An interpretable machine learning model for predicting forest fire danger based on Bayesian optimization

Zhiyang Liu[1], Kuibin Zhou[1*], Qichao Yao[2] and Pedro Reszka[3]

[1] *College of Safety Science and Engineering, Nanjing Tech University, Nanjing 211816, Jiangsu, China*
[2] *National Institute of Natural Hazards, Haidian, Beijing 100085, China*
[3] *Faculty of Engineering and Sciences, Universidad Adolfo Ibáñez, Santiago 7941169, Chile*
* Corresponding author, E-mail: kbzhou@njtech.edu.cn

Abstract

As global warming increases forest fire frequency, early prevention and effective management become crucial. This requires models that are both accurate and easily understood. However, traditional machine learning models, which typically use preset parameters, are often inaccurate and hard to interpret. Therefore, this study introduces an enhanced approach using data from 2000 to 2019 in the Sichuan and Yunnan provinces of China, incorporating 18 driving factors. Bayesian optimization algorithms, i.e., the Gaussian Process (GP) and Tree-structured Parzen Estimator (TPE) probabilistic proxy models, were used to optimize the hyperparameters for LightGBM, Random Forest (RF), and Support Vector Machine (SVM), respectively. Finally, forest fire danger prediction models were constructed to draw forest fire danger maps, and the performance was compared between different models. In detail, the model's predictive performance was evaluated using metrics like accuracy, recall, precision, Balanced F Score (F1), and area under curve (AUC). The evaluation demonstrated that the TPE-LightGBM exhibited remarkable accuracy (AUC = 0.962). The forest fire danger map categorizes the study area into five danger levels. The TPE-LightGBM effectively classifies 62.58% of the study area as low-danger level and 5.33% as high-danger Level V. The Shapley additive explanation (SHAP) model interpretation of TPE-LightGBM highlights daily the average relative humidity, sunshine hours, elevation, daily average air pressure, and daily maximum ground surface temperature as the primary influential factors, followed by the human activity indexed by the gross domestic product (GDP) and the distance to the nearest railway.

Introduction

In the context of climate change and global warming, forest fire occurrence increases threat to life, property, forest resources, and the environment[1]. As given by the National Bureau of Statistics of China[2], a total of 7,301 forest fires occurred and burned an area of 48,000 hectares from 2018 to 2022. Therefore, the development of accurate and interpretable forest fire danger models is crucial for early warning and emergency response.

Forest fires involve the interaction of multiple factors at different spatial and temporal scales, including vegetation, topography, meteorology, and human activities[3−5]. Early studies of forest fires mainly explored the temporal and spatial distribution. They estimated the spatial clustering characteristics of fire occurrence[6], but they were limited to judging the macroscopic distribution of forest fires. The remote sensing technology coupled with Geographic Information Systems (GIS) facilitates extensive data acquisition, which in turn supports the application of logistic regression models, Geographically Weighted Logistic Regression[7], Poisson models[8], and various other statistical methods for the analysis of factor interrelationships. However, statistical methods assume that the interactions between factors are linear, leading to poor prediction accuracy of the developed models[9].

Many studies recently utilized the 'black box' approach of machine learning to address the complex relationships among factors. It has been demonstrated that machine learning models are adept at handling the complex nonlinear relationships inherent among meteorological, topographical, anthropogenic, and vegetative factors, thereby enabling the precise mapping of forest fire danger. Van Beusekom et al.[10] conducted a study in Puerto Rico, utilizing meteorological data and human activities as predictors. They applied RF to analyze the correlation between fire occurrences. In another study, Yue et al.[11] focused on Nanning City, incorporating meteorology, topography, human activities, and vegetation as predictors. They employed LightGBM, Classification and Regression Tree (CART), RF, and XGBoost to develop a susceptibility prediction model. Their findings indicated that the XGBoost model outperformed others, particularly in identifying high-danger areas within a specific region of Nanning. Wang et al.[12], in their research on Yunnan Province, selected 16 predictors encompassing meteorological, topographical, vegetative data, and measures such as the distance between vegetation and rivers or roads, as well as population density. They employed Logistic Regression (LR), SVM, Artificial Neural Network (ANN), RF, Gradient Boosting Decision Tree (GBDT), and LightGBM models for analysis. Their analysis revealed that LightGBM was the most accurate model, which was subsequently utilized to construct susceptibility models for forest fire and to map associated danger areas.

Although machine learning models have achieved good performance in forest fire danger assessment, choosing model

parameters is crucial for achieving high classification accuracy and effective danger mapping. The 'black box' nature poses an additional challenge, making the interpretation of machine learning model results less transparent. To address this issue, there is a need for models that are not only accurate but also understandable, which helps to interpret what causes forest fires and why the model predicts what it does. Optimization algorithms can be instrumental in fine-tuning the hyperparameters of machine learning models, thereby enhancing their predictive performance[13,14]. Furthermore, interpretable artificial intelligence (AI) offers solutions to the 'black box' dilemma, with the SHAP model being a notable example. It provides insights into the output results, objectively quantifying the impact and contribution of each factor[15–17]. It is noteworthy that previous studies have often relied on Gaussian Process (GP) models as probabilistic proxies for hyperparameter optimization[12,18]. However, the potential of tree-structured Parzen estimator (TPE) models as probabilistic proxies has been somewhat overlooked. Further research is needed to compare the advantages and disadvantages of TPE for predicting forest fires.

In this study, an interpretable machine learning model is developed to predict forest fire danger based on GP and TPE optimization. The fire occurrence data from 2000−2019 in Sichuan and Yunnan provinces, China were utilized for analysis. Eighteen factors, encompassing vegetation, topography, meteorology, and human activities, were selected to interpret the temporal and spatial distribution of forest fires. Six optimal machine learning models were developed, after using GP and TPE probabilistic proxy models within a Bayesian optimization framework to fine-tune the hyperparameters of LightGBM, RF, and SVM, respectively. Comparative analyses were conducted

for the six models, using Accuracy, Precision, Recall, Balanced F Score (F1), and area under curve (AUC) indexes. The SHAP model was used to interpret the optimal machine learning models, providing insights into the contribution and influence of each factor. Finally, a forest fire danger map was produced to serves as a scientific foundation for forest fire likelihood prediction and early warning systems in Sichuan and Yunnan.

Data and methods

Study area

Sichuan and Yunnan Provinces in China, covering 880,100 km^2, were chosen for this study (Fig. 1). The two provinces have complex topography and landscape dominated by mountains and plateaus. Sichuan Province has three main climate zones: the Central Subtropical Humid zone, the Southwest Mountain Semi-Humid zone, and the Northwest Alpine Plateau zone. Yunnan province belongs to the Subtropical Plateau Monsoon type. The overall climate features include a slight annual temperature difference and an extensive daily temperature difference. Precipitation distribution across seasons and regions is uneven, showing characteristics of 'east wet and west dry'. Additionally, the study area has diverse vegetation, including approximately 73.87 million hectares of forest and about 2.25 billion cubic meters of forest reserves. The area often has many forest fires in China. Therefore, mapping the fire susceptibility in this region can help effectively predict the likelihood of such occurrences[19].

Figure 2 displays the uneven distribution of forest fires in Sichuan and Yunnan provinces from 2000 to 2019. In 2010, the number of forest fires reached a maximum value of 3606. This
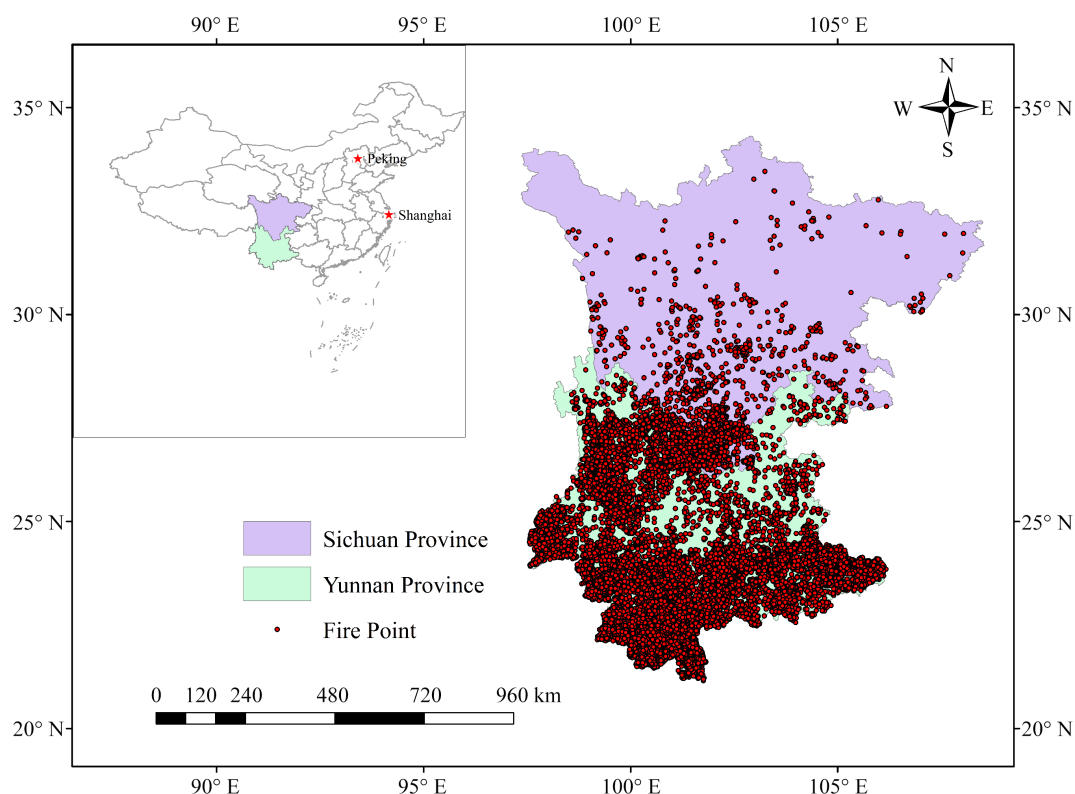

Fig. 1 Overview of the study region.

Count of TIME by year

Fig. 2 Inter-annual variability of forest fires in Sichuan and Yunnan provinces.
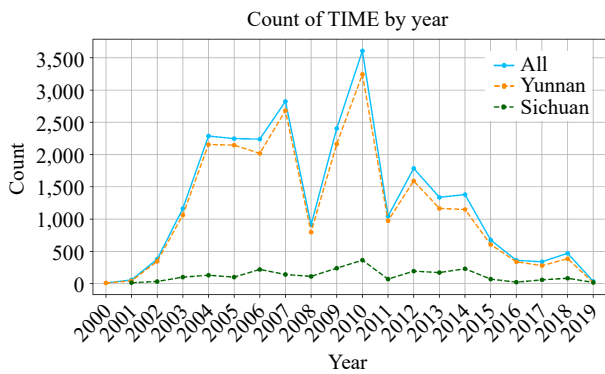
Count of TIME by month

Fig. 3 Inter-monthly variation of forest fires in Sichuan and Yunnan provinces.

was followed by 2,287 and 2,823 fires in 2004 and 2007, respectively. The number of forest fires declined sharply from 2010 to 2011, dropping from 3,606 to 1,045 fires. Figure 3 indicates that most forest fires happen from January to May, peaking in May with 7,891 fires. Fires are much fewer from June to December, making up less than 10% of the yearly total.

Data sources

Fire database

The dependent variable in this study was whether forest fires occurred or not. The National Institute of Natural Hazards of the Ministry of Emergency Management provided fire point data for Sichuan and Yunnan from 2000 to 2019, including information on the longitude, latitude, and date of occurrence of fire points.

The data were corrected to avoid duplication, records with inconsistent data were deleted, and only those with the location type of forest land were retained. A total of 25,591 fire point records are used in this study. As shown in Fig. 1, each fire pixel represents a fire point.

Non-fire points were also considered to construct a dichotomous forest fire model, which was randomly generated by ArcGIS 10.8 software at a scale of 1:1.5 in the study area. For analysis, fire points were assigned a value of 1, and non-fire points were assigned a value of 0.

Based on GlobeLand30, i.e. a 30-meter global surface coverage dataset from the National Catalogue Service for Geographic Information of China (www.webmap.cn), the extent of the forested areas in Sichuan and Yunnan Provinces were extracted. To differentiate between non-fire and fire points in time and space, a circular buffer with 1,000-m diameter was established around each fire point[20]. Then the buffer zone was subtracted from the extent of the forested areas in Sichuan and Yunnan Provinces to define the range for non-fire points. Non-fire points were assigned random dates using Python to ensure temporal randomness[21,22].

Fire triggering factors

Many factors contribute to forest fires, which can be categorized into meteorological, topographical, vegetation, and human activity factors[23,24]. Especially, 21 factors affecting the forest fire occurrences in Sichuan and Yunnan were identified and detailed in Supplementary Table S1.

Meteorological factors influence the likelihood of fires and impact the combustion characteristics of fuels[25]. Meteorological data are derived from the 'China Surface Climatic Data Daily
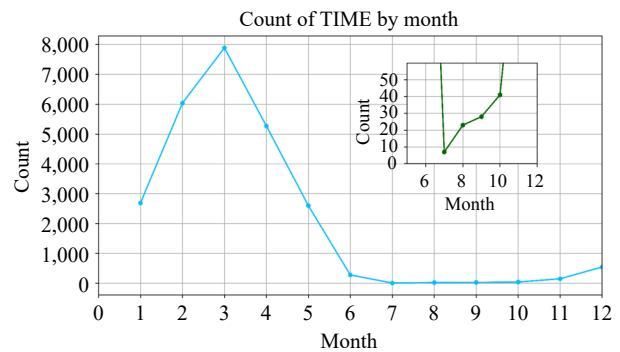
Value Dataset (V3.0)' in the China Meteorological Data Network (https://data.cma.cn). These data include the daily average temperature, daily maximum temperature, daily minimum temperature, cumulative precipitation from 20:00 to 20:00, daily average relative humidity, daily average wind speed, daily maximum wind speed, daily average air pressure, sunshine hours, daily average ground surface temperature, and daily maximum ground surface temperature. Daily meteorological data for both fire points and non-fire points were sourced from the nearest weather station. The Thiessen polygon method in ArcGIS 10.8 was used to associate each sample point with its nearest meteorological station. Python was then used to correlate the daily meteorological records for these sample points over the period from 2000 to 2019[4].

Topographic factors indirectly influence the occurrence of forest fires by affecting climate, vegetation, and other factors[26,27]. The topographic data were obtained from the Geospatial Data Cloud of the Computer Network Information Center of the Chinese Academy of Sciences (www.gscloud.cn) using the ASTER GDEM V3.0 elevation model. The elevation, slope, aspect, and topographic wetness index (TWI) of the study area were extracted, and TWI is expressed by[28,29].

$$TWI = \ln\left(\frac{SCA}{\tan\alpha}\right) \tag{1}$$

where, the SCA represents the contributing area per unit contour length at any point along the slope gradient, and α is the slope.

During the modeling process, topographic factor values for each sample point were extracted to categorize the slope direction into eight cardinal and intercardinal directions: North, Northeast, East, Southeast, South, Southwest, West, and Northwest. These directional categories were then assigned codes for the purpose of classification.

Only areas of land covered by forest were considered. Forest vegetation data were from the 1:1,000,000 Vegetation Atlas of China that can be downloaded from the Resource and Environmental Science Data Platform of the Chinese Academy of Sciences (www.resdc.cn). The vegetation was categorized into eight distinct types: coniferous forests, mixed coniferous and broad-leaved forests, Broad-leaved forests, Shrublands, Grasslands, Meadows, Alpine vegetation, and Cultivated vegetation. The vegetation types at the locations of the sample points were identified using ArcGIS 10.8 software.

Human activities, especially construction, road building, and outdoor activities, greatly affect where and how often forest fires happen[30–32]. Data on human activities are sourced from the 1:250,000 National Basic Geographic Database available on

the National Catalogue Service for Geographic Information (www.webmap.cn), encompassing roads, railways, and settlements. The Resource and Environment Science Data Platform of the Chinese Academy of Sciences (www.resdc.cn) provides the gross domestic product (GDP) and population density data in 2000, 2005, 2010, 2015, and 2019. Utilizing ArcGIS 10.8 software, Euclidean distances from roads, railways, and settlements, along with average population density and GDP for 2000-2019, were calculated for the sample points[33].

To standardize the satellite imagery for modeling purposes, given the variability in resolution and dimensions, the data were uniformly transformed into a consistent projection coordinate system. Furthermore, each factor was uniformly resampled to achieve a uniform resolution of 30 m × 30 m, as illustrated in Supplementary Fig. S1.

Technology route

In this study, historical fire data was used to analyze the temporal and spatial distribution of fire points in Sichuan and Yunnan provinces. Then, the correlation between each factor was assessed through the multicollinearity analysis, and the data scale was standardized *via* normalization. Subsequently, the data was randomly split into a training set and a test set in a 7:3 ratio[34,35]. Three machine-learning models were trained using the dataset. Two probabilistic proxy models with Bayesian optimization were employed to fine-tune the hyperparameters of the three models. The models' performance was evaluated using the test set with metrics such as Accuracy, Precision, Recall, AUC, and F1 scores. The trained models were used to predict the fire danger across the study area. Finally, model interpretation was conducted using SHAP. The experiments were carried out in a Jupyter Notebook environment

using Python 3.11.5 and ArcGIS 10.8 software, on a system equipped with a COREi5 processor and a 16GB NVIDIA GeForce RTX 3060 graphics card. The detailed workflow is depicted in Fig. 4 and Supplementary Fig. S2.

Multicollinearity analysis

To prevent high covariance between factors that could bias the results and reduce model accuracy, the Variance Inflation Factor (*VIF*) was used to check for multicollinearity[36]. The *VIF* was calculated by:

$$VIF = \frac{1}{1 - R^2} \quad (2)$$

where, R^2 is the coefficient of complex determination.

Data normalization

To standardize the data and mitigate discrepancies in their impact on the model due to varying scales, the data were normalized to the interval of [0,1][37]. This process is illustrated by:

$$X_i^* = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \quad (3)$$

where, X_i^* is the normalized data, x_i is the original data, $x_{\max}$ and $x_{\min}$ are the highest and lowest values of the full original data, respectively.

Equation (3) cannot be used for the normalization of slope and daily average relative humidity. The slope is normalized by:

$$x_\alpha = \sin \alpha \quad (4)$$

where, α is the slope angle.

The daily average relative humidity is normalized by:

$$x_\beta = \frac{\beta}{100} \quad (5)$$

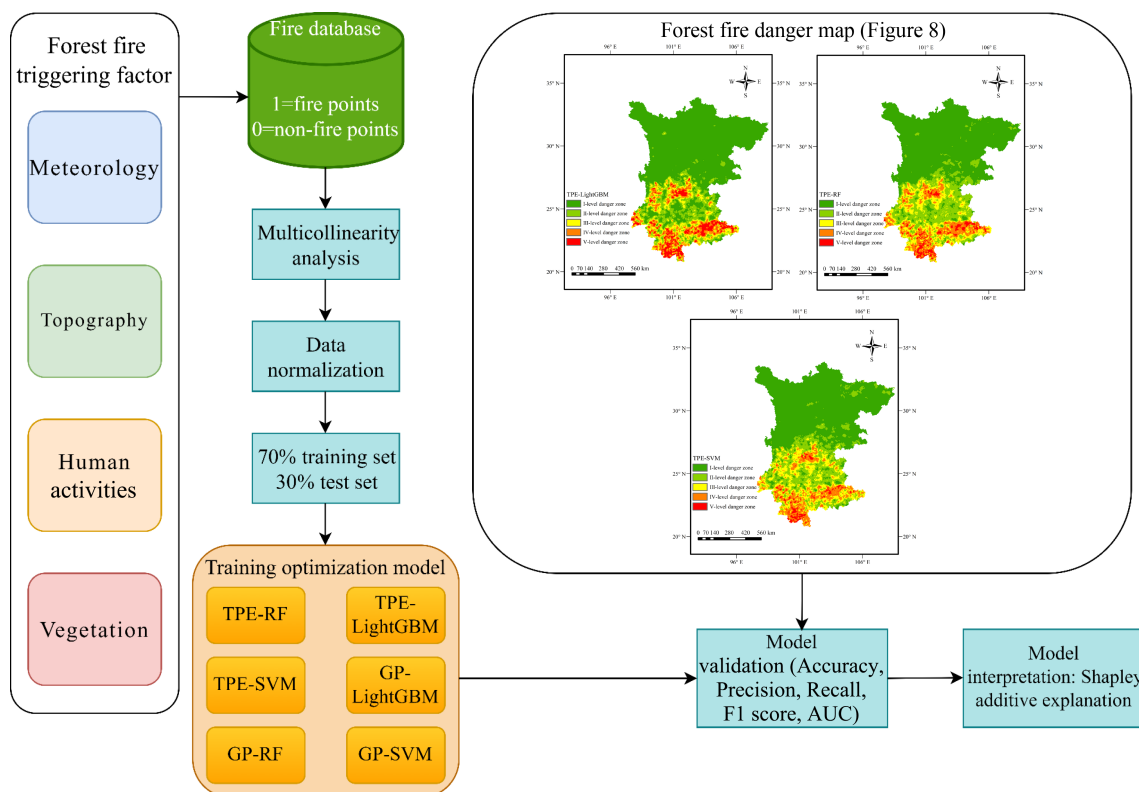where, β is the humidity value.

Fig. 4 Technology route.

Fire danger assessment model construction

Bayesian optimization algorithm

In machine learning models, the choice of hyperparameter values substantially affect performance and predictive accuracy[13,14]. This study employs two categories of Bayesian optimization algorithm to fine-tune the hyperparameters of the model, as depicted by:

$$x^* = \arg\max_{x \in X} f(x, T) \qquad (6)$$

where, x^* is the set of hyperparameters that can yield the highest score; x is the hyperparameter combination of the machine learning model; X is the hyperparameter search range; f is the acquisition function; T is the proxy model. In Eqn (6), argmax is an operator used to find the point at which a given function attains its maximum value.

The Bayesian optimization algorithm contains two key components, i.e. a probabilistic proxy model and an acquisition function. The former is used to fit the probability distribution of the sampled points, and the latter evaluates the potential of each distribution point. Adaptively scaling the parameter search space enables handling high-dimensional hyperparameter optimization tasks and facilitates finding the globally optimal solution in as few iterations as possible. The computational formulas refer to the study by Bergstra et al.[38]. GP and TPE are two distinct methodologies for modeling and optimizing hyperparameters within the realm of Bayesian optimization. The GP approach is centered on employing probabilistic models to seize the smoothness of the objective function. It posits that the variations of the objective function across the hyperparameter space is smooth, thereby constructing a probabilistic distribution that describes the behavior of the objective function. This method is particularly adept at scenarios where the objective function exhibits gradual changes, providing uncertainty estimates about the objective function that are instrumental in guiding the selection of subsequent hyperparameters. Conversely, TPE adopts a tree-based structure to more nimbly manage intricate high-dimensional hyperparameter spaces. TPE simulates the hyperparameter selection process by constructing a decision tree, leveraging historical data to assess the performance of various hyperparameter combinations, and endeavoring to identify the configuration that maximizes the objective function. The strength of TPE lies in its capacity to address hyperparameter spaces rife with uncertainty and complexity, especially when interactions between hyperparameters are present.

This study uses both GP and TPE, which are probabilistic proxy-based models, for modeling purposes. The GP model was selected for its capability to capture the smoothness of the objective function, while the TPE model was chosen for its flexibility in dealing with complex hyperparameter spaces. By integrating the two methods, we aim to inspect the hyperparameter space more comprehensively with the expectation of identifying the optimal hyperparameter configuration, thereby enhancing model performance. The study will assess how well these methods work in different scenarios and discuss their complementarity and applicability in Bayesian optimization. The framework for model hyperparameter optimization is shown in Supplementary Fig. S2.

LightGBM

LightGBM is a framework based on the Gradient Boosting Decision Tree (GBDT) algorithm. It was developed by Microsoft[39] in 2017 to improve the efficiency and calculation speed of the GBDT algorithm when dealing with extensive or high-dimensional data. Unlike GBDT that uses the Level-wise algorithm, LightGBM adopts a leaf growth strategy, specifically one that incorporates depth limitation.

$$F_r(x) = \sum_{k-1}^{r} f_k(X) \qquad (7)$$

where, $F_r(x)$ is the model comprising a set of r decision trees, and $f_k(X)$ is the kth decision tree.

The objective function consists of the loss function and the regularization term. The loss function formula is:

$$L(y_q, y_q') = \frac{1}{A} \sum_{q=1}^{A} (y_q \log P_q + (1 - y_q) \log(1 - P_q)) \qquad (8)$$

where, y_q is the type of recognition after X_q; A is the sample size; P_q is the probability of recognizing X_q as a one after it is entered into the model.

The regularization controls the splitting of leaf nodes to reduce overfitting in the model. The objective function formula is:

$$O = L(y_q, y_q') + \gamma Z + \frac{1}{2}\lambda \sum_{v=1}^{z} (W_v)^2 \qquad (9)$$

where, O is the objective function; Z is the number of leaf nodes; W_v is the output value of the v-th leaf node; γ and λ are set parameters.

LightGBM enhances performance by refining several key algorithms[39]. It utilizes an improved Histogram decision tree algorithm that discretizes data eigenvalues into a total of k bins to identify optimal split points, thereby maximizing gain and boosting computational efficiency. The one-sided gradient sampling (GOSS) algorithm prioritizes samples with higher gradients and randomly samples those with lower gradients, ensuring consistency with the original data distribution and maintaining model accuracy. The mutually exclusive feature bundling (EFB) algorithm tackles the sparsity common in high-dimensional datasets by bundling mutually exclusive features, reducing dimensionality, and enhancing computational efficiency by creating new composite features. Lastly, the Leaf-wise decision tree growth strategy selects the leaf node with the highest potential for split gain, which helps prevent overfitting and minimizes model loss.

Random forest

Random forest (RF) is an ensemble learning model that constructs multiple decision trees during training[40]. Each tree in the ensemble is learned from a different part of the data, leading to diverse classifications. The final classification is achieved by a majority vote of the individual tree predictions, as illustrated in Fig. 5. To boost model robustness, each tree is trained on a bootstrap sample of the data, with one-third of the data held out as Out-Of-Bag (OOB) samples for internal validation and to prevent overfitting.

RF excels at handling large, multivariate datasets, making it suitable to model the high-dimensional, nonlinear aspects of forest fires[41]. Meteorological, topographic, vegetation, and human activity facts significantly influence the occurrence of forest fires, and RF's ability to handle such complexities contributes to its effectiveness in this domain.

Support vector machines

As a supervised learning algorithm, Support Vector Machine (SVM) can classify data either linearly or non-linearly[42]. As
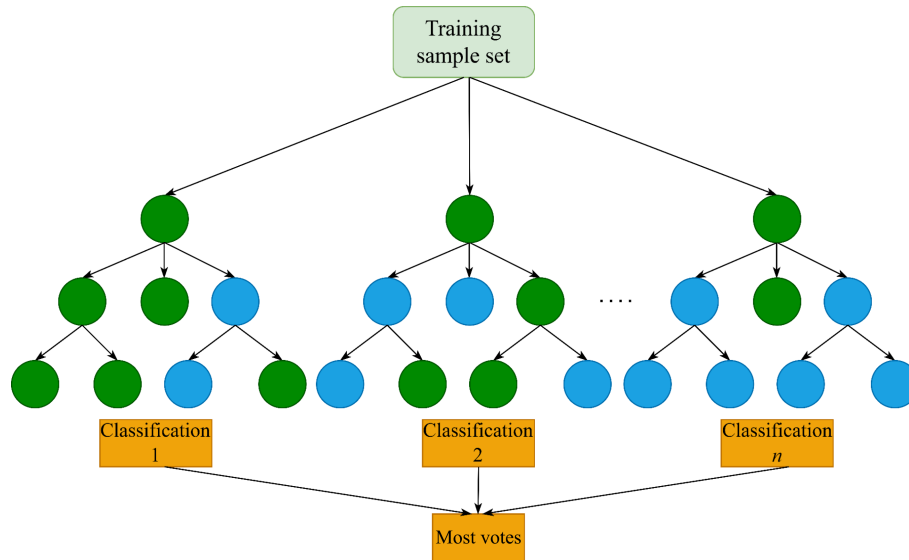
Fig. 5 Schematic diagram of the RF.

depicted in Fig. 6, The main goal of SVM is to find the best hyperplane in n-dimensional space that can separate the data into different categories, like 'fire' and 'no fire'.

$$\omega^T x + b = 0 \qquad (10)$$

where, $\omega = \{\omega_1, \omega_2 ..., \omega_n\}$ is the normal vector to the decision plane and b is the intercept term.

Separating the categories of fire and no fire based on the principle of maximum margin is equivalent to solving a convex optimization problem, as calculated by：

$$\max_{\omega, b} \frac{2}{\|\omega\|}, \quad s.t. y_i(\omega^T x_i + b) \geqslant 1, i = 1, 2, ..., m \qquad (11)$$

where, $\frac{2}{\|\omega\|}$ is the classification interval.

To handle nonlinear classification problems, Vapnik[42] introduced a nonlinear kernel function that maps the data into a higher dimensional space, facilitating the discovery of hyperplanes.

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j) \qquad (12)$$

The Radial Basis Function (RBF) is a widely-used nonlinear kernel function, and performs better in danger assessment[43]. In this study, RBF is used to develop the SVM model, as illustrated by:

$$K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2} \qquad (13)$$

Model performance evaluation

Accuracy, precision, recall, F1 score, and AUC are key performance metrics commonly used in machine learning to assess the effectiveness of a model. Generally, higher values of these five indicators suggest superior model performance. The formulas for these metrics are as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \qquad (14)$$

$$Recall = \frac{TP}{TP + FN} \qquad (15)$$

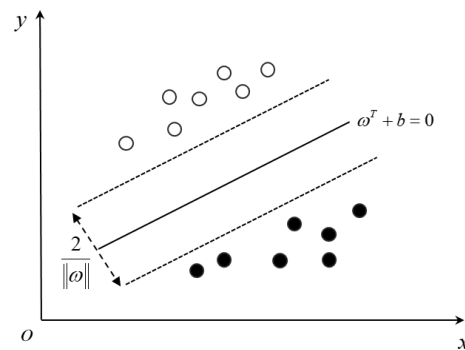$$Precision = \frac{TP}{TP + FP} \qquad (16)$$

Fig. 6 Schematic diagram of the SVM.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (17)$$

where, true positives (TP) mean the model correctly finds positive cases; false negatives (FN) means it misses positive cases; conversely, false positives (FP) means it wrongly says negatives are positives, while true negatives (TN) means it correctly identifies negatives.

The AUC of the receiver operating characteristic (ROC) curve is a definitive metric for model evaluation. The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1−specificity), across various threshold settings. In the term '1−specificity', specificity is the rate at which the model correctly identifies true negatives.

SHAP model

Machine learning models often achieve high prediction accuracy, yet they can lack interpretability regarding how input features contribute to the calculation outcomes. To address this, the SHAP (SHapley Additive exPlanation) framework was introduced to provide insights into the workings of machine learning models concerning their output results.

SHAP is grounded in cooperative game theory and measures each feature's contribution to the prediction by calculating the Shapley value[44,45]. The Shapley value for a feature, in the context of a given model and input sample is defined as the average of that feature's marginal contributions across all

possible combinations in the dataset. For a given model (f) and input sample (x), the Shapley value of feature i is defined as:

$$\varphi_i(f,x) = \sum_{S \subseteq N\{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!}(f_x(S \cup \{i\}) - f_x(S)) \quad (18)$$

where, N is the set of all features; S denotes any subset that does not contain features i; $|S|$ is the size of the set S; $|N|$ is the total number of all features; $f_x(S \cup \{i\}) - f_x(S)$ is the cumulative contribution of the features, and i denotes the cumulative contribution value of the features.

The SHAP model builds an explanatory model $g(x)$ instead of the machine learning model $f(x)$, as expressed by:

$$g(x) = \varphi_0 + \sum_{j=1}^{p} \varphi_j \quad (19)$$

where, p is the number of features; φ_0 is the predicted mean value of all training samples; φ_j is the contribution of the input feature j to the predicted value, i.e., the SHAP value of the feature. Bigger SHAP values mean the feature has a bigger impact on the model's prediction.

Results and discussion

Multicollinearity analysis: selection of forest fire drivers

Before modeling, it is essential to check for multicollinearity among all factors using SPSS 17 software to ensure the results are accurate and reliable. A VIF greater than 10 indicates strong covariance among factors, whereas a VIF less than 10 suggests no significant covariance[46]. The analysis clarified high VIF values for daily average temperature, daily maximum temperature, daily minimum temperature, daily average ground surface temperature, and daily maximum ground surface temperature. By removing the daily maximum temperature, daily minimum temperature, and daily average ground surface temperature, the VIF values were reduced to below 10 for the remaining variables. This reduction is due to the elimination of factors that were highly correlated with the daily average temperature and daily maximum ground surface temperature, which in turn decreased the overall covariance in the model. The factors that were ultimately selected are presented in Table 1.

Construction of the fire danger model

Models including TPE-LightGBM, TPE-RF, TPE-SVM, GP-Light-GBM, GP-RF, and GP-SVM were developed and evaluated using metrics such as Accuracy, Precision, Recall, F1 scores, and AUC. The performance results are detailed in Table 2 and visualized in Fig. 7. The optimal hyperparameter combinations for these models are listed in Supplementart Table S2.

In terms of overall performance valuation metrics, the TPE optimization outperforms the GP optimization. Specifically, TPE improves the accuracy, precision, recall, and F1 score of the LightGBM algorithm by 1.6%, 2.5%, 1.2%, and 1.9%, respectively. For the RF algorithm, these metrics are improved by 0.9%, 1.6%, 0.3%, and 1%, respectively. For the SVM algorithm, these metrics are improved by 0.7%, 0.7%, 1%, and 0.8%, respectively. Among the TPE-optimized models, TPE-LightGBM demonstrates the best predictive performance with the highest values in all evaluated metrics, followed closely by TPE-RF. The ROC curve analysis indicates that TPE-optimized LightGBM achieves the highest AUC score of 0.962, with TPE-RF at 0.958, GP-LightGBM at 0.953, GP-RF at 0.951, TPE-SVM at 0.930, and GP-SVM at 0.927.

Table 1. Results of the multicollinearity analysis.

No.	Factor	VIF value before eliminating factor	VIF value after eliminating factor
1	Da_AVGTEM	142.109	3.859
2	Da_MINTEM	51.681	–
3	Da_MAXTEM	30.345	–
4	Da_PRE	1.245	1.242
5	Da_AVGRH	3.440	2.220
6	Da_AVGWIN	2.609	2.420
7	Da_MAXWIN	2.603	2.536
8	Da_AVGPRS	3.999	3.855
9	SSD	3.424	2.577
10	Da_AVGGST	40.164	–
11	Da_MAXGST	11.016	4.639
12	Elevation	3.902	3.876
13	Slope	1.000	1.304
14	Aspect	1.001	1.001
15	TWI	1.040	1.190
16	Dis_to_railway	1.450	1.400
17	Dis_to_road	1.382	1.390
18	Dis_to_sett	1.458	1.459
19	Den_pop	4.882	4.871
20	GDP	3.811	3.806
21	Forest	1.104	1.108

Table 2. Performance metrics for model evaluation.

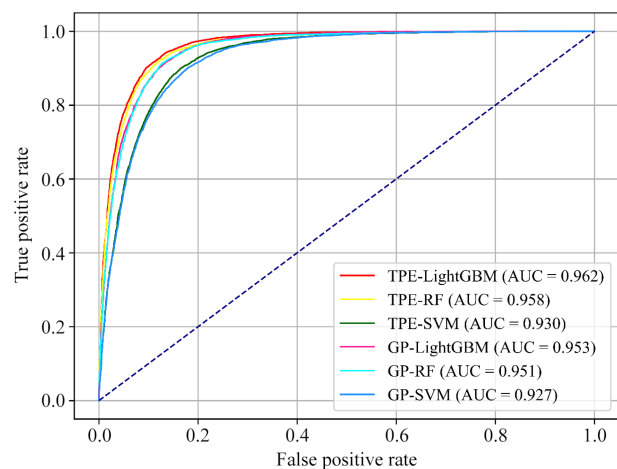| Model parameters	TPE-LightGBM	TPE-RF	TPE-SVM	GP-LightGBM	GP-RF	GP-SVM
TP	5779	5727	5570	5705	5709	5505
TN	8695	8633	8254	8505	8511	8213
FP	917	979	1358	1105	1101	1399
FN	633	685	842	707	703	907
ACC (%)	90.3	89.6	86.3	88.7	88.7	85.6
Precision (%)	86.3	85.4	80.4	83.8	83.8	79.7
Recall (%)	90.1	89.3	86.8	88.9	89.0	85.9
F1 (%)	88.2	87.3	83.5	86.3	86.3	82.7

Fig. 7 ROC curve and AUC of LightGBM, RF, and SVM models, with parameter optimization performed using Bayesian optimization techniques: GP and TPE.

In summary, both TPE-LightGBM and TPE-RF models exhibit strong potential and commendable performance, with TPE-LightGBM providing the optimal fit. TPE surpasses GP in

probabilistic proxy models for several reasons: (1) TPE is adept at managing large-scale datasets, which is characteristic of this study, by efficiently searching through the probability distributions of p(x|y) and p(y)[47]. (2) The optimization strategy of TPE effectively identifies hyperparameter combinations that meet the targeted accuracy levels. It generates diverse density functions based on historical observations and refines these through iterative feedback, offering informed suggestions for subsequent configurations[48]. (3) The inverse factorization of p(x|y) in TPE may offer greater precision than that in GP. TPE introduces some uncertainty during the exploration process, and this uncertainty helps to better search for the globally optimal solution and explore new possibilities[38].

Forest fire danger map

Forest fire danger mapping was conducted, after using the TPE probabilistic proxy model to optimize the hyperparameters and fitting models for each factor. After model validation, probability values were assigned to each pixel within the study area, yielding forest fire danger maps for Sichuan and Yunnan Provinces. The danger of forest fires is categorized into five levels, corresponding to the following probability ranges: 0−0.2, 0.2−0.4, 0.4−0.6, 0.6−0.8, and 0.8−1. These levels are designated as I, II, III, IV, and V-level danger zones[4,20,49], respectively, as detailed in Table 3.

As shown in Fig. 8, the three maps hold a similar distribution of forest fire danger in space, with fires predominantly happening in the south-central and central parts of Sichuan Province and the northwestern and southern parts of Yunnan Province. However, the three maps exhibit variations in the area ratio of each danger level relative to the entire region, as detailed in Fig. 9. the TPE-LightGBM model assigns danger zones as follows: I and II-level danger zones represent 62.58% and 13.76% of the area, respectively. The III, IV, and V-level danger zones account for 10.08%, 8.25%, and 5.33%, respectively. In contrast, the TPE-RF model allocates 54.51% and 18.94% to I and II-level danger zones, with III, IV, and V-level zones at 13.18%, 10.33%, and 3.04%, respectively. The TPE-SVM model shows I and II-level danger zones at 54.52% and 19.14%, with III, IV, and V-level zones comprising 14.68%, 9.70%, and 2.00%, respectively. Notably, the I and V-level danger zones have the highest proportion across all models, while II, III, and IV-level danger zones have the lowest.

The results of TPE-LightGBM model, as seen in Figs 8 and 9, indicate a pronounced spatial distribution. In detail, the area occupancy ratio shows a high-low bipolar distribution, which helps to classify areas as highly likely to occur fire and areas with virtually no fire. In addition, the TPE-LightGBM model exhibits strong predictive capabilities, as clarified by the four Performance metrics in Table 2 and the AUC in Fig. 7. Accordingly, the TPE-LightGBM model is a highly reliable tool for forest fire prediction in Sichuan and Yunnan Provinces.

Table 3. Criteria for the classification of forest fire danger levels.

No.	Forest fire occurrence probability	Fire danger level	Description of fire
1	0−0.2	I	Virtually no fire
2	0.2−0.4	II	Unlikely to occur
3	0.4−0.6	III	Possible to occur
4	0.6−0.8	IV	Prone to occur
5	0.8−1	V	Highly likely to occur

Interpretive analysis based on the SHAP model

The TPE-LightGBM model, as stated in Section 3.3, exhibits the best performance among the developed models. Accordingly, the SHAP interpretation provided in this section focuses exclusively on the TPE-LightGBM model. Figure 10a presents a SHAP scatterplot that illustrates the impact of various factors on the model's output. Each dot on the scatterplot corresponds to a SHAP value for a specific factor and sample. The SHAP values are plotted on the x-axis, where values above and below zero indicate a positive and negative contribution to the model output, respectively. The y-axis represents the different factors, and the color gradient from red to blue signifies the magnitude of the value of each factor, with red and blue indicating high and low values, respectively. Figure 10b features a SHAP bar chart that serves as a summary for ranking the importance of factors. It represents the average absolute value of SHAP for each factor, which helps determine their relative impact on the model's output. The SHAP analysis reveals that the most influential factors affecting the model output are, in descending order of impact: daily average relative humidity, sunshine hours, elevation, daily average air pressure, and daily maximum ground surface temperature.

Importance of meteorology factors

Figure 10 illustrates the correlation among factors such as daily average relative humidity, sunshine hours, daily average air pressure, daily maximum ground surface temperature, daily average temperature, daily precipitation, daily maximum wind speed, and daily average wind speed. These factors can promote the forest fire occurrence by reducing the moisture content in combustibles, effectively drying out fuels and increasing their flammability[50]. Contrary to the common assumption that higher temperatures exacerbate fire danger, the SHAP values unexpectedly indicate that lower daily average temperatures correlate with an increased danger of forest fires. This contradiction could be attributed to the rise in human activities in the study area as the daily temperature drops.

While extreme weather is acknowledged to precipitate forest fires, particularly in Sichuan and Yunnan Provinces, human activities are identified as the predominant cause. Chen & Di[51] reported that about 90% of forest fire events in China are attributable to human activities. Similarly, Ying et al.[52] asserted that in Yunnan Province, human activities are the chief contributors to forest fires. Wang et al.[53], through spatial analysis of fire sources, concluded that in Sichuan Province, human activities cause most forest fires, with natural factors being less frequent culprits.

Importance of topography factors

Elevation significantly influences the output of the model. It determines the temperature, with higher altitudes typically experiencing lower temperatures. Additionally, high-altitude areas are often devoid of human presence, which reduces the likelihood of forest fires. Consequently, as elevation increases, the SHAP values decrease, exerting a negative effect on the model. The contributions of vegetation type and slope aspect to the model are relatively minor. Many high and low value feature points are intermingled because these factors are categorical variables, and their encoded values represent categories rather than magnitudes of influence. The SHAP scatter plot indicates that the TWI does not make a significant
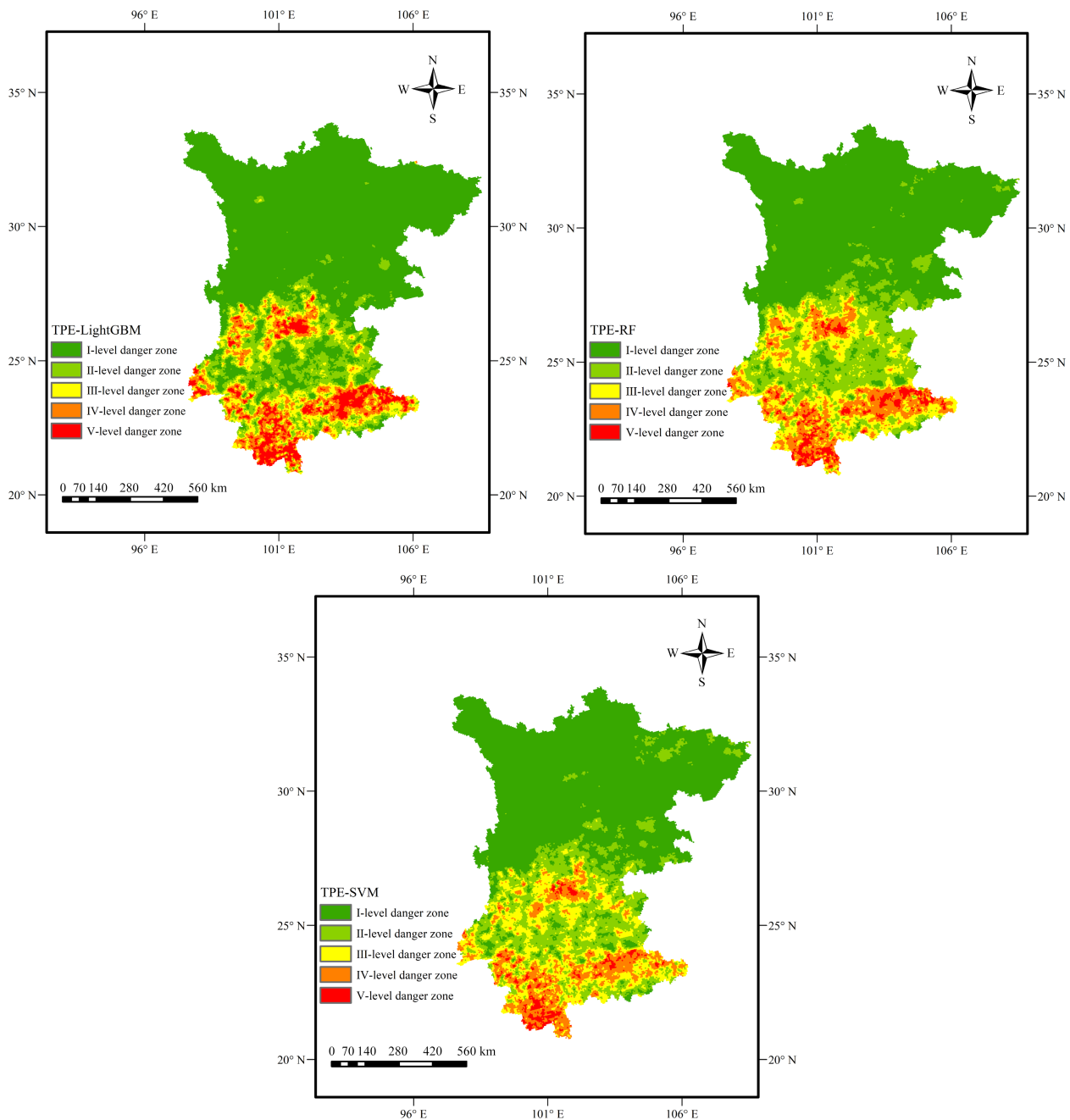
Fig. 8   Forest fire danger maps for the three models.

contribution to the fire occurrence, aligning with the findings of Eskandari et al.[27].

Importance of human activity factors

As the values of GDP and the distance to the nearest road increase, their influences on the model are positive and negative, respectively. This can reflect the growth in socioeconomic activities in Sichuan and Yunnan provinces since 2000, where the enhancement of human activity causes more forest fires.

The contribution of the model is directly proportional to the distance from the nearest railway. The closer to the railway, the less likely a fire is to occur. This is due to the rapid progress in infrastructure and the modernization of the railway system, which has led to strict safety regulations such as the prohibition of open flames in enclosed train carriages. These

regulations have significantly reduced the danger of fires caused by improper handling of cigarette butts or other flammable materials, effectively lowering the possibility of fire occurrence.

A higher population density and a shorter distance to the nearest residential area positively influences the model[16]. However, this study reveals an opposite trend, as the population in the research area is highly concentrated in urbanized, highly developed towns with low forest coverage, where the likelihood of forest fires is relatively low.

Conclusions

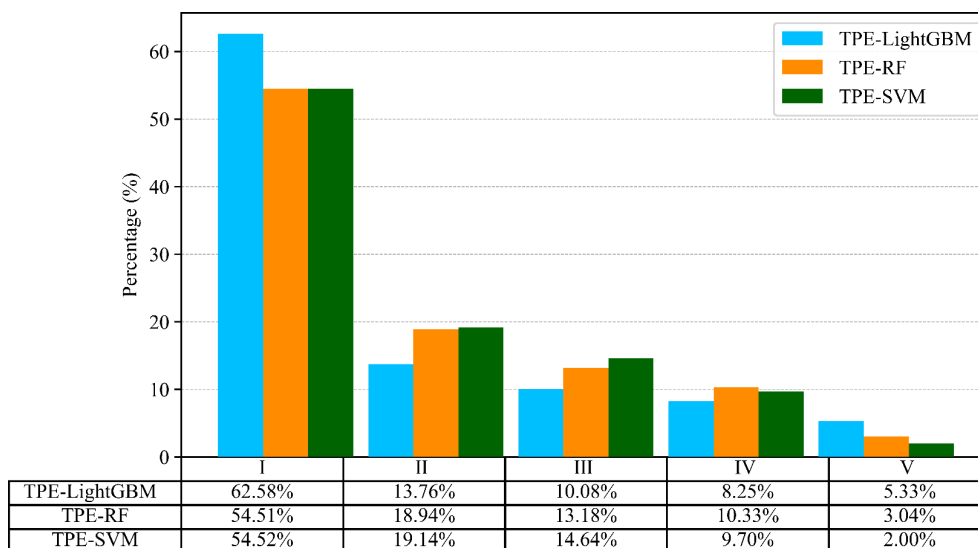This study utilizes forest fire data from Sichuan and Yunnan provinces for the period of 2000 to 2019 as the research

Fig. 9 Classification of forest fire danger levels.

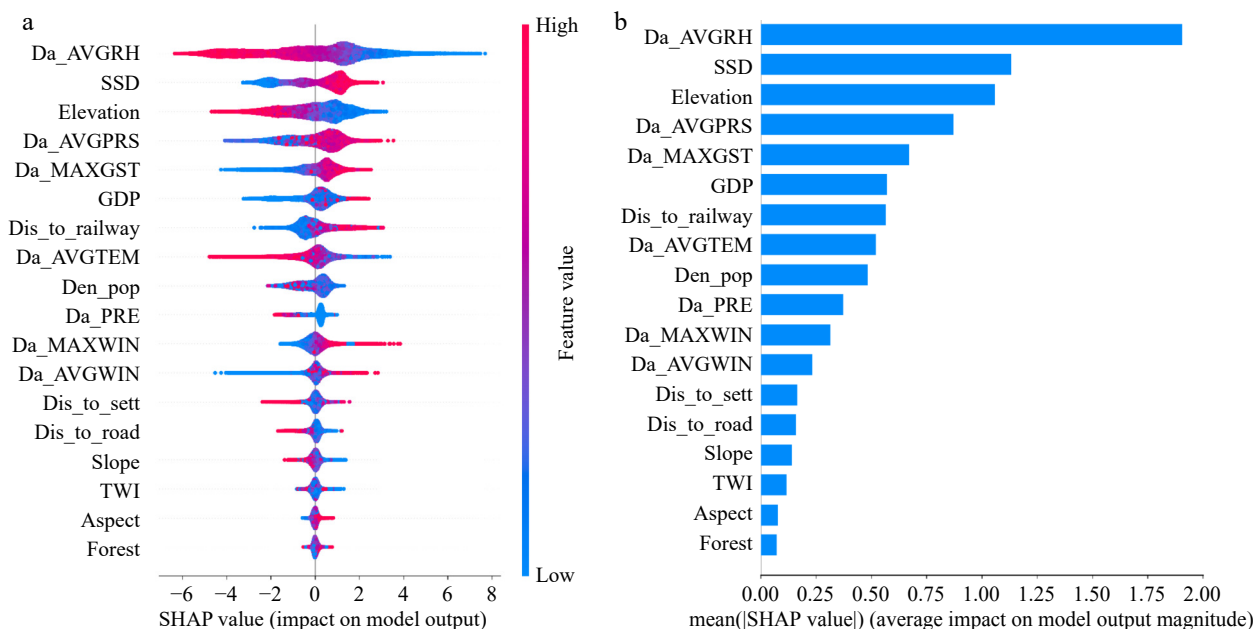|	I	II	III	IV	V
TPE-LightGBM	62.58%	13.76%	10.08%	8.25%	5.33%
TPE-RF	54.51%	18.94%	13.18%	10.33%	3.04%
TPE-SVM	54.52%	19.14%	14.64%	9.70%	2.00%

Fig. 10 Summary graphs of global factors: (a) SHAP scatter plot, (b) SHAP bar graph.

sample, conducting a spatiotemporal analysis of forest fires and selecting 18 forest fire factors. On this foundation, three machine learning models are optimized by the GP and TPE probability proxy models within the Bayesian framework, yielding TPE-LightGBM, TPE-RF, TPE-SVM, GP-LightGBM, GP-RF, and GP-SVM. Model performance is validated using evaluation metrics, with the optimal model being selected. Forest fire danger maps for Sichuan and Yunnan provinces are created. Finally, the model is interpreted using the SHAP method. The major conclusions include:

(1) Temporally, there is significant variation in the annual number of forest fires from 2000 to 2019, with a highly uneven distribution and an overall decline in forest fires after 2010. In terms of monthly variations, forest fires are predominantly concentrated between January and May. Spatially, forest fires during 2000−2019 exhibit a clustered distribution, primarily in the central and southern parts of Sichuan Province and the northwestern and southern parts of Yunnan Province.

(2) In the multicollinearity analysis, three factors, i.e. the daily maximum temperature, daily minimum temperature, and daily average ground surface temperature were excluded, leading to the selection of 18 forest fire driving factors, including daily average temperature and daily average relative humidity.

(3) Models optimized with TPE hold higher predictive accuracy than those optimized with GP, for TPE can handle large-scale datasets more effectively. In addition, TPE utilizes historical observations to generate density functions that provide new hyperparameter configuration suggestions to achieve the desired accuracy.

(4) Utilizing the TPE-optimized model, the forest fire danger map reveals similar spatial distributions across the three maps. The forest fire danger map generated by TPE-LightGBM

effectively delineates fire danger areas into levels I and V, with a clearer distinction between areas prone to fires and those not prone to fires.

(5) A global explanatory analysis of the TPE-LightGBM model provides a ranking of feature importance, identifying daily average relative humidity, sunshine hours, elevation, the daily average air pressure, and daily maximum ground surface temperature as the most significant factors.

The vegetation factors examined in this study were limited to classifying the types of vegetation. However, additional factors, such as the water content of forest fuels, should also be taken into account within the model. Moreover, while this study primarily concentrates on forest fires that occur under natural conditions, it is important to recognize that some fires are the result of human activities, including slash-and-burn practices, burning paper at graves, smoking, and arson, among others. In future research, we will aim to incorporate a broader range of human factors to improve the accuracy and applicability of model.

Author contributions

The authors confirm contribution to the paper as follows: study conception and design: Zhou K, Yao Q; data collection: Liu Z, Yao Q; analysis and interpretation of results: Liu Z, Zhou K, Reszka P; draft manuscript preparation: Liu Z, Zhou K. All authors reviewed the results and approved the final version of the manuscript.

Data availability

Meteorological data are available from https://data.cma.cn. Topographic data are available from www.gscloud.cn. Vegetation, GDP, and population density data are available from www.resdc.cn. Data on human activities are available from www.webmap.cn. The fire database analyzed during the current study are not publicly available due to the restriction of the National Institute of Natural Hazards of the Ministry of Emergency Management.

Acknowledgments

Conflict of interest

The authors declare that they have no conflict of interest.

Supplementary information accompanies this paper at (https://www.maxapress.com/article/doi/10.48130/emst-0024-0026)

Dates

References

1. Cilli R, Elia M, D'Este M, Giannico V, Amoroso N, et al. 2022. Explainable artificial intelligence (XAI) detects wildfire occurrence in the Mediterranean countries of Southern Europe. *Scientific Reports* 12:16349
2. National Bureau of Statistics of China. 2023. Statistical Communique of the People's Republic of China on National Economic and Social Development 2018−2022. https://www.stats.gov.cn/
3. Kondylatos S, Prapas I, Ronco M, Papoutsis I, Camps-Valls G, et al. 2022. Wildfire danger prediction and understanding with deep learning. *Geophysical Research Letters* 49(17):e2022GL099368
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
5. Liu N, Lei J, Gao W, Chen H, Xie X. 2021. Combustion dynamics of large-scale wildfires. *Proceedings of the Combustion Institute* 38(1):157−98
6. Li S, Wu Z, Liang Y, He H. 2017. The temporal and spatial clustering characteristics of forest fires in the great Xing'an Mountains. *Chinese Journal of Ecology* 36(1):198−204
7. Rodrigues M, de la Riva J. 2014. An insight into machine-learning algorithms to model human-caused wildfire occurrence. *Environmental Modelling & Software* 57:192−201
8. Mandallaz D, Ye R. 1997. Prediction of forest fires with Poisson models. *Canadian Journal of Forest Research* 27(10):1685−94
9. Tien Bui D, Bui QT, Nguyen QP, Pradhan B, Nampak H, et al. 2017. A hybrid artificial intelligence approach using GIS-based neural-fuzzy inference system and particle swarm optimization for forest fire susceptibility modeling at a tropical area. *Agricultural and Forest Meteorology* 233:32−44
10. Van Beusekom AE, Gould WA, Monmany AC, Khalyani AH, Quiñones M, et al. 2018. Fire weather and likelihood: characterizing climate space for fire occurrence and extent in Puerto Rico. *Climatic Change* 146(1):117−31
11. Yue W, Ren C, Liang Y, Guo Y, Zhang S. 2024. Study of wildfire hazard susceptibility in Nanning based on interpretable machine learning model. *Science Technology and Engineering* 24(2):858−70
12. Wang Z, Wang K, Li Y, Li G. 2023. Research on forest fire prediction in Yunnan province based on LightGBM and SHAP. *Fire Science and Technology* 42(11):1567−71
13. Huang CL, Dun JF. 2008. A distributed PSO−SVM hybrid system with feature selection and parameter optimization. *Applied Soft Computing* 8(4):1381−91
14. Putatunda S, Rama K. 2018. A Comparative Analysis of Hyperopt as Against Other Approaches for Hyper-Parameter Optimization of XGBoost. *SPML '18: Proceedings of the 2018 International Conference on Signal Processing and Machine Learning, Shanghai, China, 2018*. New York, NY, USA: Association for Computing Machinery. pp. 6−10. doi: 10.1145/3297067.3297080
15. Abdollahi A, Pradhan B. 2023. Explainable artificial intelligence (XAI) for interpreting the contributing factors feed into the wildfire susceptibility prediction model. *Science of The Total Environment* 879:163004
16. Iban MC, Sekertekin A. 2022. Machine learning based wildfire susceptibility mapping using remotely sensed fire data and GIS: A case study of Adana and Mersin Provinces, Turkey. *Ecological Informatics* 69:101647
17. Al-Bashiti MK, Naser MZ. 2022. Machine learning for wildfire classification: Exploring blackbox, eXplainable, symbolic, and SMOTE methods. *Natural Hazards Research* 2(3):154−65
18. Xie L, Zhang R, Zhan J, Li S, Shama A, et al. 2022. Wildfire risk assessment in Liangshan prefecture, China based on an integration machine learning algorithm. *Remote Sensing* 14(18):4592
19. Li Y, Li G, Wang K, Wang Z, Chen Y. 2023. Forest fire risk prediction based on stacking ensemble learning for Yunnan Province of China. *Fire* 7(1):13

20. Ma W, Feng Z, Cheng Z, Wang F. 2020. Study on driving factors and distribution pattern of forest fires in Shanxi province. *Journal of Central South University of Forestry & Technology* 40(9):57−69

21. Guo F, Wang G, Su Z, Liang H, Wang W, et al. 2016. What drives forest fire in Fujian, China? Evidence from logistic regression and Random Forests. *International Journal of Wildland Fire* 25(5):505−19

22. Catry FX, Rego FC, Bação FL, Moreira F. 2009. Modeling and mapping wildfire ignition risk in Portugal. *International Journal of Wildland Fire* 18(8):921−31

23. Chang Y, Zhu Z, Bu R, Chen H, Feng Y, et al. 2013. Predicting fire occurrence patterns with logistic regression in Heilongjiang Province, China. *Landscape Ecology* 28(10):1989−2004

24. Oliveira S, Oehler F, San-Miguel-Ayanz J, Camia A, Pereira JMC. 2012. Modeling spatial patterns of fire occurrence in Mediterranean Europe using Multiple Regression and Random Forest. *Forest Ecology and Management* 275:117−29

25. Syphard AD, Radeloff VC, Keuler NS, Taylor RS, Hawbaker TJ, et al. 2008. Predicting spatial patterns of fire on a southern California landscape. *International Journal of Wildland Fire* 17(5):602−13

26. Maingi JK, Henry MC. 2007. Factors influencing wildfire occurrence and distribution in eastern Kentucky, USA. *International Journal of Wildland Fire* 16:23−33

27. Eskandari S, Pourghasemi HR, Tiefenbacher JP. 2020. Relations of land cover, topography, and climate to fire occurrence in natural regions of Iran: Applying new data mining techniques for modeling and mapping fire danger. *Forest Ecology and Management* 473:118338

28. Moore ID, Grayson RB, Ladson AR. 1991. Digital terrain modelling: A review of hydrological, geomorphological, and biological applications. *Hydrological Processes* 5(1):3−30

29. Zhang C, Yang Q, Li R. 2005. Advancement in topographic wetness index and its application. *Progress in Geography* 24:116−23

30. Cardille JA, Ventura SJ, Turner MG. 2001. Environmental and social factors influencing wildfires in the upper Midwest, United States. *Ecological Applications* 11(1):111−27

31. Garcia CV, Woodard PM, Titus SJ, Adamowicz WL, Lee BS. 1995. A logit model for predicting the daily occurrence of human caused forest-fires. *International Journal of Wildland Fire* 5(2):101−11

32. Oliveira S, Pereira JMC, San-Miguel-Ayanz J, Lourenço L. 2014. Exploring the spatial patterns of fire density in Southern Europe using Geographically Weighted Regression. *Applied Geography* 51:143−57

33. Guo F, Su Z, Wang G, Sun L, Lin F, et al. 2016. Wildfire ignition in the forests of southeast China: Identifying drivers and spatial distribution to predict wildfire likelihood. *Applied Geography* 66:12−21

34. Pham BT, Jaafari A, Avand M, Al-Ansari N, Dinh Du T, et al. 2020. Performance evaluation of machine learning methods for forest fire modeling and prediction. *Symmetry* 12(6):1022

35. Shi C, Zhang F. 2023. A forest fire susceptibility modeling approach based on integration machine learning algorithm. *Forests* 14(7):1506

36. Gigović L, Pourghasemi HR, Drobnjak S, Bai S. 2019. Testing a new ensemble model based on SVM and random forest in forest fire susceptibility assessment and its mapping in serbia's Tara National Park. *Forests* 10(5):408

37. Peng W, Wei Y, Chen G, Lu G, Ye Q, et al. 2023. Analysis of wildfire danger level using logistic regression model in Sichuan Province, China. *Forests* 14(12):2352

38. Bergstra J, Bardenet R, Bengio Y, Kégl B. 2011. Algorithms for hyper-parameter optimization. *NIPS'11: Proceedings of the 24th International Conference on Neural Information Processing Systems, Granada, Spain, 2011*. Red Hook, NY, USA: Curran Associates Inc. pp. 2546−54

39. Ke G, Meng Q, Finley T, Wang T, Chen W, et al. 2017. LightGBM: a highly efficient gradient boosting decision tree. *NIPS 2017: 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 2017*. Red Hook, NY, USA: Curran Associates Inc. pp. 3149−57.

40. Breiman L. 2001. Random Forests. *Machine Learning* 45:5−32

41. Jing X, Li X, Zhang D, Liu W, Zhang W, Zhang Z. 2024. Forecast zoning of forest fire occurrence: A case study in southern China. *Forests* 15(2):265

42. Vapnik VN. 2000. *The Nature of Statistical Learning Theory*. New York: Springer. doi: 10.1007/978-1-4757-3264-1

43. Yao X, Tham LG, Dai FC. 2008. Landslide susceptibility mapping based on Support Vector Machine: A case study on natural slopes of Hong Kong, China. *Geomorphology* 101(4):572−82

44. Lundberg S, Lee SI. 2017. A unified approach to interpreting model predictions. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, 2017*. Red Hook, NY, United States: Curran Associates Inc. pp. 4768−77

45. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, et al. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2(1):56−67

46. Bui DT, Ngo PTT, Pham TD, Jaafari A, Minh NQ, et al. 2019. A novel hybrid approach based on a swarm intelligence optimized extreme learning machine for flash flood susceptibility mapping. *CATENA* 179:184−96

47. Abbas F, Zhang F, Ismail M, Khan G, Iqbal J, et al. 2023. Optimizing machine learning algorithms for landslide susceptibility mapping along the karakoram highway, Gilgit Baltistan, Pakistan: A comparative study of baseline, Bayesian, and metaheuristic hyperparameter optimization techniques. *Sensors* 23(15):6843

48. Liang M, An B, Li K, Du L, Deng T, et al. 2022. Improving genomic prediction with machine learning incorporating TPE for hyperparameters optimization. *Biology* 11(11):1647

49. Zhang G, Wang M, Liu K. 2019. Forest fire susceptibility modeling using a convolutional neural network for Yunnan Province of China. *International Journal of Disaster Risk Science* 10(3):386−403

50. Wang SSC, Qian Y, Leung LR, Zhang Y. 2021. Identifying key drivers of wildfires in the contiguous US using machine learning and game theory interpretation. *Earth's Future* 9(6):e2020EF001910

51. Chen J, Di XY. 2015. Forest fire prevention management legal regime between China and the United States. *Journal of Forestry Research* 26(2):447−55

52. Ying L, Cheng H, Shen Z, Guan P, Luo C, et al. 2021. Relative humidity and agricultural activities dominate wildfire ignitions in Yunnan, Southwest China: patterns, thresholds, and implications. *Agricultural and Forest Meteorology* 307:108540

53. Wang J, Li D, Chen F, Wang S, Niu S. 2012. Study on spatial distribution and trend change of forest fires in Sichuan Province. *Journal of Wildland Fire Science* (2):26−30