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

Polash Banerjee (banerjee.polash@gmail.com)
Sikkim Manipal University of Health Medical and Technological Sciences: Sikkim Manipal University
https://orcid.org/0000-0002-2187-9347

Research Article

Keywords: Forest fire, Prediction map, algorithm, statistical learning, GIS

Posted Date: August 31st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-750123/v1

Version of Record: A version of this preprint was published at Natural Hazards on August 15th, 2021. See the published version at https://doi.org/10.1007/s11069-021-04973-6.

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

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, 14 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 (GLM), Support Vector Machine (SVM), 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 & 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. 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 & Tambe, 2012, Sharma & Thapa, 2021). ecosystem (Banerjee et al., 2020, Dong et al., 2017; Kumar, 2012; Pradhan & 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 & Tambe, 2012; 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 (S. Sharma, Joshi, and Chhetri 2014). To date hardly any study has 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), mobilize 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 in the volatilization of soil nutrients and soil erosion; 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 & 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 & Keeley, 2019). In some cases, mixed-severity wildfires can be beneficial to some species. For instance, the wildfire induced opening of habitat patches have promoted foraging by the spotted owls (*Strix occidentalis*) and with 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 & Harrington, 1988; Guo et al., 2016; Jaafari et al., 2018; T. Kim et al., 2015; 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. 2020; Gheshlaghi, Feizizadeh and Blaschke 2020; Goleiji 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 is 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., 2016a). 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; S. J. Kim et al., 2019; Leuenberger et al., 2018; Massada et al., 2013; Sachdeva et al., 2018; Tehrany et al., 2019; Xie & 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., 2016b). 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 & 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, Garg and Srivastava 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 characterised 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 metres, Eastern Himalayan subalpine conifer forests grow from 3500 to 5000 metres; and Eastern Himalayan alpine shrub and meadows in the higher elevations (O'Neill 2019; O'Neill et al. 2020). 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 (**Figure 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

196 anthropogenic features (Devisscher et al., 2016; Ghorbanzadeh et al., 2019; Joseph et al., 2009)
197 (**Table 1**).

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

220 deforestation for firewood and timber. Also, population density partly explains illegal forest
221 exploitation due to unemployment (Sharma et al., 2014).

222 The target variable in this study was the presence or absence of wildfires. The historical
223 wildfires dataset of Sikkim Himalaya was prepared using two data sources. The ground
224 truthing-based dataset was procured from the Forest Environment and Wildlife Management
225 Department (FEWMD), Govt. of Sikkim. The timeframe of FEWMD ranged from 2016 to
226 2018. The remote sensing-based fire events dataset was accessed from the Moderate Resolution
227 Imaging Spectroradiometer (MODIS) and Visible Infrared Imaging Radiometer Suite (VIIRS)
228 data archive at the Fire Information for Resource Management System (FIRMS) site (FIRMS,
229 2020). The coarse resolution of 1km of MODIS is complemented by the relatively finer
230 resolution of 375m of VIIRS. The fire dataset of Sikkim Himalaya accessed from FIRMS
231 ranged in the timeframe from 2000 to 2019. However, the dataset of FIRMS does not
232 distinguish wildfire from other sources of fire. To extract the wildfires from otherwise, a
233 historical forest coverage raster was prepared by the union of all forest extents in the timeframe
234 from 2007 to 2010 (Shimada et al., 2014) (**Supplement Figure 1**). The FIRMS fire dataset was
235 intersected with the forest cover raster to include the fire events restricted within the forest
236 cover area. In this way, only the wildfires were identified from the FIRMS dataset.
237 Furthermore, to prevent double-counting from the data sources, wildfires of FIRMS within 200
238 m proximity from FEWMD were dropped from the dataset. The ground truthing-based dataset
239 and modified FIRMS dataset were combined to prepare the final wildfire dataset of 754 events.

240 All the feature rasters and the wildfire dataset were projected from the geographic projection
241 system of GCS-WGS-1984 to the plane projection system of WGS-1984-UTM-Zone-45N in
242 the GIS framework. The latter provides an appropriate measurement in the metric system for
243 India. The cell size and extent of all the feature rasters were standardised to be the same. Next,
244 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):

$$z_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (1)$$

where the pixel value z_i is the normalized value of a raster ranging from 0 to1 at the *i*th location of the study area extent. x_i is the pixel value of the *i*th location before the normalization. $x_{min}$is the lowest value, while $x_{max} - x_{min}$ 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, Generalised 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 characterised 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 & Nelder, 1989):

$$E(Y) = g^{-1}(X\beta) \tag{2}$$

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 β. 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:

$$P(y|x) = \frac{1}{1 + exp(-\beta^T x)} \tag{3}$$

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

$$c(\beta) = \begin{cases} -log\big(P(y|x)\big) & if\ y = 1 \\ -log\big(1 - P(y|x)\big) & if\ y = 0 \end{cases} \tag{4}$$

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

$$j(\beta) = -\frac{1}{M}\sum_{i=1}^{M}\left[y^{(i)} log\big(P^{(i)}\big) + \big(1 - y^{(i)}\big) log\big(1 - P^{(i)}\big)\right] \tag{5}$$

where M is the size of the training dataset. $P^{(i)} = P\big(y^{(i)}|x^{(i)}\big)$ 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 β 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:

$$f(x) = h(x)^T\beta + \beta_0 \tag{6}$$

Where $h(x)^T$ is a vector of basis functions defined as $(h_1(x_i),\ h_2(x_i), \dots, h_M(x_i))^T$ over *M* features and *N* instances. A basis function explains the nonlinear relationship of the feature vector x_i with the regression model $f(x)$. β is a vector of coefficients of the corresponding basis function and β_0 is the intercept. The decision boundary that maximizes the margin of the partition was considered for regression. Eq(6) can be amended to include a kernel trick *K(·)* that solves for the hypothesis $\hat{f}(x)$ by using the Lagrange dual function:

$$\hat{f}(x) = \sum_{i=1}^{N} \hat{\alpha}_i y_i K(x, x_i) + \hat{\beta}_0 \tag{7}$$

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 $\langle h(x), h(x_i)\rangle$. 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:

$$K(x, x') = exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2}\right) \tag{8}$$

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 σ parameter of RBF controls the Gaussian distribution of RBF. The hypothesis was directed towards the solution by minimizing a cost function:

$$\sum_{i=1}^{N}[1 - y_i f(x_i)] + \frac{\lambda}{2}\|\beta\|^2 \quad (9)$$

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 λ 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 $R_m,\ m \in (1, \cdots, M)$ 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:

$$f(x) = \{\bar{y}_m | x \in R_m\} \quad (10)$$

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 Θ as the parameterised form of the expected loss function. Furthermore, a loss function $L\big(y_i, f(x_i)\big)$ 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 $T(x; \Theta_t)$ from

343 negative of the gradient of the loss function g_{im} for the *i*th instance in the *m*th region at each
344 iterative stage:

$$\Theta_m = \sum_{i=1}^{N} \left(-g_{im} - T(x_i; \Theta)\right)^2 \quad (11)$$

345 Where,

$$g_{im} = \left[\frac{\partial L\left(y_i, f(x_i)\right)}{\partial f(x_i)}\right]_{f(x_i)=f_{m-1}(x_i)} \quad (12)$$

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

$$f_m(x) \leftarrow f_{m-1}(x) + \rho_i T(x_i; \Theta) \quad (13)$$

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

$$f_M(x) = \sum_{i=1}^{M} T(x; \Theta_i) \quad (14)$$

351 Thus, GBM starts the iterative ensemble process by initially constructing a weak learner
352 regression tree and compares it with the target. This is followed by the stepwise construction
353 of another such tree. But this time the focus of the algorithm is more towards the residual error.
354 This is achieved by giving more weightage to the residual error. In this way, the progressive
355 learner becomes relatively better than its predecessor. A loss function is used to track the
356 learning success of the trees. GBM fits the residual with the steepest gradient of the loss
357 function to boost the learning process. Eventually, as the learning process approaches a certain
358 tolerance level, a weighted sum of all the weak learners yields a strong learner for testing and
359 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 & 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 minimises the impurity or cost function. The feature space is dichotomised 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:

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

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 Θ. 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):

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

$$MAE\ (X,f) = \frac{1}{m_i}\sum_{i=1}^{m}\left|f\left(x^{(i)}\right) - y^{(i)}\right| \tag{17}$$

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:

$$OAC = \frac{TP + TN}{TP + FP + TN + FN} \tag{16}$$

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:

$$sensitivity = \frac{TP}{TP + FN} \tag{17}$$

$$specificity = \frac{TN}{TN + FP} \tag{18}$$

$$precision = \frac{TP}{TP + FP} \tag{19}$$

$$kappa = \frac{P_{obs} - P_{exp}}{1 - P_{exp}} \tag{20}$$

$$F1\ Score = 2 \times \left(\frac{precision \times sensitivity}{precision + sensitivity}\right) \tag{21}$$

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).

417 A comparative analysis of the performance of the machine learning methods was done using
418 the box and whisker plot and Scatter plot Matrices. To do this, 30 samples from the dataset
419 were selected using the cross-validation method. These plots provided visualization of the
420 similarities between the performances of the machine learning methods.

421 **2.9.Wildfire likelihood model**

422 In this study, several r-packages were used for data pre-processing and machine learning (Hunt,
423 2020; Kuhn, 2020; Liaw & Wiener, 2002; Robin et al., 2011). These packages were used for
424 training and testing the machine learning algorithms. Furthermore, machine learning
425 algorithms also provided the importance of the feature variables used in the model. Features
426 with no importance to the model were dropped from the dataset and algorithms were rerun.
427 This was followed by stacking of the feature rasters. The stack was used by the machine
428 learning algorithms for predicting the wildfire probability over the entire study area. The matrix
429 of predicted values of the entire study area was exported as a GeoTiff raster, that stores the
430 predicted values along with their respective latitude and longitude values. The rasters of
431 predicted values of wildfire were imported into the GIS framework for further analysis that
432 included the categorization of the raster into areas of very low, low, medium, high and very
433 high probability of wildfire (**Figure 2**).

434 **3. Results**

435 The wildfire inventory dataset was used to analyse whether wildfires were on the rise in Sikkim
436 Himalaya. Time series data of wildfires over the years 2000 to 2019 indicated that there was a
437 growing trend. The picture became clearer by forecasting the wildfires using Holt's forecast
438 model. The model predicted that the wildfires were likely to increase from 82 in 2019 to 96 in
439 2022 with an uncertainty of ± 62.343 events (**Figure 3**). To identify the relevant environmental
440 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 (**Figure 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^{o} -220^{o}, covering southeast to southwest direction. Also, wildfires were more common over steep slopes (29^{o} - 32^{o}). 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, 2013; 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^{o} - 24^{o}). 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 (400m – 630m away from water bodies). Regarding the in-situ features, wildfire events were mostly confined to low carbon content 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 Figure 2, Figure 5)**.

3.3.Model summary and model performance

3.3.1. Generalized linear model

The GLM model was run with 14 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. 10-fold 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 wildfire in Sikkim Himalaya. Also, features such as proximity to water bodies, slope and average ambient temperature were partly accountable for wildfires. Interestingly, distance from human habitations had an inverse effect on wildfire occurrences. Low soil carbon and drier soil promoted wildfires. Also, low tree cover encouraged the chances of wildfires. 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. 10-fold 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 (σ), epsilon (ε) and cost C settled at 0.106, 0.1 and 1, respectively. The ε tunes SVM by determining the number of support vectors to be considered for regression. C, which is similar to λ 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 Figure 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 '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 Figure 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 Figure 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 (**Figure 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 (**Figure 7**). The ROC curves indicated that all the models performed satisfactorily, while RF outperformed all the other models (**Figure 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 (**Figure 9**).

3.5. Wildfire prediction maps

All the feature maps were projected to the plane coordinate system of WGS-1984-UTM-Zone-45N, 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 (**Figure 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 (**Figure 10b**). GBM devoted more areas to wildfire than GLM and SVM (**Figure 10c**). RF gave more importance to valley areas than GBM, although the spatial distribution of wildfire probability showed high similarity with GBM (**Figure 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 (**Figure 11**). Compared to high likelihood categories, very high likelihood of wildfire category had a relatively larger area (**Figure 12**)

4. Discussion

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

562 **4.1.Comparison between machine learning methods and their implications**

563 In this study instead of just one algorithm, four algorithms were considered. This was mainly
564 to identify the algorithm that performs best in the wildfire prediction out of the popular
565 algorithms considered. Contrary to previous studies, RF outperformed other machine learning
566 methods in wildfire predictions (Ogutu et al., 2011; Tehrany et al., 2019; Xie & Peng, 2019).
567 This observation was in harmony with studies performed by other authors (Guo et al., 2016; E.
568 Kim et al., 2015; Massada et al., 2013). The better performance of RF in comparison to GLM
569 and SVM can be because RF uses the ensemble method of learning instead of linear or kernel-
570 based learning. In the ensemble method, the average output of several decision trees is
571 considered. This process increases the chances of correct prediction. Also, contrary to SVM,
572 RF is good at handling datasets with many outliers (Andreas, 2013). As observed from the
573 dataset, the histograms of several environmental features in this study were skewed. Perhaps,
574 this was another reason for the better performance of RF.

575 The comparative analysis of the models was based on samples extracted from the wildfire
576 dataset through the cross-validation method. It showed that GLM had a much wider range of
577 accuracy and kappa values. This can be explained by the limited number of feature variables
578 considered in the GLM model in comparison to other models. Furthermore, the smallest range
579 of MAE of GLM indicated that the possible reason for the wide ranges of accuracy and kappa
580 values can be due to a large set of outliers in the wildfire dataset (Géron, 2017). The higher
581 correlation between GLM and SVM as well as that of GBM and RF in terms of accuracy and

582 MAE showed that out of these pairs of models only one should be considered while making an
583 ensemble of models to improve the prediction capacity of wildfire events (Brownlee, 2016).

584 **4.2.Importance of feature variables**

585 Consistent with previous studies, this study suggested that meteorological features like wind
586 speed and to some extent ambient temperature were important determinants of wildfires. The
587 low wind speed and warm temperature of the valley areas are features of sub-tropical Sal and
588 Oak deciduous forests prone to wildfires in Sikkim Himalaya. The anthropogenic feature like
589 distance from the roadways on average was the strongest predictor of wildfires. To lesser
590 extent proximity to human habitations also contributed to the predictions of wildfires. These
591 observations second the previous studies on wildfires of Sikkim (Sharma et al., 2014). The
592 ecological feature like the fraction of tree cover and in-site features like soil carbon were
593 moderate predictors. Compared to other features, topographical features were not very good
594 predictors of wildfires (Arpaci et al., 2014; Estes et al., 2017; Flannigan & Harrington, 1988;
595 Guo et al., 2016; Jaafari et al., 2018; T. Kim et al., 2015; Ljubomir et al., 2019; Sachdeva et
596 al., 2018; Tien Bui et al., 2019; Yathish et al., 2019). Contrary to the MCDA-based study,
597 namely using AHP, on forest fire risk zones of Sikkim (Laha et al., 2020), the present study
598 gave limited importance to the aspect, except for the SVM model. However, indirect measures
599 of human population density, namely, proximity to human settlements and roadways supported
600 the observations of Laha et al. (2020). Like the observations by Banerjee (2021), this study
601 showed that proximity to roadways was the most important determinant of wildfire in Sikkim.
602 However, contrary to Banerjee (2021) average wind speed has been given more weight in this
603 study than average ambient temperature. Looking at the correlation matrix these two
604 meteorological variables had a significant negative correlation. However, their collinearity in
605 terms of VIF was within acceptable limits. Thereby, they were considered as independent

606 variables in this study. Furthermore, tree cover fraction has been considered as an important
607 factor in wildfire prediction in both the studies.

608 **4.3.Future risks of wildfire**

609 The study showed that wildfires were predominantly distributed in the lower altitudes and
610 valley areas of Sikkim Himalaya. Few observations can be made about these areas. The
611 meteorological conditions of these areas were identified as having relatively warmer ambient
612 temperatures and low wind speed. Also, the road network of these areas closely follows the
613 river network. Steep slope facing southeast to southwest aspect with low TWI explained most
614 of the wildfires of these areas (Graham et al., 2004; Jo et al., 2000; Mhawej et al., 2015). Low
615 soil carbon and water content areas had more incidents of wildfire. The role of human activities
616 in the occurrence of wildfires was evident from the study. These observations were similar to
617 previous studies (Arpaci et al., 2014; S. J. Kim et al., 2019). However, contrary to previous
618 studies, proximity to settlements as a feature had a contradictory role in this study as the bulk
619 of the wildfires were on average 2.5 km away from the human habitations (S. J. Kim et al.,
620 2019; Massada et al., 2013; Nami et al., 2018; Vilar et al., 2016). This may be since the land
621 use around the settlements was mainly non-forest lands like agrarian or fallow lands. Thereby
622 areas of Sikkim bordering the state of West Bengal, district borders of West and South Sikkim
623 and populated valleys of North Sikkim are more prone to wildfires.

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

628 This study is probably the first attempt to systematically prepare the WLM of Sikkim Himalaya
629 using multiple machine learning models. In line with studies done in other locations, this study

630 indicated that anthropogenic and meteorological factors were the most prominent descriptors
631 of wildfires. Also, this study highlighted that machine learning methods were reliable means
632 of preparing hazard maps. However, the reliability of the predictions heavily depends on the
633 wildfire inventory. This can be achieved by pruning instances with incorrect target variable or
634 incomplete instances. Usually, a large and representative inventory leads to better predictions.
635 Also, the engineering of features like normalization and removal of multicollinear features are
636 essential steps for dataset preparation. Regarding the choice of algorithms, consideration of the
637 nature of the dataset, in terms of whether the target variable is binomial, multinomial,
638 categorical, or continuous is important. Moreover, skewness of the features has an important
639 role in the choice of machine learning methods. Cross-validation and choice of
640 hyperparameters for the regularization are essential steps towards reliable algorithm-based
641 predictions.

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

647 ## 5. Conclusion

648 Applications of machine learning in geospatial analysis is progressively expanding. One of the
649 prominent niches of this new branch of science is the predictive modelling of natural hazards.
650 This study presents the wildfire prediction map of Sikkim Himalaya using four machine
651 learning methods. These methods were run over the wildfire dataset involving several
652 environmental features encompassing, meteorological, topographical, ecological, in-situ and
653 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.

References

679 Al_Janabi, S., Al_Shourbaji, I., & Salman, M. A. (2018). Assessing the suitability of soft
680 computing approaches for forest fires prediction. Applied Computing and Informatics,
681 14(2), 214–224. https://doi.org/10.1016/j.aci.2017.09.006

682 Ali, S. A., Parvin, F., Vojteková, J., Costache, R., Linh, N. T. T., Pham, Q. B., Vojtek, M.,
683 Gigović, L., Ahmad, A., & Ghorbani, M. A. (2021). GIS-based landslide susceptibility
684 modeling: A comparison between fuzzy multi-criteria and machine learning algorithms.
685 Geoscience Frontiers, 12(2), 857–876. https://doi.org/10.1016/j.gsf.2020.09.004

686 Andreas, M. (2013). Re: Is random forest better than support vector machines? Retrieved from:
687 https://www.researchgate.net/post/Is_random_forest_better_than_support_vector_mach
688 ines/52b4159dd4c1185d468b460d/citation/download.

689 Argyrous, G. (2011). Statistics for Research: With a Guide to SPSS (3 edition). SAGE
690 Publications Ltd.

691 Arpaci, A., Malowerschnig, B., Sass, O., & Vacik, H. (2014). Using multi variate data mining
692 techniques for estimating fire susceptibility of Tyrolean forests. Applied Geography, 53,
693 258–270. https://doi.org/10.1016/j.apgeog.2014.05.015

694 Arrawatia, M. L., & Tambe, S. (2011). Biodiversity of Sikkim: Exploring and Conserving a
695 Global Hotspot. Gangtok: Sikkim:Information and Public Relations Department.
696 http://dspace.cus.ac.in/jspui/handle/1/3028

697 ASTER Mount Gariwang image. (2018). MOD13Q1.006 Terra Vegetation Indices 16-Day
698 Global 250m; NASA EOSDIS Land Processes Distributed Active Archive Center (LP
699 DAAC). USGS Earth Resources Observation and Science (EROS) Center, Sioux Falls,
700 South Dakota. https://doi.org/10.5067/MODIS/MOD13Q1.006

701 Banerjee, P. (2021). Maximum entropy-based forest fire likelihood mapping: Analysing the
702 trends, distribution, and drivers of forest fires in Sikkim Himalaya. Scandinavian Journal
703 of Forest Research, 0(0), 1–14. https://doi.org/10.1080/02827581.2021.1918239

704 Banerjee, P., Mrinal K. Ghose, M.K. & Pradhan, R. (2020) Analytic hierarchy process based
705 spatial biodiversity impact assessment model of highway broadening in Sikkim
706 Himalaya, Geocarto International, 35:5, 470-493, DOI:
707 10.1080/10106049.2018.1520924

708 Banks, S. C., Knight, E. J., McBurney, L., Blair, D., & Lindenmayer, D. B. (2011). The Effects
709 of Wildfire on Mortality and Resources for an Arboreal Marsupial: Resilience to Fire
710 Events but Susceptibility to Fire Regime Change. PLoS ONE, 6(8).
711 https://doi.org/10.1371/journal.pone.0022952

712 Behrooz, F., Mariun, N., Marhaban, M. H., Mohd Radzi, M. A., & Ramli, A. R. (2018). Review
713 of Control Techniques for HVAC Systems—Nonlinearity Approaches Based on Fuzzy
714 Cognitive Maps. Energies, 11(3), 495. https://doi.org/10.3390/en11030495

715 Boutaba, R., Salahuddin, M. A., Limam, N., Ayoubi, S., Shahriar, N., Estrada-Solano, F., &
716 Caicedo, O. M. (2018). A comprehensive survey on machine learning for networking:

Evolution, applications and research opportunities. Journal of Internet Services and Applications, 9(1), 16. https://doi.org/10.1186/s13174-018-0087-2

Bowd, E. J., Banks, S. C., Strong, C. L., & Lindenmayer, D. B. (2019). Long-term impacts of wildfire and logging on forest soils. Nature Geoscience, 12(2), 113–118. https://doi.org/10.1038/s41561-018-0294-2

Brownlee, J. (2016, February 25). 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, J. J., & 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, W. E. (2018). Wildland fire smoke and human health. Science of The Total Environment, 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, R. E., Olsson, H., & Corona, P. (2013). Stochastic gradient boosting classification trees for forest fuel types mapping through airborne laser scanning and IRS LISS-III imagery. International Journal of Applied Earth Observation and Geoinformation, 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, L. O., Aragão, L. E. O. C., 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 S. Dong, J. Bandyopadhyay, & S. Chaturvedi (Eds.), Environmental Sustainability from the Himalayas to the Oceans: Struggles and Innovations in China and India (pp. 127–143). Springer International Publishing. https://doi.org/10.1007/978-3-319-44037-8_6

Estes, B. L., Knapp, E. E., Skinner, C. N., Miller, J. D., & Preisler, H. K. (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, R. A., Morgan, G. G., Johnston, F. H., Pouliot, G., & Rappold, A. G. (2018). The health impacts and economic value of wildland fire episodes in the

U.S.: 2008–2012. Science of The Total Environment, 610–611, 802–809. https://doi.org/10.1016/j.scitotenv.2017.08.024

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology, 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, M. D., & Harrington, J. B. (1988). A Study of the Relation of Meteorological Variables to Monthly Provincial Area Burned by Wildfire in Canada (1953–80). Journal of Applied Meteorology, 27(4), 441–452. https://doi.org/10.1175/1520-0450(1988)027<0441:ASOTRO>2.0.CO;2

Flannigan, M. D., Stocks, B. J., & Wotton, B. M. (2000). Climate change and forest fires. Science of The Total Environment, 262(3), 221–229. https://doi.org/10.1016/S0048-9697(00)00524-6

Garcia-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. Applied Soft Computing, 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, H. A., Feizizadeh, B., & Blaschke, T. (2020). GIS-based forest fire risk mapping using the analytical network process and fuzzy logic. Journal of Environmental Planning and Management, 63(3), 481–499. https://doi.org/10.1080/09640568.2019.1594726

Ghorbanzadeh, O., Kamran, K. V., & Blaschke, T. (2019). Spatial Prediction of Wildfire Susceptibility Using Global NASA MODIS Fire Products and Machine Learning Approaches. https://uni-salzburg.elsevierpure.com/en/publications/spatial-prediction-of-wildfire-susceptibility-using-global-nasa-m

Gillett, N. P., Weaver, A. J., Zwiers, F. W., & Flannigan, M. D. (2004). Detecting the effect of climate change on Canadian forest fires. Geophysical Research Letters, 31(18). https://doi.org/10.1029/2004GL020876

Graham, R. T., McCaffrey, S., & Jain, T. B. (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. International Journal of Wildland Fire, 25(5), 505–519. https://doi.org/10.1071/WF15121

Haque, M. K., Azad, M. A. K., Hossain, M. Y., Ahmed, T., Uddin, M., & Hossain, M. M. (2021). Wildfire in Australia during 2019-2020, Its Impact on Health, Biodiversity and Environment with Some Proposals for Risk Management: A Review. Journal of Environmental Protection, 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, J. E., Miller, C., Sharples, J. J., & Sullivan, A. L. (2017). Curvature effects in the dynamic propagation of wildfires. International Journal of 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, E. K., & Pham, B. T. (2018). Wildfire spatial pattern analysis in the Zagros Mountains, Iran: A comparative study of decision tree based classifiers. Ecological Informatics, 43, 200–211. https://doi.org/10.1016/j.ecoinf.2017.12.006

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

Jenks, G. (1967). The Data Model Concept in Statistical Mapping. International Yearbook of Cartography, 7, 186–190.

Jo, M. H., Lee, M. B., Lee, S. Y., Jo, Y. W., & Baek, S. R. (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, M. S. R. (2009). Forest fire in India: A review of the knowledge base. Journal of Forest Research, 14(3), 127–134. https://doi.org/10.1007/s10310-009-0116-x

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

Kim, E., Jha, M. K., & Kang, M.-W. (2015). A Sensitivity Analysis of Critical Genetic Algorithm Parameters: Highway Alignment Optimization Case Study. International Journal of Operations Research and Information Systems (IJORIS), 6(1), 30–48. https://doi.org/10.4018/ijoris.2015010103

Kim, S. J., Lim, C.-H., Kim, G. S., 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 Sensing, 11(1), 86. https://doi.org/10.3390/rs11010086

827 Kim, T., Lim, C. H., Song, C., & Lee, W. K. (2015). Estimation of Wild Fire Risk Area based
828 on Climate and Maximum Entropy in Korean Peninsular. AGU Fall Meeting Abstracts,
829 31, NH31A-1880.

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

834 Krueger, E., Ochsner, T., Engle, D., Carlson, J. D., Twidwell, D., & Fuhlendorf, S. (2015). Soil
835 Moisture Affects Growing-Season Wildfire Size in the Southern Great Plains. Soil
836 Science Society of America Journal, 79. https://doi.org/10.2136/sssaj2015.01.0041

837 Kuhn, M. (2019). 15 Variable Importance. The caret Package.
838 https://topepo.github.io/caret/variable-importance.html

839 Kuhn, M. (2020). caret: Classification and Regression Training. https://CRAN.R-
840 project.org/package=caret

841 Kulig, J. C., Dabravolskaj, J., Kulig, J. C., & Dabravolskaj, J. (2019). The psychosocial impacts
842 of wildland fires on children, adolescents and family functioning: A scoping review.
843 International Journal of Wildland Fire, 29(2), 93–103. https://doi.org/10.1071/WF18063

844 Kumar, P. (2012). Assessment of impact of climate change on Rhododendrons in Sikkim
845 Himalayas using Maxent modelling: Limitations and challenges. Biodiversity and
846 Conservation, 21(5), 1251–1266. https://doi.org/10.1007/s10531-012-0279-1

847 Laha, A., Sinha, R., & B, N. (2020). Forest Fire Risk Assessment for Sikkim using Earth
848 Observation (EO) Datasets and Multi Criteria Decision Making Technique. 2020,
849 NH033-0001.

850 Lee, D. E. (2018). Spotted Owls and forest fire: A systematic review and meta-analysis of the
851 evidence. Ecosphere, 9(7), e02354. https://doi.org/10.1002/ecs2.2354

852 Leuenberger, M., Parente, J., Tonini, M., Pereira, M. G., & Kanevski, M. (2018). Wildfire
853 susceptibility mapping: Deterministic vs. stochastic approaches. Environmental
854 Modelling & Software, 101, 194–203. https://doi.org/10.1016/j.envsoft.2017.12.019

855 Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. R News, 2(3),
856 18–22.

857 Ljubomir, G., Pamučar, D., Drobnjak, S., & Pourghasemi, H. R. (2019). 15—Modeling the
858 Spatial Variability of Forest Fire Susceptibility Using Geographical Information Systems
859 and the Analytical Hierarchy Process. In H. R. Pourghasemi & C. Gokceoglu (Eds.),
860 Spatial Modeling in GIS and R for Earth and Environmental Sciences (pp. 337–369).
861 Elsevier. https://doi.org/10.1016/B978-0-12-815226-3.00015-6

862 Lowe, P. O., Ffolliott, P. F., Dieterich, J. H., & Patton, D. R. (1978). Determining Potential
863 Wildlife Benefits from Wildfire in Arizona Ponderosa Pine Forests. 18.

Luo, G., Stone, B. L., Johnson, M. D., Tarczy-Hornoch, P., Wilcox, A. B., Mooney, S. D., Sheng, X., Haug, P. J., & Nkoy, F. L. (2017). Automating Construction of Machine Learning Models with Clinical Big Data: Proposal Rationale and Methods. JMIR Research Protocols, 6(8), e175. https://doi.org/10.2196/resprot.7757

Massada, A. B., Syphard, A. D., Stewart, S. I., & Radeloff, V. C. (2013). Wildfire ignition-distribution modelling: A comparative study in the Huron–Manistee National Forest, Michigan, USA. International Journal of Wildland Fire, 22(2), 174–183. https://doi.org/10.1071/WF11178

McCullagh, P., & Nelder, J. A. (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, J. M., Sexton, D. M. H., Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M., & Stainforth, D. A. (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, M. H., Jaafari, A., Fallah, M., & Nabiuni, S. (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(2), 373–384. https://doi.org/10.1007/s13762-017-1371-6

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in Neurorobotics, 7. https://doi.org/10.3389/fnbot.2013.00021

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

Pastro, L. A., Dickman, C. R., & Letnic, M. (2011). Burning for biodiversity or burning biodiversity? Prescribed burn vs. wildfire impacts on plants, lizards, and mammals. Ecological Applications, 21(8), 3238–3253. https://doi.org/10.1890/10-2351.1

Paul, A., Khan, M. L., Arunachalam, A., & Arunachalam, K. (2005). Biodiversity and conservation of rhododendrons in Arunachal Pradesh in the Indo-Burma biodiversity hotspot. Current Science, 89(4), 623–634. JSTOR.

Pausas, J. G., & Keeley, J. E. (2019). Wildfires as an ecosystem service. Frontiers in Ecology and the Environment, 17(5), 289–295. https://doi.org/10.1002/fee.2044

Pham, B. T., Jaafari, A., Avand, M., Al-Ansari, N., Dinh Du, T., Yen, H. P. H., Phong, T. V., Nguyen, D. H., Le, H. V., Mafi-Gholami, D., Prakash, I., Thi Thuy, H., & Tuyen, T. T. (2020). Performance Evaluation of Machine Learning Methods for Forest Fire Modeling and Prediction. Symmetry, 12(6), 1022. https://doi.org/10.3390/sym12061022

902 Pham, B. T., Jaafari, A., Phong, T. V., Yen, H. P. H., Tuyen, T. T., Luong, V. V., Nguyen, H.
903 D., Le, H. V., & Foong, L. K. (2021). Improved flood susceptibility mapping using a best
904 first decision tree integrated with ensemble learning techniques. Geoscience Frontiers,
905 12(3), 101105. https://doi.org/10.1016/j.gsf.2020.11.003

906 Pourghasemi, Hamid Reza. (2014). Re: How to interpret the negative and positive values of
907 profile and plan curvature map in GIS?. Retrieved from:
908 https://www.researchgate.net/post/How_to_interpret_the_negative_and_positive_value
909 s_of_profile_and_plan_curvature_map_in_GIS/53d206f4d2fd64b8118b464d/citation/d
910 ownload.

911 Pradhan, B. K., & Badola, H. K. (2015). Swertia chirayta, a Threatened High-Value Medicinal
912 Herb: Microhabitats and Conservation Challenges in Sikkim Himalaya, India. Mountain
913 Research and Development, 35(4), 374–381. https://doi.org/10.1659/MRD-JOURNAL-
914 D-14-00034.1

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

918 Sachdeva, S., Bhatia, T., & Verma, A. K. (2018). GIS-based evolutionary optimized Gradient
919 Boosted Decision Trees for forest fire susceptibility mapping. Natural Hazards, 92(3),
920 1399–1418. https://doi.org/10.1007/s11069-018-3256-5

921 Sannigrahi, S., Pilla, F., Basu, B., Basu, A. S., Sarkar, K., Chakraborti, S., Joshi, P. K., Zhang,
922 Q., Wang, Y., Bhatt, S., Bhatt, A., Jha, S., Keesstra, S., & Roy, P. S. (2020). Examining
923 the effects of forest fire on terrestrial carbon emission and ecosystem production in India
924 using remote sensing approaches. Science of The Total Environment, 725, 138331.
925 https://doi.org/10.1016/j.scitotenv.2020.138331

926 Satir, O., Berberoglu, S., & Donmez, C. (2016). Mapping regional forest fire probability using
927 artificial neural network model in a Mediterranean forest ecosystem. Geomatics, Natural
928 Hazards and Risk, 7(5), 1645–1658. https://doi.org/10.1080/19475705.2015.1084541

929 Sexton, J. O., Song, X.-P., Feng, M., Noojipady, P., Anand, A., Huang, C., Kim, D.-H., Collins,
930 K. M., Channan, S., DiMiceli, C., & Townshend, J. R. (2013). Global, 30-m resolution
931 continuous fields of tree cover: Landsat-based rescaling of MODIS vegetation
932 continuous fields with lidar-based estimates of error. International Journal of Digital
933 Earth, 6(5), 427–448. https://doi.org/10.1080/17538947.2013.786146

934 Sharma, K., & Thapa, G. (2021). Analysis and interpretation of forest fire data of Sikkim.
935 Forest and Society, 261–276.

936 Sharma, S., Joshi, V., & Chhetri, R. (2014). Forest fire as a potential environmental threat in
937 recent years in Sikkim, Eastern Himalayas, India. Climate Change and Environmental
938 Sustainability, 2, 55. https://doi.org/10.5958/j.2320-642X.2.1.006

939 Shimada, M., Itoh, T., Motooka, T., Watanabe, M., Shiraishi, T., Thapa, R., & Lucas, R.
940 (2014). New global forest/non-forest maps from ALOS PALSAR data (2007–2010).
941 Remote Sensing of Environment, 155, 13–31. https://doi.org/10.1016/j.rse.2014.04.014

942 Taylor, S. W., Woolford, D. G., Dean, C. B., & Martell, D. L. (2013). Wildfire Prediction to
943 Inform Fire Management: Statistical Science Challenges. Statistical Science, 28(4), 586–
944 615. https://doi.org/10.1214/13-STS451

945 Tehrany, M. S., Jones, S., Shabani, F., Martínez-Álvarez, F., & Tien Bui, D. (2019). A novel
946 ensemble modeling approach for the spatial prediction of tropical forest fire susceptibility
947 using LogitBoost machine learning classifier and multi-source geospatial data.
948 Theoretical and Applied Climatology, 137(1), 637–653. https://doi.org/10.1007/s00704-
949 018-2628-9

950 Tien Bui, D., Hoang, N.-D., & Samui, P. (2019). Spatial pattern analysis and prediction of
951 forest fire using new machine learning approach of Multivariate Adaptive Regression
952 Splines and Differential Flower Pollination optimization: A case study at Lao Cai
953 province (Viet Nam). Journal of Environmental Management, 237, 476–487.
954 https://doi.org/10.1016/j.jenvman.2019.01.108

955 Tien Bui, D., Le, H. V., & Hoang, N.-D. (2018). GIS-based spatial prediction of tropical forest
956 fire danger using a new hybrid machine learning method. Ecological Informatics, 48,
957 104–116. https://doi.org/10.1016/j.ecoinf.2018.08.008

958 Tien Bui, D., Le, K.-T. T., Nguyen, V. C., Le, H. D., & Revhaug, I. (2016b). Tropical Forest
959 Fire Susceptibility Mapping at the Cat Ba National Park Area, Hai Phong City, Vietnam,
960 Using GIS-Based Kernel Logistic Regression. Remote Sensing, 8(4), 347.
961 https://doi.org/10.3390/rs8040347

962 Tomislav Hengl, & Ichsani Wheeler. (2018). Soil organic carbon content in x 5 g / kg at 6
963 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution [Data set]. Zenodo.
964 https://doi.org/10.5281/zenodo.2525553

965 Tomislav Hengl, & Surya Gupta. (2019). Soil water content (volumetric %) for 33kPa and
966 1500kPa suctions predicted at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250
967 m resolution [Data set]. Zenodo. https://doi.org/10.5281/zenodo.2784001

968 Trouvé, R., Oborne, L., & Baker, P. J. (2021). The effect of species, size, and fire intensity on
969 tree mortality within a catastrophic bushfire complex. Ecological Applications, n/a(n/a),
970 e02383. https://doi.org/10.1002/eap.2383

971 Tuyen, T. T., Jaafari, A., Yen, H. P. H., Nguyen-Thoi, T., Phong, T. V., Nguyen, H. D., Van
972 Le, H., Phuong, T. T. M., Nguyen, S. H., Prakash, I., & Pham, B. T. (2021). Mapping
973 forest fire susceptibility using spatially explicit ensemble models based on the locally
974 weighted learning algorithm. Ecological Informatics, 63, 101292.
975 https://doi.org/10.1016/j.ecoinf.2021.101292

976 Vicars, W. C., Sickman, J. O., & Ziemann, P. J. (2010). Atmospheric phosphorus deposition at
977 a montane site: Size distribution, effects of wildfire, and ecological implications.

978 Atmospheric Environment, 44(24), 2813–2821.
979 https://doi.org/10.1016/j.atmosenv.2010.04.055

980 Vilar, L., Gómez, I., Martínez-Vega, J., Echavarría, P., Riaño, D., & Martín, M. P. (2016).
981 Multitemporal Modelling of Socio-Economic Wildfire Drivers in Central Spain between
982 the 1980s and the 2000s: Comparing Generalized Linear Models to Machine Learning
983 Algorithms. PLOS ONE, 11(8), e0161344.
984 https://doi.org/10.1371/journal.pone.0161344

985 Williams, A. P., Abatzoglou, J. T., Gershunov, A., Guzman-Morales, J., Bishop, D. A., Balch,
986 J. K., & Lettenmaier, D. P. (2019). Observed Impacts of Anthropogenic Climate Change
987 on Wildfire in California. Earth's Future, 7(8), 892–910.
988 https://doi.org/10.1029/2019EF001210

989 Xie, Y., & Peng, M. (2019). Forest fire forecasting using ensemble learning approaches. Neural
990 Computing and Applications, 31(9), 4541–4550. https://doi.org/10.1007/s00521-018-
991 3515-0

992 Yathish, H., Athira, K. V., Preethi, K., Pruthviraj, U., & Shetty, A. (2019). A Comparative
993 Analysis of Forest Fire Risk Zone Mapping Methods with Expert Knowledge. Journal of
994 the Indian Society of Remote Sensing, 47(12), 2047–2060.
995 https://doi.org/10.1007/s12524-019-01047-w

996 Zhang, G., Wang, M., & Liu, K. (2019). Forest Fire Susceptibility Modeling Using a
997 Convolutional Neural Network for Yunnan Province of China. International Journal of
998 Disaster Risk Science, 10(3), 386–403. https://doi.org/10.1007/s13753-019-00233-1

999 Zhang, G., Wang, M., & Liu, K. (2021). Deep neural networks for global wildfire susceptibility
1000 modelling. Ecological Indicators, 127, 107735.
1001 https://doi.org/10.1016/j.ecolind.2021.107735

Figures

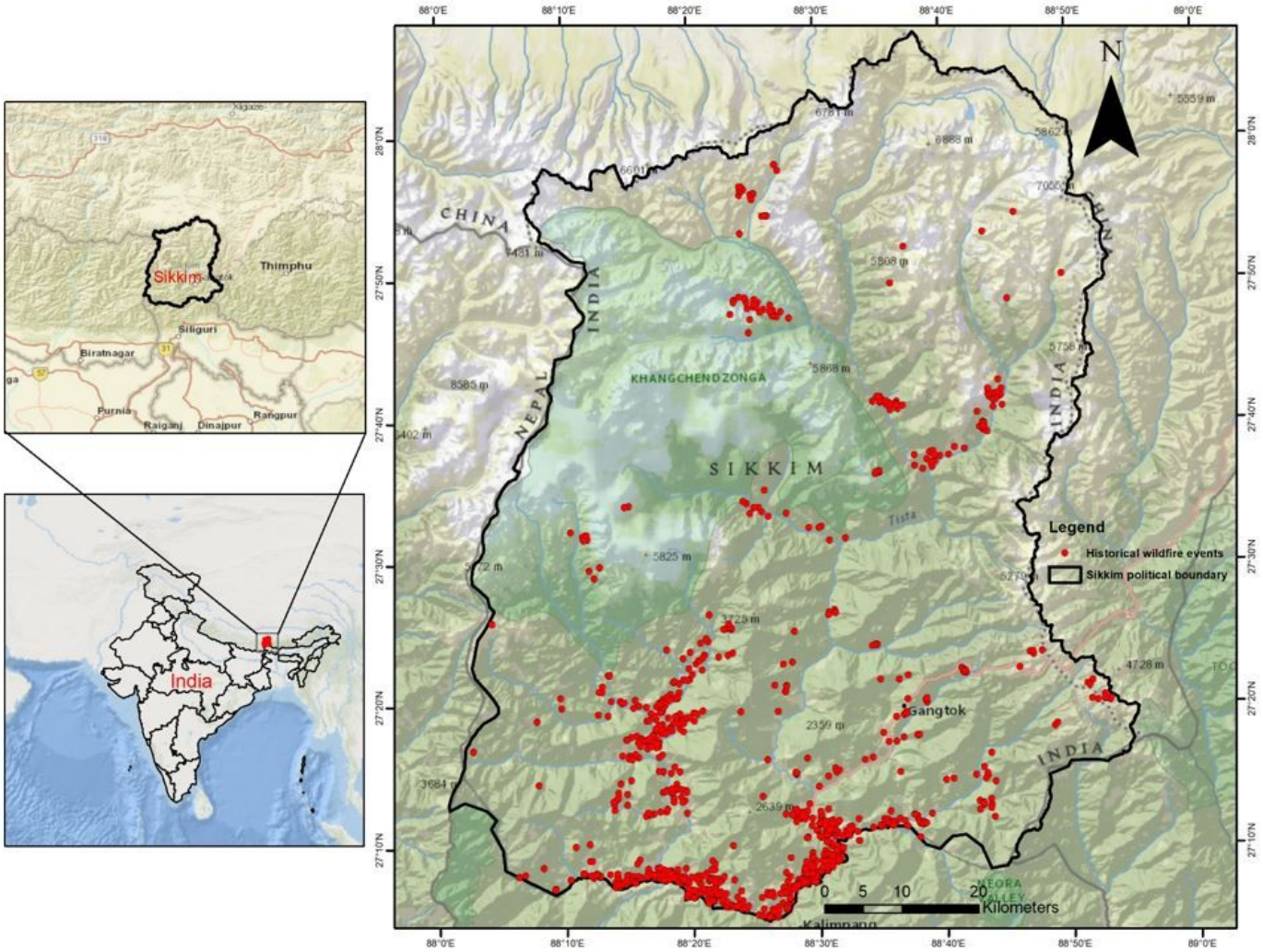

Figure 1

Figure 1: Study area. [Courtesy: ESRI]

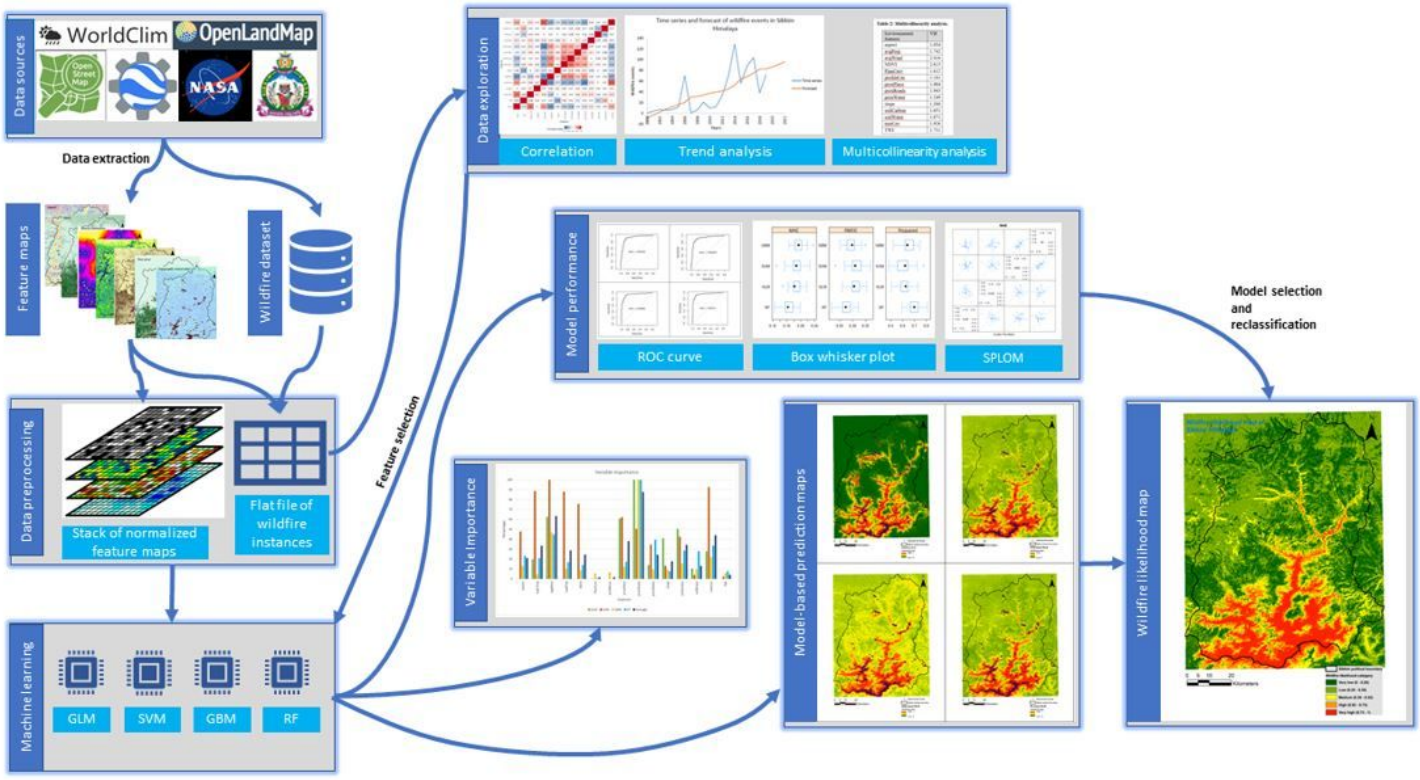

Figure 2

Methodology of the preparation of Wildfire likelihood map. (Source of raster stack image: https://i.stack.imgur.com/whXlL.png)

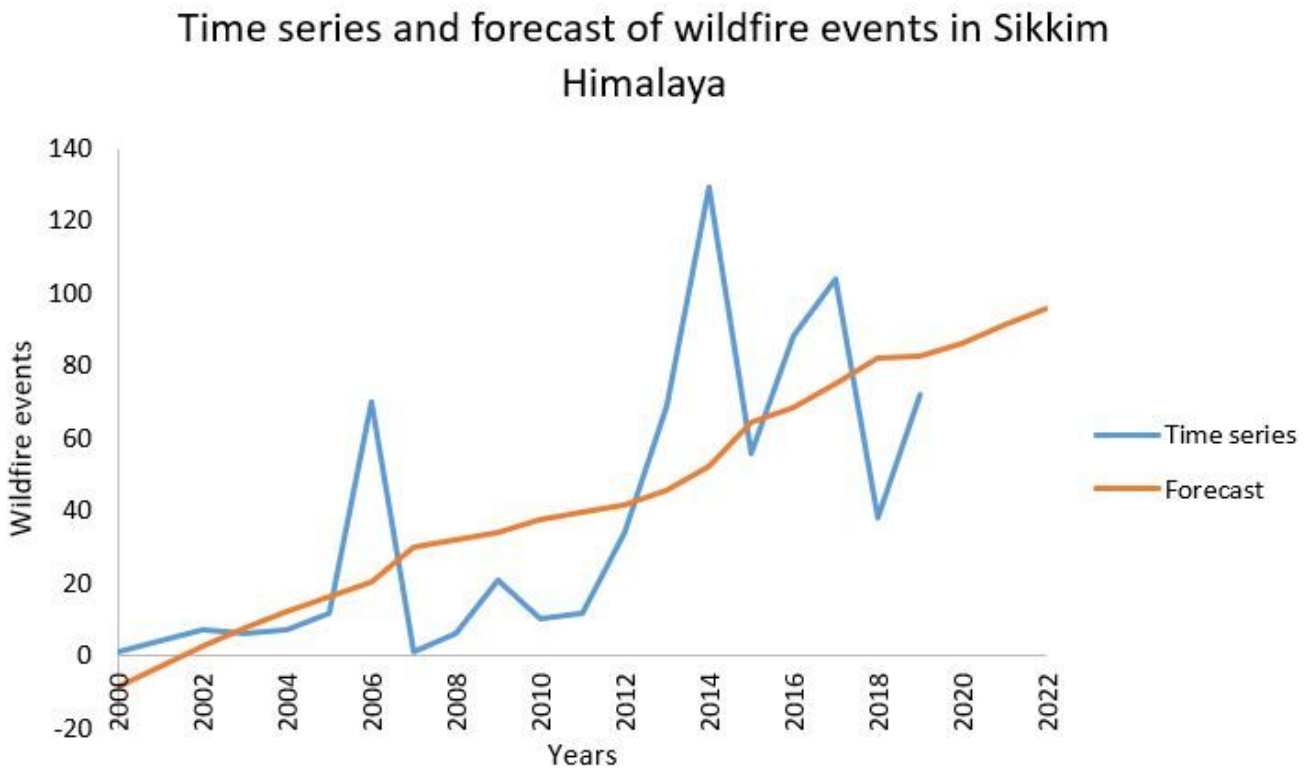

Figure 3

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.

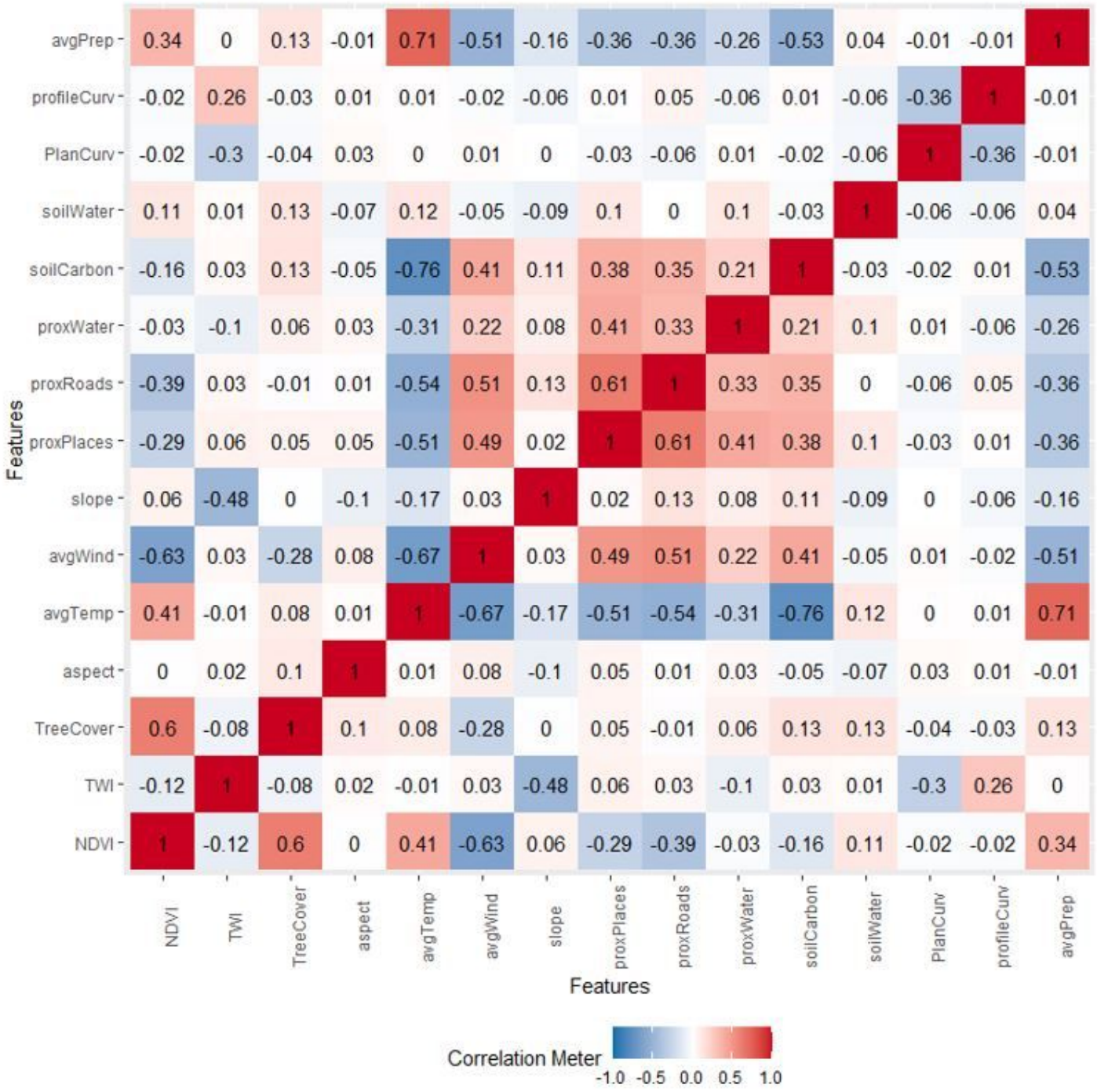

Figure 4

Correlation matrix of feature variables.

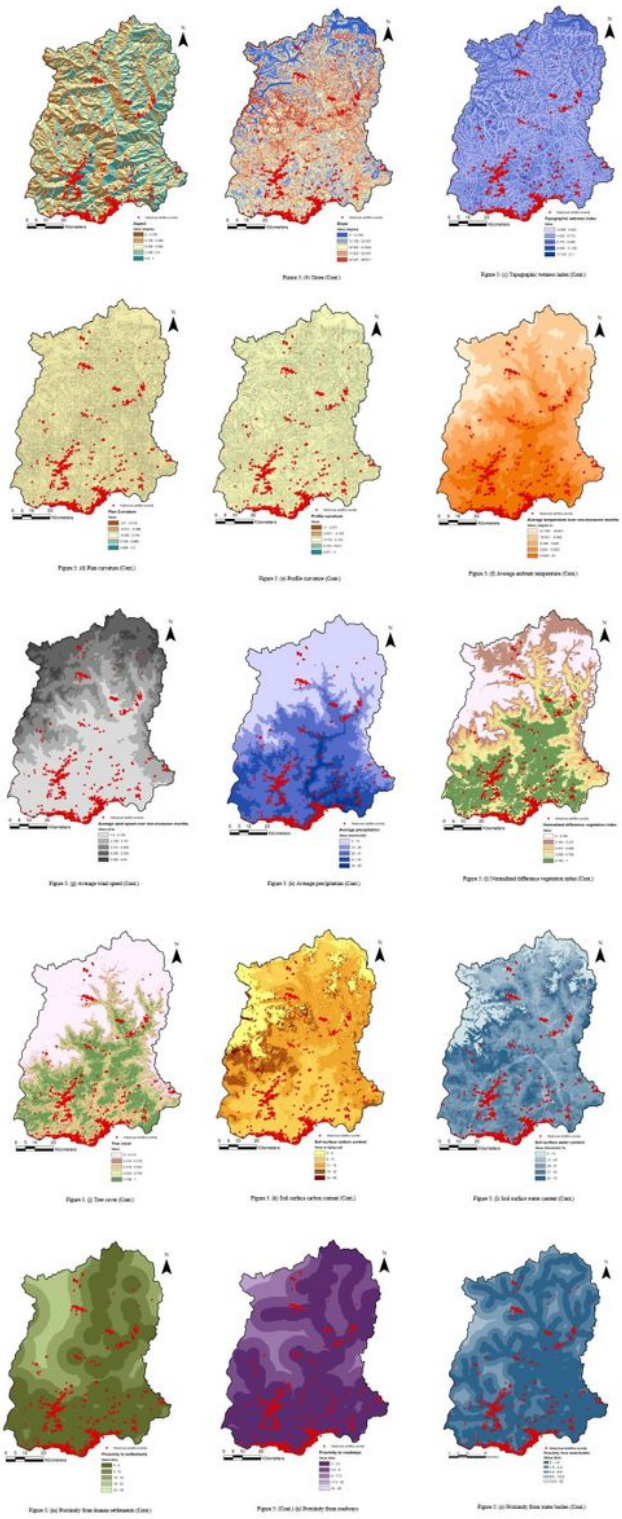

Figure 5

Environmental features. All the maps have been reclassified using Jenks natural breaks method. The natural breaks method minimizes variance within categories while maximizing the variance between categories. This leads to an increase in the quality of the classification (Jenks, 1967) (a) Aspect (Cont.)

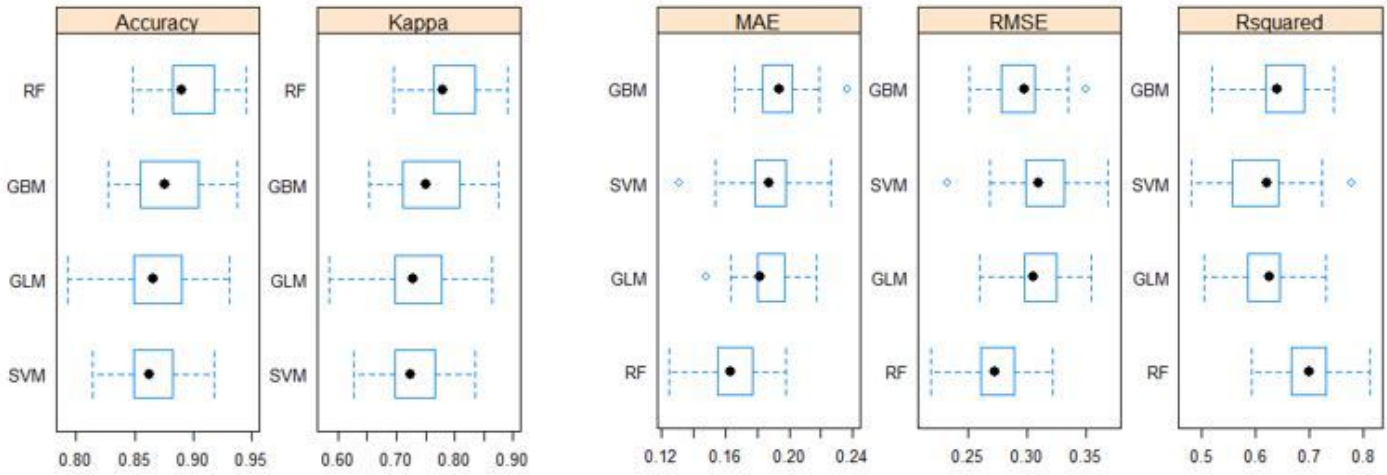

Figure 6

Box and whisker plots of model performance indices.

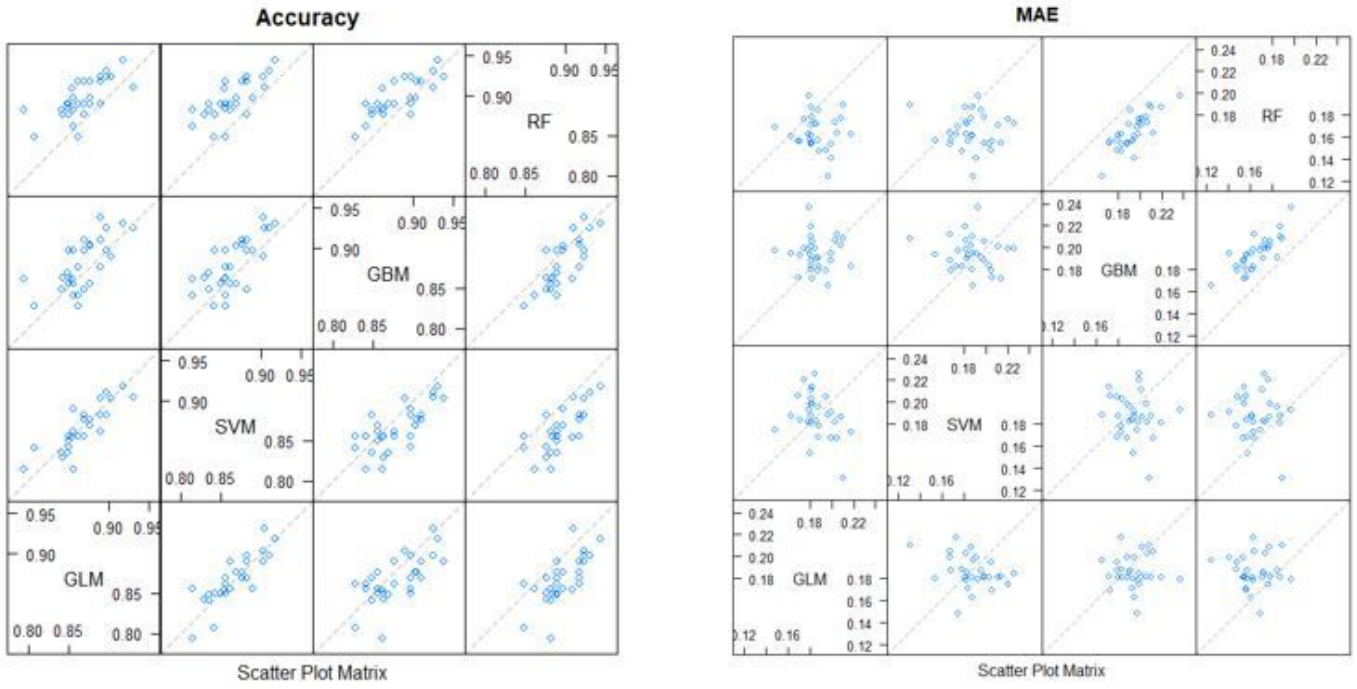

Figure 7

Scatter plot matrix of accuracy and MAE

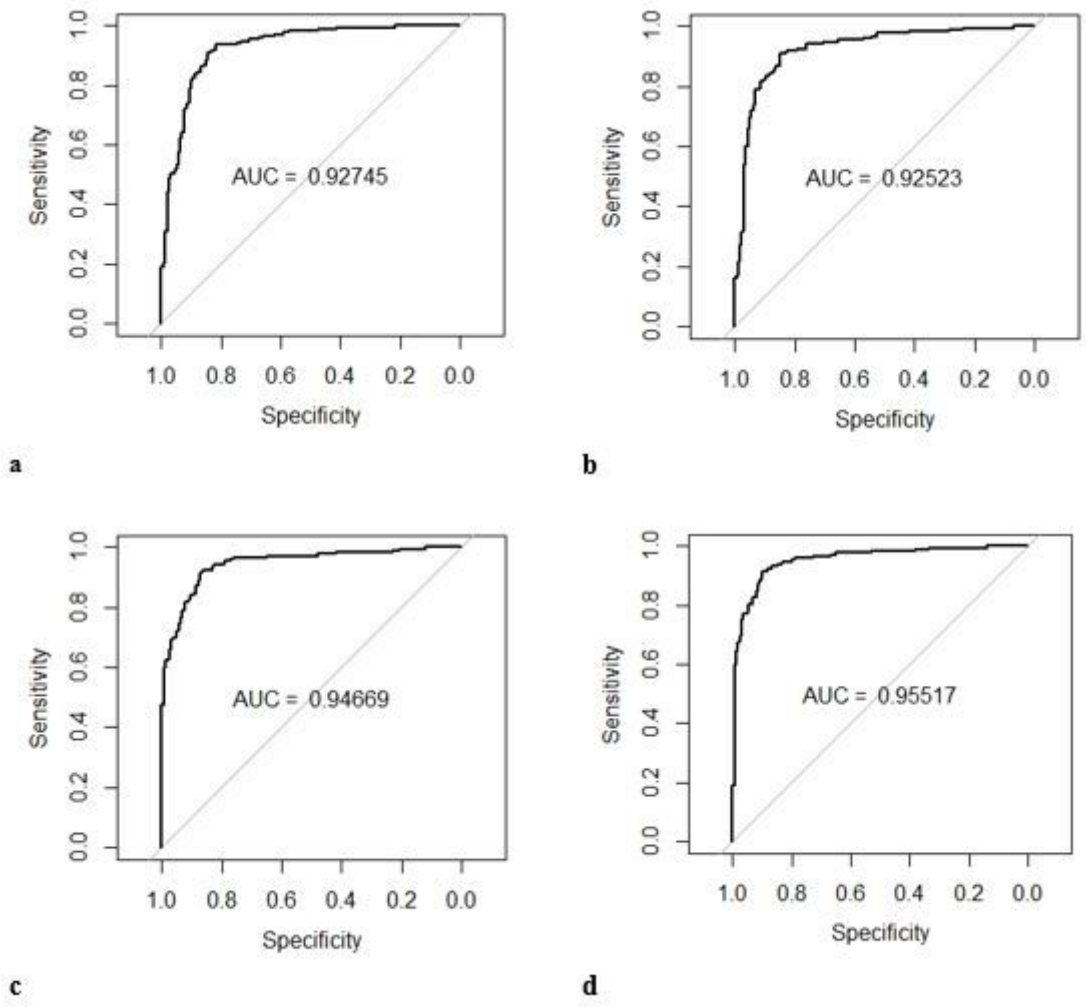

Figure 8

ROC curve of (a) GLM, (b) SVM, (c) GBM, (d) RF.

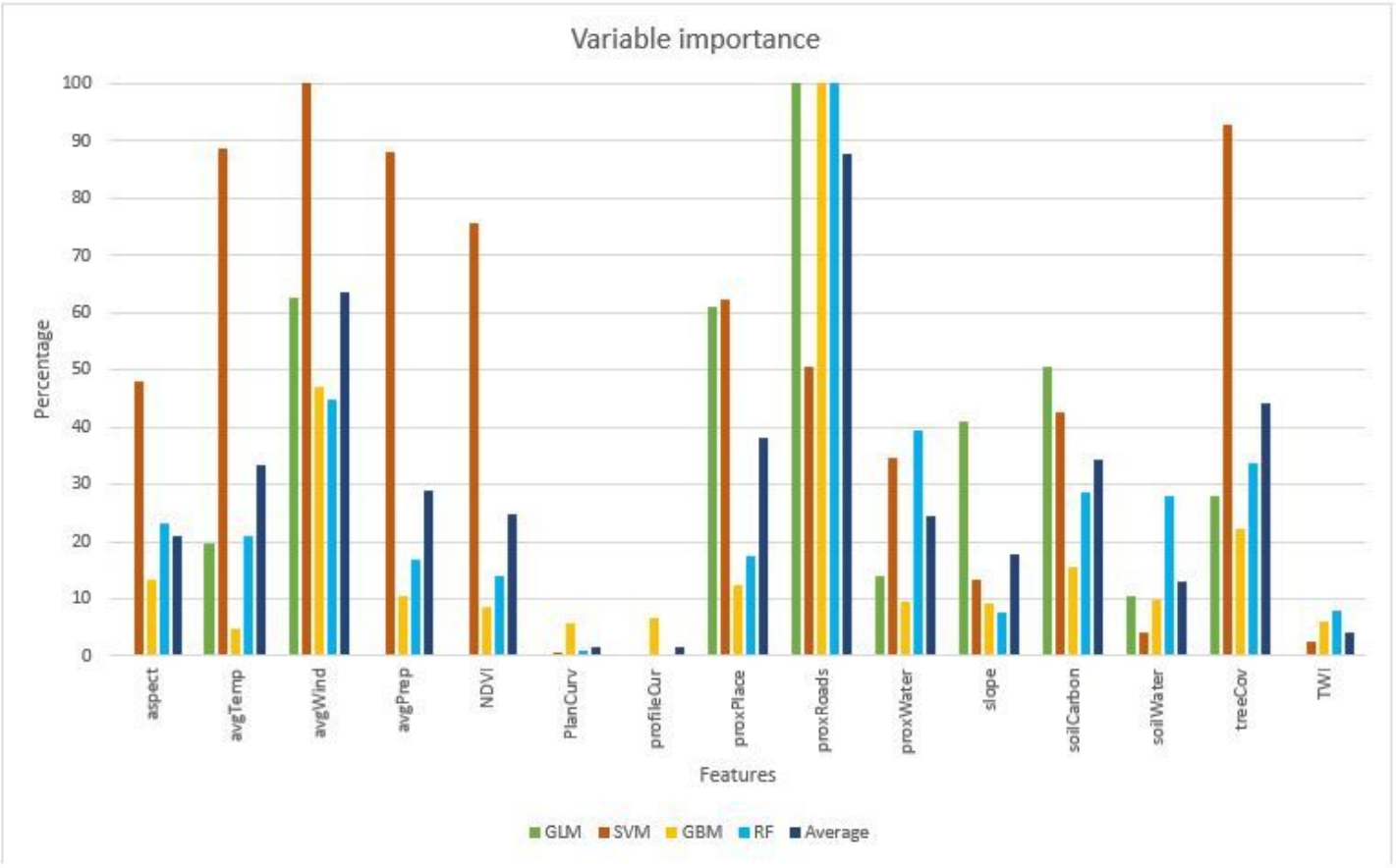

Figure 9

Importance or influence of the environmental features on the prediction models. For RF and GBM, variable importance was calculated by estimating the Mean Squared Error (MSE) of the out-of-box sample by shuffling the dataset. Loess r-squared method was used for estimating the variable importance of SVM. For GLM, the absolute value of the t-statistic of the model parameters was used to estimate the variable importance (Kuhn, 2019).

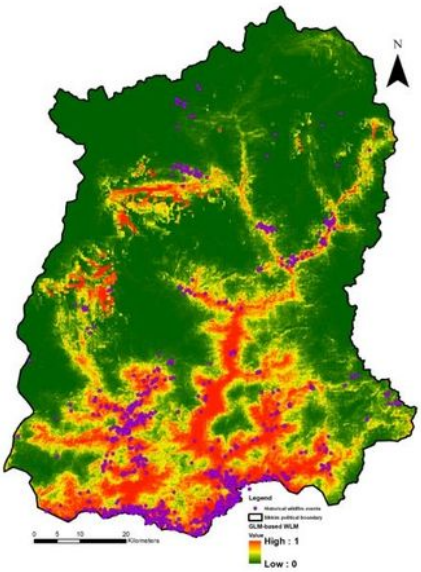

Figure 10: Wildfire likelihood map of Sikkim Himalaya based on the prediction of (a) GLM (Cont.)

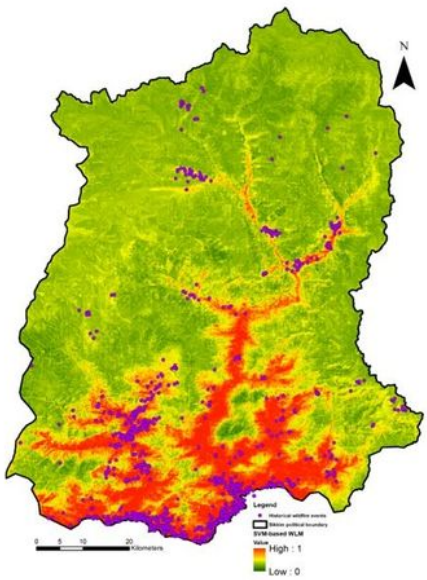

Figure 10: (b) SVM (Cont.)

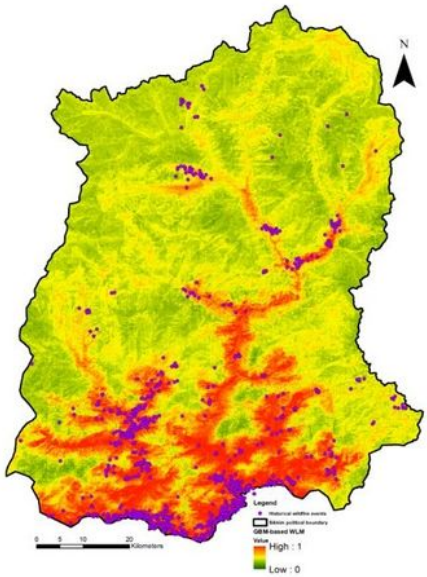

Figure 10: (c) GBM (Cont.)

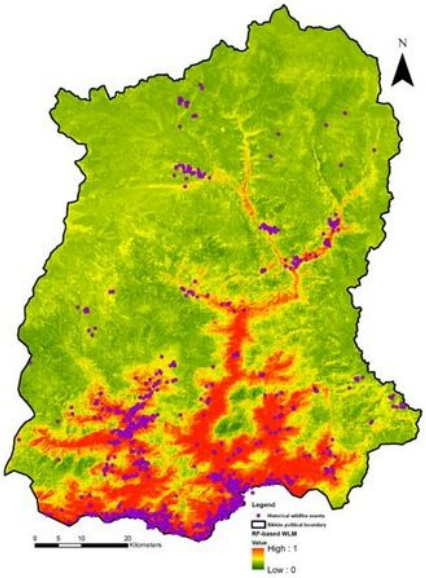

Figure 10: (Cont.) (d) RF

Figure 10

Wildfire likelihood map of Sikkim Himalaya based on the prediction of (a) GLM (Cont.)

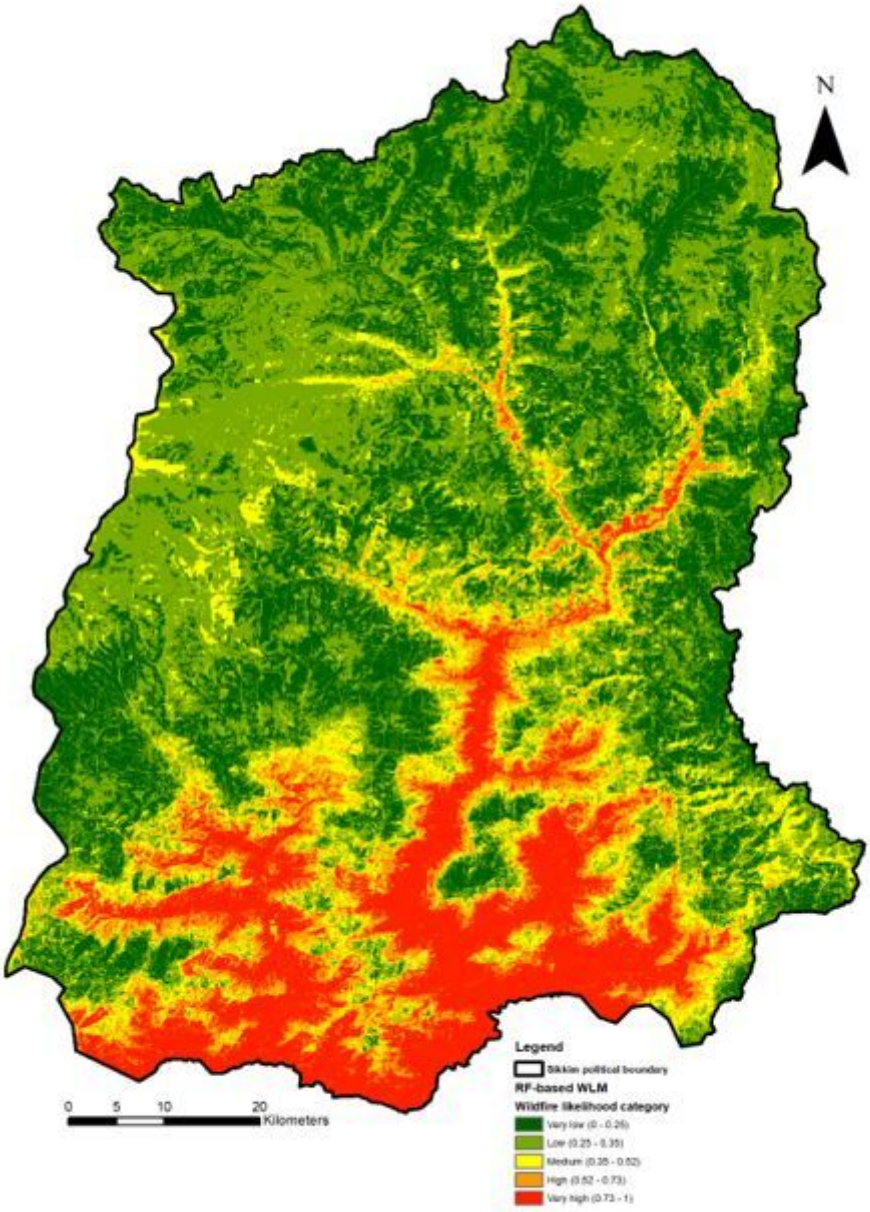

Figure 11

Wildfire likelihood map of Sikkim Himalaya showing various categories of likelihood of wildfire.

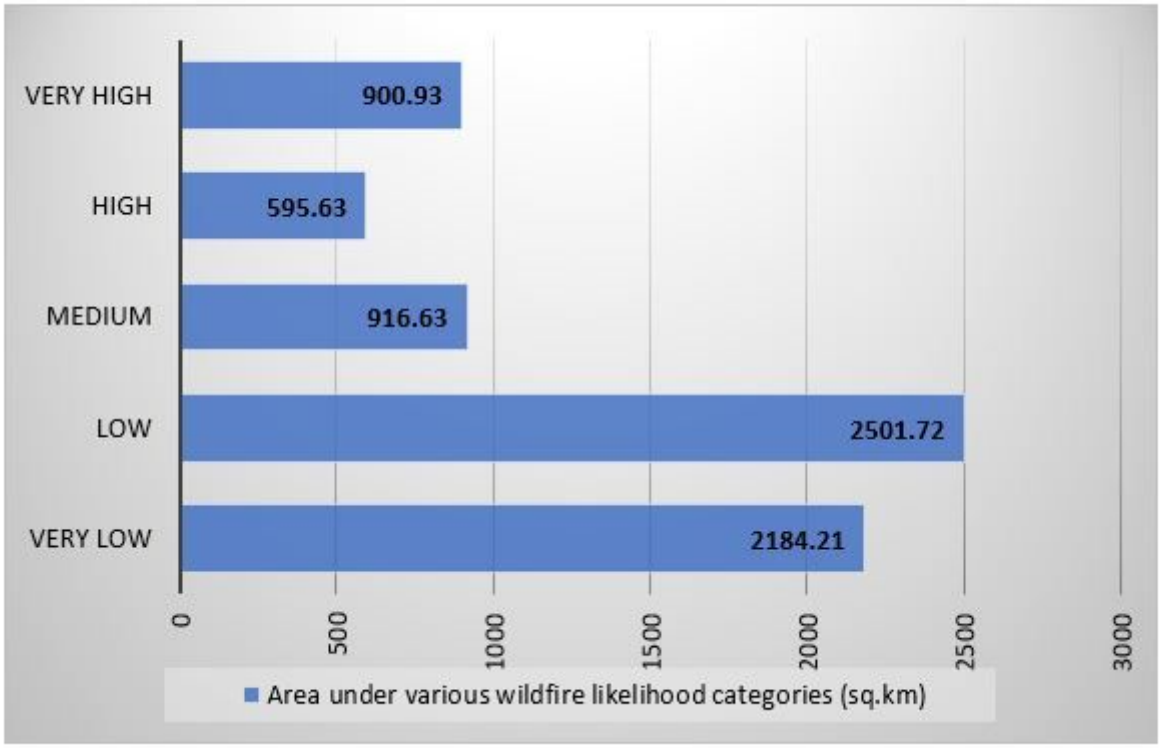

Figure 12

Areas under the various wildfire likelihood categories.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- WLMMLSupplement.docx