Shahzad *et al. Fire Ecology* (2024) 20:57
https://doi.org/10.1186/s42408-024-00289-5

Fire Ecology

ORIGINAL RESEARCH

Open Access

Comparing machine learning algorithms to predict vegetation fire detections in Pakistan

Fahad Shahzad[1†], Kaleem Mehmood[4,5†], Khadim Hussain[3], Ijlal Haidar[4,5], Shoaib Ahmad Anees[6], Sultan Muhammad[5], Jamshid Ali[4], Muhammad Adnan[7], Zhichao Wang[1*] and Zhongke Feng[1,2*]

Abstract

Vegetation fires have major impacts on the ecosystem and present a significant threat to human life. Vegetation fires consists of forest fires, cropland fires, and other vegetation fires in this study. Currently, there is a limited amount of research on the long-term prediction of vegetation fires in Pakistan. The exact effect of every factor on the frequency of vegetation fires remains unclear when using standard analysis. This research utilized the high proficiency of machine learning algorithms to combine data from several sources, including the MODIS Global Fire Atlas dataset, topographic, climatic conditions, and different vegetation types acquired between 2001 and 2022. We tested many algorithms and ultimately chose four models for formal data processing. Their selection was based on their performance metrics, such as accuracy, computational efficiency, and preliminary test results. The model's logistic regression, a random forest, a support vector machine, and an eXtreme Gradient Boosting were used to identify and select the nine key factors of forest and cropland fires and, in the case of other vegetation, seven key factors that cause a fire in Pakistan. The findings indicated that the vegetation fire prediction models achieved prediction accuracies ranging from 78.7 to 87.5% for forest fires, 70.4 to 84.0% for cropland fires, and 66.6 to 83.1% for other vegetation. Additionally, the area under the curve (AUC) values ranged from 83.6 to 93.4% in forest fires, 72.6 to 90.6% in cropland fires, and 74.2 to 90.7% in other vegetation. The random forest model had the highest accuracy rate of 87.5% in forest fires, 84.0% in cropland fires, and 83.1% in other vegetation and also the highest AUC value of 93.4% in forest fires, 90.6% in cropland fires, and 90.7% in other vegetation, proving to be the most optimal performance model. The models provided predictive insights into specific conditions and regional susceptibilities to fire occurrences, adding significant value beyond the initial MODIS detection data. The maps generated to analyze Pakistan's vegetation fire risk showed the geographical distribution of areas with high, moderate, and low vegetation fire risks, highlighting predictive risk assessments rather than historical fire detections.

Keywords Machine learning, Forest fire, Crop fire, Other vegetation fire, Prediction models

Resumen

Los fuegos de vegetación tienen grandes impactos en los ecosistemas y presentan una amenaza significativa para la vida humana. En este estudio, los fuegos de vegetación comprenden fuegos forestales, en cultivos, y otros fuegos de vegetación. Al presente, hay un limitado número de investigaciones sobre la predicción a largo plazo de los fuegos

†Fahad Shahzad and Kaleem Mehmood contributed equally to this work.

*Correspondence:
Zhichao Wang
zhichao@bjfu.edu.cn
Zhongke Feng
zhongkefeng@bjfu.edu.cn
Full list of author information is available at the end of the article

de vegetación en Pakistán. El efecto exacto de cada factor en la frecuencia de los fuegos de vegetación es poco claro cuando se usan análisis estándar. Esta investigación utilizó la alta eficiencia de los algoritmos del aprendizaje automático (i. e. Machine Learning algorithms), para combinar datos de diversas fuentes, incluyendo datos del MODIS Global Fire Atlas, y datos topográficos, de condiciones climáticas, y de diferentes tipos de vegetación adquiridos entre 2001 y 2022. Probamos muchos algoritmos y finalmente elegimos cuatro modelos para procesar formalmente los datos. Su selección fue basada en la performance de sus medidas, como la exactitud, eficiencia computacional, y los resultados preliminares de estas pruebas. El modelo de regresión logística, bosque al azar (random forest), un algoritmo de aprendizaje supervisado (support vector machine), y una técnica de potenciación de gradiente extremo (extreme Gradient Boosting) fueron usados para identificar y elegir los nueve factores clave en fuegos forestales y en cultivos y, en caso de otro tipo de vegetación, siete factores clave que causan incendios en Pakistán. Los resultados indican que los modelos de predicción alcanzaron exactitudes que variaron entre 78,7 y el 87,5% para los fuegos forestales, el 70,4 al 84,0% en el caso de los fuegos en cultivos, y del 66,6 al 83,1% para otro tipo de vegetación. Adicionalmente, el área de los valores bajo la curva (AUC) variaron del 83,6 al 93,4% para fuegos forestales, del 72,6 al 90,6% para los cultivos, y del 74,2 al 90,7% para otro tipo de vegetación. El modelo Random Forest fue quien presentó la mayor exactitud –87,5% en fuegos forestales, 84,0% en cultivos, y 83.1% en otro tipo de vegetación–, y también el AUC más alto (93,4%) para fuegos forestales, (90,6%) en cultivos, y 90,7 en otro tipo de vegetación, lo que probó ser el modelo más óptimo. Los modelos proveyeron de perspectivas predictivas en condiciones específicas y susceptibilidades regionales a la ocurrencia de incendios, adicionando un valor significativo más allá de los datos iniciales de detección por MODIS. Los mapas generados para analizar el riesgo de incendio de la vegetación de Pakistán mostraron áreas de distribución geográfica con riesgo alto, moderado y bajo, señalando determinaciones predictivas más que detecciones históricas de fuegos.

Introduction

Wildfires represent a critical ecological and environmental challenge, impacting ecosystems and human communities globally. This study narrows its focus on the scope of wildfires, particularly vegetation fires, highlighting their frequency, spread, and management strategies. Forest loss and degradation, the emission of significant gasses and aerosols, etc., and the decrease in biodiversity have been identified as significantly contributing to increased vulnerability to fires (Albar et al. 2018). The global occurrence of wildfires shows considerable variation, with estimates suggesting they annually affect between 300 and 400 million hectares, varying significantly by geographic intensity and local conditions (van Lierop et al. 2015; Attri et al. 2020). Over 80% of global wildfires occur in savannahs and grasslands, mainly in South America, Australia, Africa, and South Asia. Forest and shrub-dominated regions account for 20% (Schultz et al. 2008). Annually, substantial funds are allocated towards fire management efforts to reduce or prevent the adverse consequences of wildfires (Thomas et al. 2017). Wildfire events lead to the death and displacement of fauna (Tien Bui et al. 2016; Bhujel et al. 2017), pose risks to the lives and livelihoods of local communities, impact soil fertility and water cycles, release harmful pollutants, including particulate matter (Shahdeo et al. 2020) that may contribute to global warming, and result in the loss of vegetation cover (Martell 2007; Usoltsev et al. 2020; Shobairi et al. 2022; Anees et al. 2022b, 2024; Akram et al. 2022; Aslam et al. 2022; Khan et al. 2024). Advancements in remote sensing technologies have contributed significantly to the monitoring and evaluating of vegetation fires (Gitas et al. 2012). Previous research has leveraged multi-temporal and multi-sensor remote sensing technologies to assess and monitor vegetation fires (Table 1).

Table 1 List of sensers used for fire monitoring

Sensor package	Source
SPOT 2A/2B (MSI)/SPOT 4–5 (VGT)/SPOT 1–7 (HRV)	Krishna and Reddy 2012
Landsat TM/ETM + /OLI	Manaswini and Sudhakar Reddy 2015
ENVISAT (MERIS)	Saranya et al. 2014; Reddy and Sarika 2022
IRS AwiFS	Reddy et al. 2017
SUOMI NPP VIIRS/Terra-Aqua (MODIS)	Chuvieco et al. 2018
NOAA 7–19 (AVHRR)/PROBA V/Sentinel-1A/1B (SAR)/IRS LISS III	Chuvieco et al. 2019

Vegetation fires result from a complex network of interactions among various natural variables, including climate and weather conditions (Andreevich et al. 2020), fuel composition, and topography. The ignition sources for these fires encompass hot surfaces, electrical sparks, flames, friction, static electricity, mechanical impacts (such as from machinery contact or falling rocks), and natural events like lightning (Vadrevu et al. 2008; Bui et al. 2017; Nami et al. 2018). Although human activities are globally recognized as predominant causes of fires, practices such as slash-and-burn for agricultural purposes are widely prevalent in South and Southeast Asia. Our study focuses on the climatic influences on fire occurrences in Pakistan. This study addresses how climatic factors, rather than direct human interventions, predominantly influence fire dynamics in Pakistan. While acknowledging the significant impact of human activities on fire occurrences as seen in regions such as the Eastern Ghats and northeast India (Vadrevu et al. 2008), Sarawak in Malaysia (Kleinman et al. 1995), and the Chittagong hill tracts in Bangladesh (Borggaard et al. 2003), our analysis focuses on how environmental variables (Anees et al. 2022b) like temperature, humidity, and solar radiation play crucial roles in the region's fire ecology. Topographic factors such as aspect, slope (Muhammad et al. 2023), and elevation are also considered for their effects on the extent of burnt areas and fire intensity based on comparisons across different studies (Nunes et al. 2016; Pan et al. 2023).

Various models have been documented in the literature, focusing on distinct phases of the fire control cycle. These include vegetation fire occurrence models (Botequim et al. 2017), vegetation fire spread models (Zhai et al. 2020), deployment and dispatch models, vegetation fire damage models, and decision and information systems as technological support platforms (Marques et al. 2012; Duff and Tolhurst 2015). The studies describing models briefly discuss prominent algorithms in each category, including supervised, unsupervised, and agent-based modeling approaches. Additionally, they included references on the fundamentals of machine learning. Supervised learning works to establish a correlation between input data that has been labeled and the corresponding known output using a continuous target factor. A constant variable of interest is used in regression analyses, with various applications including fire vulnerability, fire occurrence, fire spread and burn area estimation, smoke and emissions prediction, and, finally, climate change assessment (Jain et al. 2020). Unsupervised learning aims to uncover patterns and relationships within data without using a specific target or outcome variable to guide the learning process. It is applicable for tasks involving clustering and dimensionality reduction. Clustering tasks in this context are used for fire mapping, fire detection, prediction of burnt areas, and fire weather prediction (Bot & Borges 2022). Some fire prediction algorithms, prominent for their computational speed and simplicity, utilize both supervised and unsupervised learning techniques to determine vegetation fire risks. These include neural networks, decision trees, random forest (Eslami et al. 2021), regression trees, and classification algorithms (Cabral et al. 2018), along with K-nearest neighbor, support vector machines, K-means clustering, self-organizing maps, autoencoders, hidden Markov models, and hard competitive learning (Arnold et al. 2014). A prominent gap exists in long-term, predictive studies integrating environmental, meteorological, and human factors, particularly across broader geographical scales (Sohail et al. 2023). This gap highlights the need for enhanced predictive modeling to inform proactive fire management strategies. In response to these gaps, our research aims to (1) compile a comprehensive dataset of historical fire incidents in Pakistan from 2001 to 2022; (2) develop a predictive model for wildfire occurrences using MODIS data, incorporating various environmental and meteorological variables to forecast spatial and temporal patterns; and (3) conduct a long-term trend analysis to evaluate the frequency, distribution, and severity of wildfires in Pakistan over the past two decades.

Materials and methods

Study area

The research focused on Pakistan, covering the period from 2001 to 2022. Pakistan is located in the western zone of South Asia, northeast of the Arabian Sea, between latitudes 24° and 37° N and longitudes 62° and 75° E (Qasim et al. 2014). Pakistan covers an area of 875,832 km2. Forests cover 2113 km2, croplands cover 176,976 km2, and other vegetation covers 261,755 km2. According to MODIS data, there were 208,943 fire events recorded in Pakistan from 2001 to 2022, including 642 in forests, 158,474 in croplands, and 31,484 in other vegetation types. Figure 1 shows classifications of forested land, cropland, and other vegetated land.

The country is known for its diverse landscapes, which include towering mountains in the north and expansive arid regions in the southwest. It has four distinct seasons: a mild and dry winter (December to February), a hot and dry spring (March to May), a rainy season (June to August), and a post-monsoon season (September to November) (Begum et al. 2011). Pakistan's forest cover is only 4.5%, a substantial concern considering the country's agricultural-driven economy and location within the South Asian Ecological Zone (Oliveira et al. 2011). Throughout the latter half of the twentieth century, evidence indicated an escalating incidence of wildfires in

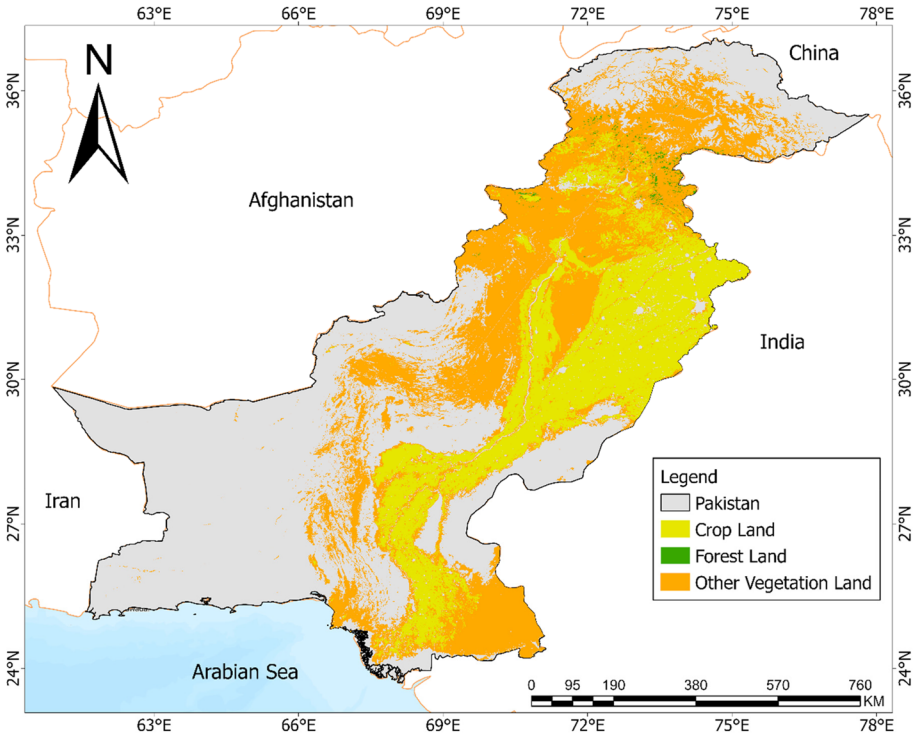

Fig. 1 Study area map along with various LULC

Pakistan, contributing to increased burn area (Rafaqat et al. 2022a, b). Characterized by its lowest elevation at sea level and vulnerability to desertification, the eastern region of Pakistan requires targeted conservation and fire prevention strategies, particularly considering the availability of remote sensing technologies and worldwide databases that provide opportunities for a more detailed identification of factors causing fires and enhanced prediction models (Rafaqat et al. 2022a, b). This region is particularly vulnerable to wildfires due to its dry environment with little rainfall and susceptibility to desertification (Kattel et al. 2019; Anees et al. 2022a).

Datasets

Handling of response variable

This study employs a comprehensive approach to analyze historical fire data, focusing on the period from January 2001 to December 2022. This study used the MODIS fire product from the Fire Information for Resource Management System (FIRMS), which gave information about active fires found by NASA's Aqua and Terra satellites' MODIS instruments (https://firms.modaps.eosdis.nasa.gov) (Zhang et al. 2021). We combined the monthly global 500 m grid product with 1 km of MODIS active fire observations to enhance the spatial analysis of the MCD64A1 Version 6 Burned Area data product (Giglio et al. 2018). This product facilitates the identification of per-pixel burned areas, detecting thermal anomalies and fire locations at a moderate resolution (Katagis and Gitas 2022). We used this data to evaluate fire regimes on a national to continental scale, identify global hot spots of fire, and monitor trends in global vegetation fire occurrences (Giglio et al. 2006; Chuvieco et al. 2008). All fire events reported with a confidence level exceeding 50% were considered for detailed analysis. The analysis followed a grid-based approach, examining each 1×1 km grid cell for vegetation fire occurrences, binary-labeled as "1" for presence and "0" for absence. In this study, analyzing land use and land cover was crucial for understanding the distribution and types of vegetation affected by fires. The International Geosphere-Biosphere Project (IGBP) classification scheme of the MODIS product MCD12Q1 was used in the study (Liang et al. 2015; Badshah et al. 2024). This product has 500-m-level data on land cover (Sulla-Menashe and Friedl 2018). The dataset available on the LP DAAC website (https://lpdaac.usgs.gov/) greatly aided in identifying the surfaces beneath various types of vegetation in the study area (Usoltsev et al. 2022; Zhao et al. 2022). The research area shown in Table 2 underwent a careful process of mosaicking and reprojection using the Hierarchical Data Format-Earth Observing System (HDF-EOS) to Grid (HEG) tools. This step was crucial for achieving an accurate and coherent spatial representation of land cover types. The study area

Table 2 Descriptions of vegetation types

Vegetation types	Classes	Source	Resolution	Unit	Format	Duration
Forest	Evergreen broadleaf forest	LP DAAC	500 m	m^2	HDF-EOS	2001–2021
	Deciduous needle leaf forest					
	Deciduous broadleaf forest					
	Mixed forests					
Other vegetation	Closed shrublands					
	Open shrublands					
	Woody savannas					
	Savannas					
	Grassland					
Crop	Cropland					
	Cropland-natural vegetation mosaics					

divided grid cells into categories based on the land cover types a vegetation fire had affected, including forest fire, other vegetation, and cropland. Five hundred twelve out of 642 forest cells, 124,179 out of 158,474 cropland cells, and 22,663 out of 31,484 vegetation cells were marked as "fire cells" and given the number "1."

During the dataset development, we created two random subsets of the actual MODIS vegetation fire ignition spots that were detected. We allocated 70% of this data for training the models and the remaining 30% for testing their performance. This division is standard practice in machine learning to validate models effectively, ensuring they can generalize well to new, unseen data. Using a 70–30 split, we aim to provide a robust dataset for training while retaining sufficient data for an accurate assessment of model performance in real-world scenarios (Rubí et al. 2023).

Selection and handling of predictor variables

This study utilized the Shuttle Radar Topography Mission's (SRTM) Digital Elevation Model (DEM) dataset to investigate the impact of elevation, slope, and aspect as shown in Fig. 2 on the vegetation fire analysis. The SRTM dataset, downloaded from the SRTM Data Portal

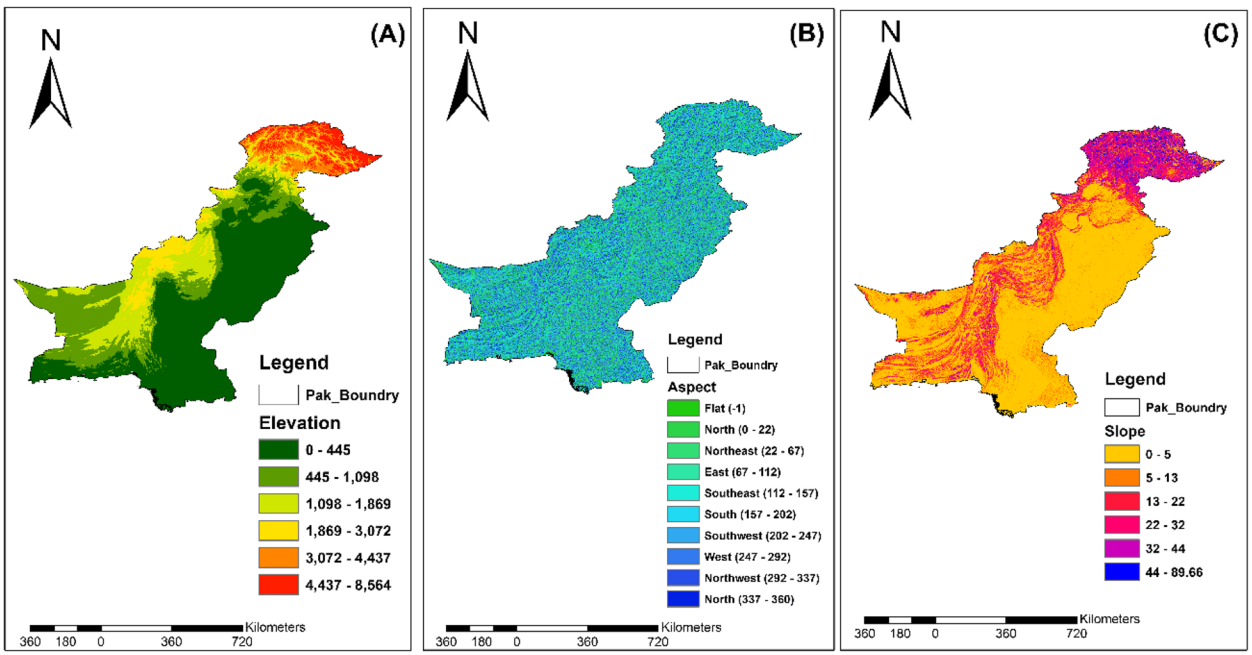

Fig. 2 Topographical factors. **A** Elevation. **B** Aspect. **C** Slope

(January 1, 2023), provide highly accurate nationwide coverage.

The historical monthly climatic data was downloaded from two different sources: WorldClim (https://www.worldclim.org/) (Barreto and Armenteras 2020) and ERA 5 climate reanalysis data (https://cds.climate.copernicus.eu/) (Zhang et al. 2021) accessed on January 1, 2023). Key climatic variables extracted from WorldClim include minimum temperature (°C), maximum temperature (°C), and precipitation (mm), presented in GeoTiff format with a spatial resolution of approximately 2.5 min (~ 21 km2). Additional climatic variables sourced from ERA 5 climate reanalysis include northward and eastward components of the 10 m wind (m/s), skin temperature (°C), surface net solar radiation (W/m2), surface net thermal radiation (W/m2), surface pressure (hPa), soil temperature (°C), and forecast albedo (unitless). These variables are provided in Netcdf format with a spatial resolution of about 9 km2. All data underwent meticulous preprocessing using RStudio, specifically employing the "raster" and "ncdf4" packages, alongside the ArcGIS software (Table 3).

Detection of violations of assumptions about independent variables

A linear regression model may encounter multicollinearity, characterized by a substantial correlation among its independent variables. This multicollinearity has the potential to distort the model's estimation and impede accurate predictions (Chang et al. 2013). The correlation matrix shown in Fig. 3 uses a color scale ranging from blue (low correlation) to red (high correlation) to identify significant correlations between variables. Each cell in the matrix represents the correlation coefficient between two variables, providing a visual aid to detect potential multicollinearity issues. Analysis of multicollinearity involves assessing variance inflation factors (VIF) and tolerance levels (TOL), which are commonly utilized to evaluate the relationships among independent variables. It is widely acknowledged that a TOL value below 0.1 and a VIF value exceeding 10 indicate the presence of multicollinearity (Bui et al. 2019; Li et al. 2022). These thresholds suggest that multicollinearity could significantly impact the reliability of regression and classification model estimates. TOL and VIF are computed as follows (Eqs. 1 and 2):

$$\mathrm{TOL} = 1 - \mathrm{R}^2 \tag{1}$$

$$VIF = \frac{1}{1 - \mathrm{R}^2} = \frac{1}{TOL} \tag{2}$$

where the coefficient of complex determination is denoted by R^2.

Mann–Kendall mutation test

The Mann–Kendall mutation test is a statistical method used to analyze temporal fluctuations and detect significant trends or "mutational changes" within time series data. These "mutational changes" refer to substantial alterations in the trend of the data, such as shifts from increasing to decreasing values or vice versa, which could indicate environmental or systemic changes. This method is valued for its straightforward implementation, high precision, broad applicability across diverse datasets, minimal human intervention, and efficient validation capabilities (Yue et al. 2002). The time series x, including

Table 3 Descriptions of independent variables

Category	Predictors	Abbreviations	Source	Resolution	Unit	Format	Duration
Topography	Slope	S	https://earthexplorer.usgs.gov/	30 m	M	Geo.Tiff	2020
	Aspect	A					
	Elevation	E					
Climatic data	Minimum temperature	Temp_min	WorldClim	21 km^2	°C	Geo.Tiff	2001–2021
	Maximum temperature	Temp_max			°C		
	Precipitation	Ppt			mm		
	Northward components of the 10 m wind	Wind U	ERA 5 climate reanalysis data	9 km^2	ms^{-1}	Netcdf	2001–2022
	Eastward components of the 10 m wind	Wind V			ms^{-1}		
	Skin temperature	Mean_temp			°C		
	Surface net solar radiation	Net_solar			jm^{-2}		
	Surface net thermal radiation	Net_termal			jm^{-2}		
	Surface pressure	SF			Pa		
	Soil temperature	soil_temp			°C		
	Forecast albedo	FA			%		

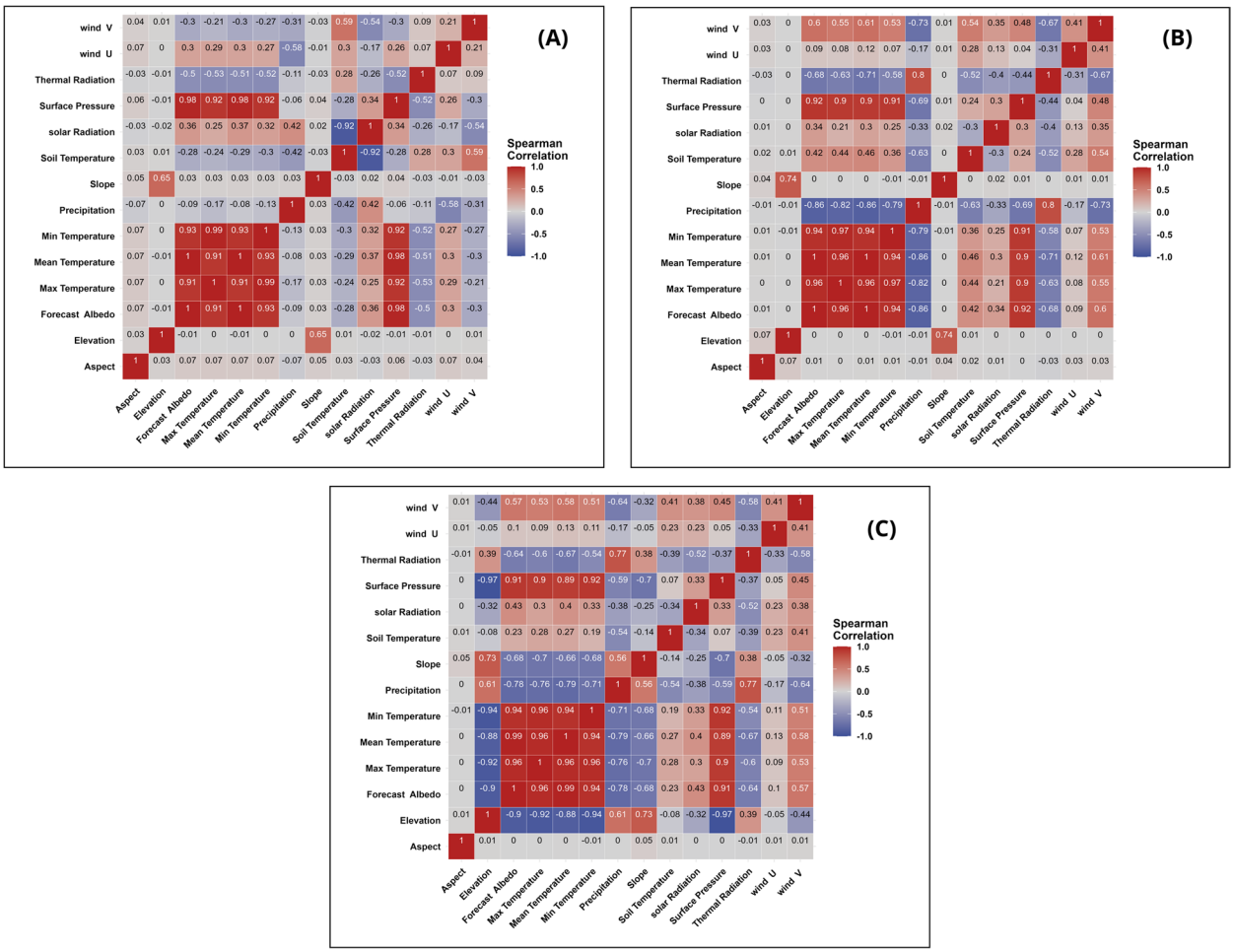

Fig. 3 The Spearman rank correlation heat maps for **a** forest, **b** crop, and **c** other vegetation

n samples, represents the fundamental temporal variations. By analyzing these patterns, it is possible to obtain knowledge of the historical evolution of the environmental system, including weather variables and MODIS-detected changes that generated the data (Mehmood et al. 2024d). The test calculates a sequence of detecting mutations according to the Eq. 3:

$$d_k = \sum_{i=1}^{k} \gamma_i (k = 2, 3 \ldots, n). \tag{3}$$

The sequence dk is a succession of independent units that adhere to the common scoring factors for calculating (dk) (Zhang et al. 2020):

$$UF(d_k) = \frac{[d_k - E(dk)]}{\sqrt{var}(dk)} \tag{4}$$

(dk) indicates the expected value, Var(dk) is the variance, and UFk is a standard distribution of values. The statistical order is determined by analyzing the time series x in the order $x_1, x_2, \ldots, x_n$. The reverse sequence of x ($x_n, x_{n-1}, \ldots, x_1$) is computed. This procedure is repeated, and the value of d_k is assessed by comparing each computed d_k to its expected statistical properties, including the mean and variance, to determine deviations that suggest trends. A UB or UF value greater than 0 indicates the presence of both positive and negative trends in the time series. When these values exceed or fall below the key threshold (significance level), the time series trends upward or downward. The area beyond the threshold line is the mutation time region of the significant line (Feng et al. 2016).

Methodological overview machine learning models

Logistic regression

The logistic regression method is a classical statistical modeling method used to model binary outputs given one or more independent variables (Balboa et al. 2024). It is effective in different geographic locations for predicting and analyzing the variables that drive fire occurrence

at different topographical levels (Garcia et al. 1995; Martínez et al. 2009). Many researchers have included model applicability (Oliveira et al. 2012; Rodrigues and De la Riva 2014). The formula for LR is:

$$\text{Logit}(p) = ln\left(\frac{p}{P-1}\right) \tag{5}$$

The equation represents the relationship between the probability of vegetation fire occurrence (P) and the number of variables (n), where ($a1$, $a2$, …, an) are the coefficients for each variable and ($\times 1, \times 2$, …, xn) are the factors that impact the rate of vegetation fires (Peng et al. 2002; Zhang et al. 2021).

Random forest

The random forest (RF) model was employed to determine the variables that drive vegetation fires and their respective influences on the probability of vegetation fires in the geographical areas of Pakistan. The RF model, presented by Breiman (2001), employs multiple decision trees to train and predict samples, rendering it a classifier (Haddouchi and Berrado 2019). RF is a machine learning method based on an ensemble of classification and regression trees (CARTs). Each tree in the RF model is built using bootstrap samples, enhancing the model's robustness against outliers and variability, which is critical for predictive accuracy in forest fire forecasting (Su et al. 2018; Zhang et al. 2022). The RF model is a fast machine-learning approach that can handle many input factors and delivers high predicted accuracy (Sarkar et al. 2024). Still, it is sensitive to the danger of overfitting (Luo et al. 2024).

$$h(x) = \frac{1}{T}\sum_{t=1}^{T} h(x, \theta_t) \tag{6}$$

Hyperparameter adjustment was critical to derive the final models (Probst et al. 2019; Mehmood et al. 2024a, b, c). The number of trees (n=1000), tree depth (maximum depth of 8), and minimum node size (minimum of 7 samples per leaf node) were optimized in the forest and crop fire prediction, but in the case of other vegetation, a minimum size of 6 for each node. The final prediction is obtained by taking the mean of each regression subtree $\{h(x, \theta_t)\}$, T represents the number of decision trees, θt represents a random vector that is independently and identically distributed, and x represents the input vector. The predictive efficacy of the model is determined by the quantity of random features and trees (Segal and Xiao 2011).

eXtreme Gradient Boosting

eXtreme Gradient Boosting (XGBoost), presented by Chen and Guestrin in 2016, is an innovative gradient-boosting decision tree (GBDT) algorithm (Chen and Guestrin 2016). It utilizes Taylor's second-order expansion to optimize the loss function, exhibiting improved computing efficiency and generalization ability compared to other machine learning algorithms (Xie et al. 2022). The XGBoost model represents:

$$\widehat{y}_i = \sum_{k=1}^{k} f_k(x_i), f_k \in F \tag{7}$$

Here, $\widehat{y}_i$ is the predicted value for the ith sample, k denotes the number of decision trees, x_i is the input data for the ith sample, $f_k(x_i)$ is the k th decision tree generated in the k th iteration, and f_k belongs to the tree collection space F (Luo et al. 2024).

The objective function for XGBoost is:

$$Obj = \sum_{i=1}^{N} l\left(y_i, \widehat{y}_i\right) + \sum_{k=1}^{k} \Omega\left(f_k\right) = \sum_{i=1}^{N} l\left[y_i, \widehat{y}_i^{t-1} + f_t(x_i)\right] + \sum_{k=1}^{k} \Omega\left(f_k\right) \tag{8}$$

In Eq. (8), the first part represents the loss function, the difference between the predicted and observed numbers. The second component is a regularization term that essentially governs the complexity of the model, guides the construction of a tree structure, and prevents overfitting (Piraei et al. 2023).

Support vector machines

Pattern classification and nonlinear regression widely utilize support vector machines (SVMs). SVMs are based on the idea of minimizing structural risk (Jodhani et al. 2024). The fundamental concept behind SVMs is to create a classification hyperplane that serves as a decision boundary. The distance between positive and negative examples achieves superior generalization accuracy(Naderpour et al. 2019). SVMs specialize in manipulating data in high-dimensional environments by effectively employing kernel functions to tackle diverse nonlinear problems(Rossi and Villa 2006). For a two-class SVM, considering a training set $T=\{(x_1, y_1), \cdots (x_1, y_1)\} \in (X \times Y)^l$, where $x_i \in X=R^n$ and $y_i \in \{1, -1\}$ for (i =1,2,…, l) which represents the feature vector. The consequence parameter C and the kernel function $K(x, x')$ are specified. The problem of optimization is then formulated and resolved in the following manner (Boubeta et al. 2015):

$$\underset{\alpha}{min}\ \tfrac{1}{2}\Sigma_{i=1}^{j}\Sigma_{j=1}^{1} y_i y_j a_i a_j k\left(\varkappa, x'\right) - \Sigma_{j=1}^{1}\alpha_j \tag{9}$$

$$s.t.\Sigma_{i=1}^{j} y_i \alpha_i = 0, 0 \le \alpha_i \le C, \dot{i} = 1, \ldots, l \quad (10)$$

The optimal solution $\alpha^* = (\alpha^*, \ldots, \alpha^*)^T$ is obtained. A positive component $\alpha^* : 0 \le \alpha_j^* \le C$ is then selected, and the threshold is computed as follows (Pang et al. 2022):

$$b^* = y_j - \sum_{i=1}^{1} y_i \alpha_i K\left(x_i - x_j\right) \quad (11)$$

Finally, the decision function is constructed:

$$\boldsymbol{f}(\boldsymbol{x}) = sgn(\sum\nolimits_{i=1}^{1} \alpha_i * y_i K(x, x_i) + b^* \quad (12)$$

Model performance evaluation methods

Accuracy serves as a metric for evaluating categorical models, representing the percentage of correctly predicted outputs by the model as follows (Shao et al. 2023):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (13)$$

TP is the percentage of true positive cases, *TN* is the proportion of true negative cases, *FP* indicates the percentage of false positive cases, and *FN* is false negative cases (Pang et al. 2022). Recall or sensitivity, also presented as part of our evaluation metrics in Table 5, measures the proportion of actual positives that are correctly identified by the model and is calculated as (Eq. 15). The F1 score, which combines precision and recall into a single metric, is particularly useful when dealing with imbalanced datasets and is computed using (Eq. 16).

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (14)$$

The F1 score, combining precision and recall, is computed as:

$$F1Score = 2.\frac{\text{Precision . Recall}}{\text{Precision+Recall}} \quad (15)$$

The kappa coefficient is an indicator of statistical significance used to assess the level of reliability in testing. The expression is given by the following (Watson and Petrie 2010):

$$Kappa = \frac{P_0 - P_E}{1 - P_E} \quad (16)$$

where *Po* is the accuracy of the prediction, and *Pe* is the probability of chance agreement, derived from the class probabilities, and is crucial in understanding the kappa calculation as it considers both the observed and expected agreements. Kappa coefficients are categorized into five categories to represent varying degrees of accuracy: 0.0 to 0.20 for extremely low accuracy, 0.21 to 0.40 for medium accuracy, 0.41 to 0.60 for high accuracy, 0.61 to 0.80 for excellent accuracy, and 0.81 to 1 for virtually perfect accuracy (Landis and Koch 1977).

The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1-specificity), illustrating the trade-offs between true positive and false positive rates across different thresholds (Carter et al. 2016). The area measures the accuracy of the results under the curve (ROC). The equations for the sensitivity and specificity are as follows (El Emam et al. 2001; Pang et al. 2022). The AUC quantifies the overall ability of the model to discriminate between classes and is discussed in terms of effectiveness (Muschelli III 2020). The area under the curve (AUC) measures the model's predictive power, categorized into four distinct groups: 0.5–0.85 denotes medium performance, 0.85 ~ 0.95 signifies high performance, and 1.0 indicates ideal performance (Yingyongyudha et al. 2016; Sun et al. 2021). Figure 4 illustrates the workflow depicted in this paper.

Results

This study examined the multicollinearity of various environmental and topographic factors; their tolerance (TOL) values are more than 0.1, and variance inflation factors (VIF) are less than 10 across different vegetation types: forest, crop, and other vegetation, as shown in Table 4. This indicates a lack of covariance among the factors that may initiate fires, suggesting that these variables can inform fire risk assessments within the defined constraints of this study area and period.

Mann–Kendall mutation

The Mann–Kendall test applied to vegetation fires in Pakistan from 2001 to 2022 reveals fluctuating but overall upward trends in fire hotspots. Specifically, from 2006 to 2007, UF values were negative, indicating a temporary decline in fire occurrences. Conversely, from 2001 to 2006 and 2008 to 2022, UF values were consistently above zero, demonstrating a rising trend in the frequency of fires. Notably, the UF curve surpasses the 0.05 confidence level (± 1.96 standard deviations), suggesting that the decline and rise in fire frequencies are statistically significant. These trends are visually detailed in Fig. 5. In Fig. 6, the temporal evolution of vegetation fires spanning the years 2001 to 2022 is depicted, with a detailed legend categorizing the data into distinct types, including forest fires, crop fires, and other vegetation fires.

The cumulative anomaly curve on the vegetation fire points in Pakistan showed negative, indicating a consistent buildup of negative anomalies from 2001 to 2022, as shown in Fig. 7. The Mann–Kendall test shows a substantial increase trend in vegetation fires, but the curve's below-zero position suggests consistent deviations from predicted values. These anomalies suggest that hotspots

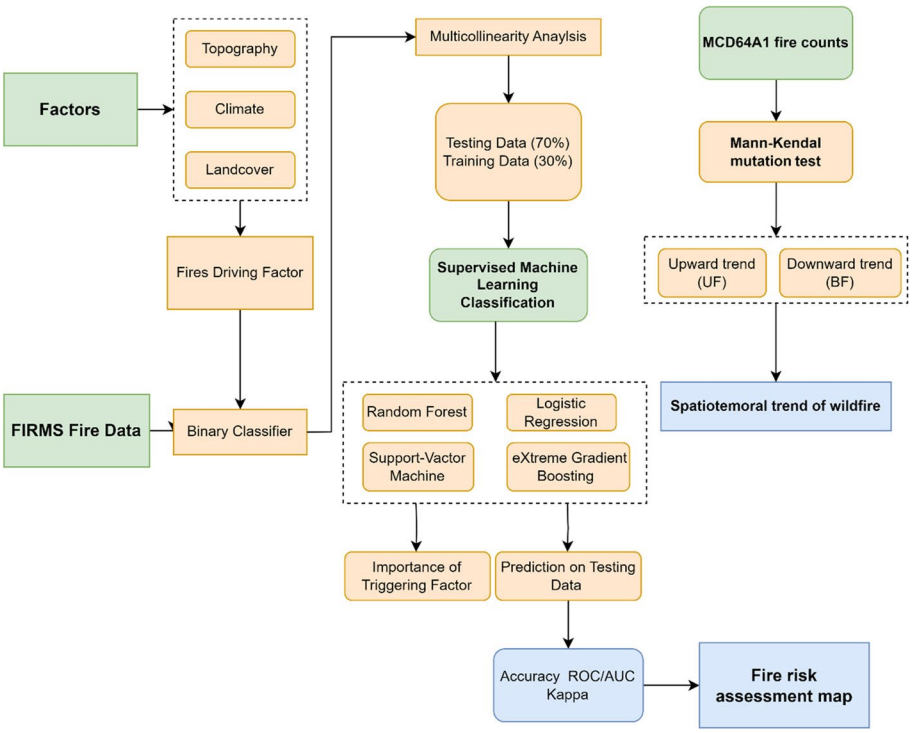

Fig. 4 Flowchart illustrating the stages involved in data processing and the outputs

Table 4 Results of multicollinearity analysis

Num	Factor	TOL for forest	TOL for crop	TOL for other vegetation	VIF for forest	VIF for crop	VIF for other vegetation
1	Slope	0.67	0.5	0.4	1.48	1.96	2.45
2	Aspect	0.98	0.9	0.99	1.01	1	1
3	Elevation	0.67	0.5	Not significant	1.48	1.96	Not significant
4	Minimum temperature	0.19	0.39	Not significant	5.07	2.5	Not significant
5	Precipitation	0.44	0.25	0.25	2.24	3.95	3.86
6	Northward components of the 10 m wind	0.47	0.77	0.73	2.11	1.28	1.35
7	Eastward components of the 10 m wind	0.64	0.47	0.56	1.54	2.12	1.77
8	Soil temperature	0.18	0.52	0.3	5.36	1.89	3.32
9	Surface net thermal	0.44	0.45	0.34	2.24	2.21	2.86

frequently go below expectations, requiring further investigation into specific time frames and environmental variables. The point at which UF and UB meet the confidence line validates its validity to detect an essential change in the number of national hotspots between 2001 and 2022.

Logistic regression

To assess prediction accuracy across different vegetation types, a logistic regression modeling approach was used. The accuracy and AUC scores for each type of vegetation are as follows: in forest fire, the model achieved 81.6% accuracy and 87.3% AUC; in crop fire, the accuracy was 70.4% and the AUC 72.6%; and in other vegetation fires, the accuracy was 66.6% with an AUC of 74.2%. These results are presented concurrently in Table 5 for forest vegetation, crop vegetation, and other vegetation. The ROC curves, which predict the rates of the four modeling approaches, are shown in Fig. 8. This figure illustrates the effectiveness of each approach in distinguishing between

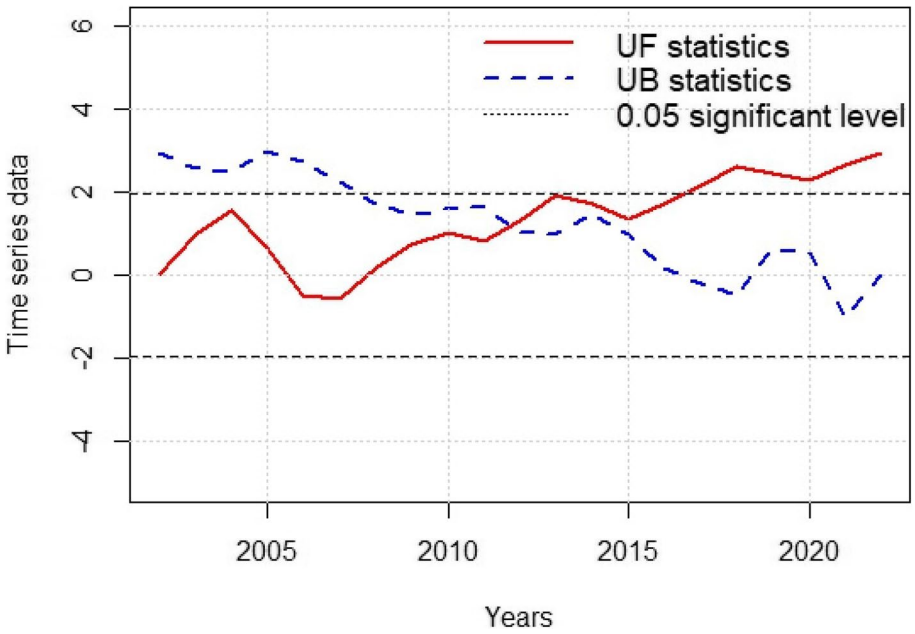

Fig. 5 Mann–Kendall mutation test curve illustrating the temporal trends of fire hotspots from 2001 to 2022

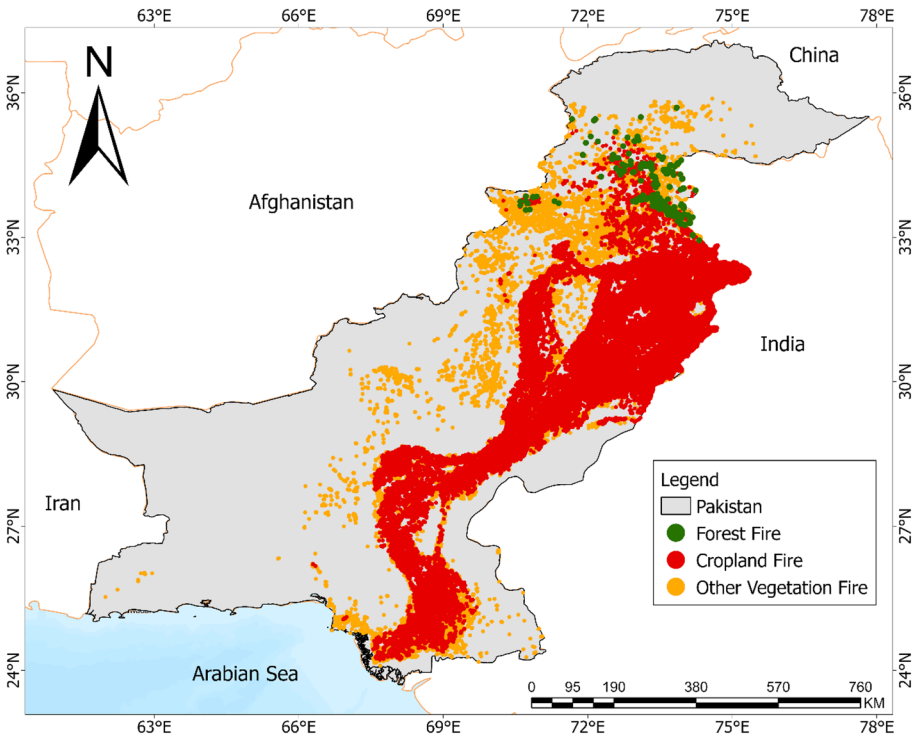

Fig. 6 The incidence of vegetation fires during the same period, categorized by fire types such as forest fires, crop fires, and other vegetation fires, highlighting spatial variations

the presence and absence of fire under various conditions. Additionally, In Fig. 9, the analysis reveals that the importance of initiating factors varies significantly across different types of vegetation. Notably, while weather-related variables tend to dominate across all categories, their impact is not uniformly distributed. For forest fires,

Fig. 7 Cumulative distance leveling curve for vegetation fires

Table 5 Results from the evaluation of the four models for different types of vegetation

Model type	Vegetation type	Accuracy (%)	AUC (%)	Precision (%)	Recall (%)	F1 score (%)
Logistic regression	Forest fire	81.6	87.3	90.0	71.0	79.4
	Crop fire	70.4	72.6	69.3	74.3	71.7
	Other vegetation	66.6	74.2	66.2	68.7	67.4
Random forest	Forest fire	87.5	93.4	83.5	86.9	90.6
	Crop fire	84.0	90.6	83.0	85.9	84.4
	Other vegetation	83.1	90.7	83.1	83.3	83.2
SVM	Forest fire	78.7	83.6	89.3	65.1	75.3
	Crop fire	74.5	80.7	71.6	81.1	76.1
	Other vegetation	68.7	74.8	64.9	81.1	72.1
XGBoost	Forest fire	86.0	92.6	89.7	81.3	85.3
	Crop fire	83.9	90.0	82.3	86.1	84.2
	Other vegetation	79.4	87.6	77.7	82.2	79.9

variables such as wind speed (wind *V*), Soil Temp, and Tmin appear as the most influential, whereas for crop fires, factors like net thermal and ppt take precedence. This variation underscores the complexity of fire risk factors and the need to tailor fire management strategies to specific vegetation types and environmental conditions.

Random forest

The present study used advanced features of tidy models and ranger packages to predict forest, crop, and other vegetation fire RF models. This investigation led to the model's configuration space and found the greatest accuracy balance shown in Table 5. When modified with these parameters, the forest fire model predicted accuracy of 87.5% and 93.4; in crop fire, the model had 84.0% accuracy and 90.6% AUC; and in other vegetation fire, the model exhibited 83.1% accuracy and 90.7% AUC. Figure 8 demonstrates that the RF model surpassed the performance of the other three modes, as evaluated with accuracy and AUC metrics. Hence, we deemed the

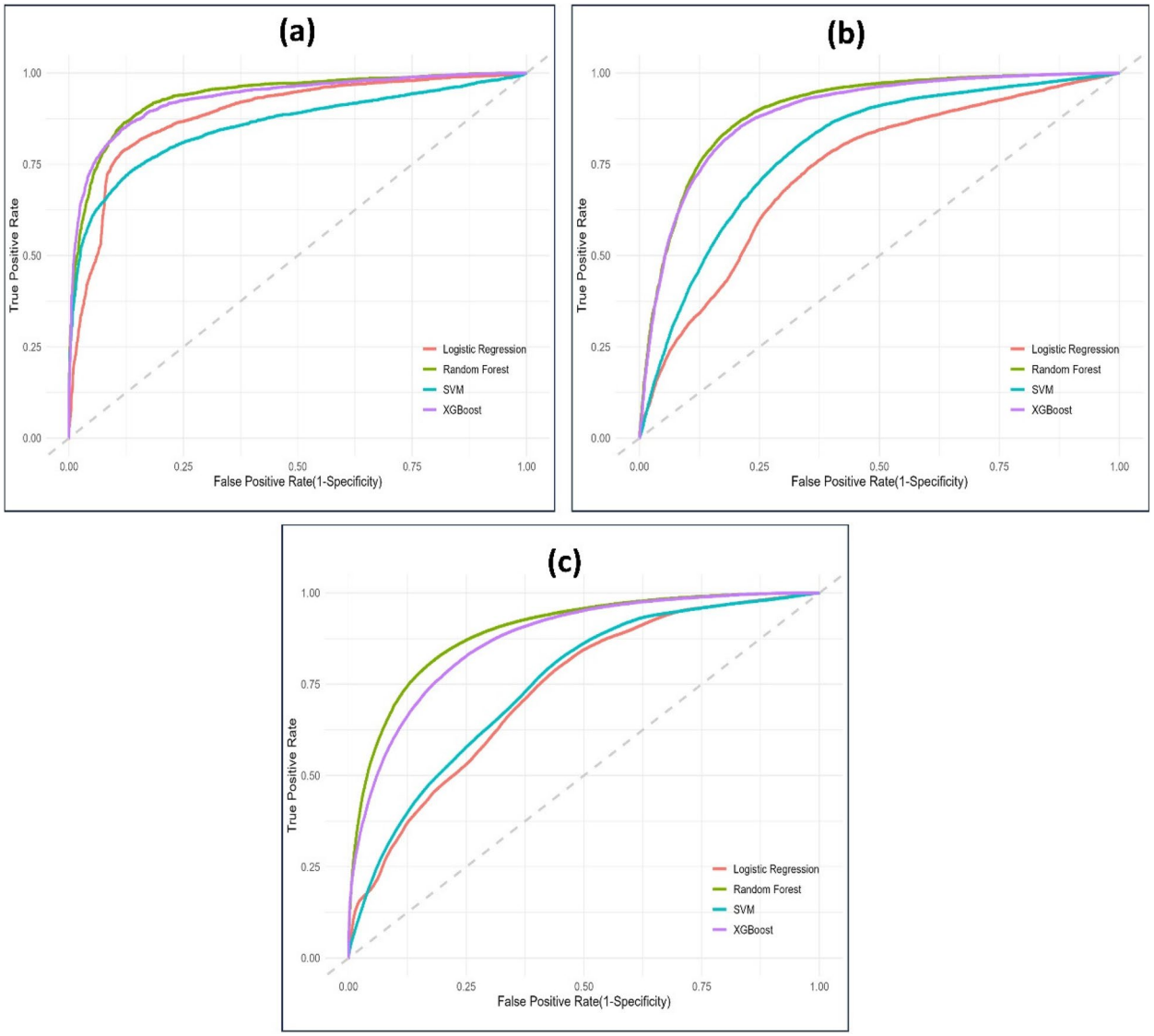

Fig. 8 AUC curve of prediction rates of four models: **a** forest, **b** crop, and **c** other vegetation

RF model as the most appropriate choice out of the four models for predicting forest fires in Pakistan. Figure 10 shows the variable importance factors of forests, crops, and other vegetation.

Support vector machine

In this section of the study, the accuracy and generalizability of the SVMs model are used to predict forest, crop, and other vegetation fires. The forest fire model predicted accuracy of 78.7% and 83.6 AUC; in crop fire, the model had 74.5% accuracy and 80.7% AUC; and in other vegetation fire, the model exhibited 68.7% accuracy and 74.8% AUC. The ROC curve of prediction rates of the SVM model is shown in Fig. 8. Overall, the SVM models provided significant predictive capability for different types of vegetation fire. These findings highlight the SVM model's robust predictive performance across various vegetation types, underscoring its potential utility in designing targeted and effective fire prevention and management strategies. Further investigation into feature influence using advanced interpretative methods could enhance the model's applicability and provide deeper insights into critical factors driving vegetation fire risks.

eXtreme Gradient Boosting

This study showed how well the XGBoost models we built can predict different types of vegetation fire. The accuracy and performance of the XGBoost model were

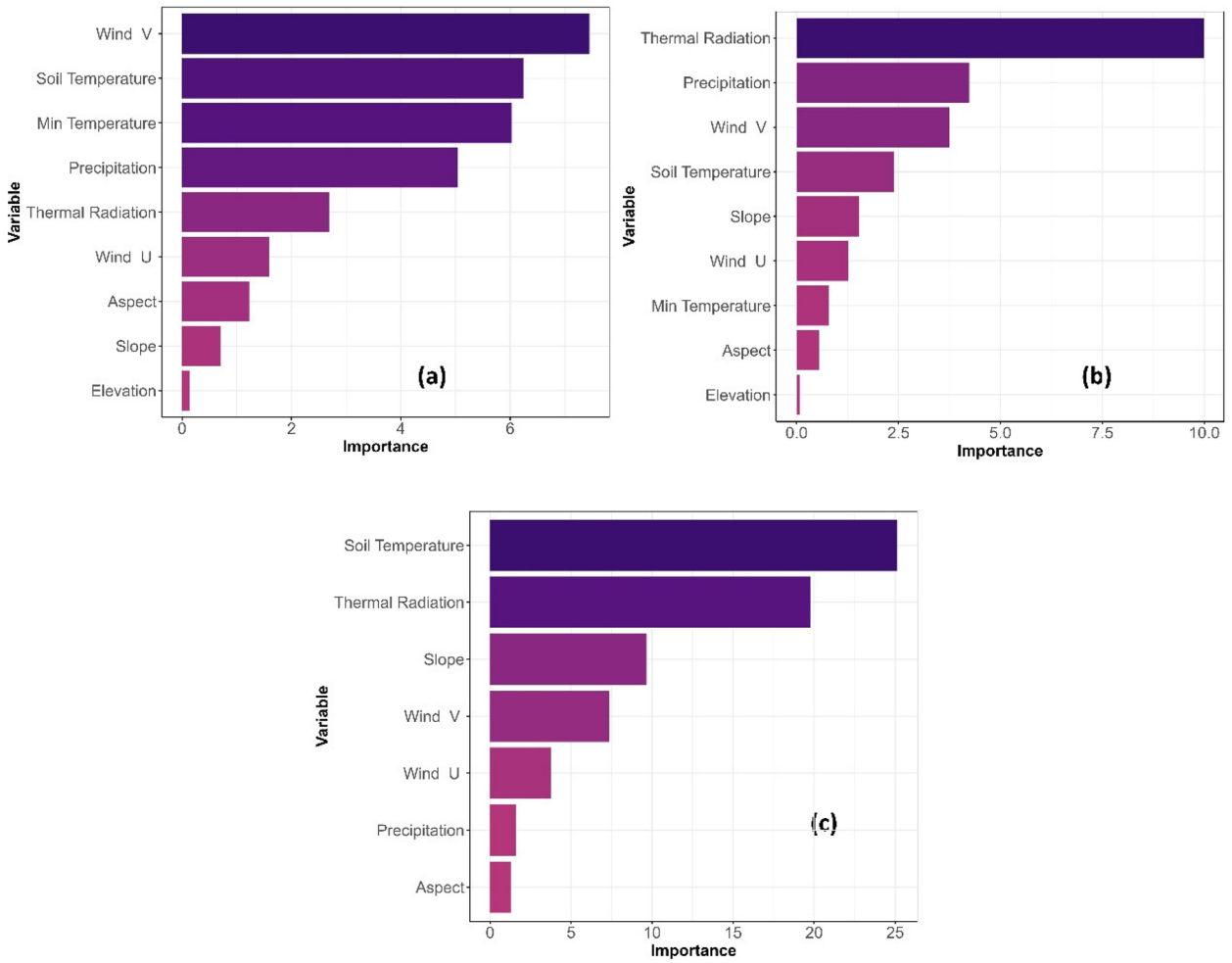

Fig. 9 The importance of initiating factor indicators in the LR model: **a** forest, **b** crop, and **c** other vegetation

constantly evaluated to ensure that, they were suitable for diverse prediction conditions. The XGBoost model's accuracy and AUC scores are shown in Table 5. The forest fire model showed an accuracy of 86.0% with an AUC of 92.6%; for crop fires, the model achieved an accuracy of 83.9% with an AUC of 90.0%; and for other vegetation fires, it recorded an accuracy of 79.4% with an AUC of 87.6%. Figure 8 displays the AUC curves, illustrating the predictive performance of the XGBoost model across different types of vegetation. The results show that XGBoost models are second best for vegetation fire prediction in Pakistan using this set of variables for fires from 2001 to 2022. The model could be used to improve management and mitigation approaches for vegetation fire.

Vegetation fire risk assessment

By assessing the precision of the four models, we selected the RF model, which had the best accuracy, to determine the likelihood of vegetation fire happening in the whole country. We used ArcGIS 10.8 to create a cartographic representation of Pakistan's potential danger of vegetation fires. The values indicated in the legends in Fig. 11 represent the expected probability of vegetation fires in Pakistan. For example, a vegetation fire has a probability of 1, showing the highest possibility of occurrence. The number of red regions ranges from 0.8 to 1, showing a high danger where vegetation fires are very likely to happen. Figure 11 illustrates that the prevalence of vegetation fires in Pakistan mainly occurs in specific regions. These regions include the northwest, covering various districts of Khyber Pakhtunkhwa (KP), such as Malakand Division, Bannu, Parachinar, Tank, and Kohat. Additionally, the northeast region, comprising Azad Jammu and Kashmir (AJK) and Gilgit-Baltistan (GB), demonstrates a high incidence of vegetation fires. The southeast region, which includes Punjab and Sindh, along with Islamabad, Dera Ghazi Khan, Multan, Karachi, Hyderabad, and Mirpur Khas, also faces many vegetation fires. Lastly, the

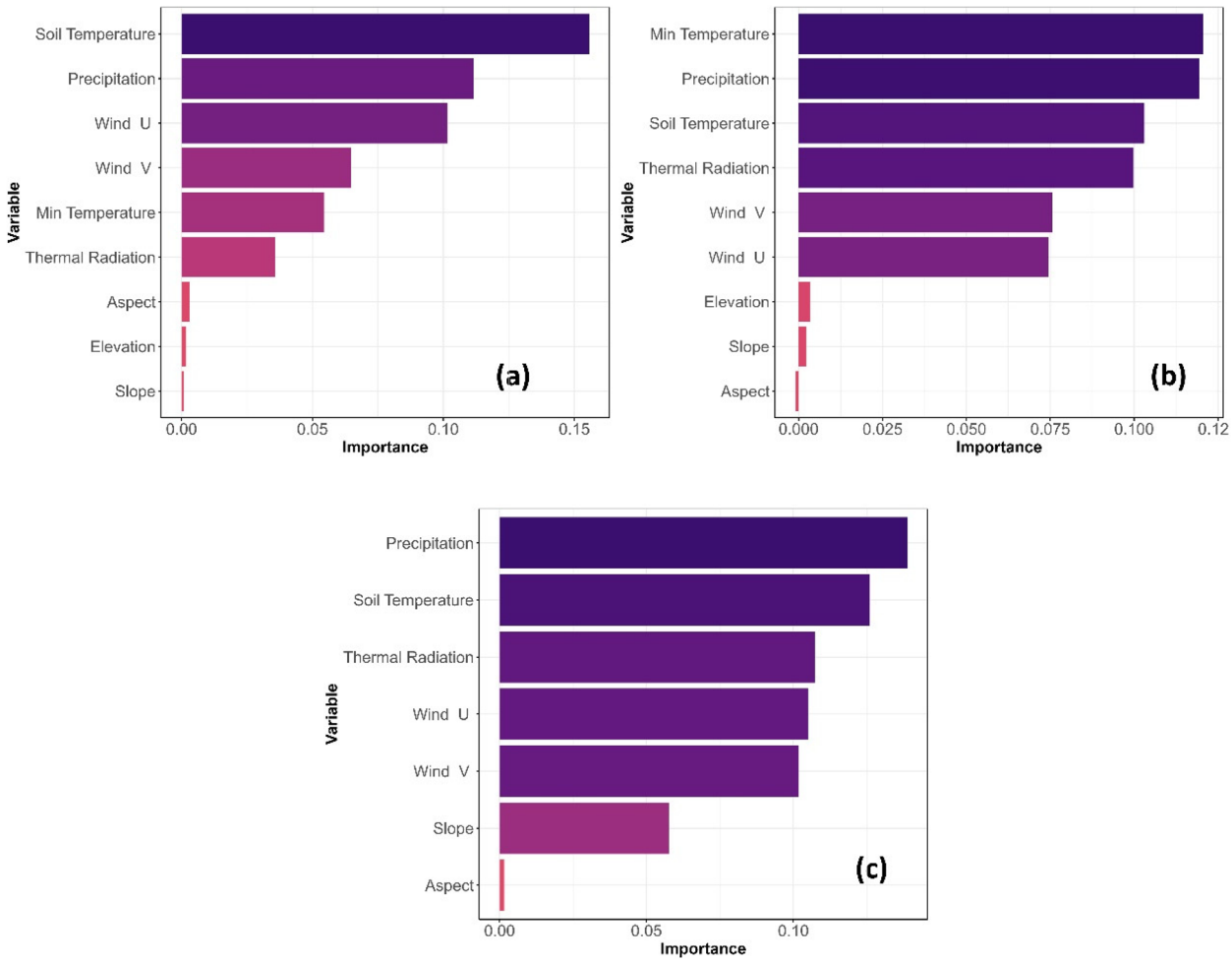

Fig. 10 Importance of initiating factor indicators in the RF model: **a** forest, **b** crop, and **c** other vegetation

southwestern region, specifically Baluchistan, including Quetta, is prone to vegetation fires. Generally, the likelihood of vegetation fires is more significant in western areas of Pakistan than in the eastern regions. Additionally, the possibility of vegetation fires is higher in southern Pakistan than in the northern areas.

Discussion

In our study, we examined the various factors influencing vegetation fire risk. Our analysis incorporates a detailed evaluation of meteorological variables such as average annual daily high temperature, annual average relative humidity, total annual precipitation, and average annual wind speed. We also considered broader climatic factors, topological features, and different vegetation types as significant determinants of fire risk (Li et al. 2022). In this study, we selected nine variables for analyzing forest and crop fires and seven for other types of vegetation based on their demonstrated association with fire occurrences and their statistical significance in preliminary models. The associated variables identified include soil temperature, minimum temperature, northward and eastward components of the 10 m wind, precipitation, surface net thermal radiation, slope, aspect, and elevation. For other vegetation types, elevation and minimum temperature were less significant in predicting fire ignition. These factors were crucial in training our machine learning models to predict vegetation fires effectively and were instrumental in the development of risk maps using the RF model. Fire factors and conditions vary by area (Abid 2021). This is primarily due to country-specific environmental and socioeconomic variables. It is also related to the investigated region and the environment of every country (Oliveira et al. 2012; Sun et al. 2023). According to Chang et al. (2013), land use intensity, precipitation, and vegetation type are the key variables affecting Durango State, Mexico fires. Fuel moisture, vegetation

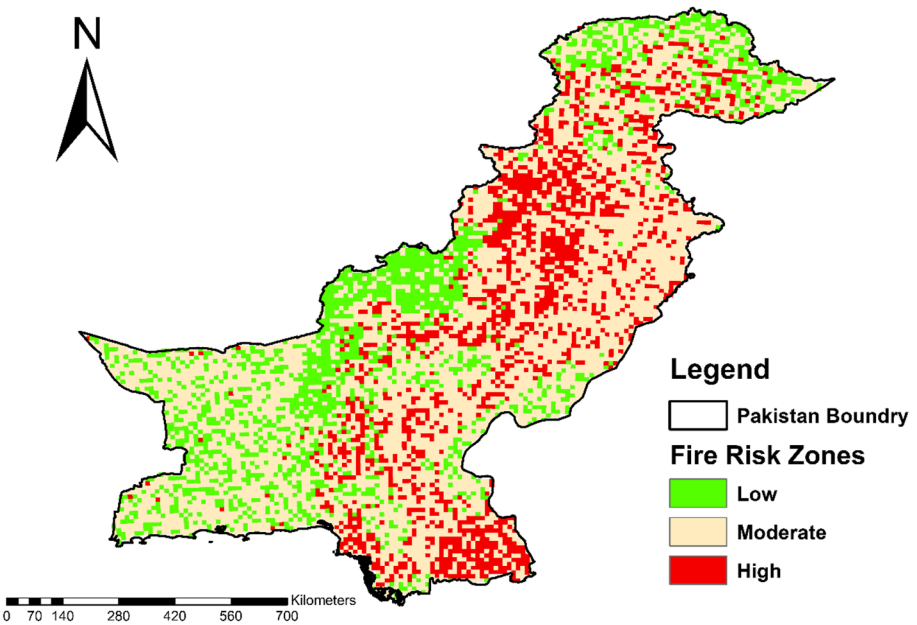

Fig. 11 Vegetation fire risk assessment map

type, and human activity in northeast China greatly influence man-made fires. In eastern Kentucky, height and slope are the high-influence variables that affect vegetation fires. The most critical factors affecting vegetation fires in Swaziland are elevation, mean annual rainfall, mean annual temperature, and land cover (Dlamini 2010).

This study tested four machine learning methods to predict fire occurrence and show each model's strengths and applications on vegetation fire in Pakistan. The classic LR models provide a good prediction with an 81.6% predicted accuracy for forest fires, 69.2 for crop fires, and 66.5 for other vegetation fires. The performance of LR in many predictive modeling situations was robust, achieving high accuracy and reliability, although it did not always outperform the more complex models. However, it may not effectively capture complex non-linear interactions compared to more advanced algorithms (Khalaji et al. 2022). The literature often acknowledges that advanced machine learning models, such as RF, SVM, and XGBoost, perform better than LR, particularly in complex prediction tasks like mapping and vulnerability of vegetation fire risk assessment. This study demonstrated that RF exhibited outstanding results, achieving an 87.5% accuracy for forest fires, 84.0% in crop fires, and 81.7% in other vegetation. This corresponds to previous studies in environmental modeling, which emphasize the tendency toward RF and similar ensemble techniques. These methods were selected for their ability to quickly analyze data with many variables and to capture a wide variety of interactions (Shmuel and Heifetz 2022). An integrated approach may show outstanding results in the context of the SVM model, which demonstrated functional flexibility in previous research (Rodrigues and De la Riva 2014).

XGBoost has a powerful technique for prediction analysis, showing outstanding results in several fields, such as vegetation fire prediction. The XGBoost models show a remarkable degree of accuracy and ROC AUC values, which aligns with the present literature that indicates their value in accurate overall classification (Mohajane et al. 2021; Mehmood et al. 2024a, b, c). The model's ability to manage insufficient data and its optimal utilization of gradient boosting make it an essential tool for assessing environmental risks. Research has identified the intellectual capacities required to develop effective approaches to managing and mitigating risks in different vegetation environments (Tehrany et al. 2019). Comparing different models reveals little complexity, understanding, and variation in predictive capabilities. While models like RF and XGBoost could show better predictive accuracy, LR provides a more understandable framework, which is essential to policy formulation and strategic decision-making (Peng et al. 2021).

Furthermore, the SVM model uses a unique kernel method, which provides a highly flexible solution for non-linear problems with the environment. Therefore, it can be highly beneficial in analyzing datasets with complex feature associations (Lopez-Martin et al. 2019). Our

research methodology also included the key foundational adjustment of hyperparameter modification, which significantly impacts model performance. Modifying the parameters of models like RF and XGBoost (e.g., number of trees, depth of trees, or minimum size for a node) significantly affects their accuracy and ability to generalize. These methods are supported by research highlighting the importance of model optimization (Jiang et al. 2022). According to the Mann–Kendall mutation test, vegetation fires showed an unstable increasing trend in Pakistan. This test is flexible and responsive, which is necessary to consistently show the temporal fluctuations and various kinds of change (Vadrevu et al. 2019; Mehmood et al. 2024a, b, c).

The result suggests that future studies should use a broader range of data sources, including remote sensing data and socio-economic aspects, to enhance the accuracy and applicability of prediction models. Moreover, it is essential to take advantage of the advancements in hybrid models, which enable the combination of various techniques to enhance prediction accuracy while maintaining accessibility. Therefore, the prediction and analysis of vegetation fires continue to be a significant area of research with substantial potential to avoid disasters and protect natural resources. The consistent and dependable performance of advanced machine learning models in the field of vegetation fire provides numerous possibilities for future research efforts and practical implementations. Both scholars and professionals could actively contribute to advancing more efficient methods in mitigating fire hazards and minimizing the impact of vegetation fire. This may be achieved through continuous improvement of these models and their integration with comprehensive data sources.

Conclusion

This research applied feature selection techniques to identify the most important variables associated with vegetation fire incidents in Pakistan. The key factors influencing the occurrence of vegetation fires were identified as meteorological and topographical, including soil temperature, minimum temperature, northward and eastward components of the 10 m wind, precipitation, surface net thermal radiation, and slope. We constructed four different types of prediction models for every kind of vegetation fire (forest, crop, and other vegetation) using the following ML algorithms: logistic regression (LR), random forest (RF), support vector machine (SVM), and eXtreme Gradient Boosting (XGBoost). The RF model demonstrated the best overall predictive capability, with an accuracy rate of 87.5% in forest fires, 84% in crop fires, and 83.1% in other vegetation fires. Hence, given its balance of computational speed and minimal variable requirements, the RF model is the most efficient choice for vegetation fire prediction in Pakistan. Using these probabilities, we created a map illustrating the annual likelihood of vegetation fires occurring throughout Pakistan during the study period. The study has significant implications for wildfire management policy and strategy. These algorithms accurately predict fires, helping governments and firefighting agencies allocate resources and devise preventative methods. The study's long-term trend analysis shows an unpredictable increase in vegetation fires in Pakistan, underscoring the importance of adaptable and flexible models to reflect temporal fluctuations and changes in fire dynamics. Further research should include remote sensing and socio-economic elements to enhance predictive model accuracy and applicability. Hybrid models, which integrate multiple machine learning methods, can improve prediction accuracy while remaining user-friendly.

Acknowledgements
We are grateful to Precision Forestry Key Laboratory of Beijing, Beijing Forestry University for providing assistance and platforms for this research.

Institutional review board statement
Not applicable.

Informed consent statement
Not applicable.

Authors' contributions
Fahad Shahzad: conceptualization, methodology, software, formal analysis, visualization, data curation, writing—original draft, investigation, validation, writing—review and editing. Kaleem Mehmood: visualization, writing—review and editing, Zhongke Feng: writing—review and editing, Supervision. Khadim Hussain: writing—review and editing. Ijlal Haidar: writing—review and editing. Shoaib Ahmad Anees: formal analysis, investigation, writing—review and editing. Sultan Muhammad: writing—review and editing. Jamshid Ali: writing—review and editing. Muhammad Adnan: writing—review and editing. Zhichao Wang: writing—review and editing, Supervision. All authors have read and agreed to the published version of the manuscript.

Funding
This study was supported by 5·5 Engineering Research & Innovation Team Project of Beijing Forestry University (BLRC2023A03) and the Natural Science Foundation of Beijing (8232038, 8234065) and the Key Research and Development Projects of Ningxia Hui Autonomous Region (2023BEG02050).

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

Declarations

Ethics approval and consent to participate
This research did not involve human or animal subjects; therefore, formal ethical approval was not required. The study strictly adheres to general ethical principles, and the authors are committed to upholding the highest standards of ethical research conduct. Any potential conflicts of interest that could have influenced the ethical conduct of this research have been declared. Informed consent was obtained from all participants involved in this study. Participants were provided with detailed information about the research objectives, procedures, potential risks, and benefits before agreeing to participate. They were assured that their participation was voluntary, and

they had the right to withdraw from the study at any time without facing any consequences.
All participants were informed about the confidentiality measures in place to protect their identity and personal information. Data collected during the study will be used solely for research purposes and will be securely stored. This study was conducted in accordance with ethical standards and guidelines, and participants were encouraged to ask questions and seek clarification at any stage of the research process. If you have any further questions or concerns regarding the consent process, please contact fahadshahzadbjfu@gmail.com.

Competing interests
The authors declare no competing interests.

Author details
[1]Precision Forestry Key Laboratory of Beijing, Beijing Forestry University, Beijing 100083, China. [2]Key Laboratory of Genetics and Germplasm Innovation of Tropical Special Forest Trees and Ornamental Plants, Ministry of Education, College of Tropical Crops Hainan University, Hainan University, Haikou 570228, China. [3]State Forestry and Grassland Administration Key Laboratory of Forest Resources and Environmental Management, Beijing Forestry University, Beijing 100083, P. R. China. [4]Key Laboratory for Silviculture and Conservation of Ministry of Education, Beijing Forestry University, Beijing 100083, P. R. China. [5]Institute of Forest Science, University of Swat, Main Campus Charbagh, Swat 19120, Pakistan. [6]Department of Forestry, The University of Agriculture Dera Ismail Khan, Dera Ismail Khan 29050, Pakistan. [7]State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China.

Received: 14 February 2024 Accepted: 26 May 2024
Published online: 25 June 2024

References

Abid, F. 2021. A survey of machine learning algorithms based forest fires prediction and detection systems. *Fire Technology* 57 (2): 559–590.
Akram, M., U. Hayat, J. Shi, and S.A. Anees. 2022. Association of the female flight ability of Asian spongy moths (Lymantria dispar asiatica) with locality, age and mating: A case study from China. *Forests* 13 (8): 1158.
Albar, I., et al. 2018. Spatio-temporal analysis of land and forest fires in Indonesia using MODIS active fire dataset. *Land-Atmospheric Research Applications in South and Southeast Asia* 105–127.
Andreevich, U.V., S.S.O. Reza, T.I. Stepanovich, A. Amirhossein, Z. Meng, S.A. Anees, and C.V. Petrovich. 2020. Are there differences in the response of natural stand and plantation biomass to changes in temperature and precipitation? A case for two-needled pines in Eurasia. *Journal of Resources and Ecology* 11 (4): 331.
Anees, S.A., et al. 2022a. Estimation of fractional vegetation cover dynamics and its drivers based on multi-sensor data in Dera Ismail Khan, Pakistan. *Journal of King Saud University-Science* 34 (6): 102217.
Anees, S.A., X. Zhang, M. Shakeel, M.A. Al-Kahtani, K.A. Khan, M. Akram, and H.A. Ghramh. 2022b. Estimation of fractional vegetation cover dynamics based on satellite remote sensing in Pakistan: A comprehensive study on the FVC and its drivers. *Journal of King Saud University-Science* 34 (3): 101848.
Anees, S.A., X. Yang, and K. Mehmood. 2024. The stoichiometric characteristics and the relationship with hydraulic and morphological traits of the Faxon fir in the subalpine coniferous forest of Southwest China. *Ecological Indicators* 159: 111636.
Arnold, J.D., S.C. Brewer, and P.E. Dennison. 2014. Modeling climate-fire connections within the great basin and upper colorado river basin, western united states. *Fire Ecology* 10: 64–75.
Aslam, M.S., P. Huanxue, S. Sohail, M.T. Majeed, S.U. Rahman and S.A. Anees. 2022. Assessment of major food crops production-based environmental efficiency in China, India, and Pakistan. *Environmental Science and Pollution Research* 1–10.
Attri, V., R. Dhiman, and S. Sarvade. 2020. A review on status, implications and recent trends of forest fire management. *Archives of Agriculture and Environmental Science* 5 (4): 592–602.
Badshah, M.T., et al. 2024. The role of random forest and Markov chain models in understanding metropolitan urban growth trajectory. *Frontiers in Forests and Global Change* 7: 1345047.
Balboa, A., et al. 2024. Logistic regression vs machine learning to predict evacuation decisions in fire alarm situations. *Safety Science* 174: 106485.
Barreto, J.S., and D. Armenteras. 2020. Open data and machine learning to model the occurrence of fire in the ecoregion of "llanos colombo–venezolanos." *Remote Sensing* 12 (23): 3921.
Begum, B.A., et al. 2011. Long–range transport of soil dust and smoke pollution in the South Asian region. *Atmospheric Pollution Research* 2 (2): 151–157.
Bhujel, K.B., R. Maskey-Byanju, and A.P. Gautam. 2017. Wildfire dynamics in Nepal from 2000–2016. *Nepal Journal of Environmental Science* 5: 1–8.
Borggaard, O.K., A. Gafur, and L. Petersen. 2003. Sustainability appraisal of shifting cultivation in the Chittagong Hill Tracts of Bangladesh. *AMBIO: A Journal of the Human Environment* 32 (2): 118–123.
Bot, K., and J.G. Borges. 2022. A systematic review of applications of machine learning techniques for wildfire management decision support. *Inventions* 7 (1): 15.
Botequim, B., et al. 2017. Modeling post-fire mortality in pure and mixed forest stands in Portugal—a forest planning-oriented model. *Sustainability* 9 (3): 390.
Boubeta, M., et al. 2015. Prediction of forest fires occurrences with area-level Poisson mixed models. *Journal of Environmental Management* 154: 151–158.
Breiman, L. 2001. Random forests. *Machine learning* 45: 5–32.
Bui, D.T., 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.
Bui, D.T., et al. 2019. A novel hybrid approach based on a swarm intelligence optimized extreme learning machine for flash flood susceptibility mapping. *CATENA* 179: 184–196.
Cabral, A.I.R., et al. 2018. Burned area estimations derived from Landsat ETM+ and OLI data: Comparing genetic programming with maximum likelihood and classification and regression trees. *ISPRS Journal of Photogrammetry and Remote Sensing* 142: 94–105.
Carter, J.V., et al. 2016. ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves. *Surgery* 159 (6): 1638–1645.
Chang, Y., et al. 2013. 'Predicting fire occurrence patterns with logistic regression in Heilongjiang Province, China. *Landscape Ecology* 28: 1989–2004.
Chen, T., and C. Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.
Chuvieco, E., L. Giglio, and C. Justice. 2008. Global characterization of fire activity: Toward defining fire regimes from Earth observation data. *Global Change Biology* 14 (7): 1488–1502.
Chuvieco, E., et al. 2018. Generation and analysis of a new global burned area product based on MODIS 250 m reflectance bands and thermal anomalies. *Earth System Science Data* 10 (4): 2015–2031.
Chuvieco, E., et al. 2019. Historical background and current developments for mapping burned area from satellite Earth observation. *Remote Sensing of Environment* 225: 45–64.
Dlamini, W.M. 2010. A Bayesian belief network analysis of factors influencing wildfire occurrence in Swaziland. *Environmental Modelling & Software* 25 (2): 199–208.
Duff, T.J., and K.G. Tolhurst. 2015. Operational wildfire suppression modelling: A review evaluating development, state of the art and future directions. *International Journal of Wildland Fire* 24 (6): 735–748.
El Emam, K., W. Melo, and J.C. Machado. 2001. The prediction of faulty classes using object-oriented design metrics. *Journal of Systems and Software* 56 (1): 63–75.
Eslami, R., et al. 2021. GIS-based forest fire susceptibility assessment by random forest, artificial neural network and logistic regression methods. *Journal of Tropical Forest Science* 33 (2): 173–184.
Feng, X., et al. 2016. Evolution of spatial pattern of county regional economy in Yangtze River economic belt. *Economic Geography* 36: 18–25.
Garcia, C.V., et al. 1995. A logit model for predicting the daily occurrence of human caused forest-fires. *International Journal of Wildland Fire* 5 (2): 101–111.

Giglio, L., et al. 2018. The Collection 6 MODIS burned area mapping algorithm and product. *Remote Sensing of Environment* 217: 72–85.
Giglio, L., I. Csiszar, and C.O. Justice. 2006. Global distribution and seasonality of active fires as observed with the Terra and Aqua Moderate Resolution Imaging Spectroradiometer (MODIS) sensors. *Journal of Geophysical Research: Biogeosciences*, 111 (G2): 1–12.
Gitas, I., et al. 2012. Advances in remote sensing of post-fire vegetation recovery monitoring—a review. *Remote Sensing of Biomass-Principles and Applications* 1: 334.
Haddouchi, M., and A. Berrado. 2019. A survey of methods and tools used for interpreting random forest. In *2019 1st International Conference on Smart Systems and Data Science (ICSSD)*, 1-6. IEEE.
Jain, P., S.C. Coogan, S.G. Subramanian, M. Crowley, S. Taylor, and M.D. Flannigan. 2020. A review of machine learning applications in wildfire science and management. *Environmental Reviews* 28 (4): 478–505.
Jiang, L., et al. 2022. Prediction of coronary heart disease in gout patients using machine learning models. *Mathematical Biosciences and Engineering* 20 (3): 4574–4591.
Jodhani, K.H., et al. 2024. Assessment of forest fire severity and land surface temperature using Google Earth Engine: A case study of Gujarat State, India. *Fire Ecology* 20 (1): 23.
Katagis, T., and I.Z. Gitas. 2022. Assessing the accuracy of MODIS MCD64A1 C6 and FireCCI51 burned area products in Mediterranean ecosystems. *Remote Sensing* 14 (3): 602.
Kattel, D.B., et al. 2019. Seasonal near-surface air temperature dependence on elevation and geographical coordinates for Pakistan. *Theoretical and Applied Climatology* 138: 1591–1613.
Khalaji, A., et al. 2022. Machine learning algorithms for predicting mortality after coronary artery bypass grafting. *Frontiers in Cardiovascular Medicine* 9: 977747.
Khan, W.R., M. Nazre, S. Akram, S.A. Anees, K. Mehmood, F.H. Ibrahim, ..., and X. Zhu. 2024. Assessing the productivity of the Matang Mangrove Forest reserve: review of one of the best-managed mangrove forests. *Forests*, 15 (5): 747.
Kleinman, P.J.A., D. Pimentel, and R.B. Bryant. 1995. The ecological sustainability of slash-and-burn agriculture. *Agriculture, Ecosystems & Environment* 52 (2–3): 235–249.
Krishna, P.H., and C.S. Reddy. 2012. Assessment of increasing threat of forest fires in Rajasthan, India using multi-temporal remote sensing data (2005–2010). *Current Science* 1288–97.
Landis, J.R., and G.G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 159–174.
Li, W., et al. 2022. Predictive model of spatial scale of forest fire driving factors: A case study of Yunnan Province, China. *Scientific Reports* 12 (1): 19029.
Liang, D., et al. 2015. Evaluation of the consistency of MODIS Land Cover Product (MCD12Q1) based on Chinese 30 m GlobeLand30 datasets: A case study in Anhui Province, China. *ISPRS International Journal of Geo-Information* 4 (4): 2519–2541.
Lopez-Martin, M., et al. 2019. Shallow neural network with kernel approximation for prediction problems in highly demanding data networks. *Expert Systems with Applications* 124: 196–208.
Luo, M., et al. 2024. Improving Forest Above-Ground Biomass Estimation by Integrating Individual Machine Learning Models. *Forests* 15 (6): 975.
Manaswini, G., and C. Sudhakar Reddy. 2015. 'Geospatial monitoring and prioritization of forest fire incidences in Andhra Pradesh, India. *Environmental Monitoring and Assessment* 187: 1–12.
Marques, S., et al. 2012. Assessing wildfire occurrence probability in Pinus pinaster Ait. stands in Portugal. *Forest Systems* 21: 111–120.
Martell, D.L., 2007. Forest fire management: current practices and new challenges for operational researchers. In *Handbook of operations research in natural resources*, eds. A Weintraub, C Romero, T Bjørndal, R Epstein, pp. 489–509. New York: Springer Science+ Business Media.
Martínez, J., C. Vega-Garcia, and E. Chuvieco. 2009. Human-caused wildfire risk rating for prevention planning in Spain. *Journal of Environmental Management* 90 (2): 1241–1252.
Mehmood, K., S.A. Anees, M. Luo, M. Akram, M. Zubair, K.A. Khan, and W.R. Khan. 2024a. Assessing chilgoza pine (Pinus gerardiana) forest fire severity: remote sensing analysis, correlations, and predictive modeling for enhanced management strategies. *Trees, Forests and People* 100521.
Mehmood, K., S.A. Anees, A. Rehman, A. Tariq, Q. Liu, et al. 2024b. Assessing forest cover changes and fragmentation in the Himalayan temperate region: implications for forest conservation and management. *Journal of Forestry Research* 35 (1): 82. https://doi.org/10.1007/s11676-024-01734-6.
Mehmood, K., S.A. Anees, A. Rehman, A. Tariq, M. Zubair. et al. 2024c. Exploring spatiotemporal dynamics of NDVI and climate-driven responses in ecosystems: Insights for sustainable management and climate resilience. *Ecological Informatics* 102532.
Mehmood, K., et al. 2024d. Analyzing vegetation health dynamics across seasons and regions through NDVI and climatic variables. *Scientific Reports* 14 (1): 11775.
Mohajane, M., et al. 2021. Application of remote sensing and machine learning algorithms for forest fire mapping in a Mediterranean area. *Ecological Indicators* 129: 107869.
Muhammad, S., K. Mehmood, S.A. Anees, M. Tayyab, F. Rabbi, K. Hussain, H.U. Rahman, M. Hayat, and U. Khan. 2023. Assessment of regeneration response of Silver Fir (Abies pindrow) to slope, aspect, and altitude in Miandam area in District Swat, Khyber-Pakhtunkhwa, Pakistan. *International Journal of Forest Sciences.* 4: 246–252.
Muschelli, J., III. 2020. ROC and AUC with a binary predictor: A potentially misleading metric. *Journal of Classification* 37 (3): 696–708.
Naderpour, M., et al. 2019. Forest fire induced Natech risk assessment: A survey of geospatial technologies. *Reliability Engineering & System Safety* 191: 106558.
Nami, M.H., et al. 2018. Spatial prediction of wildfire probability in the Hyrcanian ecoregion using evidential belief function model and GIS. *International Journal of Environmental Science and Technology* 15: 373–384.
Nunes, A.N., L. Lourenço, and A.C.C. Meira. 2016. Exploring spatial patterns and drivers of forest fires in Portugal (1980–2014). *Science of the Total Environment* 573: 1190–1202.
Oliveira, S.L.J., J.M.C. Pereira, and J.M.B. Carreiras. 2011. Fire frequency analysis in Portugal (1975–2005), using Landsat-based burnt area maps. *International Journal of Wildland Fire* 21 (1): 48–60.
Oliveira, S., et al. 2012. Modeling spatial patterns of fire occurrence in Mediterranean Europe using multiple regression and random forest. *Forest Ecology and Management* 275: 117–129.
Pan, S.A., S.A. Anees, X. Li, X. Yang, X. Duan, and Z. Li. 2023. Spatial and temporal patterns of non-structural carbohydrates in Faxon fir (Abies fargesii var. faxoniana), subalpine mountains of Southwest China. *Forests* 14 (7): 1438.
Pang, Y., et al. 2022. Forest fire occurrence prediction in China based on machine learning methods. *Remote Sensing* 14 (21): 5546.
Peng, C.-Y.J., K.L. Lee, and G.M. Ingersoll. 2002. An introduction to logistic regression analysis and reporting. *The Journal of Educational Research* 96 (1): 3–14.
Peng, J., et al. 2021. An explainable artificial intelligence framework for the deterioration risk prediction of hepatitis patients. *Journal of Medical Systems* 45: 1–9.
Piraei, R., S.H. Afzali, and M. Niazkar. 2023. Assessment of XGBoost to estimate total sediment loads in rivers. *Water Resources Management* 37 (13): 5289–5306.
Probst, P., M.N. Wright, and A. Boulesteix. 2019. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9 (3): e1301.
Qasim, M., S. Khlaid, and D.F. Shams. 2014. Spatiotemporal variations and trends in minimum and maximum temperatures of Pakistan. *J Appl Environ Biol Sci* 4 (8S): 85–93.
Rafaqat, W., M. Iqbal, R. Kanwal, and S. Weiguo. 2022a. Evaluation of wildfire occurrences in Pakistan with global gridded soil properties derived from remotely sensed data. *Remote Sensing* 14 (21): 5503.
Rafaqat, W., M. Iqbal, R. Kanwal, and W. Song. 2022b. Study of driving factors using machine learning to determine the effect of topography, climate, and fuel on wildfire in Pakistan. *Remote Sensing* 14 (8): 1918.
Reddy, C.S., et al. 2017. Nationwide assessment of forest burnt area in India using Resourcesat-2 AWiFS data. *Current Science* 1521–1532.
Reddy, C.S., and N. Sarika. 2022. Monitoring trends in global vegetation fire hot spots using MODIS data. *Spatial Information Research* 30 (5): 617–632.
Rodrigues, M., and J. De la Riva. 2014. An insight into machine-learning algorithms to model human-caused wildfire occurrence. *Environmental Modelling & Software* 57: 192–201.
Rossi, F., and N. Villa. 2006. Support vector machine for functional data classification. *Neurocomputing* 69 (7–9): 730–742.

Rubí, J.N.S., P.H.P. de Carvalho, and P.R.L. Gondim. 2023. Application of machine learning models in the behavioral study of forest fires in the Brazilian Federal District region. *Engineering Applications of Artificial Intelligence* 118: 105649.
Saranya, K.R.L., et al. 2014. Decadal time-scale monitoring of forest fires in Similipal Biosphere Reserve, India using remote sensing and GIS. *Environmental Monitoring and Assessment* 186: 3283–3296.
Sarkar, M.S., et al. 2024. Ensembling machine learning models to identify forest fire-susceptible zones in Northeast India. *Ecological Informatics* 81: 102598.
Schultz, M.G., et al. 2008. Global wildland fire emissions from 1960 to 2000. *Global Biogeochemical Cycles* 22(2): 1–17.
Segal, M., and Y. Xiao. 2011. Multivariate random forests. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1 (1): 80–87.
Shahdeo, Ananya, et al. 2020. Wildfire prediction and detection using random forest and different color models. *International Research Journal of Engineering and Technology* 7 (06): 7326–7332.
Shao, Y., et al. 2023. An ensemble model for forest fire occurrence mapping in China. *Forests* 14 (4): 704.
Shmuel, A., and E. Heifetz. 2022. Global wildfire susceptibility mapping based on machine learning models. *Forests* 13 (7): 1050.
Shobairi, S.O.R., H. Lin, V.A. Usoltsev, A.A. Osmirko, I.S. Tsepordey, Z. Ye, and S.A. Anees. 2022. A comparative pattern for Populus spp. and Betula spp. stand biomass in Eurasian climate gradients. *Croatian Journal of Forest Engineering: Journal for Theory and Application of Forestry Engineering* 43 (2): 457–467.
Sohail, M., S. Muhammad, K. Mehmood, S.A. Anees, F. Rabbi, M. Tayyab, K. Hussain, M. Hayat, and U. Khan. 2023. Tourism, threat, and opportunities for the forest resources: A case study of Gabin Jabaa, District Swat, Khyber-Pakhtunkhwa, Pakistan. *International Journal of Forest Sciences* 3 (3): 194–203.
Su, Z., 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.
Sulla-Menashe, D., and M.A. Friedl. 2018. User guide to collection 6 MODIS land cover (MCD12Q1 and MCD12C1) product. *Usgs: Reston, Va, Usa* 1: 18.
Sun, D., et al. 2021. Assessment of landslide susceptibility mapping based on Bayesian hyperparameter optimization: A comparison between logistic regression and random forest. *Engineering Geology* 281: 105972.
Sun, L., et al. 2023. The development of a set of novel low cost and data processing-free measuring instruments for tree diameter at breast height and tree position. *Forests* 14 (5): 891.
Tehrany, M.S., et al. 2019. A novel ensemble modeling approach for the spatial prediction of tropical forest fire susceptibility using LogitBoost machine learning classifier and multi-source geospatial data. *Theoretical and Applied Climatology* 137: 637–653.
Thomas, D., et al. 2017. The costs and losses of wildfires. *NIST Special Publication* 1215 (11): 1–72.
Tien Bui, D., et al. 2016. Tropical forest fire susceptibility mapping at the Cat Ba National Park Area, Hai Phong City, Vietnam, using GIS-based kernel logistic regression. *Remote Sensing* 8 (4): 347.
Usoltsev, V.A., B. Chen, S.O.R. Shobairi, I.S. Tsepordey, V.P. Chasovskikh, and S.A. Anees. 2020. Patterns for Populus spp. stand biomass in gradients of winter temperature and precipitation of Eurasia. *Forests* 11 (9): 906.
Usoltsev, V.A., H. Lin, S.O.R. Shobairi, I.S. Tsepordey, Z. Ye, and S.A. Anees. 2022. The principle of space-for-time substitution in predicting Betula spp. Biomass change related to climate shifts. *Applied Ecology and Environmental Research* 20 (4): 3683–3698.
Vadrevu, K.P., K.V.S. Badarinath, and E. Anuradha. 2008. Spatial patterns in vegetation fires in the Indian region. *Environmental Monitoring and Assessment* 147: 1–13.
Vadrevu, K.P., et al. 2019. Trends in vegetation fires in south and southeast Asian countries. *Scientific Reports* 9 (1): 7422.
van Lierop, P., et al. 2015. Global forest area disturbance from fire, insect pests, diseases and severe weather events. *Forest Ecology and Management* 352: 78–88.
Watson, P.F., and A. Petrie. 2010. Method agreement analysis: A review of correct methodology. *Theriogenology* 73 (9): 1167–1179.
Xie, L., et al. 2022. Wildfire risk assessment in Liangshan Prefecture, China based on an integration machine learning algorithm. *Remote Sensing* 14 (18): 4592.
Yingyongyudha, A., et al. 2016. The Mini-Balance Evaluation Systems Test (Mini-BESTest) demonstrates higher accuracy in identifying older adult participants with history of falls than do the BESTest, Berg Balance Scale, or Timed Up and Go Test. *Journal of Geriatric Physical Therapy* 39 (2): 64–70.
Yue, S., P. Pilon, and G. Cavadias. 2002. Power of the Mann-Kendall and Spearman's rho tests for detecting monotonic trends in hydrological series. *Journal of Hydrology* 259 (1–4): 254–271.
Zhai, C., et al. 2020. Learning-based prediction of wildfire spread with real-time rate of spread measurement. *Combustion and Flame* 215: 333–341.
Zhang, L., et al. 2020. Analysis of drought evolution in the Xilin River basin based on standardized precipitation evapotranspiration index. *Arid Zone Research* 37: 819–829.
Zhang, Z., et al. 2021. Spatiotemporal analysis of active fires in the Arctic region during 2001–2019 and a fire risk assessment model. *Fire* 4 (3): 57.
Zhang, F., et al. 2022. Performance of multiple machine learning model simulation of process characteristic indicators of different flood types. *Progress in Geography* 41: 1239–1250.
Zhao, Y., et al. 2022. Temporal and spatial patterns of biomass burning fire counts and carbon emissions in the Beijing–Tianjin–Hebei (BTH) region during 2003–2020 based on GFED4. *Atmosphere* 13 (3): 459.

Publisher's Note