MODIS-FIRMS and ground-truthing-based wildfire likelihood mapping of Sikkim Himalaya using machine learning algorithms

Polash Banerjee

Received: 6 October 2020 / Accepted: 6 August 2021 / Published online: 15 August 2021 © The Author(s), under exclusive licence to Springer Nature B.V. 2021

Abstract
Wildfires in limited extent and intensity can be a boon for the forest ecosystem. However, recent episodes of wildfires of 2019 in Australia and Brazil are sad reminders of their heavy ecological and economical costs. Understanding the role of environmental factors in the likelihood of wildfires in a spatial context would be instrumental in mitigating it. In this study, 15 environmental features encompassing meteorological, topographical, ecological, in situ and anthropogenic factors have been considered for preparing the wildfire likelihood map of Sikkim Himalaya. A comparative study on the efficiency of machine learning methods like Generalized Linear Model, Support Vector Machine, Random Forest (RF) and Gradient Boosting Model (GBM) has been performed to identify the best performing algorithm in wildfire prediction. The study indicates that all the machine learning methods are good at predicting wildfires. However, RF has outperformed, followed by GBM in the prediction. Also, environmental features like average temperature, average wind speed, proximity to roadways and tree cover percentage are the most important determinants of wildfires in Sikkim Himalaya. This study can be considered as a decision support tool for preparedness, efficient resource allocation and sensitization of people towards mitigation of wildfires in Sikkim.

Keywords Forest fire · Prediction map · Algorithm · Statistical learning · GIS

1 Introduction

The forests of Sikkim are part of the Eastern Himalayan biodiversity hotspot. It is home to a variety of rare and endemic species of flora and fauna (Arrawatia and Tambe 2011; Paul et al. 2005). The pressure of climate change, deforestation, development and overgrazing is a growing challenge for the conservation of this fragile ecosystem. These pressures have a direct impact on the wildfire regime of Sikkim Himalaya. For instance, from 2004 to 2009 in Sikkim, the drought-like conditions and index of hotness showed an increasing trend,
while the number of rainy days was below average. These trends indicate the impact of climate change in Sikkim (Arrawatia and Tambe 2012; Sharma and Thapa 2021; Banerjee et al. 2020; Dong et al. 2017; Kumar 2012; Pradhan and Badola 2015). Moreover, the drier winter season converts the deadwood and forest litter into potential fuel for wildfires. Studies indicate that wildfires are most common in the low elevation Sal forests of Sikkim, followed by temperate sub-alpine and coniferous forests (Arrawatia and Tambe 2011; Sharma et al. 2014).

Lightning is the main cause of wildfires in the sparsely populated North Sikkim (Sharma et al. 2014). In contrast, the high incidents of wildfire in the Oak and Sal forests of East, South and West districts of Sikkim are mainly due to human activities. Slash and burn farming, forest fire to deter wild animals from entering settlements and bonfires in the forest areas are the intentional causes of wildfires in Sikkim. On the other hand, car sparks, power transformers in forest areas, use of the traditional torch called Ranku, throwing live cigarettes/bidis are the accidental causes of wildfires (Sharma et al. 2014). To date, very limited studies have been done to map the likelihood of wildfires in Sikkim Himalaya. Identifying the wildfire hotspots and factors contributing to it is the first step towards disaster mitigation. A Multi-Criteria Decision-Making Technique showed that all the districts of Sikkim except for North Sikkim district are at higher risk of wildfire. Furthermore, dense forests of Sikkim are prone to wildfires due to human activities and the aspect of the area. The model accuracy was 82.36% (Laha et al. 2020). MaxEnt machine learning-based prediction of wildfires in Sikkim Himalaya indicates that proximity to roads, the fraction of tree cover and meteorological conditions were the major determinants of wildfire events. The model accuracy was 95% (Banerjee 2021).

Wildfires can be considered a double-edged sword. Uncontrolled, large-scale and frequent wildfires can inflict colossal damage to a forest ecosystem. For instance, wildfire destroys the wildlife habitats (Pastro et al. 2011; Haque et al. 2021), causes atmospheric phosphorus deposition in the local water bodies (Vicars et al. 2010), mobilizes heavy metals (Campos et al. 2015), promotes leaching of soil nutrients (Murphy et al. 2004), interferes in the mobility of wildlife (Banks et al. 2011), promotes the loss of soil biota, increase the volatilization of soil nutrients, soil erosion, and causes decline in biodiversity and forest biomass (Chandra and Bhardwaj 2015; Parashar and Biswas 2003). Long-term impacts of ecosystem disturbances including wildfire events can substantially change the nutrient composition of the soil. This can have profound ecological and functional impacts (Bowd et al. 2019). Wildfires have a differential impact on the mortality of plants depending on the species and size of the vegetation (Trouvé et al. 2021). Perhaps, such a wildfire-induced selection process can have a long-term impact on the species composition and overall health of the forest ecosystem. The increasing trend of wildfires in the Himalayan forests induced by climate change is acting as a positive feedback loop as they are accountable for the large-scale emission of greenhouse gases (Sannigrahi et al. 2020). Smoke generated from the wildfires can be a major health hazard. This health impact can have a wide geographic cover depending on the spread of the wildfire, population distribution and quality of health services (Cascio 2018). Health hazards associated with exposure to wildfires have high economic impacts in the form of public health liability in terms of premature deaths and respiratory diseases (Fann et al. 2018). Furthermore, a likely association between wildfire and the psychosocial health of children, adolescents and family have been observed (Kulig et al. 2019).

In contrast, periodic and relatively smaller scale wildfires do benefit the forest ecosystem. The release of nutrients from the burnt biomass into the soil improves the fertility of the vegetation. This is reflected in the increase in the abundance of grazers, rodents and
birds in the forest. Also, wildfires increase standing biomass over the years in an undisturbed forest ecosystem (Lowe et al. 1978). Fire and Landscape Ecology Assessment Tool (FLEAT), a modelling tool to assess whether wildfires benefit or harm an ecosystem, suggests that wildfire has great ecological benefits (Keane and Karau 2010). Some of these benefits are echoed as ecological services to mankind. For instance, wildfires help in pest control, enhances flowering, pollination and germination (Pausas and Keeley 2019). In some cases, mixed-severity wildfires can be beneficial to some species. For instance, the wildfire-induced opening of habitat patches has promoted foraging by the spotted owls (Strix occidentalis) that has led to a significant increase in their abundance in the USA (Lee 2018).

Wildfire is governed by a wide range of environmental features. Studies indicate that meteorological features like atmospheric temperature, wind, precipitation, humidity and lightning events are good predictors of wildfire. Topographic features like elevation, slope, aspect, topographic wetness index, topographic roughness index and plan curvature have been widely used in preparing Wildfire Likelihood Map (WLM). In-situ features like soil type, soil moisture, land surface temperature, land use, potential evapotranspiration and soil carbon content have also been used in wildfire prediction. Ecological features like vegetation type, normalized difference vegetation index (NDVI), tree cover fraction, standing biomass, fuel biomass and proximity to water bodies are some of the accepted ones in preparing a WLM. Wildlife prediction studies have also used proximity to roads, proximity to settlements and population density as anthropogenic factors for wildfire occurrences (Arpaci et al. 2014; Estes et al. 2017; Flannigan and Harrington 1988; Guo et al. 2016; Jaafari et al. 2018; Kim et al. 2015a, b; Ljubomir et al. 2019; Sachdeva et al. 2018; Sharma et al. 2014; Tien Bui et al. 2019; Yathish et al. 2019).

Effective allocation of resources is a crucial step towards appropriate wildfire mitigation (Gheshlaghi et al. 2020). This issue becomes even more pressing with the increasing trend of wildfires due to climate change (Flannigan et al. 2000; Gillett et al. 2004; Williams et al. 2019). In most cases, authorities are effective in controlling wildfire. However, a small fraction of wildfires does get accidentally overlooked by the authorities. These wildfires can inflict substantial damage to the forest ecosystem as well as the local economy unless prior knowledge of the likelihood of wildfires is available to the stakeholders (Taylor et al. 2013). WLM provides the spatial probability of the occurrence of wildfires over a study area. Such a likelihood map can be prepared by the criteria-based overlay of environmental feature maps that influence wildfires.

Expert opinion-based multicriteria decision analyses like Analytic Hierarchy Process (AHP) and Analytical Network Process (ANP) have been used to prepare WLP. These methods rely on constructing a hierarchal structure of the model criteria and alternatives. Pairwise comparison of the criteria at each hierarchal level and that of the alternatives yield the criteria weight and the importance of alternatives in the context of the model (Banerjee et al. 2020; Yathish et al. 2019; Ljubomir et al. 2019; Regodic et al. 2018; Gheshlaghi et al. 2020; Tien Bui et al. 2017). Other decision-based methods like fuzzy logic and fuzzy AHP have also been widely applied in preparing WLPs (Garcia-Jimenez et al. 2017; Gheshlaghi et al. 2020). Hybrid methods involving analytical network process and fuzzy logic have been applied in WLMs with fair success (Gheshlaghi et al. 2020). However, expert opinion and fuzzy logic-based methods have an innate limitation of subjectivity in the decision process. Moreover, such methods cannot handle a relatively large number of criteria as well as logical conditions. This is primarily because the comparison of criteria and logical conditions inflate rapidly with the increase in the criteria set. Also, unlike machine learning, expert opinion-based decision methods do not learn from the dataset (Behrooz et al. 2018).
Machine learning is a subdiscipline of artificial intelligence, that can learn from the dataset available, provided the data are sufficient and representative of the population under consideration (Géron 2017; Mitchell 1997). A range of popular machine learning methods has been widely applied in wildfire likelihood mapping (Banerjee 2021). For instance, logistic regression, a relatively simple machine learning method, has been widely successful in predicting wildfires in several studies (Guo et al. 2016; Tien Bui et al. 2016). Decision tree-based methods like Random Forest (RF) and Gradient Boosting Method (GBM) have been equally successful in wildfire predictions. The success of both methods is due to their simple approach of iterative dichotomization of the feature space with a tuning criterion that minimizes the cost of false prediction of the target variable (Arpaci et al. 2014; Chirici et al. 2013; Guo et al. 2016; Kim et al. 2019; Leuenberger et al. 2018; Massada et al. 2013; Sachdeva et al. 2018; Tehrany et al. 2019; Xie and Peng 2019). Support Vector Machine (SVM) is another machine learning method that has fared well in wildfire prediction in several GIS-based studies. SVM performs prediction by attempting to maximize the margin between the clusters of elements of the target variable in the features space. Here, feature space represents the hypervolume of environmental features used for predicting wildfire while the target variable is a binary set of presence and absence of wildfire events (Al_Janabi et al. 2018; Tehrany et al. 2019; Tien Bui et al. 2016). Brain mimicking machine learning methods like multilayer perceptron, deep learning and convolutionary neural network have been applied in wildfire prediction with high success. Collectively these methods are called Artificial Neural Network (ANN). ANN uses various neural architectures that iteratively adjusts the weight of the nodes that represent neurons of the simple brain. The nodes explain a certain aspect of the model features. ANN performs predictions by adjusting the weights of the nodes in a way that minimizes the cost of false prediction (Al_Janabi et al. 2018; Satir et al. 2016; Tien Bui et al. 2018, 2019; Xie and Peng 2019; Zhang et al. 2019). The popularity of machine learning is based on its ability to automate the process of prediction. Also, it progressively improves upon its learning through successive exposure to new datasets. Furthermore, it objectively identifies trends and patterns. These merits come at a cost of high computational time and the data-greedy nature of machine learning. Moreover, methods like ANN are difficult to interpret due to the inner complexity of their architecture.

In this study, an attempt has been made to prepare the WLM of Sikkim Himalaya using various machine learning methods. The overarching objectives of this study include the preparation of WLM of the study area based on a comparative analysis of the machine learning methods. Secondly, the identification of the environmental features influencing wildfires in the study area. This study is most likely the first attempt to prepare the WLM of Sikkim Himalaya using a comparative analysis of machine learning methods. It presents a high-resolution WLM of Sikkim Himalaya that can significantly facilitate in the identification of wildfire hotspots. Accordingly, a robust wildfire management system can be developed by the state as well as the civic authority towards efficient resource allocation, early warning systems and awareness programmes. The study shows that roadways are the most important determinant of wildfires in Sikkim followed by meteorological factors like wind speed and ambient temperature.

2 Materials and methods

2.1 Study area

Sikkim is the second smallest state of India situated in the north-eastern hills of Himalaya. It is neighboured by Tibet in the north, Bhutan in the east, West Bengal in the south and
Nepal in the west. Sikkim is richly endowed by nature in terms of rugged mountainous terrain and a wide variety of endemic flora and fauna. Also, Sikkim is home to a vibrant collection of indigenous cultures and tribal communities.

Sikkim extends from 27° 00′ 46″ N to 28° 07′ 48″ N in latitude and 88° 00′ 58″ E to 88° 55′ 25″ E in longitude. Altitude-wise Sikkim varies from 280 m in the south to 8586 m in the north. The north of Sikkim is covered by the Great Himalayan range soaring to the world’s third-highest peak, Mt. Kangchenjunga. The two main rivers of Sikkim include the Teesta River and its tributary, the Rangeet (Shukla et al. 2018).

The climate of Sikkim in addition to having winter, summer, spring and fall seasons, has a monsoon season that lasts from June to September. The dry period of Sikkim lasts from December to March. It is characterized by cold, dry and windy conditions. Much of the wildfires in Sikkim occur during this dry period. Overall, Sikkim has a subtropical climate in the south and a tundra climate in the north.

Different types of vegetation ecotype populate Sikkim, such as the Himalayan subtropical broadleaf forests dominate the lower elevations, Eastern Himalayan broadleaf forests populate the temperate zone above the altitude of 1500 m, Eastern Himalayan subalpine conifer forests grow from 3500 to 5000 m; and Eastern Himalayan alpine shrub and meadows in the higher elevations (O’Neill 2019). In terms of human presence, the bulk population of Sikkim reside in the southern and eastern parts with an average population density of 86 km\(^{-1}\) (COI 2011). This becomes evident by realizing that much of the road network and agrarian land dominate these areas of Sikkim (Fig. 1).

### 2.2 Data sources and data processing

Supervised machine learning algorithms require a dataset, often in the form of a table for learning. The table consists of columns and rows. The columns represent attributes or features based on which the predictions are made by the algorithm. Apart from the feature columns, a target or response variable column is also included. The target variable is the output that the algorithm learns to predict. The rows of the table, known as the instances, represent a set of features and their corresponding target variable. In this study, several environmental attributes were considered encompassing meteorological, topographical, ecological, in situ and anthropogenic features (Devisscher et al. 2016; Ghorbanzadeh et al. 2019; Joseph et al. 2009) (Table 1).

Meteorological factors like above-ground air temperature, precipitation and wind speed are important determinants of wildfire. Higher temperature, low precipitation and high wind speed facilitate wildfire occurrences (Flannigan and Harrington 1988; Mhawej et al. 2015). Topographical factors like elevation influence meteorological factors. At higher elevations, the wind speed increases while precipitation and temperature tend to decrease. Also, a steeper slope facilitates the spread of wildfire. Aspect influences the amount of insolation and the humidity of the biomass. Studies have shown that wildfires are more common in the south to the southwest direction (Graham et al. 2004; Jo et al. 2000; Mhawej et al. 2015). Topographic Wetness Index (TWI) is a measure of potential soil wetness (Krawchuk et al. 2016). A low TWI indicates drier soil, a potential facilitator of wildfires. Also, the curvature of the terrain influences convection, radiation and the transport of burning material (Hilton et al. 2017). All the topological features in this study were prepared from the DEM using ArcGIS. Ecological features like NDVI indicate the health of vegetation and potential fuel biomass. Areas with moderate tree cover are more vulnerable to wildfires (Zhang et al. 2019). Proximity to water
bodies indicates the level of soil moisture and human disturbances. In situ conditions like low soil moisture promotes wildfire (Krueger et al. 2015). Also, soil surface carbon content can be considered as a representative of the litter content of the forest floor. Dry litter acts as fuel to wildfire. Anthropogenic factors like proximity to the road network, settlements and population density were considered in this study. These features substitute for human-induced activities such as sparks of vehicle engines. Other wildfires inducing activities are, slash and burn farming and burning of forest vegetation to deter wildlife from entering settlements. Population density has a strong relation with recreational activities in the forested areas such as bonfires, deforestation for firewood and timber. Also, population density partly explains illegal forest exploitation due to unemployment (Sharma et al. 2014).

The target variable in this study was the presence or absence of wildfires. The historical wildfires dataset of Sikkim Himalaya was prepared using two data sources. The ground-truthing-based dataset was procured from the Forest Environment and Wildlife Management Department (FEWMD), Govt. of Sikkim. The timeframe of FEWMD ranged from 2016 to 2018. The remote sensing-based fire events dataset was accessed from the Moderate Resolution Imaging Spectroradiometer (MODIS) and Visible Infrared Imaging Radiometer Suite (VIIRS) data archive at the Fire Information for Resource Management System (FIRMS) site (FIRMS 2020). The coarse resolution of 1 km of MODIS is complemented by the relatively finer resolution of 375 m of VIIRS. The fire dataset of Sikkim Himalaya accessed from FIRMS ranged in the timeframe from 2000 to 2019. However, the dataset of FIRMS does not distinguish wildfire from
| Factors                      | Abbreviation | Feature name     | Unit      | Resolution | Description                                                                 | Source                                      |
|------------------------------|--------------|------------------|-----------|------------|------------------------------------------------------------------------------|---------------------------------------------|
| Meteorological factors       | avgTemp      | Average temperature | °C        | 30 s       | Average pixel value calculated over the dry period                           | Worldclim2 (Fick and Hijmans 2017)          |
|                              | avgWind      | Average wind speed | ms⁻¹      |            |                                                                               |                                             |
|                              | avgPrep      | Average precipitation | mm/month |            |                                                                               |                                             |
| Topographical factors        | aspect       | Aspect degree    | degree    | 90 m       | All the topographic features were derived from the DEM of Shuttle Radar Topography Mission (SRTM). Data collection timeframe was from 2000–02-11 to 2000–02-22 | SRTM (Jarvis et al. 2008)                   |
|                              | slope        | Slope degree     | degree    |            |                                                                               |                                             |
|                              | PlanCurv     | Plan curvature    | Radians meter⁻¹ |            |                                                                               |                                             |
|                              | profileCurv  | Profile curvature | Radians meter⁻¹ |            |                                                                               |                                             |
|                              | TWI          | TWI Ratio        |           |            |                                                                               |                                             |
| Ecological factors           | NDVI         | NDVI Ratio       | 250 m     |            | Accessed from the Advanced Spaceborne Thermal Emission and Reflection Radiometer data. Data collection timeframe was from 2000–02-18 to 2020–07-02 | ASTER (ASTER Mount Gariwang image 2018)     |
|                              | treeCov      | Per cent tree cover | %         | 30 m       | Accesses from the Global Landsat Tree Cover Continuous Fields, prepared from three-year epochs of 2000, 2005 and 2010 assessing the woody vegetation greater than 5 m in height | Sexton et al. 2013                          |
|                              | proxWater    | Proximity to waterbodies | m         | 30.7 m     | Prepared by generating raster of Euclidian distance from the vector polylines | © OpenStreetMap                             |
| In situ factors              | soilCarbon   | Soil carbon      | 0 cm depth | 250 m      | Soil carbon content in × 5 g / kg at the soil surface                        | OpenLandMap (Hengl and Wheeler 2018)       |
|                              | soilWater    | Soil water       | 0 cm depth | 250 m      | Soil water content (volumetric %) for 33 kPa and 1500 kPa suctions predicted at the soil surface | OpenLandMap (Hengl and Gupta 2019)          |
| Factors            | Abbreviation | Feature name                  | Unit | Resolution | Description                                                                 | Source               |
|--------------------|--------------|-------------------------------|------|------------|----------------------------------------------------------------------------|----------------------|
| Anthropogenic factors | proxRoads    | Proximity to road network     | m    | 30.7 m     | Prepared by generating raster of Euclidian distance from the vector polylines | © OpenStreetMap      |
| proxPlace          | Proximity to human habitations | m          |      | 30.7 m     | Prepared by generating raster of Euclidian distance from the vector points   |                      |
other sources of fire. To extract the wildfires from otherwise, a historical forest coverage raster was prepared by the union of all forest extents in the timeframe from 2007 to 2010 (Shimada et al. 2014) (Supplement Fig. 1). The FIRMS fire dataset was intersected with the forest cover raster to include the fire events restricted within the forest cover area. In this way, only the wildfires were identified from the FIRMS dataset. Furthermore, to prevent double-counting from the data sources, wildfires of FIRMS within 200 m proximity from FEWMD were dropped from the dataset. The ground-truthing-based dataset and modified FIRMS dataset were combined to prepare the final wildfire dataset of 754 events.

All the feature rasters and the wildfire dataset were projected from the geographic projection system of GCS-WGS-1984 to the plane projection system of WGS-1984-UTM-Zone-45N in the GIS framework. The latter provides an appropriate measurement in the metric system for India. The cell size and extent of all the feature rasters were standardized to be the same. Next, the rasters were exported as GeoTiff rasters from the GIS framework and imported into the R-programming framework for feature extraction and machine learning. The feature rasters were transformed into unitless rasters by normalization method (Chang 2017):

\[ z_i = \frac{x_i - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \]  

(1)

where the pixel value \( z_i \) is the normalized value of a raster ranging from 0 to 1 at the \( i \)th location of the study area extent. \( x_i \) is the pixel value of the \( i \)th location before the normalization. \( x_{\text{min}} \) is the lowest value, while \( x_{\text{max}} - x_{\text{min}} \) is the range of the raster.

Apart from the wildfire-presence dataset, an equal number of the wildfire-absence dataset was prepared within the study area extent using random sampling. The two datasets were combined to prepare the presence-absence dataset. Next, the feature rasters were stacked and the presence-absence dataset was used to extract the feature values from the feature stack to populate the presence-absence dataset with the features (Supplement 1). The presence-absence dataset was used to train various machine learning methods. For this, the target variable was treated as a categorical variable while the features were treated as continuous variables. The dataset was split into training and testing subsets with 75% devoted to training and the rest for testing the prediction model (Boutaba et al. 2018; Géron 2017; Luo et al. 2017). There is no hard and fast rule of splitting the dataset for training and testing. Usually, the percentage splitting of the dataset can be arrived at by a trial-and-error method based on the total size of the dataset or by following the prevailing practices in machine learning research. In this study, the latter option was adopted. Furthermore, the training dataset was split into ten subsets for cross-validation. During cross-validation, repeatedly, one of the subsets of the ten training subsets was used to test, while the others were used to train the machine learning algorithm. The average of the errors generated by the repeated tests made it possible to adjust the parameters of the model for better ML performance.

Four machine learning methods were considered for this study, namely, generalized linear model (GLM), SVM, RF and GBM. The selection was based on their prevalence in the disaster prediction mapping literature, relatively higher efficiency and efficacy in predictions and better model interpretability.
2.3 Multicollinearity analysis

Studies suggest that a wide range of environmental features can influence wildfire depending on the area of interest. However, identifying the environmental features specific to the present study area was the key step for robust prediction modelling. For this, correlation matrix and multicollinearity analysis were performed (Argyrous 2011). In multicollinearity analysis, the Variance Inflation Factor (VIF) exceeding 10 are used to identify multicollinear variables. This indicated data redundancy of the dataset. The quality of the dataset was improved by dropping one or more of such feature variables and rechecking multicollinearity criteria. Once the multicollinearity criterion was met the dataset was fit for further analysis.

2.4 Generalized linear model

GLM is an umbrella term used for linear regression models that are characterized by certain properties. Firstly, the dependent or the target variable of the regression model belongs to the exponential family of the probability distribution. In other words, the predictor behaves nonlinearly with the change in the model parameter. Secondly, the predictor itself is linear. Thirdly, there exists a link function that yields the predictor value when provided with appropriate covariate features of the model. GLM is expressed as (McCullagh and Nelder 1989):

\[ E(Y) = g^{-1}(X\beta) \]  

where \( E(\cdot) \) is the expected value of the matrix of dependent variables \( Y \). \( g \) is the link function that takes \( X \) as the matrix of covariate features and vector of parameters \( \beta \). The predictors, as well as the covariates in GLM, can be continuous or categorical. GLM includes regression models like linear regression, ANOVA, logistic regression, log-linear regression, Poisson regression and multinomial response regression. In the case of a binomial-dependent variable, like \( Y = \{0, 1\} \), GLM takes the form of the logistic regression model. The sigmoidal function used for logistic regression generates an S-shaped probability distribution between 0 and 1:

\[ P(y|x) = \frac{1}{1 + \exp(-\beta^T x)} \]  

where \( \beta = [\beta_0, \beta_1, \ldots, \beta_n] \) is the vector of model parameters and \( x = [1, x_1, \ldots, x_n] \) is the vector of predictor features of an instance from the dataset. The superscript \( T \) represents the transpose of a vector. For training of the model, a cost function was used as (Géron 2017):

\[ c(\beta) = \begin{cases} -\log(P(y|x)) & \text{if } y = 1 \\ -\log(1 - P(y|x)) & \text{if } y = 0 \end{cases} \]  

The model was fit to the training dataset using the maximum likelihood method by minimizing the convex log loss function:

\[ j(\beta) = -\frac{1}{M} \sum_{i=1}^{M} [y^{(i)} \log(P^{(i)}) + (1 - y^{(i)}) \log(1 - P^{(i)})] \]  

where \( M \) is the size of the training dataset. \( P^{(i)} = P(y^{(i)}|x^{(i)}) \) as given in Eq. (3) and superscript \( (i) \) is the index of the instance of training. The partial derivative of the log loss
function over all the parameters $\beta$ gives the gradient descent for the model to reach the solution.

2.5 Support vector machine

In SVM a vector of features $x = (x_1 \ldots x_M)$ and their corresponding target $y$, pair up to make the instances $(x,y)$. SVM is a supervised machine learning method that partitions the $M$-dimensional feature space by a decision boundary. The decision boundary is generated by the inner products of a set of feature vectors called support vectors that defines the partition. The decision boundary can be expressed as:

$$f(x) = h(x)^T \beta + \beta_0$$

where $h(x)^T$ is a vector of basis functions defined as $(h_1(x_i), h_2(x_i), \ldots, h_M(x_i))^T$ over $M$ features and $N$ instances. A basis function explains the nonlinear relationship of the feature vector $x_i$ with the regression model $f(x)$. $\beta$ is a vector of coefficients of the corresponding basis function and $\beta_0$ is the intercept. The decision boundary that maximizes the margin of the partition was considered for regression. Equation (6) can be amended to include a kernel trick $K(\cdot)$ that solves for the hypothesis $\hat{f}(x)$ by using the Lagrange dual function:

$$\hat{f}(x) = \sum_{i=1}^{N} \hat{\alpha}_i y_i K(x, x_i) + \hat{\beta}_0$$

where $\hat{\alpha}_i$ is a positive constraint. The kernel $K(x, x_i)$ is the generalized form of the inner product of a feature point and support vectors expressed as $(h(x), h(x_i))$. Depending on the nature of the SVM algorithm, the kernel type is selected. In this study, the radial basis function (RBF) kernel has been considered:

$$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$$

Considering a two-class regression, like the wildfires, if the feature point $x$ in the feature space occurs close to a support vector, then the RBF will be approaching one. This will lead Eq. (7) to the prediction of the likely value of $\hat{f}(x) = \hat{y}$. The $\sigma$ parameter of RBF controls the Gaussian distribution of RBF. The hypothesis was directed towards the solution by minimizing a cost function:

$$\sum_{i=1}^{N} [1 - y_i f(x_i)] + \frac{\lambda}{2} \|\beta\|^2$$

Considering $y_i = 1$ and correct prediction made by the hypothesis, i.e. $f(x_i) = 1$ during the training of SVM, Eq. (9) reduces to minimizing the term $\frac{\lambda}{2} \|\beta\|^2$. This second term of Eq. (9) was mainly to maximize the margin of the decision boundary. The $\lambda$ hyperparameter controls the smoothness of the decision boundary.

Thus, through iterations, the SVM constructed a decision boundary in the feature space using a kernel function and by minimizing the cost function of misclassification. A test instant was regressed, based on its location with respect to the decision boundary and proximity to the support vectors. A detailed account of this discussion can be found in Hastie et al. (2017)
2.6 Gradient boosting model

GBM uses ensemble learning that progressively does weighted stage-wise addition of weak learners in the form of regression trees to compensate for the limitation of the existing weak learners. This process eventually generates a strong learner. A regression tree partitions the feature space into regions $R_m, m \in {1, \ldots, M}$ based on a split criterion that minimizes the sum of squares of the mean of the region from its corresponding target value. The splitting criterion can be the minimization of entropy of information or the Gini index. The outcome of the partitioning is expressed as:

$$f(x) = \{\bar{y}_m | x \in R_m\}$$

(10)

where $\bar{y}_m$ is the mean of the target of the $R_m$. Consider $T(x; \Theta)$ to be a tree learner, defined by an instance $x$ and $\Theta$ as the parameterized form of the expected loss function. Furthermore, a loss function $L(y_i, f(x_i))$ is constructed to estimate the deviation of $f(x_i)$ from $y_i$. GBM approaches the solution by minimizing the squared sum of the tree learner $T(x; \Theta)$ from negative of the gradient of the loss function $g_{im}$ for the $i$th instance in the $m$th region at each iterative stage:

$$\Theta_m = \sum_{i=1}^{N} (-g_{im} - T(x_i; \Theta))^2$$

(11)

where

$$g_{im} = \left[ \frac{\partial L(y_i, f(x_i))}{df(x_i)} \right]_{f(x_i) = f_{m-1}(x_i)}$$

(12)

Equation (12) expresses the slope of the gradient descent of the loss function over the hypothesis space. The GBM solves by taking the steepest gradient descent $g_{im}$ using a learning rate $\rho_i$. This is followed by updating $f_m(x)$ giving more weight to trees that give higher $\Theta$, as given in Eq. (13). As a result, the boosting process focuses more on residual errors:

$$f_m(x) \leftarrow f_{m-1}(x) + \rho_i T(x_i; \Theta)$$

(13)

The final boosted tree is the weighted sum of all the trees:

$$f_M(x) = \sum_{i=1}^{M} T(x; \Theta_i)$$

(14)

Thus, GBM starts the iterative ensemble process by initially constructing a weak learner regression tree and compares it with the target. This is followed by the stepwise construction of another such tree. But this time the focus of the algorithm is more towards the residual error. This is achieved by giving more weightage to the residual error. In this way, the progressive learner becomes relatively better than its predecessor. A loss function is used to track the learning success of the trees. GBM fits the residual with the steepest gradient of the loss function to boost the learning process. Eventually, as the learning process approaches a certain tolerance level, a weighted sum of all the weak learners yields a strong learner for testing and regression. In this study, a stochastic GBM was applied that
used the Gaussian loss function for training. The stochastic GBM takes a subsample of the training instances. This prevents overfitting, speeds up the computation and helps in the regularization of the model. A more detailed discussion on this topic can be found in Hastie et al. (2017) and Natekin and Knoll (2013).

### 2.7 Random forest model

Random forest is another form of ensemble-type machine learning method that uses regression trees for the solution set. In this method, several regression trees are constructed by selecting a subsample from the dataset by bagging method as discussed in GBM. Furthermore, instead of taking the entire set of features, a subset of the same is randomly selected by the algorithm, usually $\sqrt{p}$, where $p$ is the number of features of the model. A small subset of the feature set helps in reducing the variance of the average of the predictions of the trees. This increases the accuracy of the prediction. After that, the best regressor feature and splitting value are selected that minimizes the impurity or cost function. The feature space is dichotomized using the splitting value. This creates two daughter nodes of the decision tree. This process is iterated till the minimum number of terminal nodes of the tree is reached and the error reduction reaches a plateau. The value of the target is estimated as:

$$\hat{f}_{ij}^B(x) = \frac{1}{B} \sum_{b=1}^{B} T(x; \Theta_b)$$  \hspace{1cm} (15)

where $B$ is the total number of regression trees constructed. $T(x; \Theta_b)$ is the tree defined by the instance $x$ and loss parameter $\Theta$. Like GBM, RF also enjoys high accuracy in predictions (Géron 2017; Hastie et al. 2017).

### 2.8 Model performance

The performance of a model was assessed by estimating how far was the prediction of the model as compared to the actual observation. In this study, several indices were considered to cover all the aspects of the model performance (Ali et al. 2021; Pham et al. 2020, 2021; Tuyen et al. 2021; Zhang et al. 2021). They include Root Mean Square Error (RMSE), Mean Absolute Error (MAE):

$$RMSE(X, f) = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (f(x^{(i)}) - y^{(i)})^2}$$  \hspace{1cm} (16)

$$MAE(X, f) = \frac{1}{m} \sum_{i=1}^{m} |f(x^{(i)}) - y^{(i)}|$$  \hspace{1cm} (17)

where $f(\cdot)$ is the hypothesis of actual observation $y$ of $i$th instance from a total of $m$-sized test set of the dataset matrix $X$. Ideal value for both the performance criteria is zero. However, RMSE is more sensitive to outliers than the MAE (Géron 2017).

Furthermore, two widely accepted model performance criteria have been used in this study, namely, confusion matrix and Receiver Operating Characteristic (ROC) Curve. The confusion matrix is a tabular representation of correctly and incorrectly classified
instances, based on the comparison of the predicted and observed values of the target variable of the test set. Typically, in a binary classification problem, the confusion matrix has four elements. The first element of the primary diagonal of this matrix holds the True Positives (TP), the presence-instances that have been correctly classified as positives. The second element of the primary diagonal holds the True Negatives (TN), the absence-instances that have been correctly classified. In contrast, the first element of the secondary diagonal is the False Negatives (FN), the presence-instances, that have been incorrectly classified as an absence. The second element of the secondary diagonal is the False Positives (FP), the absence-instances, that have been incorrectly classified as presence. The overall accuracy (OAC) of the model is estimated as:

$$OAC = \frac{TP + TN}{TP + FP + TN + FN}$$  \hfill (18)

An accuracy value not less than 0.7 is usually considered a good prediction. Other relevant performance indices include Sensitivity, Specificity, Precision, Cohen’s Kappa and F1 score. Sensitivity is the measure of the proportion of presence-instances that have been classified as positives (Eq. 17). Specificity is the measure of the proportion of absence-instances that have been classified as negatives (Eq. 18). Precision is the measure of the accuracy of positive predictions of the model (Eq. 19). Cohen’s Kappa is the measure of the deviation of the relative predicted agreement by the model from the hypothetical probability of chance agreement (Eq. 20). F1 score is the harmonic mean of precision and sensitivity values (Eq. 21). It holds a higher value only if both precision and sensitivity are high:

$$sensitivity = \frac{TP}{TP + FN}$$  \hfill (19)

$$specificity = \frac{TN}{TN + FP}$$  \hfill (20)

$$precision = \frac{TP}{TP + FP}$$  \hfill (21)

$$kappa = \frac{P_{obs} - P_{exp}}{1 - P_{exp}}$$  \hfill (22)

$$F1 Score = 2 \times \left( \frac{precision \times sensitivity}{precision + sensitivity} \right)$$  \hfill (23)

where $P_{obs}$ is the proportion of pixels correctly classified as wildfires or as non-wildfires, and $P_{exp}$ is the proportion of pixels for which the amount of agreement is expected by chance only.

The ROC curve is used to visualize the change in sensitivity over the specificity of the prediction model. A perfect prediction by a model should yield an ideal tuple of (1,1) implying perfect sensitivity and specificity. Usually, a good model generates a concave ROC curve with respect to the diagonal connecting (1,0) to (0,1) for the tuple (sensitivity, specificity), with a high value of area under the curve (AUC).

A comparative analysis of the performance of the machine learning methods was done using the box and whisker plot and scatter plot matrices. To do this, 30 samples from the
2.9 Wildfire likelihood model

In this study, several R-packages were used for data pre-processing and machine learning (Hunt 2020; Kuhn 2020; Liaw and Wiener 2002; Robin et al. 2011). These packages were used for training and testing the machine learning algorithms. Furthermore, machine learning algorithms also provided the importance of the feature variables used in the model. Features with no importance to the model were dropped from the dataset and algorithms were rerun. This was followed by stacking of the feature rasters. The stack was used by the machine learning algorithms for predicting the wildfire probability over the entire study area. The matrix of predicted values of the entire study area was exported as a GeoTiff raster, that stores the predicted values along with their respective latitude and longitude values. The rasters of predicted values of wildfire were imported into the GIS framework for further analysis that included the categorization of the raster into areas of very low, low, medium, high and very high probability of wildfire (Fig. 2).

3 Results

The wildfire inventory dataset was used to analyse whether wildfires were on the rise in Sikkim Himalaya. Time series data of wildfires over the years 2000 to 2019 indicated that there was a growing trend. The picture became clearer by forecasting the wildfires using Holt’s forecast model. The model predicted that the wildfires were likely to increase from 82 in 2019 to 96 in 2022 with an uncertainty of ±62.343 events (Fig. 3). To identify the
relevant environmental features contributing to wildfires, a multicollinearity analysis was performed.

### 3.1 Multicollinearity analysis

Based on the prevailing literature, initially, 16 environmental features were considered for the prediction. However, multicollinearity analysis brought the number of explanatory environmental features down to 15 (Fig. 4, Table 2).

### 3.2 Impact of environmental features on wildfires

Wildfires showed more propensity over certain intervals of the ranges of environmental features. Except for TWI, they were mostly normally distributed over the topographical feature. For instance, in aspect, wildfires were more common over the interval 140°–220°, covering southeast to southwest direction. Also, wildfires were more common over steep slopes (29°–32°). Wildfire events over plan and profile curvatures showed normal distributions over certain ranges. In the case of plan curvature, all the wildfires occurred in convex curvatures, while the same occurred in concave curvatures for profile curvature (Pourghasemi 2014). In contrast, wildfires showed skewness over lower TWI interval (5–6.5). Wildfires also showed more skewness over the meteorological features. Temperature-wise wildfires were very common in warmer areas of Sikkim, mainly the lower altitude areas with high average temperature (19°–24°). Moreover, wildfires are more concentrated over low average wind speed (about 1.6 ms⁻¹). Regarding ecological features, wildfires were more common over moderate NDVI values (about 0.6) and moderate tree cover (43–49%). Also, wildfires were more clustered near the water bodies (400–630 m away from water bodies). Regarding the in situ features, wildfire events were mostly confined to low carbon content.

![Time series and forecast of wildfire events in Sikkim Himalaya](image_url)

**Fig. 3** Time series of wildfire events in Sikkim Himalaya from 2000 to 2019. The Holt’s forecast model indicates an increasing trend of wildfire in Sikkim Himalaya when projected to the year 2022. The forecast has an average boundary of ±62.343 wildfire events from 2020 onwards.
soil (25–40 g/kg of soil) and moderately humid soil surfaces (30–45 volumetric %). Considering the anthropogenic features, wildfires showed moderate skewness towards areas close to settlements (1 km to 2.5 km from settlements). While wildfires were high in the areas close to roadways (≤ 400 m from the roads) (Supplement Figs. 2, 5).

3.3 Model summary and model performance

3.3.1 Generalized linear model

The GLM model was run with 15 environmental features as explanatory variables. However, features like Aspect, plan and profile curvatures, and TWI were found to be not significant. Thereby, they were dropped from the model and the model was rerun (Table 2). Hence, the GLM model-based prediction included ten explanatory variables and 1090 instances. Tenfold cross-validation was performed for model tuning. From Table 2, it is evident that proximity to roadways and low wind speed were the strongest determinants of

---

**Fig. 4** Correlation matrix of feature variables
The model was able to explain 62% of the predictions (Table 3). The model performance was satisfactory, with low RMSE and MAE, while high AUC, Accuracy, Kappa, Sensitivity, Specificity, Precision, F1 Score, and Goodness of fit ($R^2$) (Table 4).

### 3.3.2 Support vector machine

The nonlinear kernel, Radial Basis function, was used in SVM for the prediction of wildfires. SVM used 727 support vectors to distinguish between the presence and absence of wildfire instances from the training dataset. Tenfold cross-validation was performed to tune the model. SVM uses several parameters known as hyper-parameters to tune the algorithm and converge to the solution. Model hyper-parameters, namely sigma ($\sigma$), epsilon ($\epsilon$) and cost C settled at 0.106, 0.1 and 1, respectively. The $\epsilon$ tunes SVM by determining the number of support vectors to be considered for regression. C, which is similar to $\lambda$ in Eq. (9), is accountable for regularization that provides a trade-off between over- and under-fitting of SVM. The objective function value of SVM settled at $-292.33$ and the training error, the convergent error of the model achieved from the training set, was 0.286. RMSE of the final iteration of SVM was the same as GLM (Supplement Fig. 3a). However, other performance indices were worse than GLM except for sensitivity, F1 score and RMSE (Table 4).

### 3.3.3 Gradient boosting model

Under GBM, Stochastic Gradient Boosting was used using the Gaussian loss function. While converging to the solution, the GBM takes smaller learning steps to reduce the effect of each additional fitted weak learner tree. This penalization reduces the chances of giving undue importance to erroneous iterations. This method is called ‘shrinkage’. The

### Table 2 Multicollinearity analysis

| Environmental features | VIF |
|------------------------|-----|
| aspect                 | 1.054 |
| avgPrep                | 1.742 |
| avgWind                | 2.416 |
| NDVI                   | 2.815 |
| PlanCurv               | 1.612 |
| profileCur             | 1.181 |
| proxPlace              | 1.904 |
| proxRoads              | 1.943 |
| proxWater              | 1.249 |
| slope                  | 1.388 |
| soilCarbon             | 1.651 |
| soilWater              | 1.071 |
| treeCov                | 1.926 |
| TWI                    | 1.731 |

avgTemp was considered as the independent variable.
‘n.minobsinnode’, another tuning parameter of GBM, is the minimum number of observations in trees at the terminal nodes. The GBM in this study used the default values of shrinkage and n.minobsinnode at 0.1 and 10, respectively. GBM converged to the solution with 150 decision trees (n.trees) with an interaction depth of 3 (Supplement Fig. 3b). Performance-wise GBM outperformed GLM and SVM, except for MAE (Table 4).

3.3.4 Random Forest model

RF was used to predict wildfires in Sikkim Himalaya using 500 decision trees. RF converged to the solution when the algorithm selected eight environmental features at random at each split (mtry) (Supplement Fig. 3c). The mean squared residual of RF was 0.082, while RF explained 67.27% variance of out-of-bag predictions of the target variable of the training set (% Var explained). RF outperformed all the other prediction models (Table 4).

3.3.5 Comparative analysis of models

The box and whisker plots of accuracy and kappa showed a very similar pattern of distribution over the 30 samples selected to form the wildfire dataset using the cross-validation method. In both cases, RF distribution was rightwards than all other models indicating a higher model performance, followed by GBM. The greater width of the box in the case of GBM indicated a greater interquartile range. This showed a generalization approach of classification by GBM which is reflected in its better classification ability than GLM and SVM. The long range of observations as whiskers in the case of GLM indicated its inability to classify the incidents efficiently. In contrast, in the case of MAE and RMSE, RF showed a distribution shifted towards the left with a compact interquartile range and the range of observations skewed towards the left. This showed that RF is a much better contender in classification than other models which provide very similar MAE values, however with varied prediction distributions. RMSE being sensitive to outliers showed a clearer picture with the lowest value for RF followed by GBM. In the case of $R^2$, RF showed a compact interquartile distribution with a high value. In contrast, GBM showed the next highest value of $R^2$; however, its mean value has been pulled leftward by the skewed distribution of its predictions (Fig. 6). Scatter plot matrix (SPLOM) showed that RF and GBM have a higher correlation in predictions. Similarly, GLM and SVM showed a strong correlation. In contrast, such a correlation was not observed in the case of MAE, except for RF and GBM (Fig. 7). The ROC curves indicated that all the models performed satisfactorily, while RF outperformed all the other models (Fig. 8).

3.4 Variable importance

Proximity to roadways got the highest importance in GLM, GBM and RF. This was followed by average wind speed that got the highest importance in SVM and GLM. Also, features such as average temperature, NDVI and tree cover were found important in SVM. Methods like RF and GBM that uses regression trees for prediction gave disproportionately high importance to proximity to roadways and average wind speed. In contrast, methods like SVM and GLM that do not rely on regression trees gave more distributed importance to all the features. Topographic features like plan curvature, profile curvature and TWI received no to low importance in all the models. In situ features received low to moderate
importance. Anthropogenic features, meteorological features and ecological features were found to be the most important determinants of predictions (Fig. 9).

3.5 Wildfire prediction maps

All the feature maps were projected to the plane coordinate system of WGS-1984-UTM-Zone-45 N, as it is appropriate for India. Furthermore, all the feature maps were resampled to 30.7 m resolution. Accordingly, all the WLMs had the same projection system and resolution. In all WLMs, the southern part and valley areas of Sikkim Himalaya were found to be at a higher risk of wildfires. In the case of GLM, wildfire probability in most of the study areas was found to be very low, except for warmer valley areas of southern parts of Sikkim Himalaya. However, GLM put more emphasis on areas with high soil carbon content. This led to the consideration of such areas as high wildfire likelihood values (Fig. 10a). In contrast, SVM gave a slightly higher probability than GLM to most of the study area, along with giving a higher probability of wildfire to a much larger fraction of the study area (Fig. 10b). GBM devoted more areas to wildfire than GLM and SVM (Fig. 10c). RF gave more importance to valley areas than GBM, although the spatial distribution of wildfire probability showed high similarity with GBM (Fig. 10d). Based on the accuracy and performance of the prediction models, the WLM of RF was considered the best for Sikkim Himalaya. The WLM of RF was classified into five categories namely very low, low, medium, high and very high likelihood of wildfire based on natural breaks in the GIS framework (Fig. 11). Compared to high likelihood categories, very high likelihood of wildfire category had a relatively larger area (Fig. 12).

4 Discussion

The overarching objective of this study was to prepare the WLM of Sikkim Himalaya based on a comparative study of machine learning methods with appropriate explanatory variables. The study yielded prediction maps with good model performance indices.

4.1 Comparison between machine learning methods and their implications

In this study instead of just one algorithm, four algorithms were considered. This was mainly to identify the algorithm that performs best in the wildfire prediction out of the popular algorithms considered. Contrary to previous studies, RF outperformed other machine learning methods in wildfire predictions (Ogutu et al. 2011; Tehrany et al. 2019; Xie and Peng 2019). This observation was in harmony with studies performed by other authors (Guo et al. 2016; Kim et al. 2015a, b; Massada et al. 2013). The better performance of RF in comparison with GLM and SVM can be because RF uses the ensemble method.
Fig. 5 (continued)
Fig. 5 (continued)
Fig. 5 (continued)
### Table 3  GLM model summary

| Explanatory variables | Coefficients | Std. error | z value | Pr(>|z|) | Significance level |
|-----------------------|--------------|------------|---------|---------|-------------------|
| (Intercept)           | 6.4769       | 1.5862     | 4.083   | 0.000   | ***               |
| avgTemp               | 4.3986       | 1.2076     | 3.642   | 0.000   | ***               |
| avgWind               | -20.9441     | 3.6801     | -5.691  | 0.000   | ***               |
| NDVI                  | -3.9417      | 1.4549     | -2.709  | 0.007   | **                |
| proxPlace             | 10.4498      | 1.865      | 5.603   | 0.000   | ***               |
| proxRoads             | -34.9035     | 4.6755     | -7.465  | 0.000   | ***               |
| proxWater             | -9.1646      | 2.7138     | -3.377  | 0.001   | ***               |
| slope                 | 4.4118       | 0.9465     | 4.661   | 0.000   | ***               |
| soilCarbon            | -11.8139     | 2.3141     | -5.105  | 0.000   | ***               |
| soilWater             | -4.6546      | 1.4511     | -3.208  | 0.001   | **                |
| treeCov               | -2.4769      | 0.6129     | -4.041  | 0.000   | ***               |

Significance level codes: 0 ‘****’ 0.001 ‘***’ 0.01 ‘*’ 0.05

### Table 4  Performance of prediction models

| Performance measures          | Indices | Ideal Values | Models | GLM | SVM | GBM | RF |
|-------------------------------|---------|--------------|--------|-----|-----|-----|-----|
| ROC-based measure             | AUC     | 1            | 0.927  | 0.925 | 0.947 | 0.955 |
| Confusion Matrix-based measures| Accuracy | 1            | 0.881  | 0.851 | 0.898 | 0.914 |
|                               | Kappa    | 1            | 0.761  | 0.701 | 0.795 | 0.828 |
|                               | Sensitivity | 1            | 0.828  | 0.833 | 0.862 | 0.879 |
|                               | Specificity | 1            | 0.931  | 0.867 | 0.931 | 0.947 |
|                               | Precision | 1            | 0.854  | 0.849 | 0.879 | 0.895 |
|                               | F1 Score | 1            | 0.840  | 0.841 | 0.871 | 0.887 |
| Statistical measures          | $R^2$    | 1            | 0.612  | 0.593 | 0.638 | 0.670 |
|                               | RMSE     | 0            | 0.332  | 0.320 | 0.301 | 0.287 |
|                               | MAE      | 0            | 0.195  | 0.201 | 0.202 | 0.178 |

Fig. 6  Box and whisker plots of model performance indices
Fig. 7 Scatter plot matrix of accuracy and MAE

Fig. 8 ROC curve of a GLM, b SVM, c GBM, d RF
of learning instead of linear or kernel-based learning. In the ensemble method, the average output of several decision trees is considered. This process increases the chances of correct prediction. Also, contrary to SVM, RF is good at handling datasets with many outliers (Andreas 2013). As observed from the dataset, the histograms of several environmental features in this study were skewed. Perhaps, this was another reason for the better performance of RF.

The comparative analysis of the models was based on samples extracted from the wildfire dataset through the cross-validation method. It showed that GLM had a much wider range of accuracy and kappa values. This can be explained by the limited number of feature variables considered in the GLM model in comparison with other models. Furthermore, the smallest range of MAE of GLM indicated that the possible reason for the wide ranges of accuracy and kappa values can be due to a large set of outliers in the wildfire dataset (Géron 2017). The higher correlation between GLM and SVM as well as that of GBM and RF in terms of accuracy and MAE showed that out of these pairs of models only one should be considered while making an ensemble of models to improve the prediction capacity of wildfire events (Brownlee 2016).

4.2 Importance of feature variables

Consistent with previous studies, this study suggested that meteorological features like wind speed and to some extent ambient temperature were important determinants of wildfires. The low wind speed and warm temperature of the valley areas are features of sub-tropical Sal and Oak deciduous forests prone to wildfires in Sikkim Himalaya. The

![Variable importance](image-url)
anthropogenic feature like distance from the roadways on average was the strongest predictor of wildfires. To lesser extent proximity to human habitations also contributed to the predictions of wildfires. These observations second the previous studies on wildfires of Sikkim Himalaya.
Sikkim (Sharma et al. 2014). The ecological feature like the fraction of tree cover and in-site features like soil carbon were moderate predictors. Compared to other features, topographical features were not very good predictors of wildfires (Arpaci et al. 2014; Estes et al. 2017; Flannigan and Harrington 1988; Guo et al. 2016; Jaafari et al. 2018; Kim et al.)

Fig. 11 Wildfire likelihood map of Sikkim Himalaya showing various categories of likelihood of wildfire
Contrary to the MCDA-based study, namely using AHP, on forest fire risk zones of Sikkim (Laha et al. 2020), the present study gave limited importance to the aspect, except for the SVM model. However, indirect measures of human population density, namely, proximity to human settlements and roadways supported the observations of Laha et al. (2020). Like the observations by Banerjee (2021), this study showed that proximity to roadways was the most important determinant of wildfire in Sikkim. However, contrary to Banerjee (2021) average wind speed has been given more weight in this study than average ambient temperature. Looking at the correlation matrix these two meteorological variables had a significant negative correlation. However, their collinearity in terms of VIF was within acceptable limits. Thereby, they were considered as independent variables in this study. Furthermore, tree cover fraction has been considered as an important factor in wildfire prediction in both the studies.

4.3 Future risks of wildfire

The study showed that wildfires were predominantly distributed in the lower altitudes and valley areas of Sikkim Himalaya. Few observations can be made about these areas. The meteorological conditions of these areas were identified as having relatively warmer ambient temperatures and low wind speed. Also, the road network of these areas closely follows the river network. Steep slope facing southeast to southwest aspect with low TWI explained most of the wildfires of these areas (Graham et al. 2004; Jo et al. 2000; Mhawej et al. 2015). Low soil carbon and water content areas had more incidents of wildfire. The role of human activities in the occurrence of wildfires was evident from the study. These observations were similar to previous studies (Arpaci et al. 2014; Kim et al. 2019). However, contrary to previous studies, proximity to settlements as a feature had a contradictory role in this study as the bulk of the wildfires were on average 2.5 km away from the human habitations (Kim et al. 2019; Massada et al. 2013; Nami et al. 2018; Vilar et al. 2016). This may be since the land use around the settlements was mainly non-forest lands like agrarian
or fallow lands. Thereby areas of Sikkim bordering the state of West Bengal, district borders of West and South Sikkim and populated valleys of North Sikkim are more prone to wildfires.

The WLMs did not effectively predict the wildfires of upper North Sikkim. This may be since in this study meteorological factors, like the occurrence of lightning was not considered. In contrast, a study done earlier does mention the role of lightning in wildfires in North Sikkim (Sharma et al. 2014).

This study is probably the first attempt to systematically prepare the WLM of Sikkim Himalaya using multiple machine learning models. In line with studies done in other locations, this study indicated that anthropogenic and meteorological factors were the most prominent descriptors of wildfires. Also, this study highlighted that machine learning methods were reliable means of preparing hazard maps. However, the reliability of the predictions heavily depends on the wildfire inventory. This can be achieved by pruning instances with incorrect target variable or incomplete instances. Usually, a large and representative inventory leads to better predictions. Also, the engineering of features like normalization and removal of multicollinear features are essential steps for dataset preparation. Regarding the choice of algorithms, consideration of the nature of the dataset, in terms of whether the target variable is binomial, multinomial, categorical, or continuous is important. Moreover, skewness of the features has an important role in the choice of machine learning methods. Cross-validation and choice of hyperparameters for the regularization are essential steps towards reliable algorithm-based predictions.

The outcomes of this study can be useful to the stakeholders for the preparedness and effective allocation of fire-retarding resources and manpower to wildfire-prone areas. Furthermore, vulnerability assessment of wildfire in Sikkim can be performed based on this study by overlaying socioeconomic and environmental cost map on the wildfire likelihood map of Sikkim. Such studies can be very helpful in wildfire mitigation and land-use policies.

5 Conclusion

Applications of machine learning in geospatial analysis are progressively expanding. One of the prominent niches of this new branch of science is the predictive modelling of natural hazards. This study presents the wildfire prediction map of Sikkim Himalaya using four machine learning methods. These methods were run over the wildfire dataset involving several environmental features encompassing, meteorological, topographical, ecological, in situ and anthropogenic factors. The methods, namely Generalized Linear Model in the form of Logistic Regression, Radial Basis Function Kernel-based Support Vector Machine, Gradient Booster Method and Random Forest, are compared using model performance criteria. Amongst these, Random Forest computes the most accurate prediction followed by Gradient Booster Method. These methods produce high values of AUC, Accuracy, Kappa, Sensitivity, Specificity, Precision, F1 Score and Goodness of fit and low values of RMSE and MAE. These decision tree-based methods marginally outcompeted SVM and GLM.

Furthermore, it is concluded that meteorological factors like ambient temperature and wind speed over the dry season, as well as anthropogenic factors like proximity to roadways, are the most important descriptors of wildfires in Sikkim Himalaya. Most of the wildfires in Sikkim are prevalent in the low altitude valley areas of the south. These observations can be internalized into the wildfire mitigation policies towards the consequences
of slash and burn farming, use of fire to discourage entry of wildlife in settlements and traffic-induced wildfires. Also, long-term policy intervention can be prepared from this study regarding the impact of climate change-induced changes in the meteorological conditions of Sikkim Himalaya.

This study shows that machine learning can be combined with GIS to produce robust geospatial models of wildfire predictions. Machine learning can be a reliable wildfire management tool. Such a tool can be further improved by integrating online learning where the prediction model can have an incremental learning from a near real-time database like MODIS FIRMS. The methodology of this study can be further extended to include more in situ and meteorological factors into the feature space. Also, other artificial intelligence methods like ANN, evolutionary algorithms and agent-based learning can be applied to the wildfire dataset to generate better and reliable prediction maps. However, such studies need to trade-off between accuracy and interpretability.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11069-021-04973-6.

Acknowledgements I acknowledge the use of data and imagery from LANCE FIRMS operated by NASA’s EarthScience Data and Information System (ESDIS) with funding provided by NASA Headquarters.

Authors’ contributions Planning, data gathering, data processing, modelling and interpretation.

Funding None.

Availability of data and materials Yes. https://doi.org/10.6084/m9.figshare.12622100.v1

Code availability Yes. https://doi.org/10.6084/m9.figshare.12622100.v1

Declarations

Conflict of interest The author declare that they have no conflict of interest.

References

Al Janabi S, Al Shourbaji I, Salman MA (2018) Assessing the suitability of soft computing approaches for forest fires prediction. Appl Comput Inform 14(2):214–224. https://doi.org/10.1016/j.aci.2017.09.006

Ali SA, Parvin F, Vojteková J, Costache R, Linh NTT, Pham QB, Vojtek M, Gigović L, Ahmad A, Ghorbani MA (2021) GIS-based landslide susceptibility modeling: a comparison between fuzzy multi-criteria and machine learning algorithms. Geosci Front 12(2):857–876. https://doi.org/10.1016/j.gsf.2020.09.004

Andreas M (2013) Re: is random forest better than support vector machines? Retrieved from: https://www.researchgate.net/post/Is_random_forest_better_than_support_vector_machines/52b4159dd4c1185d468bb460d/citation/download

Argyrous G (2011) Statistics for research: with a guide to SPSS (3 edition). SAGE Publications Ltd

Arpaci A, Malowerschnig B, Sass O, Vacik H (2014) Using multi variate data mining techniques for estimating fire susceptibility of Tyrolean forests. Appl Geogr 53:258–270. https://doi.org/10.1016/j.apgeog.2014.05.015

Arrawatia ML, Tambe S (2011) Biodiversity of Sikkim: exploring and conserving a global Hotspot. Gangtok: Sikkim:Information and Public Relations Department. http://dspace.cus.ac.in/jspui/handle/1/3028

Arrawatia ML, Tambe S (2012) Climate Change in Sikkim—Patterns, Impacts and Initiatives. Information and Public Relations Department, Government of Sikkim. http://www.sikkimforest.gov.in/clima
te-change-in-sikkim/climate%20change%20in%20sikkim%20-%20patterns%20impacts%20and%20initiatives.htm
ASTERO Mount Gariwang image (2018) MOD13Q1.006 Terra Vegetation Indices 16-Day Global 250m; NASA EOSDIS Land Processes Distributed Active Archive Center (LP DAAC), USGS Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota. https://doi.org/10.5067/MODIS/MOD13Q1.006
Banerjee P (2021) Maximum entropy-based forest fire likelihood mapping: analysing the trends, distribution, and drivers of forest fires in Sikkim Himalaya. Scand J For Res. https://doi.org/10.1080/0287581.2021.1918239
Banerjee P, Ghose MK, Pradhan R (2020) Analytic hierarchy process based spatial biodiversity impact assessment model of highway broadening in Sikkim Himalaya. Geocarto Int 35(5):470–493. https://doi.org/10.1080/10106049.2018.1520924
Banks SC, Knight EJ, McBurney L, Blair D, Lindenmayer DB (2011) The effects of wildfire on mortality and resources for an arboreal marsupial: resilience to fire events but susceptibility to fire regime change. PLoS ONE 6(8). https://doi.org/10.1371/journal.pone.0022952
Behrooz F, Mariun N, Marhaban MH, Mohd Radzi MA, Ramli AR (2018) Review of control techniques for HVAC systems—nonlinearity approaches based on fuzzy cognitive maps. Energies 11(3):495. https://doi.org/10.3390/en11030495
Boutaba R, Salahuddin MA, Limam N, Ayoubi S, Shahriar N, Estrada-Solano F, Caicedo OM (2018) A comprehensive survey on machine learning for networking: evolution, applications and research opportunities. J Internet Serv Appl 9(1):16. https://doi.org/10.1186/s13174-018-0087-2
Bowl EJ, Banks SC, Strong CL, Lindenmayer DB (2019) Long-term impacts of wildfire and logging on forest soils. Nat Geosci 12(2):113–118. https://doi.org/10.1038/s41561-018-0294-2
Brownlee J (2016) Compare the performance of machine learning algorithms in R. Machine Learning Mastery. https://machinelearningmastery.com/compare-the-performance-of-machine-learning-algorithms-in-r/
Campos I, Vale C, Abrantes N, Keizer JJ, Pereira P (2015) Effects of wildfire on mercury mobilisation in eucalypt and pine forests. CATENA 131:149–159. https://doi.org/10.1016/j.catena.2015.02.024
Cascio WE (2018) Wildland fire smoke and human health. Sci Total Environ 624:586–595. https://doi.org/10.1016/j.scitotenv.2017.12.086
Chang K-T (2017) Introduction to geographic information systems (4 edition). McGraw Hill Education
Chirici G, Scotti R, Montaghi A, Barbati A, Cartisano R, Lopez G, Marchetti M, McRoberts RE, Olsson H, Corona P (2013) Stochastic gradient boosting classification trees for forest fuel types mapping through airborne laser scanning and IRS LISS-III imagery. Int J Appl Earth Obs Geoinf 25:87–97. https://doi.org/10.1016/j.jag.2013.04.006
COI (2011) Provisional population totals paper 1 of 2011: Sikkim [Office of the Registrar General & Census Commissioner, India Ministry of Home Affairs, Government of India]. http://censusindia.gov.in/2011-prov-results/prov_data_products_sikkim.html
Devisscher T, Anderson LO, Aragão LEOC, Galván L, Malhi Y (2016) Increased wildfire risk driven by climate and development interactions in the Bolivian Chiquitania. Southern Amazonia PLOS ONE 11(9):e0161323. https://doi.org/10.1371/journal.pone.0161323
Dong S, Chettri N, Sharma E (2017) Himalayan biodiversity: trans-boundary Conservation Institution and Governance. In: Dong S, Bandyopadhayaj, Chaturvedi S (eds) Environmental sustainability from the Himalayas to the Oceans: struggles and innovations in China and India (pp 127–143). Springer, Berlin. https://doi.org/10.1007/978-3-319-44037-8_6
Estes BL, Knapp EE, Skinner CN, Miller JD, Preisler HK (2017) Factors influencing fire severity under moderate burning conditions in the Klamath Mountains, northern California, USA. Ecosphere 8(5):e01794. https://doi.org/10.1002/ecs2.1794
Fann N, Alman B, Broome RA, Morgan GG, Johnston FH, Pouliot G, Rappold AG (2018) The health impacts and economic value of wildland fire episodes in the US: 2008–2012. Sci Total Environ 610–611:802–809. https://doi.org/10.1016/j.scitotenv.2017.08.024
Fick SE, Hijmans RJ (2017) WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. Int J Climatol 37(12):4302–4315. https://doi.org/10.1002/joc.5086
FIRMS (2020) Active fire data | Earthdata. https://earthdata.nasa.gov/earth-observation-data/near-real-time/firms/active-fire-data/
Flannigan MD, Harrington JB (1988) A study of the relation of meteorological variables to monthly provincial area burned by wildfire in Canada (1953–80). J Appl Meteorol 27(4):441–452. https://doi.org/10.1175/1520-0450(1988)027%3c0441:ASOTRO%3e2.0.CO;2
Flannigan MD, Stocks BJ, Wotton BM (2000) Climate change and forest fires. Sci Total Environ 262(3):221–229. https://doi.org/10.1016/S0048-9697(00)00524-6
García-Jimenez S, Jurio A, Pagola M, De Miguel L, Barrenechea E, Bustince H (2017) Forest fire detection: a fuzzy system approach based on overlap indices. Appl Soft Comput 52:834–842. https://doi.org/10.1016/j.asoc.2016.09.041

Géron A (2017) Hands-on machine learning with scikit-learn and TensorFlow: concepts, tools, and techniques to build intelligent systems (1 edition). O'Reilly Media

Gheshlaghi HA, Feizizadeh B, Blaschke T (2020) GIS-based forest fire risk mapping using the analytical network process and fuzzy logic. J Environ Planning Manage 63(3):481–499. https://doi.org/10.1080/09640568.2019.1594726

Ghorbanzadeh O, Kamran KV, Blaschke T (2019) Spatial prediction of wildfire susceptibility using global NASA MODIS fire products and machine learning approaches. https://uni-salzburg.eisvoum.com/en/publications/spatial-prediction-of-wildfire-susceptibility-using-global-nasa-m

Gillett NP, Weaver AJ, Zwiers FW, Flannigan MD (2004) Detecting the effect of climate change on Canadian forest fires. Geophys Res Lett. https://doi.org/10.1029/2004GL020876

Goleiji E, Hosseinli SM, Khorasani N, Monavari SM (2017) Forest fire risk assessment—an integrated approach based on multicriteria evaluation. Environmental Monitoring and Assessment 189(12):612. https://doi.org/10.1007/s10661-017-6225-7

Graham RT, McCaffrey S, Jain TB (2004) Science basis for changing forest structure to modify wildfire behavior and severity. Gen Tech Rep RMRS-GTR-120. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 43 p, 120. https://doi.org/10.2737/RMRS-GTR-120

Guo F, Wang G, Su Z, Liang H, Wang W, Lin F, Liu A (2016) What drives forest fire in Fujian, China? Evidence from logistic regression and Random Forests. Int J Wildland Fire 25(5):505–519. https://doi.org/10.1071/WF15121

Haque MK, Azad MAK, Hossain MY, Ahmed T, Uddin M, Hossain MM (2021) Wildfire in Australia during 2019–2020, its impact on health, biodiversity and environment with some proposals for risk management: a review. J Environ Prot 12(6):391–414. https://doi.org/10.4236/jep.2021.126024

Hastie T, Tibshirani R, Friedman J (2017) The elements of statistical learning: data mining, inference, and prediction, second edition (2nd ed. 2009, Corr. 9th printing 2017 edition). Springer

Hilton JE, Miller C, Sharples JJ, Sullivan AL (2017) Curvature effects in the dynamic propagation of wildfires. Int J Wildland Fire 25(12):1238–1251. https://doi.org/10.1071/WF16070

Hunt T (2020) ModelMetrics: rapid calculation of model metrics. https://CRAN.R-project.org/package=ModelMetrics

Jaafari A, Zenner KE, Pham BT (2018) Wildfire spatial pattern analysis in the Zagros Mountains, Iran: a comparative study of decision tree based classifiers. Eco Inform 43:200–211. https://doi.org/10.1016/j.ecoinf.2017.12.006

Jarvis A, Reuter HI, Nelson A, Guevara E (2008) Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m Database. http://srtm.cgiar.org

Jenks G (1967) The data model concept in statistical mapping. Int Yearbook Cartography 7:186–190

Jo MH, Lee MB, Lee SY, Jo YW, Baek SR (2000) The development of forest fire forecasting system using internet GIS and satellite remote sensing. 21st Asian Conference on Remote Sensing, Taipei, Taiwan, 1161–1166

Joseph S, Anitha K, Murthy MSR (2009) Forest fire in India: a review of the knowledge base. J for Res 14(3):127–134. https://doi.org/10.1007/s10310-009-0116-x

Keane RE, Karau E (2010) Evaluating the ecological benefits of wildfire by integrating fire and ecosystem simulation models. Ecol Model 221(8):1162–1172. https://doi.org/10.1016/j.ecolmodel.2010.01.008

Kim E, Jha MK, Kang M-W (2015a) A sensitivity analysis of critical genetic algorithm parameters: highway alignment optimization case study. Int J Oper Res Inf Syst 6(1):30–48. https://doi.org/10.4018/ijoris.20151010103

Kim SJ, Lim C-H, Kim GS, Lee J, Geiger T, Rahmati O, Son Y, Lee W-K (2019) Multi-temporal analysis of forest fire probability using socio-economic and environmental variables. Remote Sens 11(1):86. https://doi.org/10.3390/rs11010086

Kim T, Lim CH, Song C, Lee WK (2015) Estimation of wild fire risk area based on climate and maximum entropy in Korean Peninsula. AGU Fall Meeting Abstracts, 31, NH31A-1880

Krawchuk MA, Haire SL, Coop J, Parisien M-A, Whitman E, Chong G, Miller C (2016) Topographic and fire weather controls of fire refugia in forested ecosystems of northwestern North America. Ecosphere 7(12):e01632. https://doi.org/10.1002/ecs2.1632

Krueger E, Ochsner T, Engle D, Carlson JD, Twidwell D, Fuhlendorf S (2015) Soil moisture affects growing-season wildfire size in the southern great plains. Soil Sci Soc Am J. https://doi.org/10.2136/sssaj2015.01.0041
Kuhn M (2019) Variable importance. The caret Package. https://topepo.github.io/caret/variable-importance.html

Kuhn M (2020) caret: Classification and regression training. https://CRAN.R-project.org/package=caret

Kulig JC, Dabravolskaj J, Kulig JC, Dabravolskaj J (2019) The psychosocial impacts of wildland fires on children, adolescents and family functioning: a scoping review. Int J Wildland Fire 29(2):93–103. https://doi.org/10.1071/WF18063

Kumar P (2012) Assessment of impact of climate change on Rhododendrons in Sikkim Himalayas using Maxent modelling: limitations and challenges. Biodivers Conserv 21(5):1251–1266. https://doi.org/10.1007/s10531-012-0279-1

Laha A, Sinha R (2020) Forest fire risk assessment for sikkim using earth observation (EO) datasets and multi criteria decision making technique. 2020, NH033-0001

Lee DE (2018) Spotted owls and forest fire: a systematic review and meta-analysis of the evidence. Ecosphere 9(7):e02354. https://doi.org/10.1002/ecs2.2354

Leuenberger M, Parente J, Tonini M, Pereira MG, Kanevski M (2018) Wildfire susceptibility mapping: Deterministic vs. stochastic approaches. Environ Model Softw 101:194–203. https://doi.org/10.1016/j.envsoft.2017.12.019

Liaw A, Wiener M (2002) Classification and regression by randomForest. R News 2(3):18–22

Ljubomir G, Pamucar D, Drobnjak S, Pourghasemi HR (2019) Modeling the spatial variability of forest fire susceptibility using geographical information systems and the analytical hierarchy process. In: Pourghasemi HR, Gokceoglu C (eds) Spatial modeling in GIS and R for earth and environmental sciences (pp 337–369). Elsevier. https://doi.org/10.1016/B978-0-12-815226-3.00015-6

Lowe PO, Fiolliott PF, Dieterich JH, Patton DR (1978) Determining potential wildlife benefits from wildfire in Arizona Ponderosa Pine Forests. 18

Luo G, Stone BL, Johnson MD, Tarczy-Hornoch P, Wilcox AB, Mooney SD, Sheng X, Haug PJ, Nkoy FL (2017) Automating construction of machine learning models with clinical big data: proposal rationale and methods. JMI Res Protocols 6(8):175. https://doi.org/10.2196/resprot.7757

Massada AB, Syphard AD, Stewart SI, Radolff VC (2013) Wildfire ignition-distribution modelling: A comparative study in the Huron-Manistee National Forest, Michigan, USA. Int J Wildland Fire 22(2):183–193. https://doi.org/10.1071/WF11178

McCullagh P, Nelder JA (1989) Generalized linear models (2 edition). Chapman and Hall/CRC

Mhawej M, Faour G, Adjizian-Gerard J (2015) Wildfire likelihood’s elements: a literature review. Challenges 6(2):282–293. https://doi.org/10.3390/challe6020282

Mitchell T (1997) Machine learning (1st edition). McGraw-Hill Education

Murphy JM, Sexton DMH, Barnett DN, Jones GS, Webb MJ, Collins M, Stainforth DA (2004) Quantification of modelling uncertainties in a large ensemble of climate change simulations. Nature 430(7001):768–772. https://doi.org/10.1038/nature02771

Nami MH, Jaafari A, Fallah M, Nabiuni S (2018) Spatial prediction of wildfire probability in the Hypercanian ecoregion using evidential belief function model and GIS. Int J Environ Sci Technol 15(2):373–384. https://doi.org/10.1007/s13762-017-1371-6

Natekin A, Knoll A (2013) Gradient boosting machines, a tutorial. Front Neurorobot. https://doi.org/10.3389/fnbot.2013.00021

Ogutu JO, Piepho H-P, Schulz-Streeck T (2011) A comparison of random forests, boosting and support vector machines for genomic selection. BMC Proc 5(Suppl 3):S11. https://doi.org/10.1186/1753-6561-5-S3-S11

O’Neill AR (2019) Evaluating high-altitude Ramsar wetlands in the Eastern Himalayas. Global Ecol Conser 20: https://doi.org/10.1016/j.gecco.2019.e00715

Pastro LA, Dickman CR, Letnic M (2011) Burning for biodiversity or burning biodiversity? Prescribed burn vs. wildfire impacts on plants, lizards, and mammals. Ecol Appl 21(8):3238–3253. https://doi.org/10.1890/10-2351.1

Paul A, Khan ML, Arunachalam A, Arunachalam K (2005) Biodiversity and conservation of rhododendrons in Arunachal Pradesh in the Indo-Burma biodiversity hotspot. Curr Sci 89(4):623–634

Pausas JG, Keeley JE (2019) Wildfires as an ecosystem service. Front Ecol Environ 17(5):289–295. https://doi.org/10.1890/1808901301

Pham BT, Jaafari A, Avand M, Al-Ansari N, Du Dinh T, Yen HPH, Phong TV, Nguyen DH, Le HV, Mafi-Gholami D, Prakash I, Thi Thuy H, Tuyen TT (2020) Performance evaluation of machine learning methods for forest fire modeling and prediction. Symmetry 12(6):1022. https://doi.org/10.3390/sym12061022

Pham BT, Jaafari A, Phong TV, Yen HPH, Tuyen TT, Luong VV, Nguyen HD, Le HV, Foong LK (2021) Improved flood susceptibility mapping using a best first decision tree integrated with ensemble learning techniques. Geosci Front 12(3):101105. https://doi.org/10.1016/j.gsf.2020.11.003
Pourghasemi HR (2014) Re: how to interpret the negative and positive values of profile and plan curvature map in GIS?. Retrieved from: https://www.researchgate.net/post/How_to_interpret_the_negative_and_positive_values_of_profile_and_plan_curvature_map_in_GIS/53d206f4d2fd64b8118b464d/citation/download

Pradhan BK, Badola HK (2015) Swertia chirayta, a threatened high-value medicinal herb: microhabitats and conservation challenges in Sikkim Himalaya, India. Mountain Res Dev 35(4):374–381. https://doi.org/10.1659/MRd-JOURNAL-D-14-00034.1

Robin X, Turek N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M (2011) PROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinform 12:77

Regodic L, Gigovic G, Jakovljevic D, Sekulovic M (2018) GIS Multi-Criteria Analysis for Identifying and Mapping Forest Fire Hazard: Nevesinje, Bosnia and Herzegovina. Tehnicki Vjesnik - Technical Gazette 25(3):891–898

Sachdeva S, Bhatia T, Verma AK (2018) GIS-based evolutionary optimized gradient boosted decision trees for forest fire susceptibility mapping. Nat Hazards 92(3):1399–1418. https://doi.org/10.1007/s11069-018-3256-5

Sannigrahi S, Pillai F, Basu B, Basu AS, Sarkar K, Chakraborti S, Joshi PK, Zhang Q, Wang Y, Bhatt S, Bhatt A, Jha S, Keesstra S, Roy PS (2020) Examining the effects of forest fire on terrestrial carbon emission and ecosystem production in India using remote sensing approaches. Sci Total Environ 725:138331. https://doi.org/10.1016/j.scitotenv.2020.138331

Satir O, Berberoglu S, Donmez C (2016) Mapping regional forest fire probability using artificial neural network model in a Mediterranean forest ecosystem. Geomat Nat Haz Risk 7(5):1645–1658. https://doi.org/10.1080/19475705.2015.1084541

Sexton JO, Song X-P, Feng M, Noojipady P, Anand A, Huang C, Kim D-H, Collins KM, Channan S, DiMiceli C, Townshend JR (2013) Global, 30-m resolution continuous fields of tree cover: Landsat-based rescaling of MODIS vegetation continuous fields with lidar-based estimates of error. Int J Digit Earth 6(5):427–448. https://doi.org/10.1080/17538947.2013.17538947.10178646

Sharma K, Thapa G (2021) Analysis and interpretation of forest fire data of Sikkim. For Soc 261–276

Sharma S, Joshi V, Chhetri R (2014) Forest fire as a potential environmental threat in recent years in Sikkim, Eastern Himalayas, India. Clim Change Environ Sustain 2:55. https://doi.org/10.1007/s11069-020-00001-9

Shimada M, Itoh T, Motooka T, Watanabe M, Shiraishi T, Thapa R, Lucas R (2014) New global forest/non-forest maps from ALOS PALSAR data (2007–2010). Remote Sens Environ 155:13–31. https://doi.org/10.1016/j.rse.2014.09.029

Shukla A, Garg PK, Srivastava S (2018) Evolution of glacial and high-altitude lakes in the Sikkim, Eastern Himalaya Over the Past Four Decades (1975–2017). Frontiers in Environmental Science, 6. https://doi.org/10.1007/j.01616-016-00081-X

Taylor SW, Woolford DG, Dean CB, Martell DL (2013) Wildfire prediction to inform fire management: statistical science challenges. Stat Sci 28(4):586–615. https://doi.org/10.1214/13-STS5451

Tehrany MS, Jones S, Shabani F, Martínez-Álvarez F, Tien Bui D (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. Theoret Appl Climatol 137(1):637–653. https://doi.org/10.1007/s00704-019-2628-9

Tien Bui D, Hoang N-D, Samui P (2019) Spatial pattern analysis and prediction of forest fire using new machine learning approach of multivariate adaptive regression splines and differential flower pollination optimization: a case study at Lao Cai province (Viet Nam). J Environ Manage 237:476–487. https://doi.org/10.1016/j.jenvman.2019.01.010

Tien Bui D, Le K-TT, Nguyen VC, Le HD, Revhaug I (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. https://doi.org/10.3390/rs8040347

Tien Bui D, Le HV, Hoang N-D (2018) GIS-based spatial prediction of tropical forest fire danger using a new hybrid machine learning method. Eco Inform 48:104–116. https://doi.org/10.1016/j.ecoinf.2018.08.008

Tien Bui D, Le K-TT, Nguyen VC, Le HD, Revhaug I (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. https://doi.org/10.3390/rs8040347

Hengl T, Wheeler I (2018) Soil organic carbon content in x 5 g / kg at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution. Zenodo. https://doi.org/10.5281/zenodo.2542735

Hengl T, Gupta S (2019) Soil water content (volumetric %) for 33kPa and 1500kPa suctions predicted at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution. Zenodo. https://doi.org/10.5281/zenodo.2784001
Trouvé R, Oborne L, Baker PJ (2021) The effect of species, size, and fire intensity on tree mortality within a catastrophic bushfire complex. Ecol Appl n/a(n/a): e02383. https://doi.org/10.1002/eap.2383

Tuyen TT, Jasafari A, Yen HPH, Nguyen-Thoi T, Phong TV, Nguyen HD, Van Le H, Phuong TTM, Nguyen SH, Prakash I, Pham BT (2021) Mapping forest fire susceptibility using spatially explicit ensemble models based on the locally weighted learning algorithm. Eco Inform 63:101292. https://doi.org/10.1016/j.ecoinf.2021.101292

Vicars WC, Sickman JO, Ziemann PJ (2010) Atmospheric phosphorus deposition at a montane site: Size distribution, effects of wildfire, and ecological implications. Atmos Environ 44(24):2813–2821. https://doi.org/10.1016/j.atmosenv.2010.04.055

Vilar L, Gómez I, Martínez-Vega J, Echavarría P, Riaño D, Martín MP (2016) Multitemporal modelling of socio-economic wildfire drivers in central Spain between the 1980s and the 2000s: comparing generalized linear models to machine learning algorithms. PLoS ONE 11(8):e0161344. https://doi.org/10.1371/journal.pone.0161344

Williams AP, Abatzoglou JT, Gershunov A, Guzman-Morales J, Bishop DA, Balch JK, Lettenmaier DP (2019) Observed impacts of anthropogenic climate change on wildfire in California. Earth’s Future 7(8):892–910. https://doi.org/10.1029/2019EF001210

Xie Y, Peng M (2019) Forest fire forecasting using ensemble learning approaches. Neural Comput Appl 31(9):4541–4550. https://doi.org/10.1007/s00521-018-3515-0

Yathish H, Athira KV, Preethi K, Pruthviraj U, Shetty A (2019) A comparative analysis of forest fire risk zone mapping methods with expert knowledge. J Indian Soc Remote Sens 47(12):2047–2060. https://doi.org/10.1007/s12524-019-01047-w

Zhang G, Wang M, Liu K (2019) Forest Fire Susceptibility modeling using a convolutional neural network for Yunnan Province of China. Int J Disaster Risk Sci 10(3):386–403. https://doi.org/10.1007/s13753-019-00233-1

Zhang G, Wang M, Liu K (2021) Deep neural networks for global wildfire susceptibility modelling. Ecol Ind 127:107735. https://doi.org/10.1016/j.ecolind.2021.107735

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.