate the relationship between covariates and dependent variables and calculate the relative importance of covariates[68][69]. RF has been applied in various fields in the past, including medicine, genetics, ecology and remote sensing. In recent years, it has been widely used in forest fire prediction and has demonstrated good predictive abilities[70][71][72][73]. The order of importance of variables can be obtained by random forest algorithm. In previous studies, the variables screened by this method have been proved to have high reliability[4][26][74][75][76]. Therefore, random forest algorithm is selected as the method of feature selection in this study. The basic idea of feature selection using RF is as follows: for the *j*th variable (X_j), the OOB error ($errOOB^j{}_t$) of each tree t is calculated, and then the value of the *j*th variable (X_j) is permuted while all others are left unchanged among OOB data and the OOB error ($err'OOB^j{}_t$) is again recalculated on this permuted dataset. RF estimates the importance of a variable by evaluating how much the prediction error increases when the OOB data for that variable are permuted. The importance score of X_j is as follows:

$$VI(X^j) = \frac{1}{ntree}\sum_t (err'OOB^j{}_t - errOOB_t) \quad (10)$$

where R is the summation of all the trees, and ntree is the number of trees in the RF[77][78]. For classification, the OOB is the misclassification probability.

A RF is a highly flexible machine learning algorithm with broad application prospects. In essence, an RF

is a classifier consisting of multiple DTs formed by random methods. These trees are not related, hence its alternative name: “random decision tree.” When the test data enter the RF, each DT is classified, and the category with the most classification results among all the DTs is taken as the final result. The flow chart of the RF algorithm is shown in Figure 4.

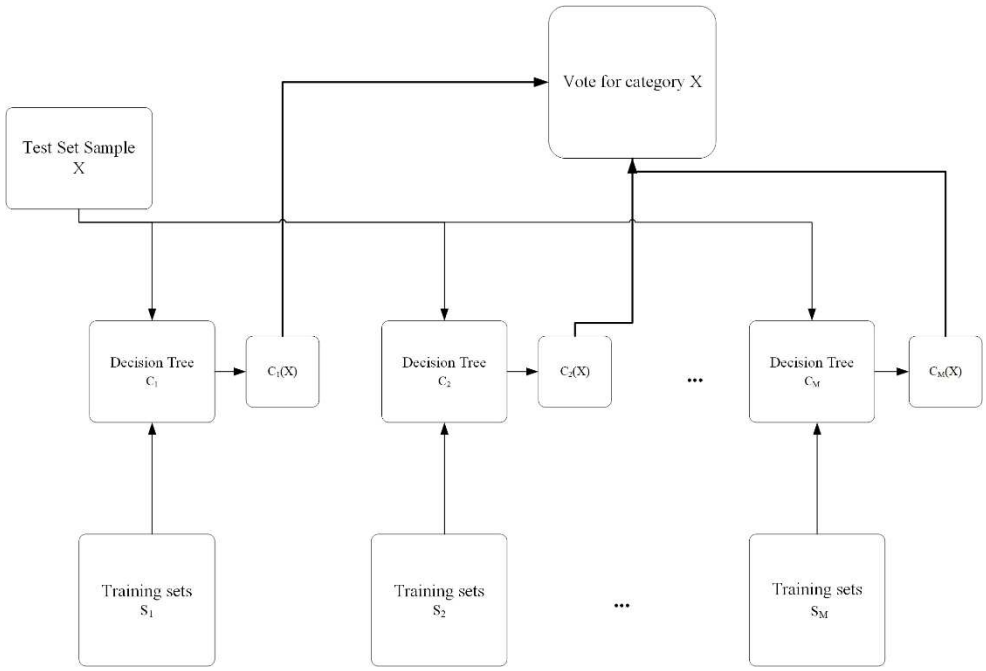


Fig. 4 The flow chart of the RF algorithm

The basic principle of the RF algorithm is as follows. Let *N* be the number of attributes of the sample. *n* is an integer greater than 0 and less than *N*. First, the bootstrap method is used for resampling, randomly generating *M* training sets S_1, S_2,...S_M. DTs A_1, A_2,...A_M corresponding to each training set are then generated. Before selecting the attribute in each non-leaf node, *n* attributes are randomly selected from the *N* attributes as the split attribute set of the current node, and the node is split in the best split mode among the *n* attributes. Each tree grows intact without pruning. For the test set sample *X*, each DT is used to test and obtain the corresponding categories: $C_1(X)$, $C_2(X)$,...$C_M(X)$. Finally, the voting method is adopted, and the category with the most output among the *M* DTs is regarded as the category to which

the test set sample *X* belongs[63].

2.3.5 Model Performance Evaluation

In this study, we used five performance indicators: accuracy, precision, recall, f1 value, and AUC to evaluate the performance of the models. Descriptions of the five indicators are given below.

1. Accuracy: the proportion of the number of samples (TP and TN) that are correctly predicted to the total number of samples. The formula is as follows:

$$P = \frac{TP+TN}{TP+FP+TN+FN} \tag{11}$$

2. Precision: characterizes the classification effect of the classifier, which is the correct frequency value predicted in the instance of the positive sample:

$$T = \frac{TP}{TP+FP} \tag{12}$$

3. Recall: characterizes the recall effect of a certain class. It is the correct frequency of prediction in the instance of the label as the positive sample:

$$R = \frac{TP}{TP+FN} \tag{13}$$

4. f1 value: the value used to measure precision and recall. It is the harmonic mean of these two values:

$$\mathrm{f1} = \frac{2TP}{2TP+FP+FN} \tag{14}$$

5. A receiver operating characteristic (ROC) curve is a method used to judge the prediction effect of the model[64]. The prediction accuracy of the model is judged by the value of the AUC, which ranges from 0.5 to 1. The larger the value is, the closer the fit of the model is.

Note: TP, FN, FP, TN in the formulas are the labels of the confusion matrix form of the output result. The form is shown in Table 5:

Table 5: Confusion matrix form

Prediction (column)/label (row)	Positive Sample	Negative Sample
Positive Sample	TP	FN
Negative Sample	FP	TN

3. Results

In this study, we used the MATLAB (MathWorks, USA, MATLAB 2019a) [61]and R Studio (JJ Allaire, RStudio-1.2.5042/R 3.6.3) programming languages to implement the algorithms. We used MATLAB to build the ANN, RBFNN, and SVM models and used R Studio to build the RF models.

To evaluate feature factors and model performance issues, the dataset was divided into two parts by randomly selecting 70% of the pre-processed sample data as the training set and 30% as the test set[62]

3.1 Feature Selection

We use the RF algorithm to select the features of all variables after pre-treatment and select the subset of features that had the greatest impact on the dependent variables for the next model construction process. We divided the whole sample according to the above proportion (70% to the training set and 30% to the test set) and repeated this process 5 times to obtain 5 training samples[4]. Then, the varSelRF package in the R language was used to select and calculate the characteristic variables of the five training samples to obtain the variable subsets of the five intermediate models, and the variables appearing more than three times in the five variable quantum sets were selected as the variables after screening. The results

317 are shown in Table 6. All variables and variables filtered by feature selection were used as input data for

318 RF modelling, and the out-of-pocket error rate (OOB) and confusion matrix were obtained as shown in

319 Table 7 below.

320 **Table 6: Results of variable selection based on RF**

No.	Variable	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Frequency
1	Lat	+	+	+	+	+	5
2	Lon	+	+	+	+	+	5
3	Avst	+	+	+	+	+	5
4	Mast	+	+		+	+	4
5	Pre	+	+		+	+	4
6	Arh	+	+	+	+	+	5
7	Suh	+	+	+	+	+	5
8	Ate	+	+	+	+	+	5
9	Mate	+	+	+	+	+	5
10	Aws						0
11	Mws						0
12	Alt	+	+	+	+	+	5
13	Slo						0
14	Asp						0
15	Set						0
16	Hig						0
17	GDP	+	+		+	+	5
18	Pop	+	+	+	+	+	5

19	NDVI	+	+	+	+	+	5
20	Sfe						0

Table 7: The results of the OOB and confusion matrix of the two samples

Total variable sample	OOB estimate of error rate		10.89%
Confusion matrix:	0	1	Classification error rate
0	20224	2716	12.3%
1	2168	20737	9.5%
Sample of screened variables	OOB estimate of error rate		10.65%
Confusion matrix:	0	1	Classification error rate
0	20038	2810	11.8%
1	2171	20717	9.5%

It can be seen from Table 6 that the error rate outside the bag after using the whole variable modelling is 10.89%, while the error rate outside the bag after using the screened variable is 10.65%, which is lower than the result of the whole variable. After feature filtering, the performance of the model is better, and the complexity of the model is reduced, providing a simpler model. Finally, variables after feature screening were taken as the main driving factors of forest fires and entered into the subsequent model fitting process[26].

The results show that the main influencing variables are longitude, latitude, average surface temperature, daily maximum surface temperature, accumulated precipitation, average relative humidity, sunshine hours, average temperature, daily maximum temperature, altitude, population, GDP, and NDVI. These variables performed subsequent model fitting. Then, the mean decrease in accuracy obtained by the RF algorithm was used to evaluate the importance of the variable. The larger the value is, the greater the

importance of the variable is. Figure 5 shows the importance of each variable in the five random training samples and the 20 feature subsets in the full sample.

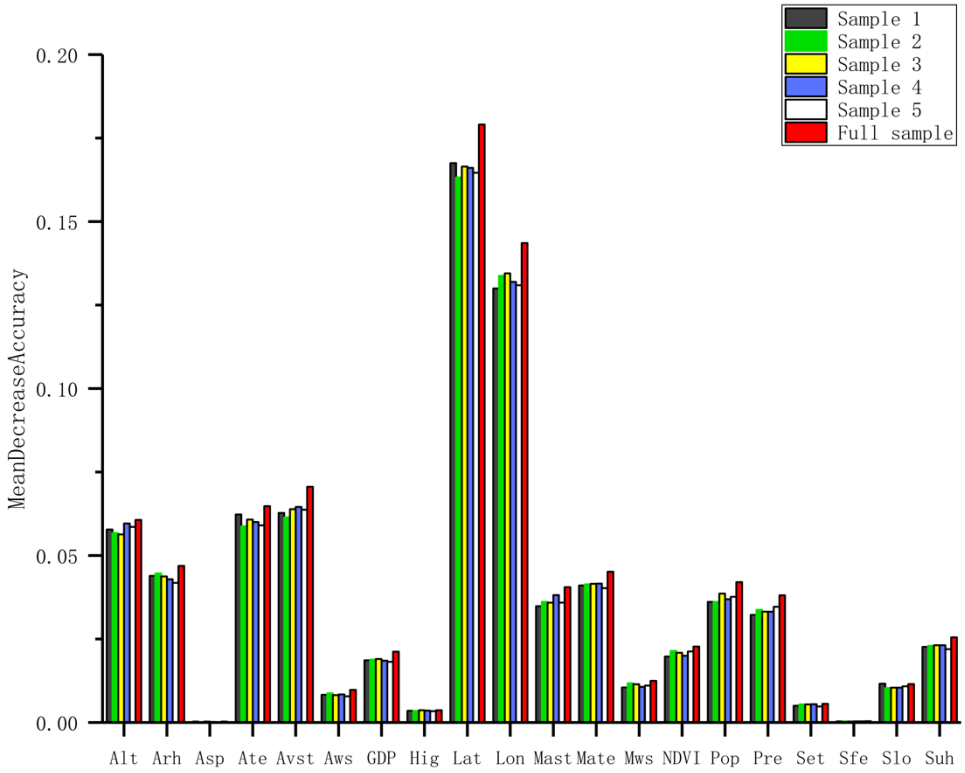


Fig. 5 Feature subset importance

3.2 Model Fitting Results

3.2.1 Artificial Neural Network

The input layer of the ANN consists of 13 neurons after feature selection: Lat, Lon, Avst, Mast, Pre, Arh, Suh, Ate, Mate, Alt, GDP, Pop, and NDVI. The output layer contains two cells (1 or 0). We use the gradient descent method to optimize the algorithm. We set the number of hidden layer cells between 1 and 50, automatically select the optimal number of hidden layer cells as the final result, and finally obtain the number of hidden layer cells as 5. The comparison between the predictive value and the actual value in the test dataset is shown in Figure 6. Note: Due to the large sample size, only a part of the sample

comparison chart is displayed. This is also the case for the following comparison charts.

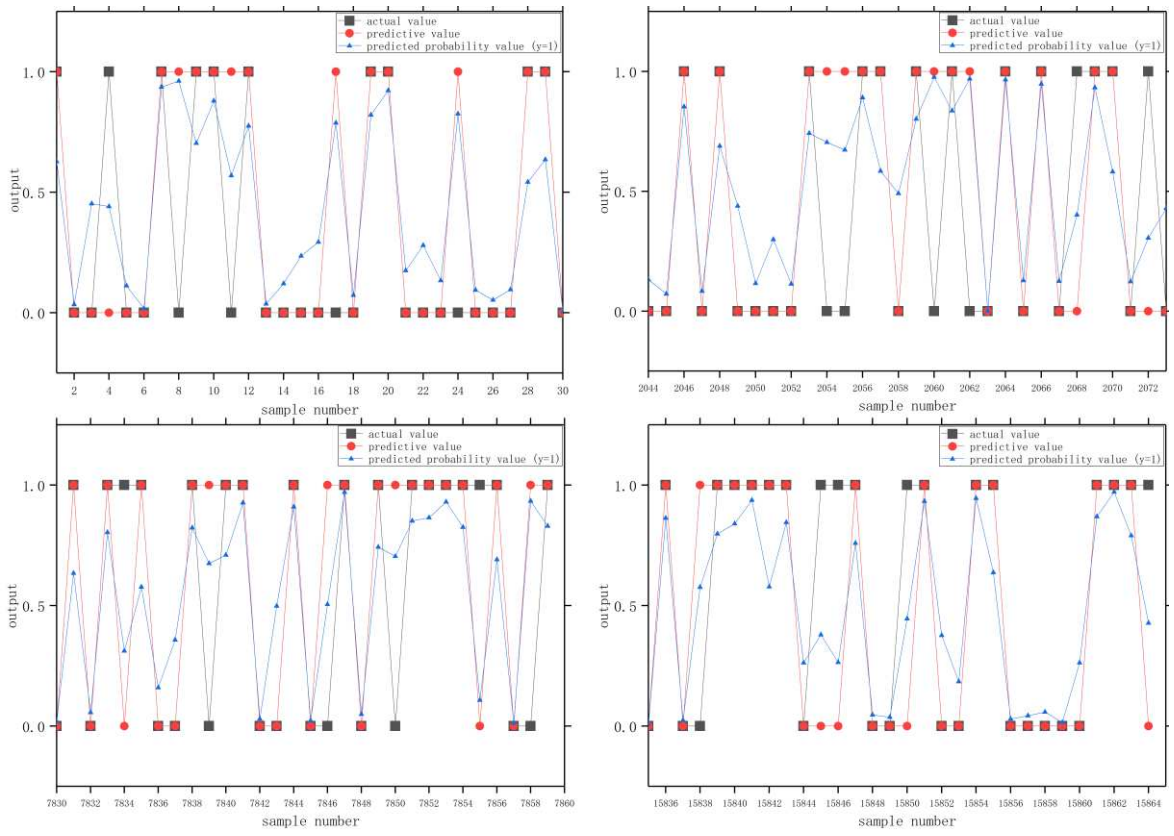


Fig. 6 Comparison charts of the predictive and actual values of the ANN

3.2.2 Radial Basis Function Neural Network

The input and output layer variables of the RBF neural network were the same as those of the ANN. The number of hidden layers and the number of cells contained are the same as those in the MPNN model, which are the optimal results automatically selected. After training, we obtained a hidden layer containing 10 units. The comparison charts of the predictive and actual values of the test set are shown in Figure 7.

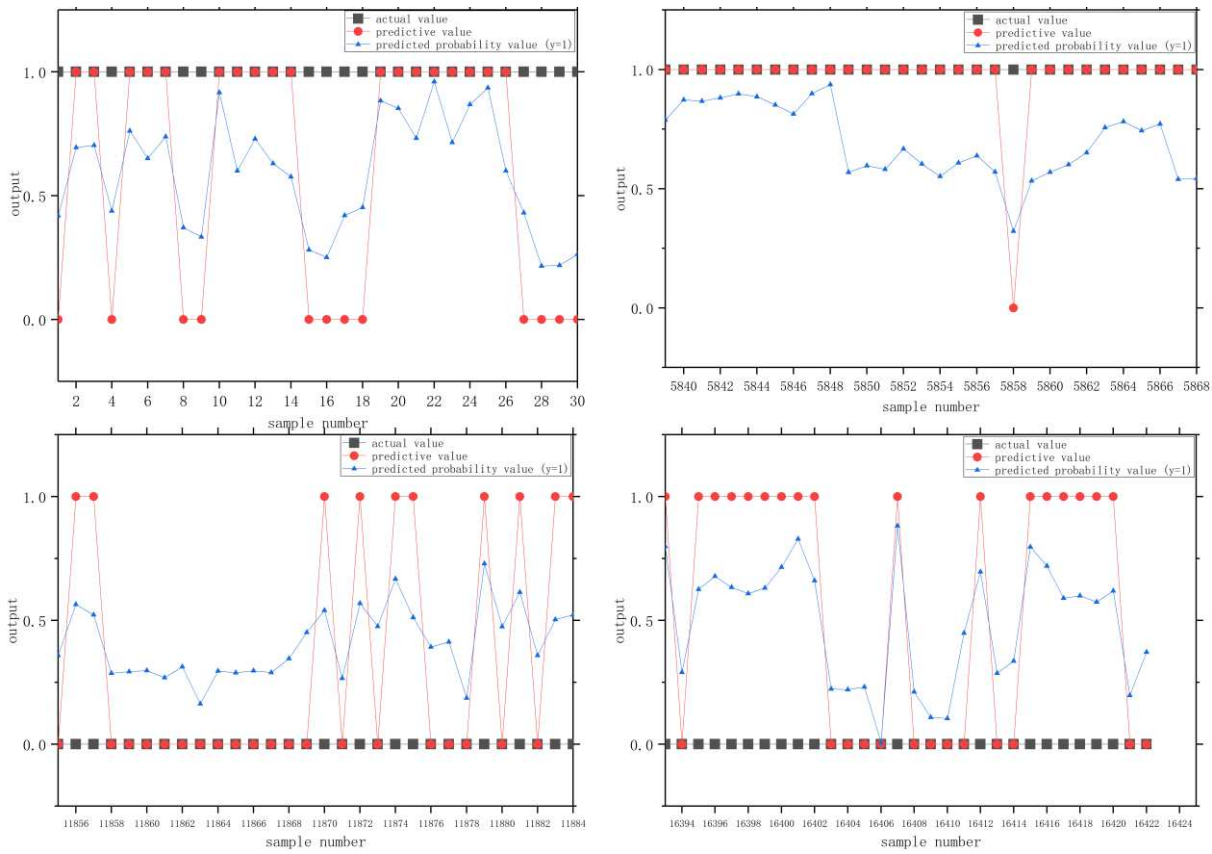
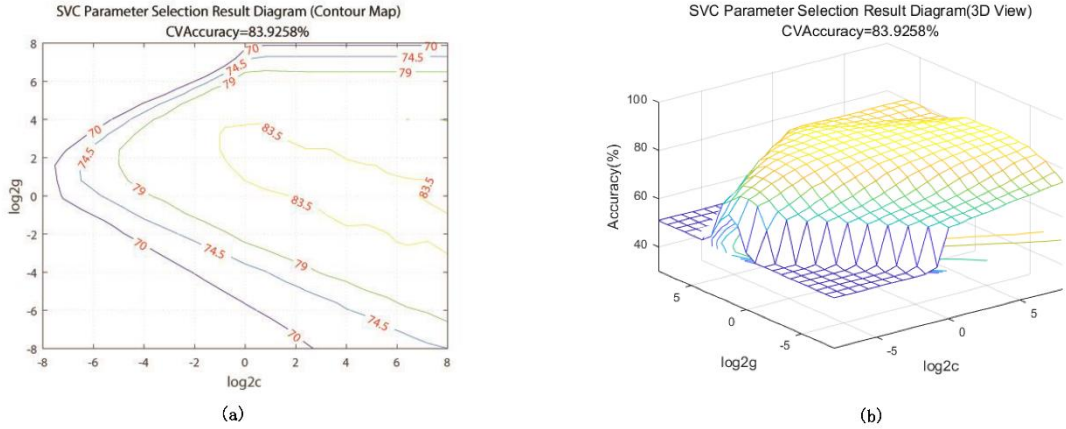


Fig. 7 Comparison charts of the predictive and actual values of the RBFNN (part of the sample)

3.2.3 Support-Vector Machine

We used the LIBSVM package of MATLAB to construct the SVM. The model was constructed using the RBF kernel function for processing nonlinear data. We used the grid search method and 10-fold cross-validation to select the parameters and determine the penalty parameter *C* and the kernel parameter *g*. Figure 8 shows a contour map and a 3D view of the result of the SVC parameter selection. After calculation, the accuracy rate of the grid search method reached 83.9%, and the accuracy rate of cross-validation reached 82.6%.



363

364 **Fig 8. SVC parameter selection result: (a) contour map (b) 3D view**

365 It can be seen from the results that the optimal values of *C* and *g* are 1.74 and 3.03, respectively. After
366 setting the parameters to the optimal values, we performed SVM modelling and obtained the predicted
367 values. Figure 9 shows the comparison charts of the actual and predicted values. After optimization, the
368 total number of support vectors was 19,460, and the number of support vectors at the boundary was
369 17,260. After model training, the accuracy rate of the training set was 86.02%, the accuracy rate of the
370 test set was 84.27%, and the performance of the model was high.

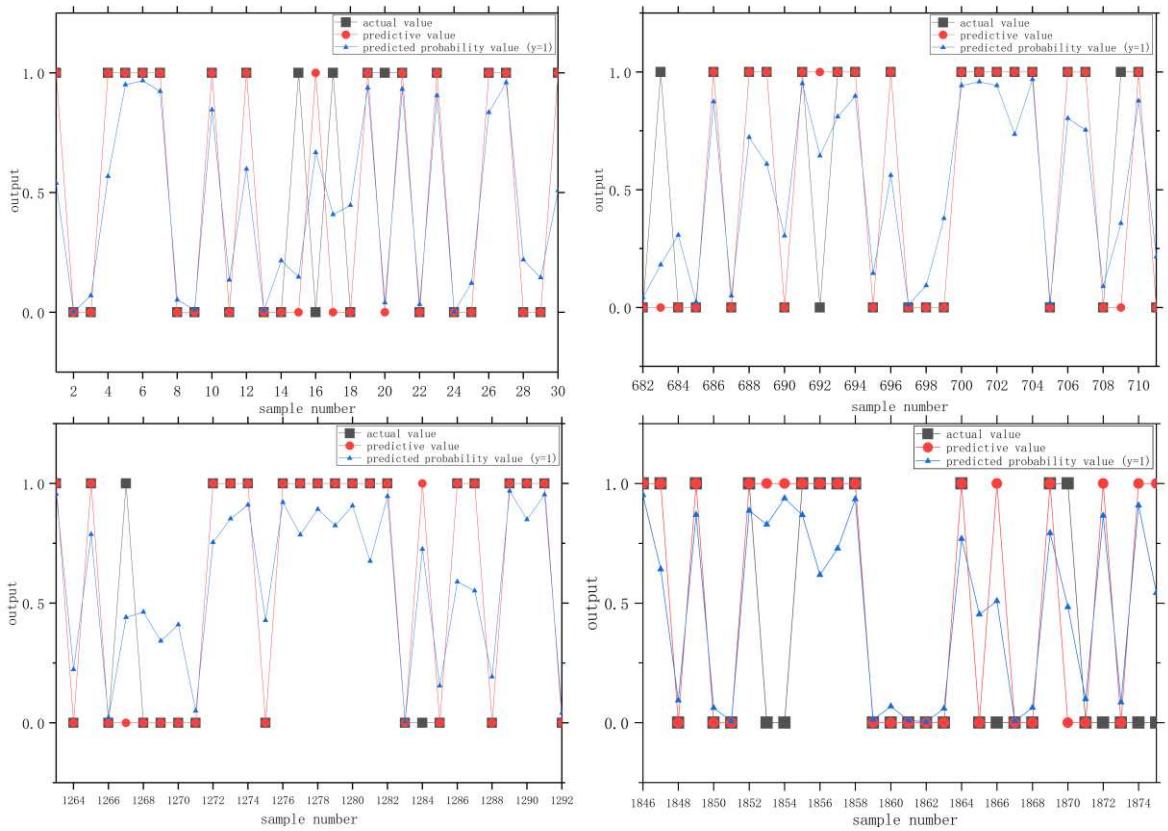


Fig. 9 Comparison charts of the predictive and actual values of the SVM (part of the sample)

3.2.4 Random Forest

We used the randomForest package in the R language to train the random training samples. We adjust the ntree (number of DTs) and mtry (the node value of the trees) parameters. We then used cross-validation to determine the optimal parameters of the model. Finally, we obtained the number of trees and the accuracy of the test and training data through cross-validation. As shown in Figure 10, when the number of DTs is 400 and the node value of the trees is 2, the accuracy tends to be stable. We used the optimal number of DTs to create comparison charts of the actual and predicted values of the test set (Figure 11) and the average accuracy decline of 13 forest fire driving factors (Figure 12). Figure 10 shows that among the main forest fire driving factors in China, the location variables that have the greatest

influence on the occurrence of forest fires are longitude and latitude. Rainfall is the variable with the smallest influence on the occurrence of forest fires.

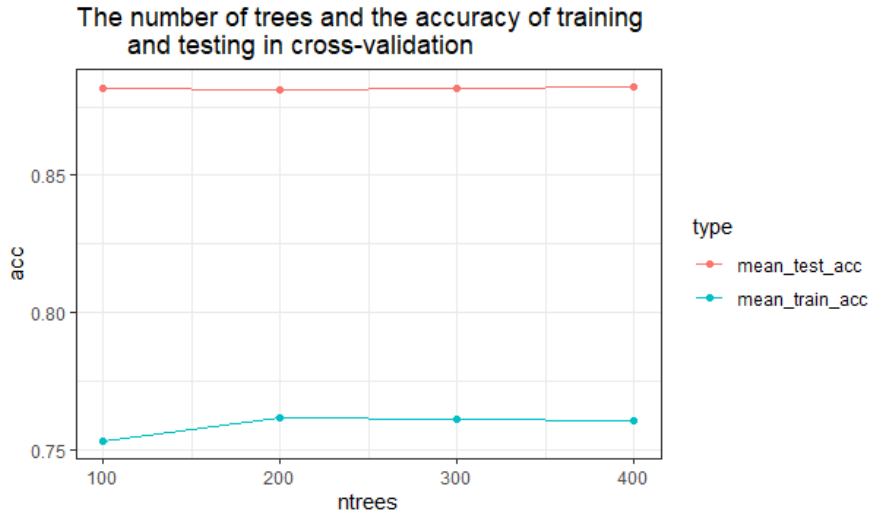


Fig. 10 The number of trees and the accuracy of training and testing in cross-validation

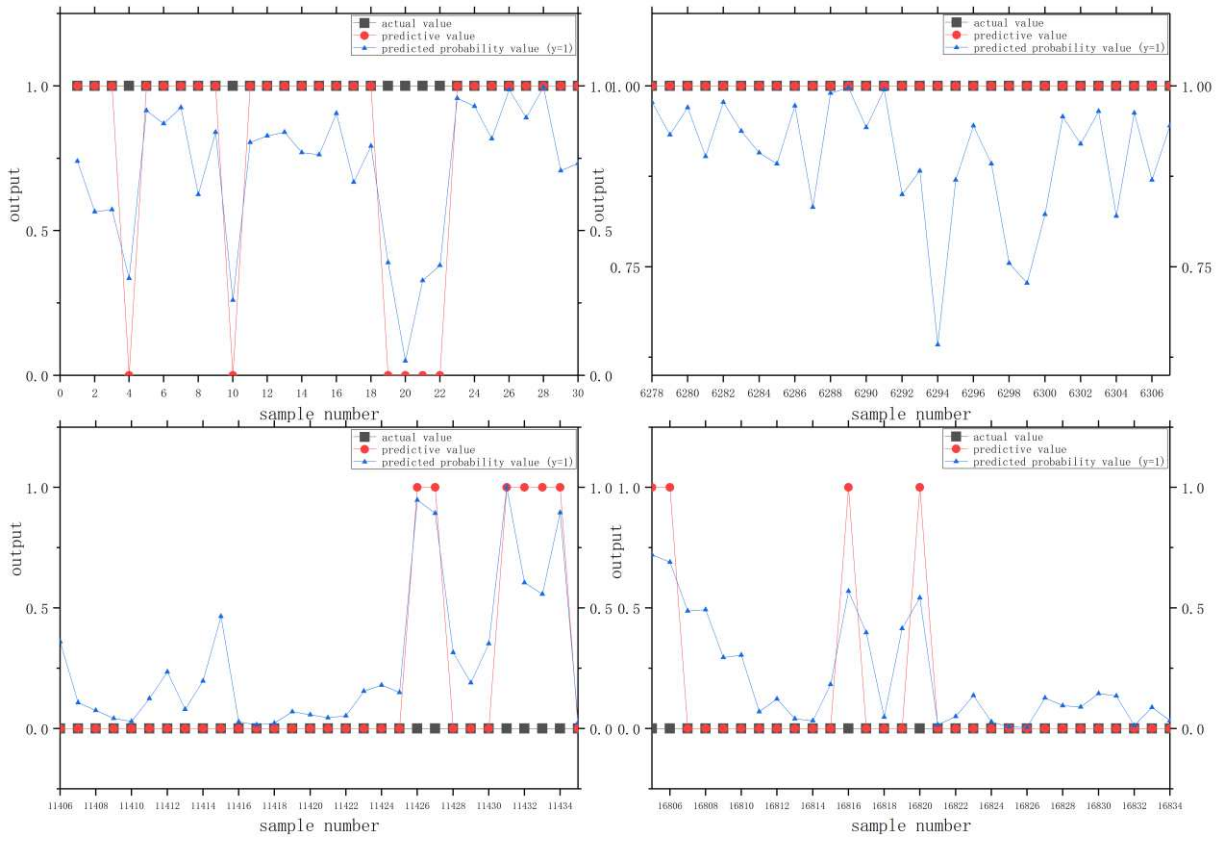
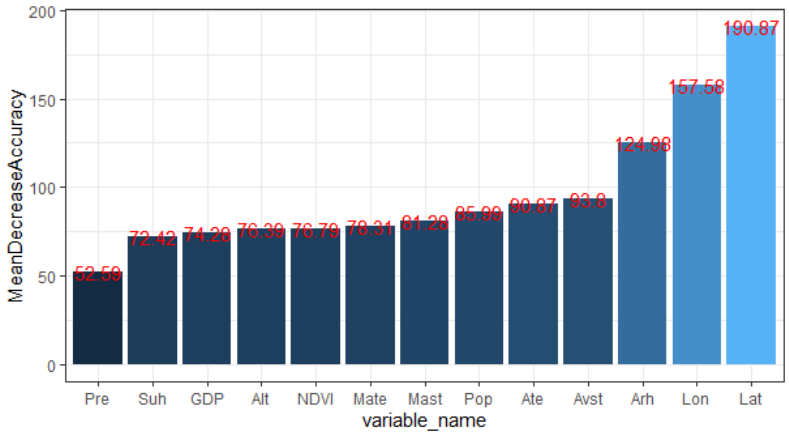


Fig. 11 Comparison charts of the predictive and actual values of the RF (part of the sample)



389

390 **Fig. 12 Mean decrease accuracy of 13 variables**

391 ## 3.3 Accuracy Evaluation

392 We used the prediction results of the four models to construct a confusion matrix to obtain the accuracy,
393 precision, recall, f1 value, and AUC value, as shown in Table 8. Figure 11 shows the ROC curves of the
394 four models. As shown in Table 6, the accuracy and f1 values of each model were more than 75%, and
395 the AUC value was more than 0.80. Thus, the performance of all four models was high. Among the four
396 models, the RF model had the highest predictive ability, with an accuracy rate of 89.2%, a f1 value of
397 89%, and the highest AUC value, reaching 0.960. Compared with the other three models, the prediction
398 ability of the RBF neural network was the lowest, with an accuracy rate of 75.8% and an AUC value of
399 0.840. As shown in Figure 13, the RF model outperformed the other three models. We therefore
400 considered the RF model to be the most suitable among the four models for forest fire prediction in China.

401 **Table 8: Evaluation results of the four models**

Model	Accuracy (%)	Precision (%)	Recall (%)	f1 value (%)	AUC
ANN	83.0	85.4	79.6	82.4	0.904

RBFNN	75.8	73.1	81.6	77.1	0.840
SVM	84.3	83.0	86.8	84.8	0.917
RF	89.2	90.2	87.9	89.0	0.960

Fig 11. Comparison charts of accuracy, precision, recall, and f1 values of the four models

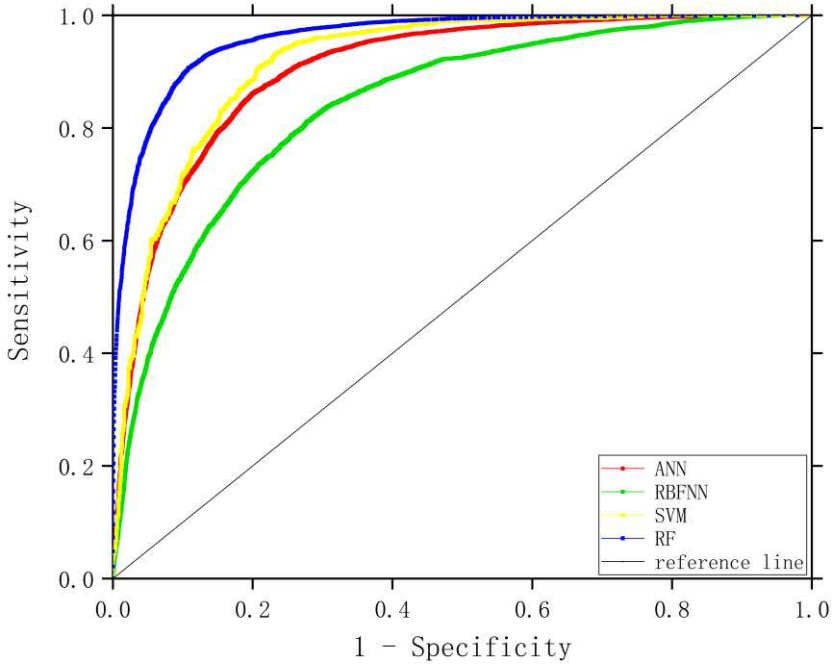


Fig. 13 ROC curves of the four models

3.4 Fire Risk Classification

After evaluating the accuracy of the four models, we used the RF model (highest accuracy) to obtain the probability of forest fire occurrence for the full sample. We used ArcGIS to draw a forest fire probability map (Figure 14) and a seasonal forest fire probability map (Figure 15) for China. The numbers in the legends in Figure 14 and Figure 15 indicate the predicted value of the probability of forest fires in China. For example, the probability of a forest fire is 1, which means that the probability of a forest fire is the greatest; the number of red areas is 0.8-1, which indicates that the area is in a high-risk state, and forest fires are very likely to occur. Figure 14 shows that the high incidence of forest fires in China is mainly

concentrated in the northeast (such as the Greater Xing'an Mountains region), the southeast (such as Guangdong, Jiangxi, and Fujian), and the southwest (such as Yunnan and Sichuan). Overall, the probability of forest fires in eastern China is higher than that in western regions, and the probability of forest fires in the north and south is higher than that in Central China. Figure 15 shows that the seasonal order of the probability of forest fires in China is, from highest to lowest, spring, winter, summer, and autumn. Spring and winter are the seasons with a high incidence of forest fires, and the fires are mainly concentrated in Northeast China (such as Heilongjiang Province) and south-eastern China (such as Fujian Province and Guangdong Province).

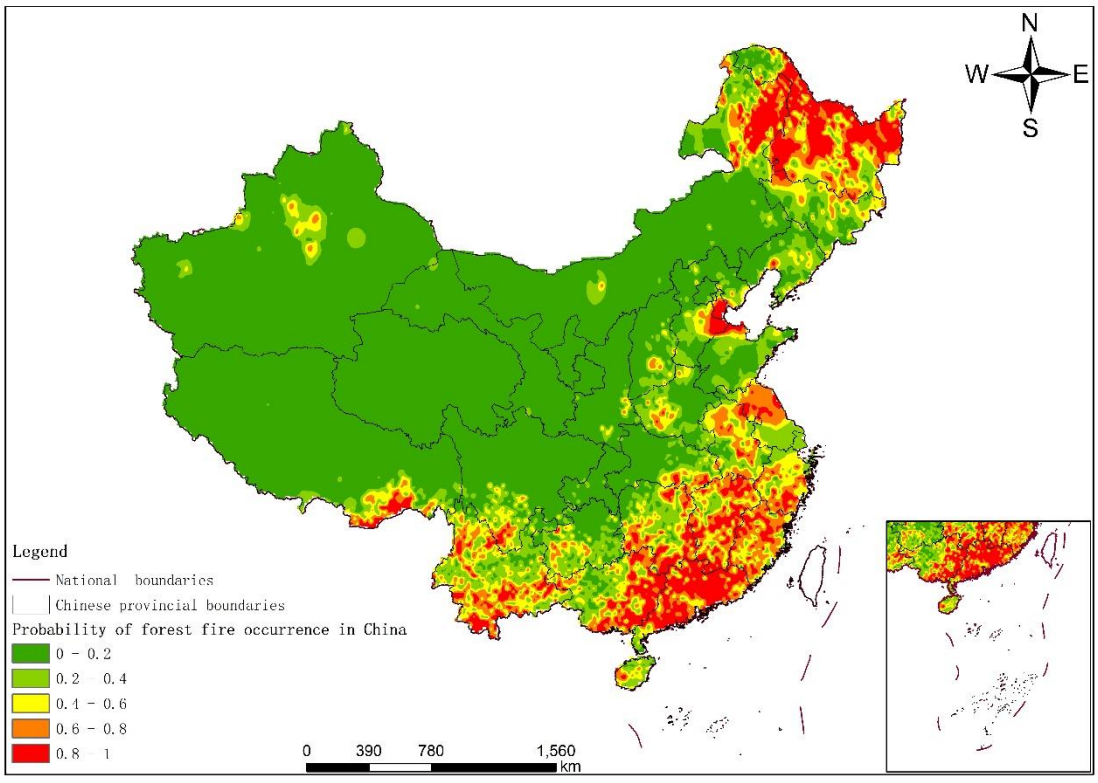


Fig. 14 Forest fire probability map for China based on the RF model

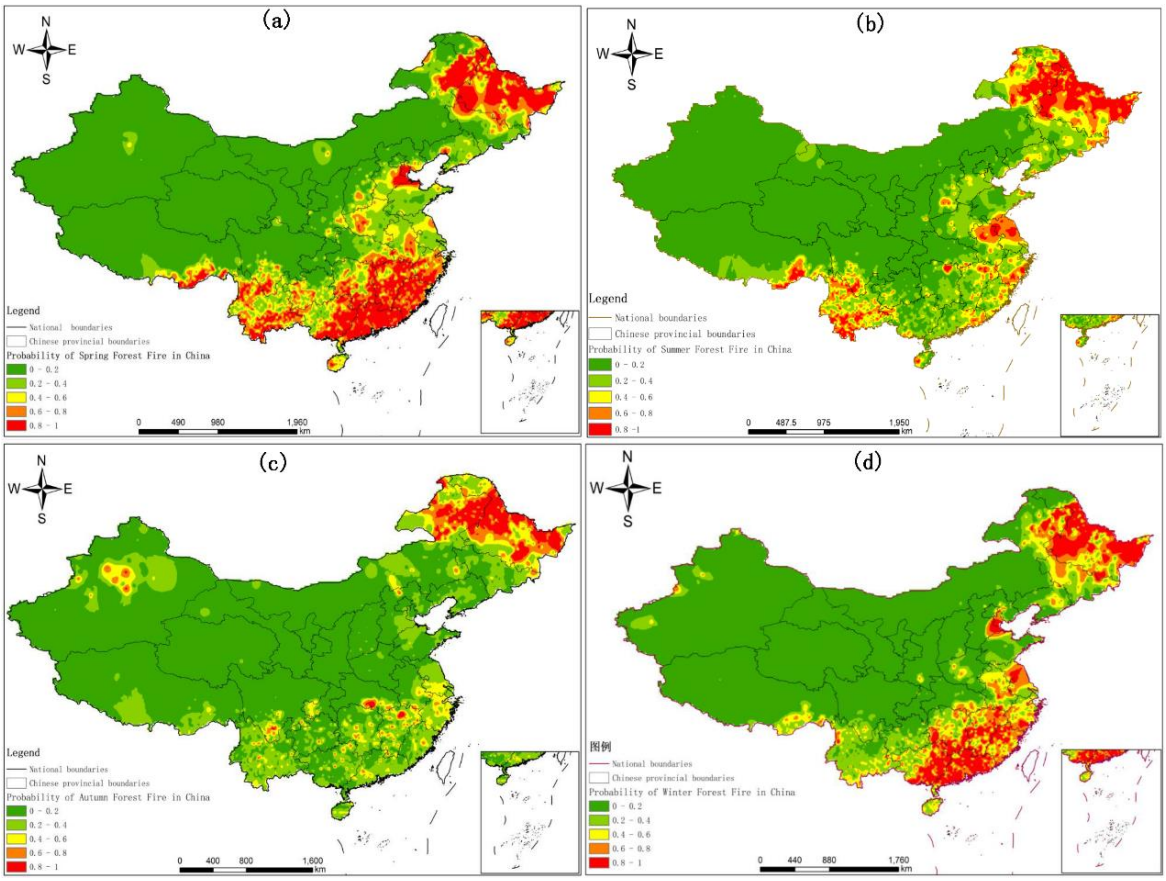


Fig. 15 Seasonal forest fire probability map for China based on the RF model: (a) spring (January, February, and March); (b) summer (April, May, and June); (c) autumn (July, August, and September); and (d) winter (October, November, and December)

4. Discussion

4.1 Major Forest Fire Driving Factors in China and Their Impacts

In this study, 21 factors influencing the occurrence of forest fires were selected. These factors could be divided into six categories: geographical location, meteorology, climate, topography, society and vegetation. Researchers have studied the influencing factors of these forest fires[79][80][81][82]. Through feature selection, the driving factors of forest fires, such as longitude, latitude, average surface temperature, daily maximum surface temperature, cumulative precipitation, average relative humidity,

sunshine duration, average temperature, daily maximum temperature, altitude, population, GDP and NDVI, were obtained. In terms of the importance of feature subsets, longitude and latitude had the greatest influence on the occurrence of forest fires. This result is because of the uneven distribution of forest resources in China. Generally, in terms of forest resources, the south has more forested area than the north, and the east has more forested area than the west. In addition, for example, many eucalyptus trees are planted in south-eastern provinces, such as Guangdong and Fujian, which are prone to cause fires. Therefore, the occurrence frequency of forest fires in China is obviously different with changes in longitude and latitude. Therefore, longitude and latitude have become the most important factors in studying the occurrence of forest fires in China. Currently, there are few studies regarding longitude and latitude as the main driving factors of forest fires. However, some researchers have confirmed that the number of forest fires increases with decreasing latitude [83][84]

Second, climatic factors have great impacts on forest fires, a result that is consistent with previous research results[85][86]. Temperature is one of the three necessary conditions for combustion. When the temperature reaches a certain level, forest fires are more likely to occur. The longer the daylight is, the higher the temperature is and the greater the likelihood of forest fires is. Rainfall and average relative humidity are other main factors affecting forest fires[56][87][88]. In addition, altitude and vegetation can affect the occurrence of forest fires. Tian et al. (2013) believed that forest fires mainly occurred in low-altitude areas [89]. Chuvieco et al. (2004) used the NDVI as a driving factor of forest fires to estimate fuel humidity[90]. The higher the NDVI value is, the higher the vegetation coverage is. The more flammable trees are, the more likely they are to cause problems associated with forest fires.

Forest fires are also driven by social and human factors (e.g., population and GDP). The larger the

population is, the greater the human activity level in the area is, and the greater the possibility of human-caused forest fires is. Catry et al. (2007) and Sepulveda (2001) reached the same conclusion[91][92]. The variables excluded in feature selection include the average wind speed, maximum wind speed, slope direction, slope, nearest distance from the fire point to the highway, nearest distance from the fire point to the residential area, and special festivals. This result indicated that in this experiment, these 7 variables had little influence on the occurrence of forest fires. However, these factors may vary with time and space. Other studies have found that these factors are the main driving factors of forest fires[93][94]. We believe that this difference may be due to the different data selected and the different feature selection methods. In future studies, multiple feature screening methods and analyses of different regions may be used to obtain more comprehensive results.

4.2 Optimal Choice of Forest Fire Prediction Model

We entered the forest fire driving factors selected by feature selection into the four models (ANN, RBFNN, SVM and RF) for training. We then evaluated them using five criteria: accuracy, precision, recall rate, f1 value, and AUC value. We selected the RF model as the optimal choice for forest fire prediction. The accuracies of all four models were above 75%, which meant that they were all reliable. The RF model, however, exhibited the greatest prediction ability. The RBFNN model had the lowest prediction performance.

Samaher et al. (2018)used a cascade correlation network, multilayer perceptron neural network, polynomial neural network, RBF, and SVM for forest fire prediction [27]. They found that the prediction performance of the SVM was the highest, and the performance of RBFNN was the lowest, which was consistent with our conclusion. Sakr et al. (2011)used an SVM and an ANN to predict fire risk in Lebanon

476 [43]. Their results showed that the performance of the SVM model was higher than that of the ANN
477 model. This finding was similar to ours. Paulo et al. (2007) used RF and SVM models to predict forest
478 fires [8]. They concluded that the performance of the SVM model was higher than that of the RF model.
479 Their findings differed from our results. A possible reason for this difference is that Paulo et al. chose
480 four types of weather factors and made predictions for small areas, whereas we chose 13 types of factors
481 and made predictions for a large area. The choice of variables and the difference in the sample size affect
482 the model training.

483 Bisquert et al. (2012)used an ANN to establish a forest fire hazard model with the highest accuracy rate
484 of 76%, which was lower than the accuracy of our model (83%)[95]. Hong (2018)used an SVM algorithm
485 to analyse Dayu County in southwestern Jiangxi Province, China [96]. The results showed that the AUC
486 value of the SVM was 0.75, which was lower than the value in our model (0.92). Pourtaghid et al.
487 (2016)used an RF to conduct forest fire sensitivity analysis with a prediction accuracy of 72.8% [51].
488 Our model reached a prediction accuracy of 89.2%.

489 The four models we selected all exhibited high predictive capabilities. The main reason for this result
490 may be that appropriate multidimensional variables have been screened out and the data sample size is
491 large, which makes the training of each model more accurate and reliable.

492 There are also differences in the characteristics of these four models. The ANN and RBFNN models can
493 be trained very quickly, and they can handle samples with a large amount of data, but their accuracy in
494 this experiment is relatively low. In subsequent studies, particle swarms or genetic algorithms could be
495 used to improve the accuracy of these models. The SVM model has a high predictive ability, but it also
496 has shortcomings. The higher the model complexity is, the lower the calculation speed is. It takes a longer

time in this model to obtain the optimal parameters when processing large amounts of sample data. We will consider using other algorithms to optimize the SVM model in the future. The RF model exhibited excellent expressive power in this experiment. It can quickly process large data samples while ensuring a high prediction accuracy.

4.3 Recommendations for Forest Fire Prevention

We produced a probability map for forest fires in China that showed that the highest incidences of forest fires were in the northeast (Heilongjiang Province and the northern Inner Mongolia Autonomous Region), the southeast (Fujian Province, Guangdong Province, and Jiangxi Province), and Yunnan Province. The pattern of forest fire points presented a spatial clustering distribution. Ma et al. obtained similar results[4]. For these high-incidence areas, watch towers and monitoring equipment should be added for monitoring and management. Moreover, the length of the forest fire barrier net should be increased to reduce the spread of fires. In addition, the number of fire brigades and fire vehicles should be increased to enhance the disaster-mitigation capabilities. Regarding seasonal forest fire risks, forest fire prevention and control should be emphasized in spring and winter. Strengthening fire-prevention management during these periods would mainly involve strengthening the management of human activities to reduce human-made forest fires and improving publicity and education, such as the addition of fire-prevention signs.

This study has some shortcomings, and there is room for improvement. One of the three elements of fire is combustible fuel. For the selection of forest fire driving factors, however, there is currently no way to obtain data on fuel load and other related factors. Thus, this experiment lacked relevant data such as the combustible load, particle size of combustible material, and combustible tree species. If possible, in future research, such data could be added to the forest fire prediction model.

This study selected four kinds of machine learning algorithms for the forest fire prediction model. Other applicable machine learning algorithms could be used in future experiments. In addition, the ability of these machine learning algorithms to analyse spatial heterogeneity is relatively weak. Subsequent research could use geographically weighted regression to build a high-precision forest fire prediction model.

5. Conclusion

This study determined the main driving factors of forest fire occurrence in China through feature selection. The main factors that affected the occurrence of forest fires to varying degrees were meteorological, topographic, man-made, and vegetation factors. We built four forest fire prediction models using the following machine learning algorithms: an ANN, an RBFNN, an SVM, and an RF. The results of the evaluation showed that the accuracy of all the models was higher than 75%. Thus, these models can be used to build forest fire prediction models. Among the four models, the RF model had the highest comprehensive predictive ability, with an accuracy of 89.25%. It was therefore the optimal choice for a forest fire prediction model in China.

We used the RF model to predict the probabilities of forest fires in China. Based on these probabilities, we drew a map of the probability of forest fire occurrence in China and a map of the probability of forest fires in China by season (spring, summer, autumn, and winter). Finally, based on these maps, we identified the high-incidence areas and areas at risk of forest fires. We then put forward fire prevention recommendations for the corresponding regions and seasons.

This research helps to understand the main forest fire driving factors in China. This study provides a reference for the selection of high-precision forest fire prediction models. In addition, it provides

suggestions regarding the optimal times and locations of forest fire prevention in China. Finally, this study provides guidance for China's forest fire prevention and control work.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Funding Statement

This research was jointly supported by the medium long-term project of "Precision Forestry Key Technology and Equipment Research" (No. 2015ZCQ-LX-01) and the National Natural Science Foundation of China (No. U1710123).

Authors' contributions

Yudong Li performed the experiment and wrote the manuscript;

Zhongke Feng contributed to the conception of the study;

Ziyu Zhao, Wenyuan Ma and Shilin Chen helped perform the analysis with constructive discussions;

Hanyue Zhang helped perform the data analysis.

Acknowledgments

We would like to acknowledge support from Beijing Key Laboratory for Precision Forestry, Beijing Forestry University, as well as all the people who have contributed to this paper.

References

[1] Bhusal, Satish, Mandal, Ram (2020) Forest fire occurrence, distribution and future risks in Arghakhanchi district, Nepal. Journal of Geography2(1):10-20. https://www.researchgate.net/publication/341701669.

[2] Singh B. K., N. Kumar, P. Tiwari (2019) Extreme Learning Machine Approach for Prediction of Forest Fires using Topographical and Metrological Data of Vietnam. 2019 Women Institute of Technology Conference on Electrical and Computer Engineering (WITCON ECE)- Dehradun Uttarakhand, India 104-112. DOI: 10.1109/WITCONECE48374.2019.9092926.

[3] D Mandallaz, R Ye (1997) Prediction of forest fires with Poisson models. Canadian Journal of Forest Research 27(10): 1685-1694. DOI: 10.1139/x97-103.

[4] Ma W., Feng Z., Cheng Z., Chen S., Wang F. (2020) Identifying Forest Fire Driving Factors and Related Impacts in China Using Random Forest Algorithm. Forests. 11(5), 507. DOI:10.3390/f11050507.

[5] Dimopoulou Maria, Giannikos Ioannis (2004) Towards an integrated framework for forest fire control. Eur. J. Oper. Res152(2):476-486. DOI: 10.1016/S0377-2217(03)00038-9.

[6] Flannigan MD, Krawchuk MA, Groot WJD, Wotton BM, et al. (2009) Implications of changing climate for global wildland fire. International Journal of Wildland Fire18(5):483–507.DOI：10.1071/WF08187.

[7] Xu ZQ, Su XY, Zhang Y (2012) Forest Fire Prediction Based on Support Vector Machine. Chinese Agricultural Science Bulletin28(13):126-131. Available online: http://en.cnki.com.cn/Article_en/CJFDTOTAL-ZNTB201213025.htm. (accessed on 24 February 2012).

[8] Paulo Cortez, Anibal Morais (2007) A Data Mining Approach to Predict Forest Fires using Meteorological Data. New Trends in Artificial Intelligence512-523.DOI：US7517766 B2.

[9] Naderpour M., Rizeei H. M., Khakzad N., Pradhan B. (2019) Forest fire induced Natech risk assessment: a survey of geospatial technologies. Reliability Engineering and System Safety 191:106558. DOI: 10.1016/j.ress.2019.106558.

[10] Miguel Boubeta, María José, Lombardía, et al. (2015) Prediction of forest fires occurrences with area-level poisson mixed models. Journal of Environmental Management 154:151-158.DOI:10.1016/j.jenvman.2015.02.009.

[11] Kanga S, Kumar S, Singh SK(2017) Climate induced variation in forest fire using remote sensing and GIS in Bilaspur district of Himachal Pradesh. International Journal Of Engineering And Computer Science 6:21695–702.DOI：10.18535/ijecs/v6i6.23.

[12] Erten E, Kurgun V, Musaoglu N(2004) Forest fire risk zone mapping from satellite imagery and GIS: a case study. XXth Congress of the International Society for Photogrammetry and Remote Sensing, Istanbul. pp: 222-230.

[13] Chuvieco E, Aguadoa I, Yebraa M(2010) Development of a framework for fire risk assessment using remote sensing and geographic information system technologies. Ecological Modelling 221:46-58.DOI：10.1016/j.ecolmodel.2008.11.017.

[14] Adab H, Devi Kanniah K, Solaimani K (2013) Modeling forest fire risk in the northeast of Iran using remote sensing and GIS techniques. Nat Hazards 65(3):1723-1743. DOI: 10.1007/s11069-012-0450-8.

[15] Sachdeva S, Bhatia T, Verma A. GIS-based evolutionary optimized gradient boosted decision trees for forest fire susceptibility mapping. Nat Hazards 2018(92):1399–418. DOI:10.1007/s11069-018-

602 3256-5.

603 [16] Dhall A., Dhasade A., Nalwade A., et al. (2020) A survey on systematic approaches in managing
604 forest fires. Applied Geography 121:102266. DOI: 10.1016/j.apgeog.2020.102266.

605 [17] Maffei C., Menenti M. (2019) Predicting forest fires burned area and rate of spread from pre-fire
606 multispectral satellite measurements. ISPRS Journal of Photogrammetry and Remote Sensing
607 158:263-278.DOI：10.1016/j.isprsjprs.2019.10.013.

608 [18] Shang C, Wulder M A, Coops N C, et al. (2020) Spatially-Explicit Prediction of Wildfire Burn
609 Probability Using Remotely-Sensed and Ancillary Data. Canadian Journal of Remote Sensing
610 46(8):1-17. DOI: 10.1080/07038992.2020.1788385.

611 [19] SU ZW, LIU AQ, et al. (2016) Driving factors and spatial distribution patteren of forest fire in
612 Fujian Province. JOURNAL OF NATURAL DISASTERS25(2):110-119. DOI:
613 10.13577/j.jnd.2016.0213.

614 [20] CV Garcia, PM Woodard, SJ Titus, et al. (1995) A Logit Model for Predicting the Daily Occurrence
615 of Human Caused Forest-Fires. International Journal of Wildland Fire5(2):101-111. DOI: DOI:
616 10.1071/WF9950101.

617 [21] Futao Guo, Selvara Selvalakshmi, Fangfang Lin, et al. (2016) Geospatial information on
618 geographical and human factors improved anthropogenic fire occurrence modeling in the Chinese
619 boreal forest. Canadian Journal of Forest Research 46(4): 582-594. DOI:10.1139/cjfr-2015-0373.

620 [22] K.Z. Liu, L.F. Shu, F.J. Zhao et al. (2017) Research on spatial distribution of forest fire based on
621 satellite hotspots data and forecasting model. Journal of Forestry Engineering2(04):128-133.
622 Available online: http://en.cnki.com.cn/Article_en/CJFDTOTAL-LKKF201704021.htm. (accessed
623 on April 2014).

[23] B.Q. LIAO, J. WEI et al. (2008) Logistic and ZIP Regression Model for Forest Fire Data. FIRE SAFETY SCIENCE3:143-149. Available online: http://en.cnki.com.cn/Article_en/CJFDTOTAL-HZKX200803002.htm. (accessed on March 2008).

[24] D Mandallaz, R Ye (1997) Prediction of forest fires with Poisson models. Canadian Journal of Forest Research27(10): 1685-1694. DOI:10.1139/x97-103.

[25] F.T. GUO, H.Q. HU et al. (2010) Relationship between forest lighting fire occurrence and weather factors in Daxing'an Mountains based on negative binomial model and zero-inflated negative binomial models. Chinese Journal of Plant Ecology21(01):159-164. Available online: http://en.cnki.com.cn/Article_en/CJFDTOTAL-ZWSB201005014.htm. (accessed on May 2010).

[26] H.L. Liang, Y.R. Lin, G. Yang et al. (2016) Application of Random Forest Algorithm on the Forest Fire Prediction in Tahe Area Based on Meteorological Factors. SCIENTIA SILVAE SINICAE.52(01):89-98. Available online: http://en.cnki.com.cn/Article_en/CJFDTotal-LYKE201601011.htm. (accessed on January 2016).

[27] Samaher Al-Janabia et al. (2018) Assessing the suitability of soft computing approaches for forest fires prediction. Applied Computing and Informatics14(2):214-224. DOI: 10.1016/j.aci.2017.09.006.

[28] Volkan Sevinca, Omer Kucukb, Merih Goltasc (2020) A Bayesian network model for prediction and analysis of possible forest fire causes. Forest Ecology and Management457:117723. DOI: 10.1016/j.foreco.2019.117723.

[29] Artés, T., Cencerrado, A., Cortés, A., and Margalef, T (2017) Time aware genetic algorithm for forest fire propagation prediction: exploiting multi-core platforms. Concurrency Computat.: Pract. Exper. 29(9): 3837. DOI:10.1002/cpe.3837.

646 [30] D.T. Bui, Q.T. Bui, Q.P. Nguyen, B. Pradhan, H. Nampak, P.T Trinh (2017) A hybrid artificial
647 intelligence approach using GIS-based neural-fuzzy inference system and particle swarm
648 optimization for forest fire susceptibility modeling at a tropical area, Agric. Agricultural and Forest
649 Meteorology. 233:32-44. DOI:10.1016/j.agrformet.2016.11.002.
650 [31] M. Denham, A. Cortés, T. Margalef, E. Luque (2008) Applying a dynamic data driven genetic
651 algorithm to improve forest fire spread prediction. International Conference on Computational
652 Science. Springer Berlin Heidelberg36-45. DOI:10.1007/978-3-540-69389-5_6.
653 [32] H. Hong, S.A. Naghibi, M.M. Dashtpagerdi, H.R. Pourghasemi, W. Chen (2017) A comparative
654 assessment between linear and quadratic discriminant analyses (LDA-QDA) with frequency ratio
655 and weights-of-evidence models for forest fire susceptibility mapping in China. Arabian Journal of
656 Geosciences10 (7):167. DOI:10.1007/s12517-017-2905-4.
657 [33] A.M. Özbayog ˘lu, R. Bozer (2012) Estimation of the burned area in forest fires using computational
658 intelligence techniques, Procedia Computer Science12:282–287. DOI:
659 10.1016/j.procs.2012.09.070.
660 [34] You Y, Lu C, Wang W, Tang C-K (2019) Relative CNN-RNN: learning relative atmospheric
661 visibility from images. IEEE Trans Image Process 28(1):45–55.
662 https://doi.org/10.1109/TIP.2018.2857219.
663 [35] Govil K, Welch ML, Ball JT, Pennypacker CR (2020) Preliminary results from a wildfire detection
664 system using deep learning on remote camera images. Remote Sensing 12(1):166.
665 DOI:10.3390/rs12010166.
666 [36] Camp A, Oliver C, Hessburg P, Everett R (1997) Predicting late-successional fire refugia pre-dating
667 European settlement in the Wenatchee Mountains. Forest Ecology and Management 95(1):63-77.

668 DOI:10.1016/S0378-1127(97)00006-6.

669 [37] Hr Pourghasemi, Beheshtirad M, Pradhan B (2016) A comparative assessment of prediction
670 capabilities of modified analytical hierarchy process (M-AHP) and Mamdani fuzzy logic models
671 using Netcad-GIS for forest fire susceptibility mapping. Geomat Nat Hazards Risk 7(2):861–885.
672 https://doi.org/10.1080/19475705.2014.984247.

673 [38] Li Z, Huang Y, Li X, Xu L (2020) Wildland Fire Burned Areas Prediction Using Long Short-Term
674 Memory Neural Network with Attention Mechanism. Fire Technology. DOI:10.1007/s10694-020-
675 01028-3.

676 [39] H. Soliman, K. Sudan and A. Mishra (2010) A smart forest-fire early detection sensory system:
677 Another approach of utilizing wireless sensor and neural networks. SENSORS, 2010 IEEE, Kona,
678 HI 1900-1904.DOI:10.1109/ICSENS.2010.5690033.

679 [40] Çetin Elmas, Yusuf Sönmez (2011) A data fusion framework with novel hybrid algorithm for multi-
680 agent Decision Support System for Forest Fire. Expert Systems with Applications.38(8):9225-9236.
681 https://doi.org/10.1016/j.eswa.2011.01.125.

682 [41] Onur Satir, Suha Berberoglu & Cenk Donmez (2016) Mapping regional forest fire probability using
683 artificial neural network model in a Mediterranean forest ecosystem, Geomatics. Natural Hazards
684 and Risk7:1645-1658. DOI:10.1080/19475705.2015.1084541.

685 [42] Maeda E. E., Formaggio A. R., Shimabukuro Y. E., et al. (2009) Predicting forest fire in the
686 Brazilian Amazon using MODIS imagery and artificial neural networks. International Journal of
687 Applied Earth Observation and Geoinformation 11(4):265–272. DOI:10.1016/j.jag.2009.03.003.

688 [43] Sakr G E, Elhajj I H, Mitri G (2011) Efficient forest fire occurrence prediction for developing
689 countries using two weather parameters. Engineering Applications of ArtificialIntelligence24(5):

888-894. DOI:10.1016/j.engappai.2011.02.017.

[44] B.C. Ko, K.H. Cheong, J.Y. Nam (2009) Fire detection based on vision sensor and support vector machines. Fire Safety Journal 44 (3): 322-329. DOI:10.1016/j.firesaf.2008.07.006.

[45] D.W. Xie, S.L. Shi (2014) Prediction for burned area of forest fires based on SVM model. Applied Mechanics and Materials513(5):4084-4089. DOI:10.4028/www.scientific.net/AMM.513-517.4084.

[46] J. Zhao, Z. Zhang, S. Han, et al. (2011) SVM based forest fire detection using static and dynamic features. Computer Science and Information Systems8(3): 821–841. DOI:10.2298/CSIS101012030Z.

[47] Cutler DR, Edwards TC, Beard KH (2007) Random forests for classification in Ecology. Ecology88(11):2783-2792. DOI:10.1890/07-0539.1.

[48] Anantha M. Prasad, Louis R. Iverson, Andy Liaw (2006) Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. Ecosystems9(2):181-199.DOI:10.1007/s10021-005-0054-1.

[49] Rodrigucs M, Dc la Riva J (2014) An insight into machines learning algorithms to model humarrcaused wildfire or currence. Environmental Modelling & Software57:192-201.DOI:10.1016/j.envsoft.2014.03.003.

[50] LIANG HL, LIN YR, YANG G, et al. (2016) Application of random forest algorithm on the forest fire prediction in Tahe area based on meteorological factors. Scientla Silvae Sinicae52(1):89-98.DOI:10.11707/j.1001-7488.20160111.

[51] Pourtaghi Z.S., Pourghasemi H. R., Aretano R., Semeraro T. (2016) Investigation of general indicators influencing on forest fire and its susceptibility modeling using different data mining techniques. Ecological Indicators. 64:72–84. DOI:10.1016/j.ecolind.2015.12.030.

712 [52] Tian X., Zhao F., Shu L., et al. (2013) Distribution characteristics and the influence factors of forest
713 fires in China. Forest Ecology and Management 310:460–467. DOI:10.1016/j.foreco.2013.08.025.
714 [53] Chang Y., Zhu Z., Bu R., et al. (2015) Environmental controls on the characteristics of mean number
715 of forest fires and mean forest area burned (1987–2007) in China. Forest Ecology and Management
716 356:13–21.DOI:10.1016/j.foreco.2015.07.012.
717 [54] Zhong M., Fan W., Liu T., Li P. (2003) Statistical analysis on current status of China forest fire
718 safety. Fire Safety Journal 38(3):257–269. DOI： 10.1016/S0379-7112(02)00079-6.
719 [55] LU A.F. (2011) Study on the relationship among forest fire, temperature and precipitation and its
720 spatial-temporal variability in China. Institute of Geographic Sciences and Natural Resources
721 Research12:1396–1400. Available online: http://en.cnki.com.cn/Article_en/CJFDTOTAL-
722 HNNT201109040.htm. (accessed on September 2011).
723 [56] Ying. L., Han. J., Du. Y., Shen Z. (2018) Forest fire characteristics in China: Spatial patterns and
724 determinants with thresholds. For. Ecol. Manag424:345–354.DOI： 10.1016/j.foreco.2018.05.020.
725 [57] F.T. Guo, Z.W. Su, G.Y. Wang et al. (2017) Understanding fire drivers and relative impacts in
726 different Chinese forest ecosystems. Science of the Total Environment605: 411-425. DOI:
727 10.1016/j.scitotenv.2017.06.219.
728 [58] GUO F, WANG G, SU Z, et al. (2016) What drives forest fire in Fujian, China? Evidence from
729 logistic regression and random forests. International Journal of Wildland Fire25(5):505-519. DOI:
730 10.1071/WF15121.
731 [59] Chang Y, Zhu Z L, Bu R C, et al. (2013) Predicting fire occurrence patterns with logistic regression
732 in Heilongjiang Province, China. Landscape Ecology28(10):1989-2004.
733 https://doi.org/10.1007/s10980-013-9935-4.

[60] XU X L. (2018) Spatial distribution data set of quarterly vegetation index (NDVI) in China. Data Registration and Publishing System of Resources and Environmental Science Data Center of Chinese Academy of Sciences (http://www.resdc.cn/DOI),2018.DOI:10.12078/2018060601.

[61] A. Subasi, (2007) EEG signal classification using wavelet feature extraction and a mixture of expert model. Expert Systems with Applications32 (4): 1084-1093.DOI：10.1016/j.eswa.2006.02.005.

[62] D. Chen (2019) Prediction of Forest Fire Occurrence in Daxing'an Mountains Based on Logistic Regression Model. FOREST RESOURCES MANAGEMENT (01):116-122. Available online: http://en.cnki.com.cn/Article_en/CJFDTotal-LYZY201901018.htm. (accessed on January 2019).

[63] X.C. Wang et al. (2013) 43 Cases of MATLAB neural network analysis. Beijing, Bei hang University Press.

[64] Yudong Li, Zhongke Feng , Shilin Chen, Ziyu Zhao, and Fengge Wang (2020) Application of the Artificial Neural Network and Support Vector Machines in Forest Fire Prediction in the Guangxi Autonomous Region, China. Discrete Dynamics in Nature and Society 2020:14. https://doi.org/10.1155/2020/5612650.

[65] Ganteaume A, Camia A, Jappiot M, et al. (2013) A review of the main driving factors of forest fire ignition over Europe. Environmental Management 51(3): 651-662.DOI:10.1007/s00267-012-9961-z.

[66] Zhu Feng, Lu Liu (2020) Estimation of forest biomass in beijing (china) using multisource remote sensing and forest inventory data. Forests 11(2): 163.DOI:10.3390/f11020163.

[67] Breiman (2001) Random forests. Machine Lear 45:15–32. DOI：10.1023/A:1010933404324.

[68] Cutler DR, Edwards TJ, Beard KH, et al. (2007) Random forests for classification in ecology. Ecology 88(11):2783–2792.DOI:10.1890/07-0539.1.

[69] Kubosova K, Brabec K, Jarkovsky J, Syrovatka V (2010) Selection of indicative taxa for river habitats: a case study on benthic macroinvertebrates using indicator species analysis and the random forest methods. Hydrobiologia 651: 101-114.DOI:10.1007/s10750-010-0280-1.

[70] Zhangwen Su, Haiqing Hu, Guangyu Wang, et al. (2018) Using GIS and Random Forests to identify fire drivers in a forest city, Yichun, China. Geomatics, Natural Hazards and Risk 9(1): 1207-1229, DOI:10.1080/19475705.2018.1505667.

[71] Oliveira S, Oehler F, San-Miguel-Ayanz J, Camia A, Pereira J. 2012. Modeling spatial patterns of fire occurrence in Mediterranean Europe using multiple regression and random forest. Forest Ecol Manage. 275:117-129.DOI: 10.1016/j.foreco.2012.03.003.

[72] Rodrigues M, Riva JDL (2014) An insight into machine-learning algorithms to model human-caused wildfire occurrence. Environ Model Softw 57:192-201. DOI: 10.1016/j.envsoft.2014.03.003.

[73] Kane VR, Lutz JA, Alina Cansler C, et al. (2015) Water balance and topography predict fire and forest structure patterns. Forest Ecol Manag 338:1-13. DOI: 10.1016/j.foreco.2014.10.038.].

[74] ZA Fang，XY B (2020). Improving land cover classification in an urbanized coastal area by random forests: the role of variable selection. Remote Sensing of Environment 251:112105. DOI: 10.1016/j.rse.2020.112105.

[75] Epifanio, I. (2017) Intervention in prediction measure: a new approach to assessing variable importance for random forests. BMC Bioinform 18:230. https://doi.org/10. 1186/s12859-017-1650-8.

[76] Chan, C. W., Desiré Paelinckx (2008). Evaluation of random forest and adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. Remote Sensing of Environment 112(6):2999-3011.DOI :

778 10.1016/j.rse.2008.02.011.

779 [77] Zhangwen Su, Haiqing Hu, Guangyu Wang, et al. (2018) Using GIS and Random Forests to identify
780 fire drivers in a forest city, Yichun, China. Geomatics, Natural Hazards and Risk 9(1): 1207-1229,
781 DOI:10.1080/19475705.2018.1505667.

782 [78] Gromping U. (2009) Variable importance assessment in regression: linear regression versus random
783 forest. The American Statistician 63(4):308–319.DOI:10.1198/tast.2009.08199.

784 [79] Ganteaume A, Camia A, Jappiot M, et al. (2013) A review of the main driving factors of forest fire
785 ignition over Europe. Environmental Management 51(3): 651-662.DOI： 10.1007/s00267-012-
786 9961-z.

787 [80] Syphard AD, Radeloff VC, Keuler NS, et al. (2008) Predicting spatial patterns of fire on a southern
788 California landscape. International Journal of Wildland Fire 17:602–613.DOI： 10.1071/WF07087.

789 [81] Pew KL, Larsen CPS (2001) GIS analysis of spatial and temporal patterns of human-caused
790 wildfires in the temperate rainforest of Vancouver Island, Canada. Forest Ecol Manag 140:1-18.DOI：
791 10.1016/S0378-1127(00)00271-1.

792 [82] Dickson BG, Prather JW, Xu Y. et al. (2006) Mapping the probability of large fire occurrence in
793 northern Arizona, USA. Landscape Ecol 21: 747–761. https://doi.org/10.1007/s10980-005-5475-x.

794 [83] Ying L., Han J., Du Y., Shen Z. (2018) Forest fire characteristics in China: Spatial patterns and
795 determinants with thresholds. Forest Ecology and Management 424:345-
796 354.DOI:10.1016/j.foreco.2018.05.020.

797 [84] Prasad A.M., Iverson L.R., Liaw A. (2006) Newer classification and regression tree techniques:
798 Bagging and random forests for ecological prediction. Ecosystems 9:181-199.
799 DOI :10.1007/s10021-005-0054-1.

800 [85] Liu Z., Yang J., Chang Y., et al. (2012) Spatial patterns and drivers of fire occurrence and its future
801 trend under climate change in a boreal forest of northeast China. Global Change Biology 18:2041-
802 2056.DOI:10.1111/j.1365-2486.2012.02649.x.
803 [86] Syphard A. D., Radeloff V. C., Keuler N. S., et al. (2008) Predicting spatial patterns of fire on a
804 southern California landscape. International Journal of Wildland Fire 17:602-
805 613.DOI:10.1071/WF07087.
806 [87] Zumbrunnen T., Pezzatti G. B., Menéndez P., et al. (2011) Weather and human impacts on forest
807 fires: 100 years of fire history in two climatic regions of Switzerland. Forest Ecology and
808 Management 261:2188-2199. DOI:10.1016/j.foreco.2010.10.009.
809 [88] Cardille J A, Ventura S J, Turner M G, et al. (2001) ENVIRONMENTAL AND SOCIAL FACTORS
810 INFLUENCING WILDFIRES IN THE UPPER MIDWEST, UNITED STATES. Ecological
811 Applications 11(1): 111-127. DOI:10.1890/1051-0761(2001)011[0111: EASFIW]2.0.CO;2.
812 [89] Tian X., Zhao F., Shu L., Wang M. (2013) Distribution characteristics and the influence factors of
813 forest fires in China. Forest Ecology and Management 310:460-
814 467.DOI:10.1016/j.foreco.2013.08.025.
815 [90] Chuvieco E., Cocero D., Riaño D., et al. (2004) Combining ndvi and surface temperature for the
816 estimation of live fuel moisture content in forest fire danger rating. Remote Sensing of Environment
817 92: 322-331. DOI:10.1016/j.rse.2004.01.019.
818 [91] Catry F. X., Damasceno P., Silva J. S., et al. (2007). Spatial distribution patterns of wildfire ignitions
819 in Portugal. Modelação espacial do risco de ignição em Portugal Continental, 8. Project: Fire
820 ecology and post-fire restoration. Available online:
821 https://www.researchgate.net/publication/240613824_Spatial_Distribution_Patterns_of_Wildfire_

822 Ignitions_in_Portugal. (accessed on January 2007).

823 [92] Avila-Flores D., Pompa-Garcia M., Antonio-Nemiga X., et al. (2010). Driving factors for forest fire
824 occurrence in Durango State of Mexico: A geospatial perspective. Chinese Geographical Science
825 20(6), 491-497. DOI:10.1007/s11769-010-0437-x.

826 [93] Maingi K J, Henry M C (2007) Factors influencing wildfire occurrence and distribution in eastern
827 Kentucky, USA. International Journal of Wildland Fire 16(1): 23-33. DOI: 10.1071/ WF06007.

828 [94] Avilaflores D Y, Pompagarcia M, Antonionemiga X, et al. (2010) Driving factors for forest fire
829 occurrence in Durango State of Mexico: A geospatial perspective. Chinese Geographical Science.
830 20(6): 491-497. DOI:10.1007/s11769-010-0437-x.

831 [95] Bisquert M, Caselles E, Sanchez J M, et al. (2012) Application of artificial neural networks and
832 logistic regression to the prediction of forest fire danger in Galicia using MODIS data. International
833 Journal of Wildland Fire 21(8): 1025-1029. DOI:10.1071/WF11105.

834 [96] Hong, H., et al. (2018) Applying genetic algorithms to set the optimal combination of forest fire
835 related variables and model forest fire susceptibility based on data mining models. The case of Dayu
836 County, China. Science of The Total Environment 630:1044-1056. DOI:
837 10.1016/j.scitotenv.2018.02.278.

Figures

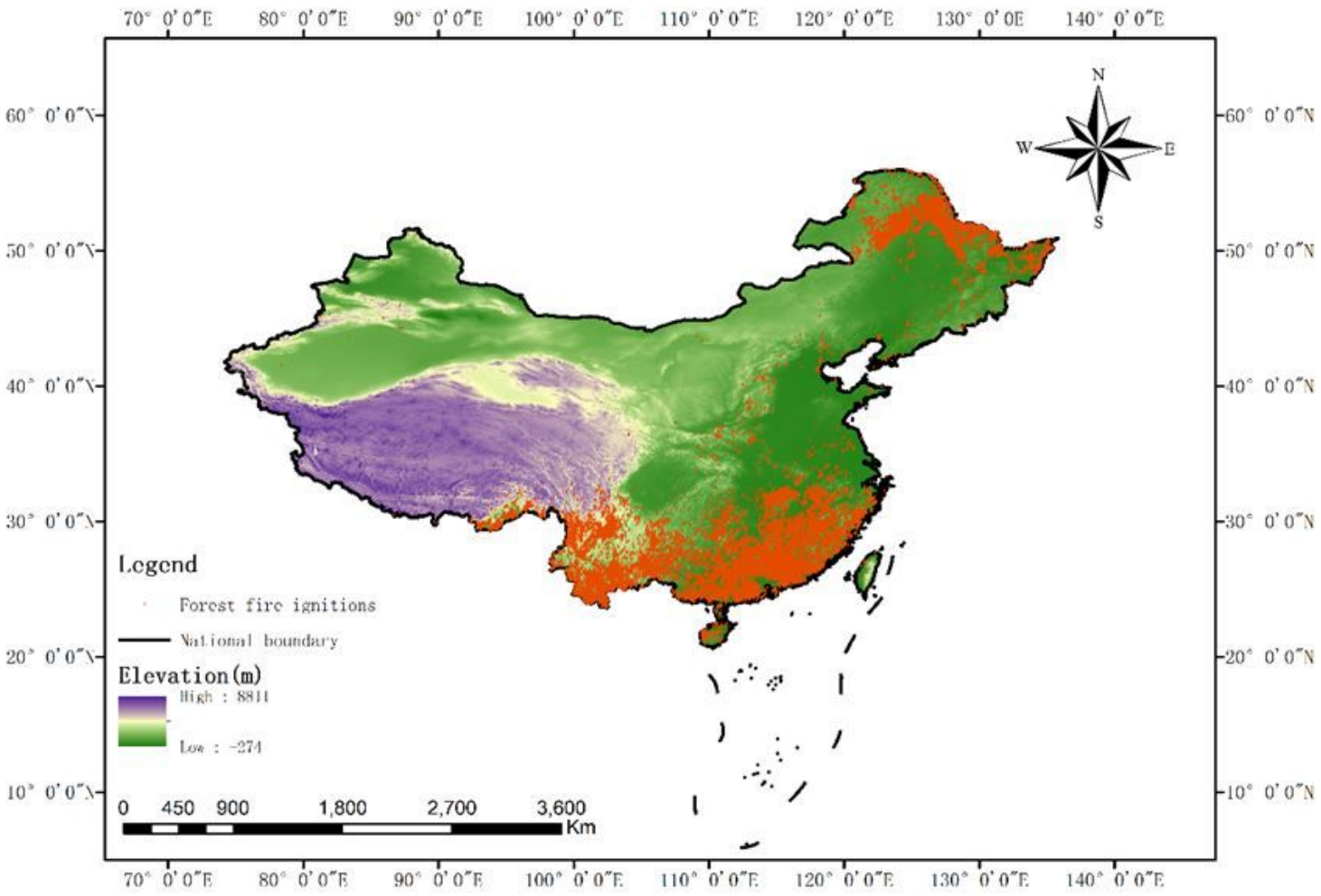


Figure 1

Map of the study area. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

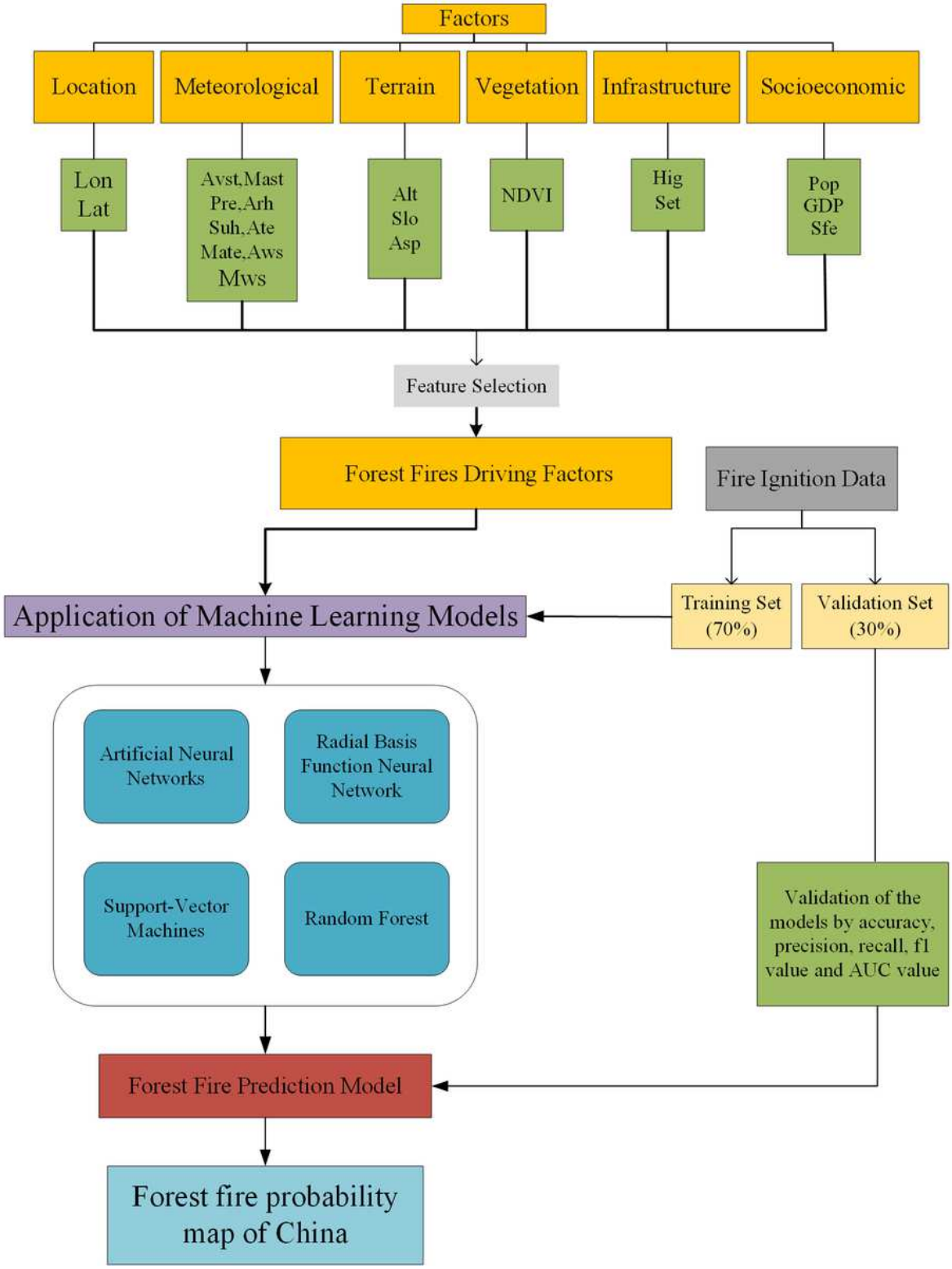


Figure 2

Flowchart of the Chinese forest fire occurrence prediction.

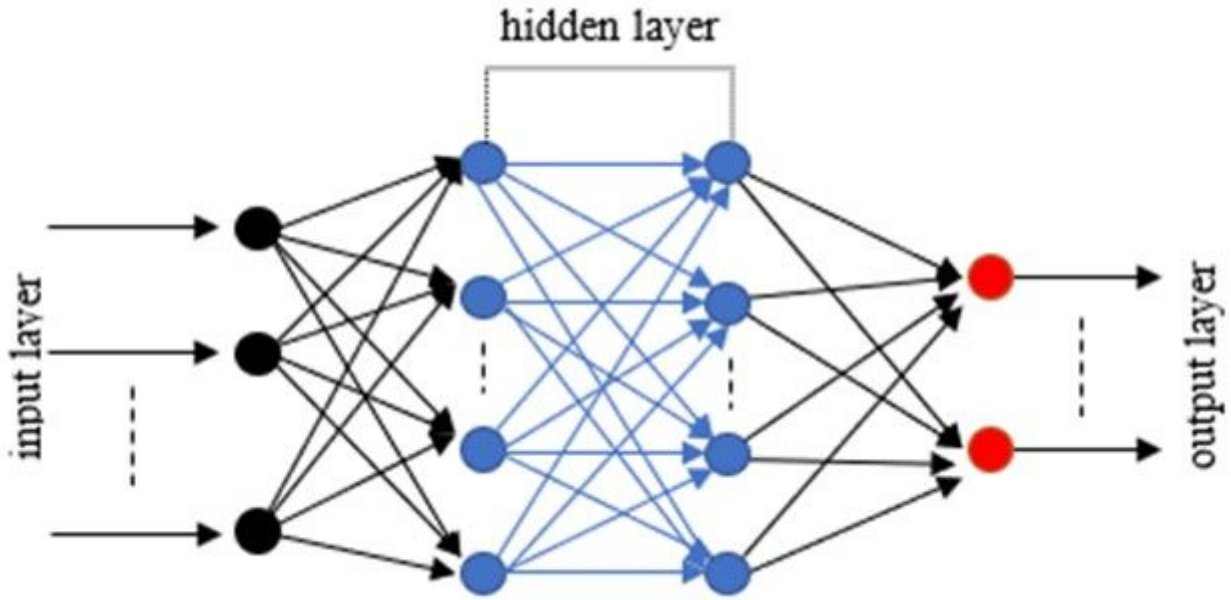


Figure 3

Diagram of the structure of an ANN.

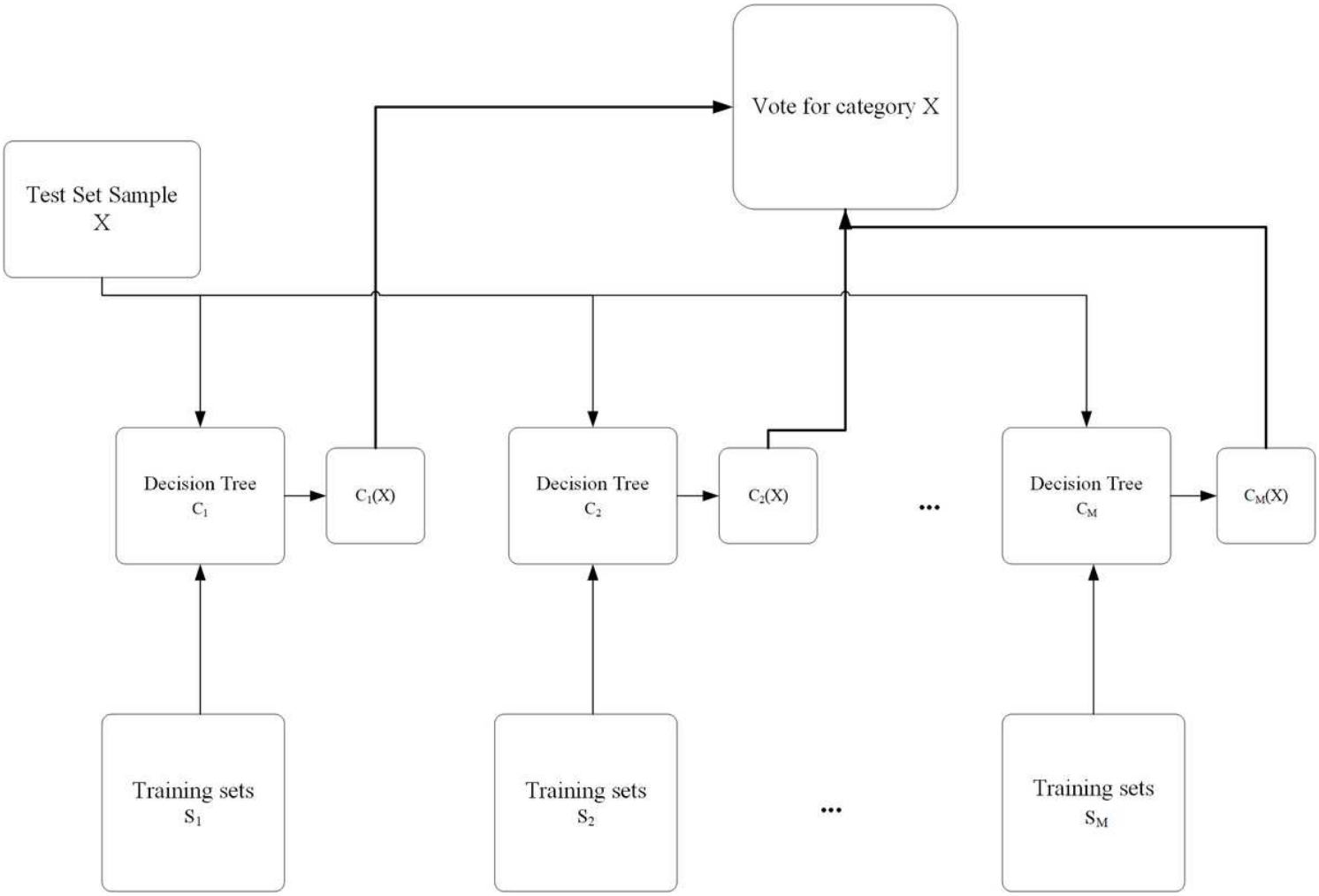


Figure 4

The flow chart of the RF algorithm

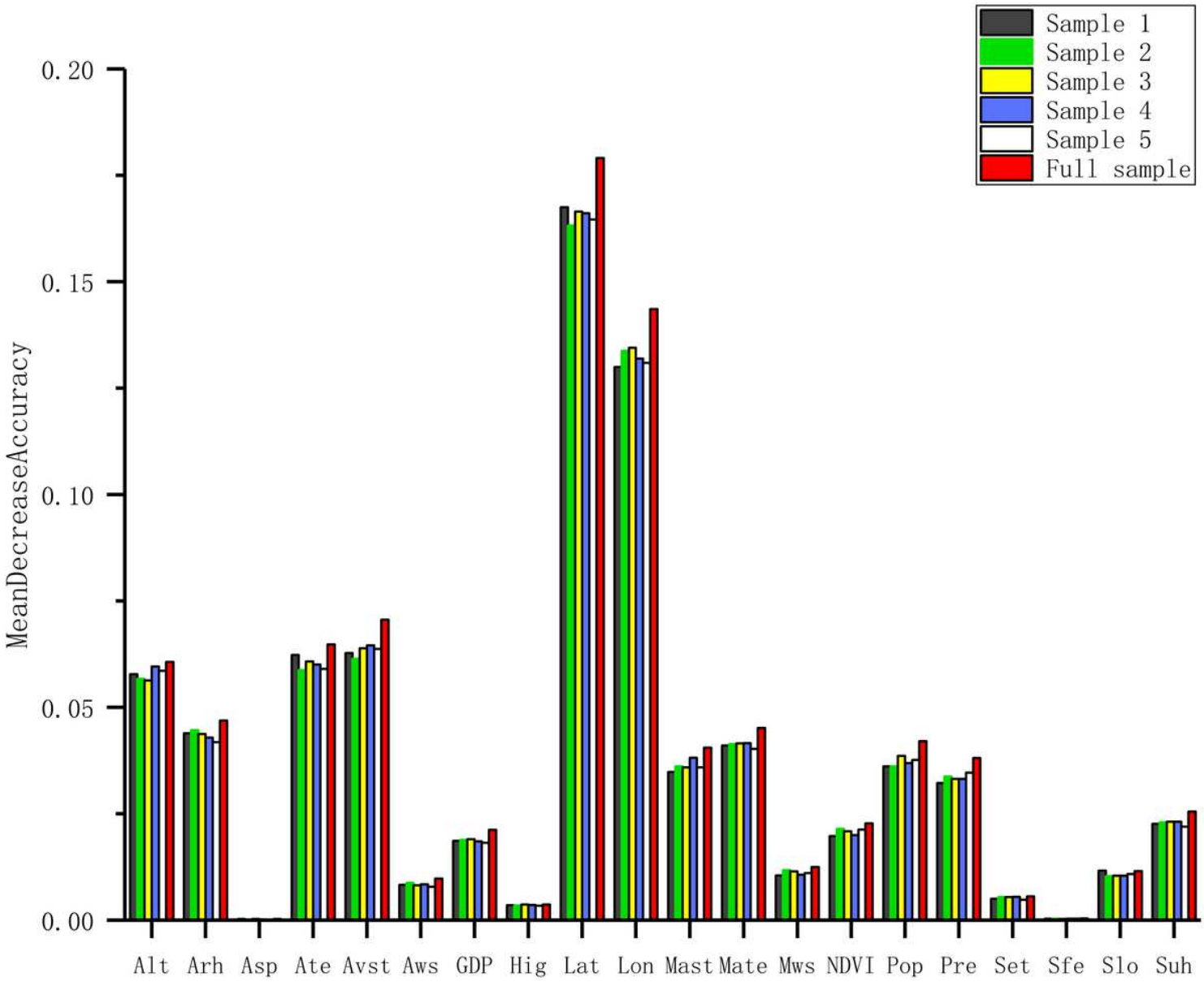


Figure 5

Feature subset importance.

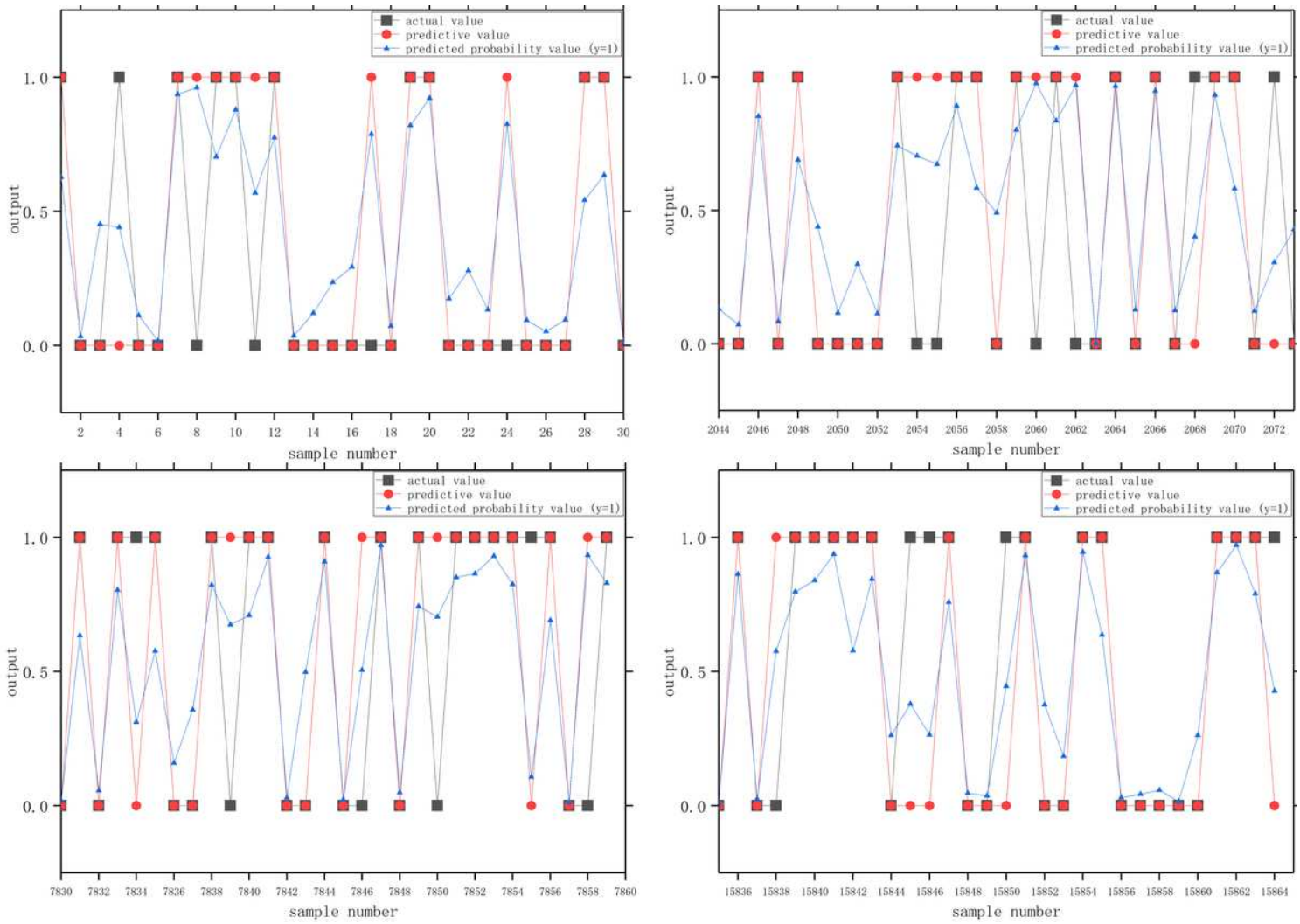


Figure 6

Comparison charts of the predictive and actual values of the ANN.

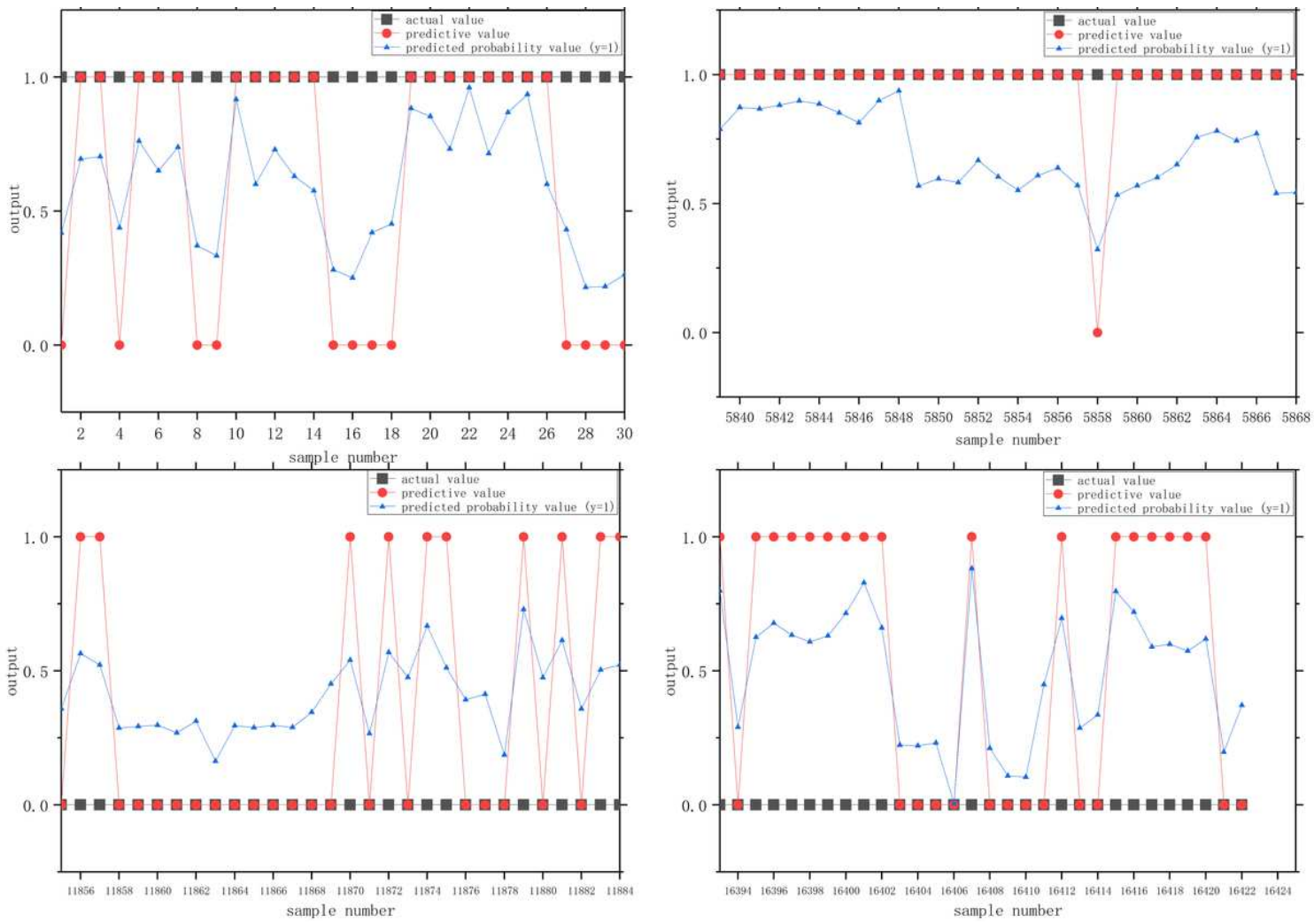


Figure 7

Comparison charts of the predictive and actual values of the RBFNN (part of the sample)

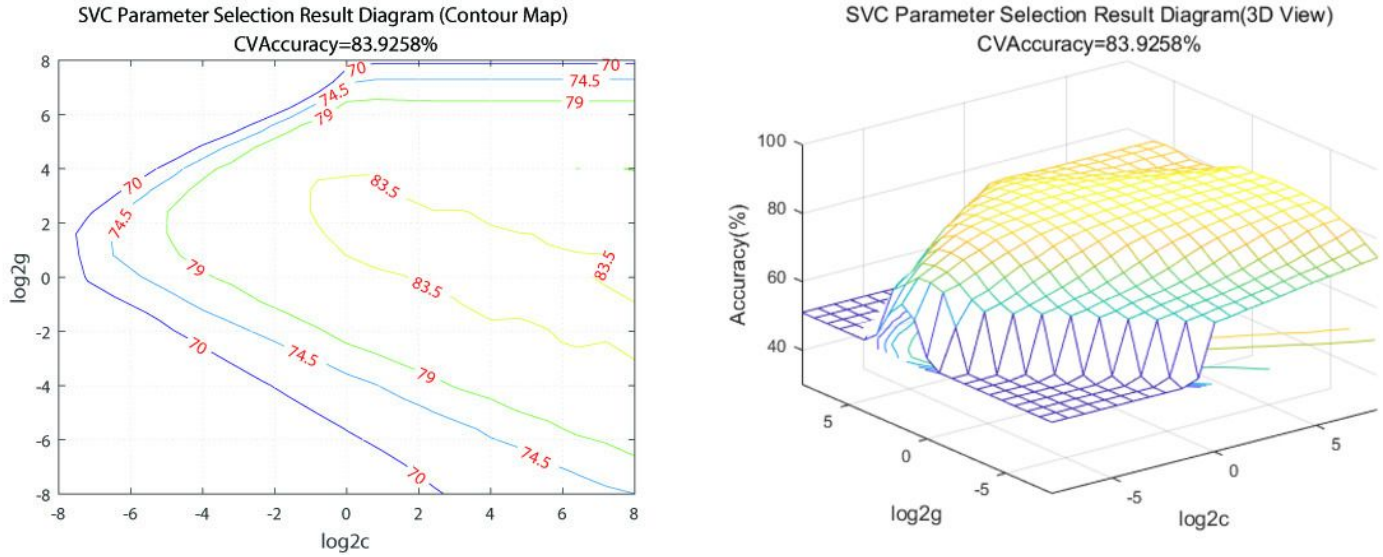


Figure 8

SVC parameter selection result: (a) contour map (b) 3D view

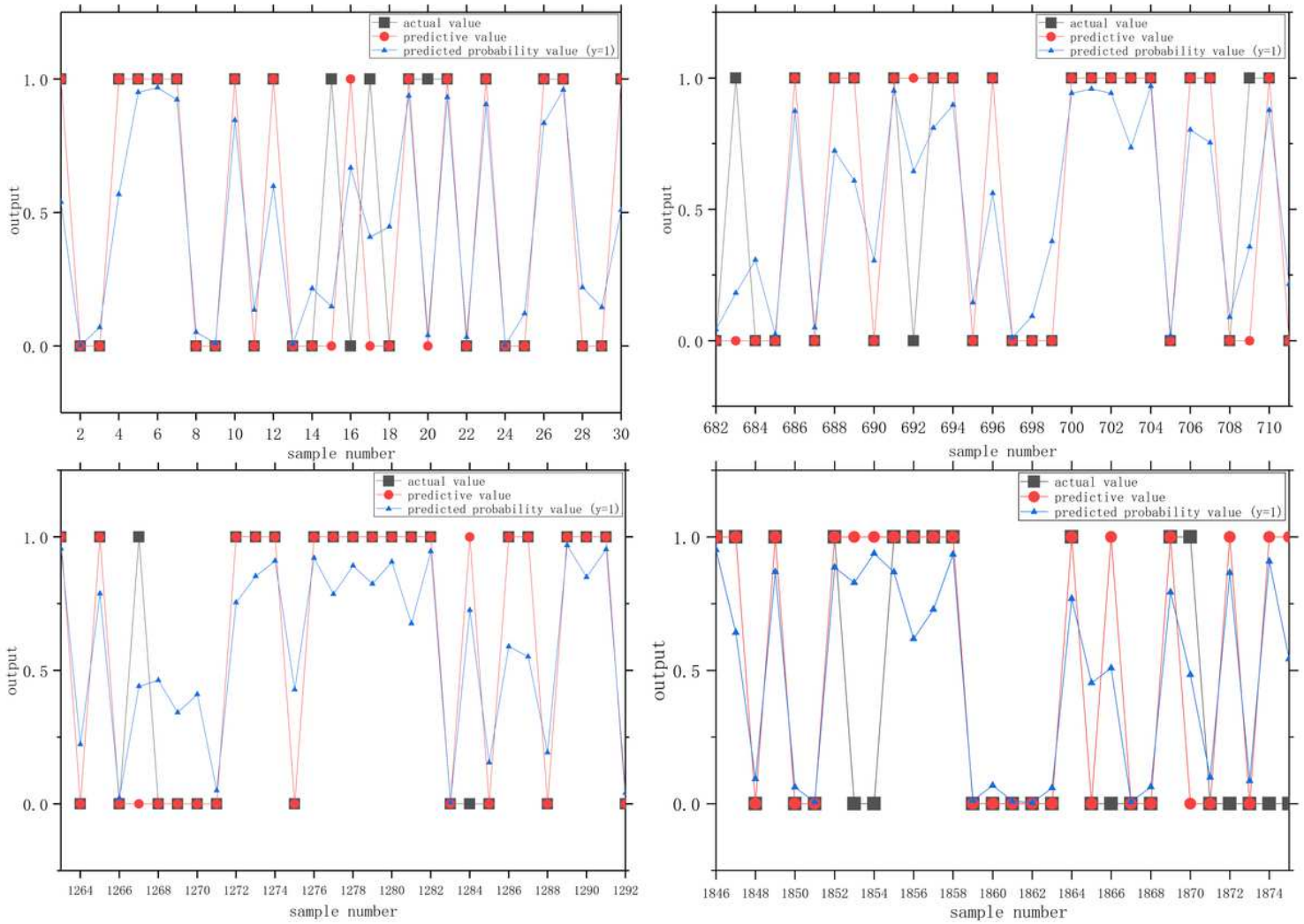


Figure 9

Comparison charts of the predictive and actual values of the SVM (part of the sample)

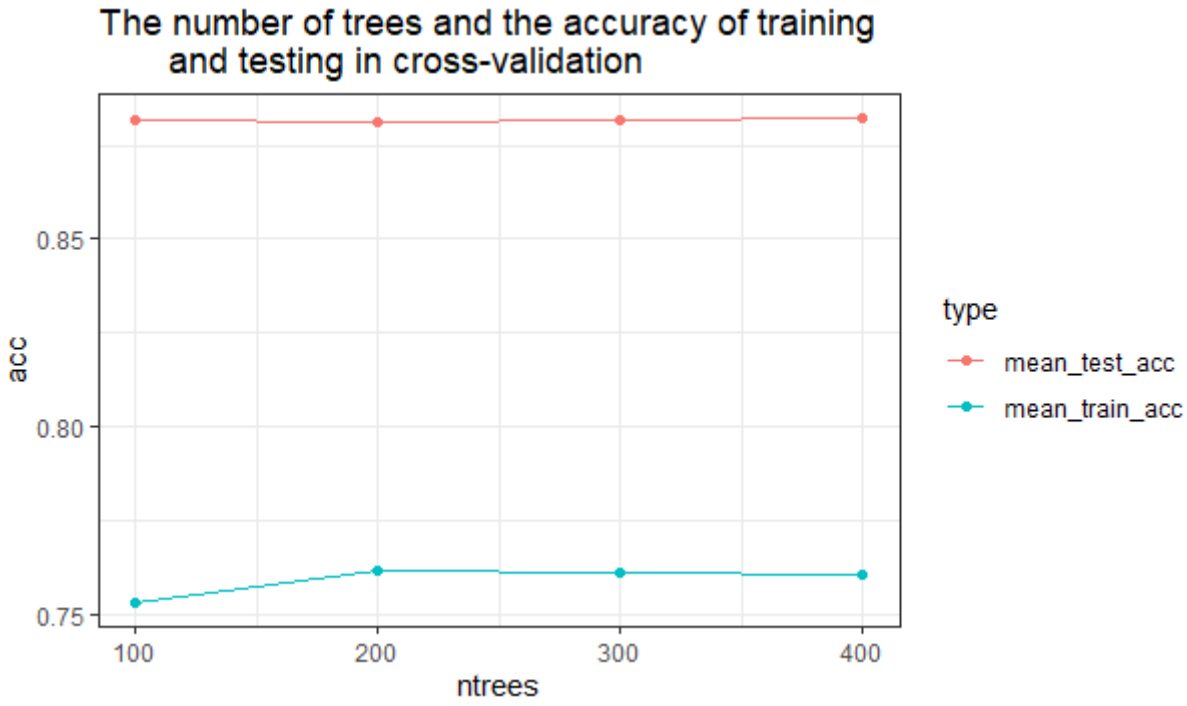


Figure 10

The number of trees and the accuracy of training and testing in cross-validation

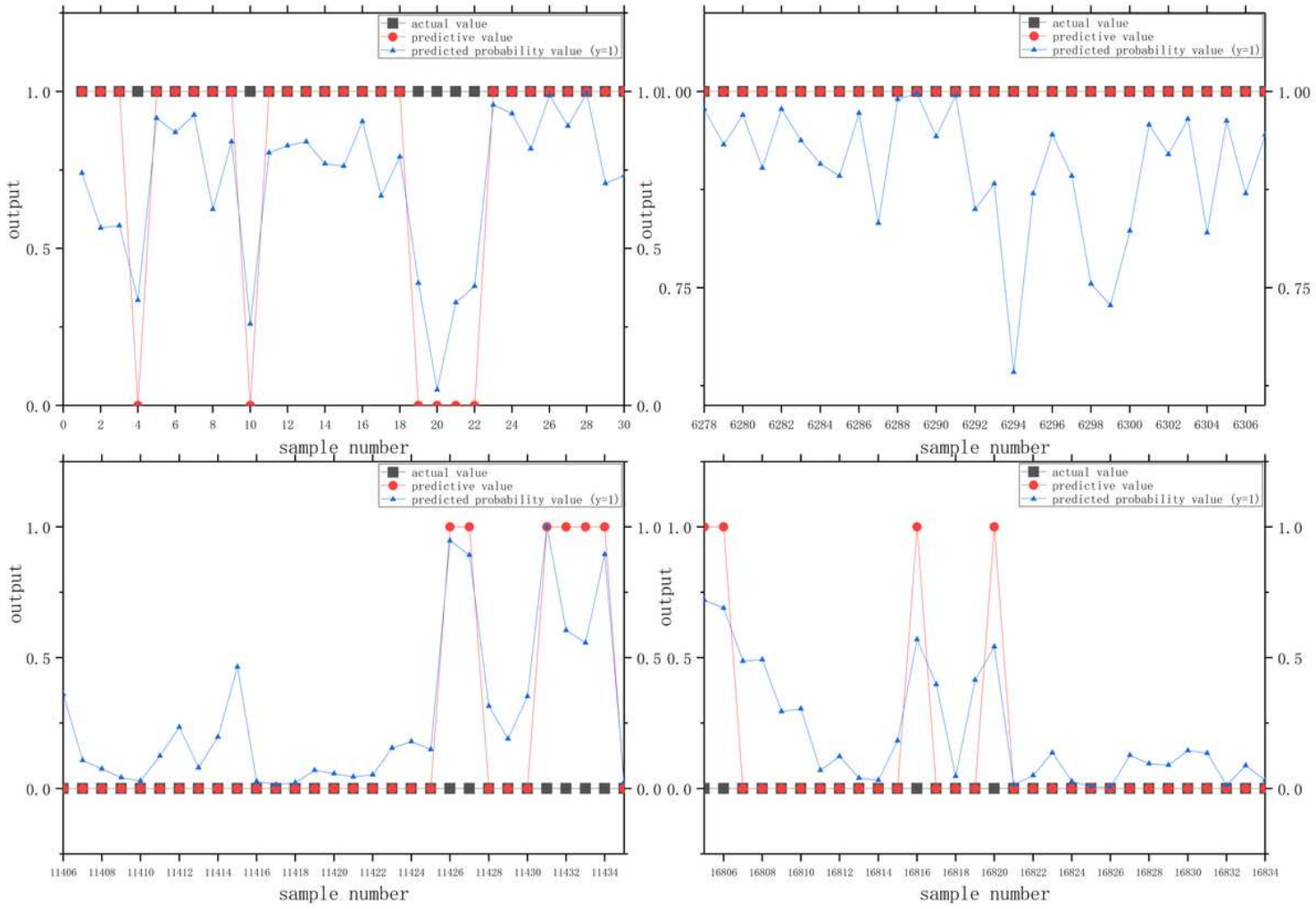


Figure 11

Comparison charts of the predictive and actual values of the RF (part of the sample)

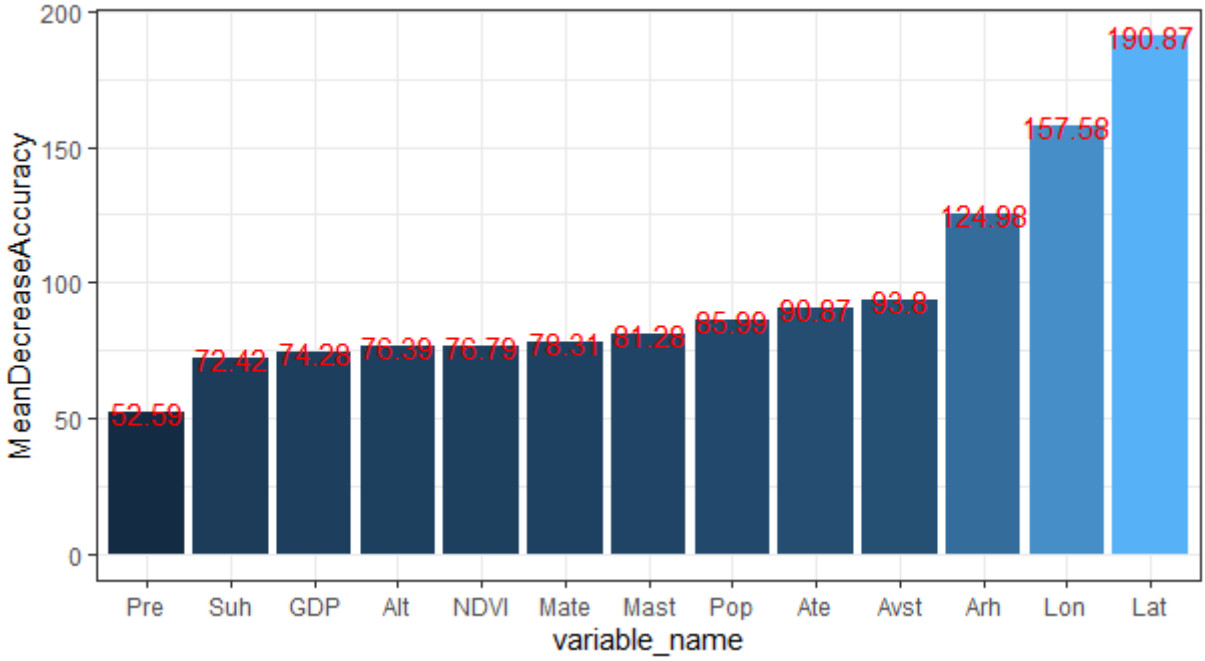


Figure 12

Mean decrease accuracy of 13 variables.

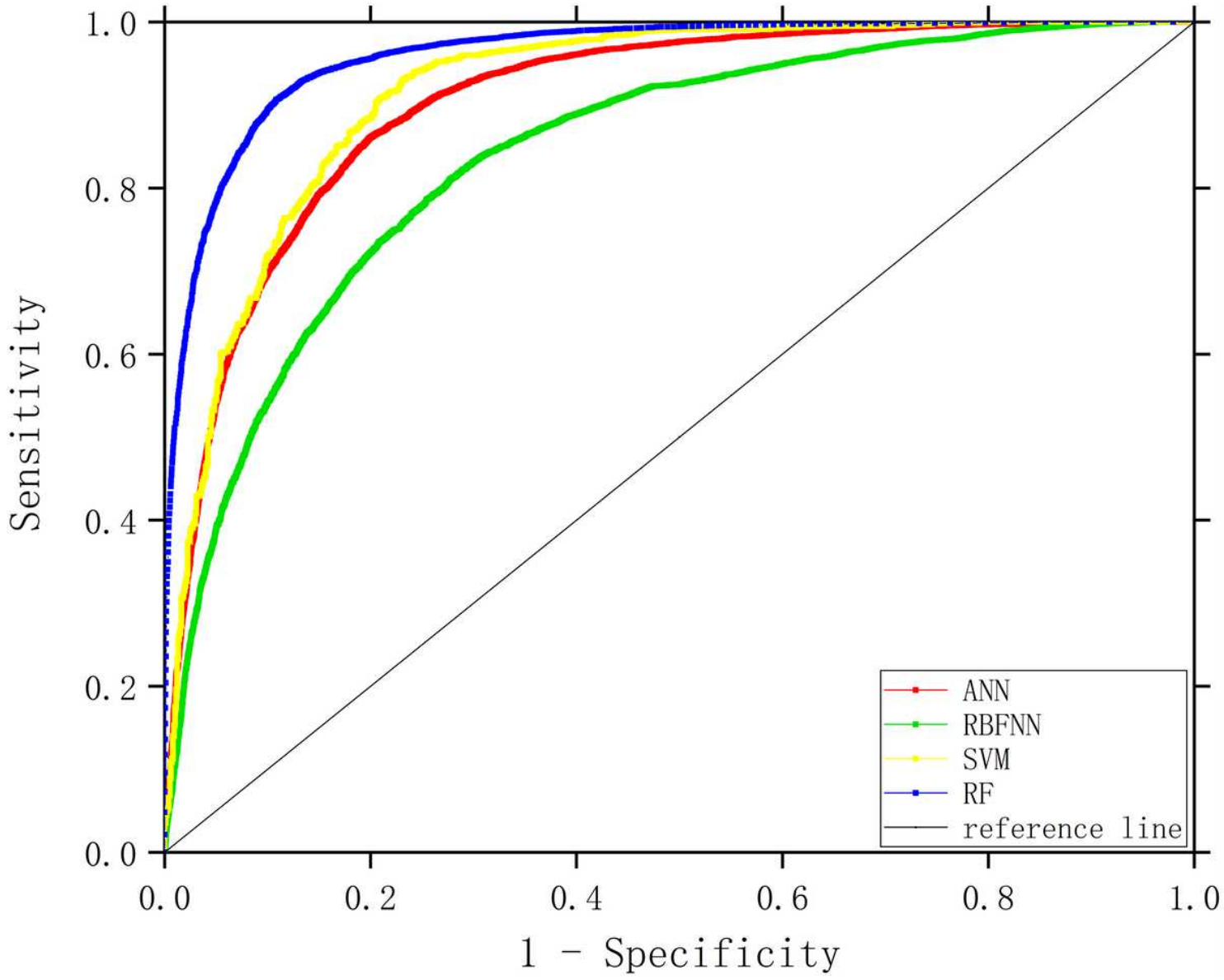


Figure 13

ROC curves of the four models.

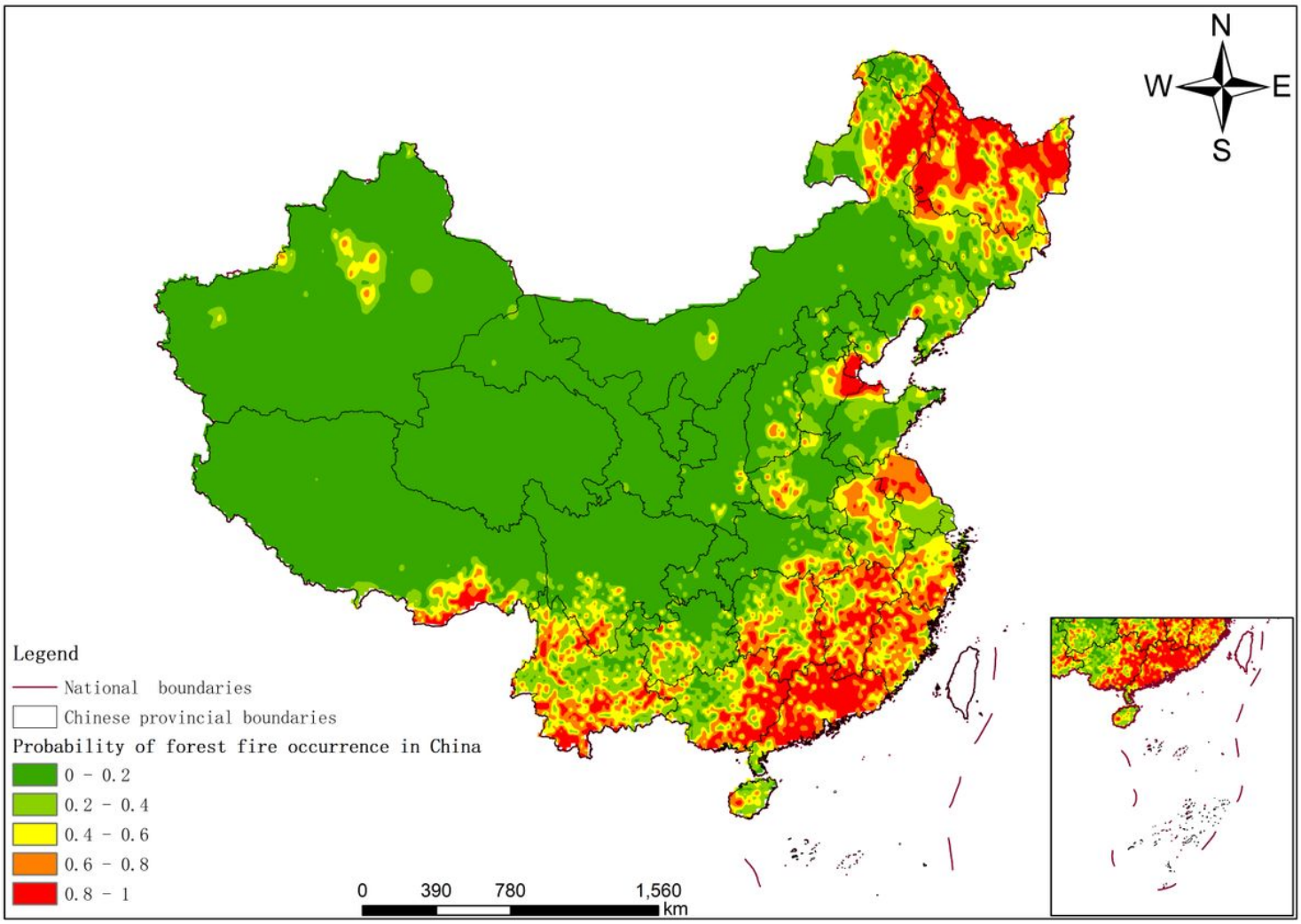


Figure 14

Forest fire probability map for China based on the RF model. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

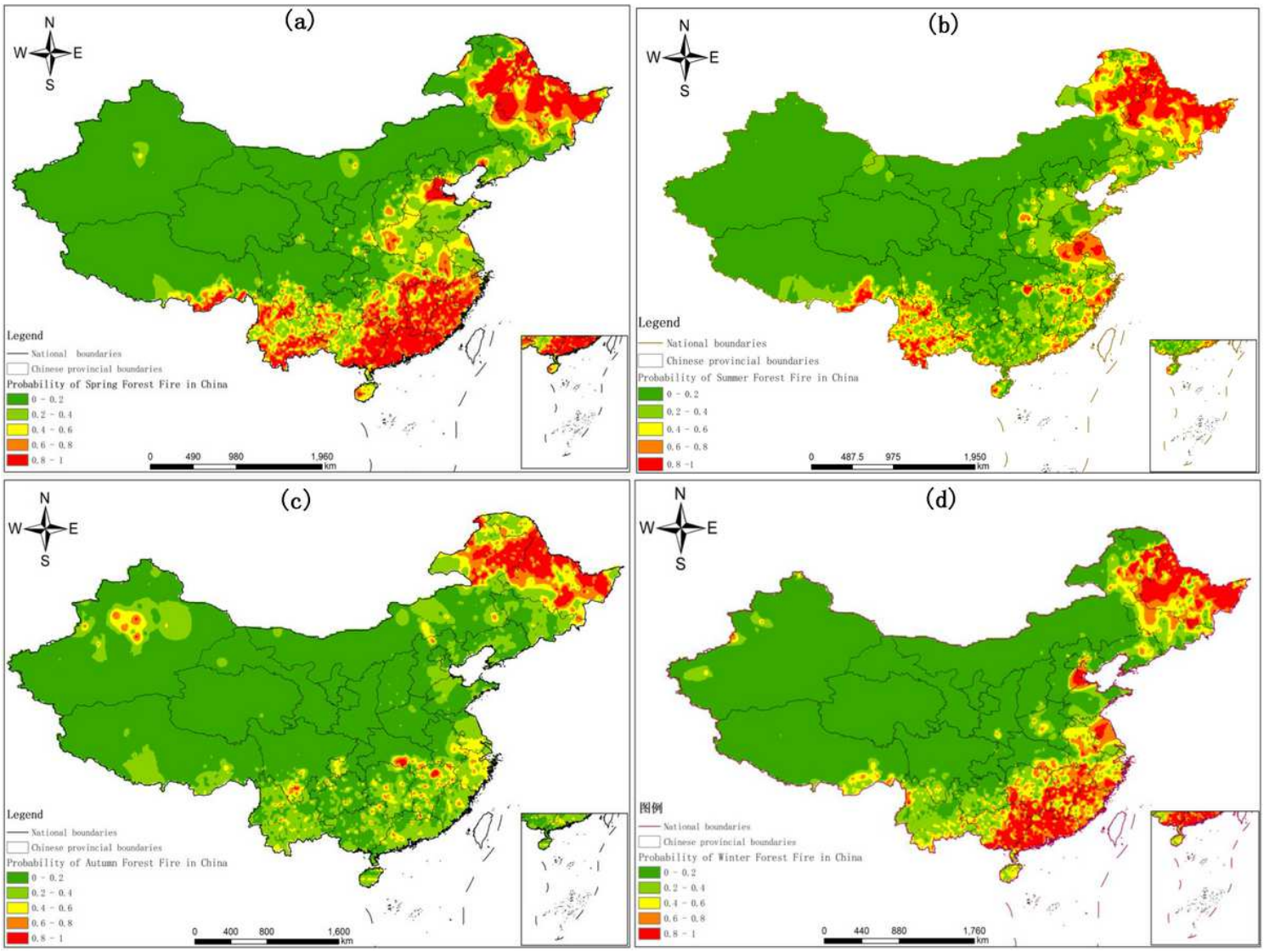


Figure 15

Seasonal forest fire probability map for China based on the RF model: (a) spring (January, February, and March); (b) summer (April, May, and June); (c) autumn (July, August, and September); and (d) winter (October, November, and December). Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.