remote sensing

MDPI

Article

Forest Fire Susceptibility Prediction Based on Machine Learning Models with Resampling Algorithms on Remote Sensing Data

Bahareh Kalantar [1,*], Naonori Ueda [1], Mohammed O. Idrees [2], Saeid Janizadeh [3], Kourosh Ahmadi [4] and Farzin Shabani [5,6]

1 RIKEN Center for Advanced Intelligence Project, Goal-Oriented Technology Research Group, Disaster Resilience Science Team, Tokyo 103-0027, Japan; naonori.ueda@riken.jp
2 Department of Surveying and Geoinformatics, Faculty of Environmental Sciences, University of Ilorin, P.M.B. 1515, 240103 Ilorin, Nigeria; mohammed.oi@unilorin.edu.ng
3 Department of Watershed Management Engineering, College of Natural Resources, Tarbiat Modares University, Tehran P.O. Box 14115-111, Iran; janizadehsaeid@modares.ac.ir
4 Department of Forestry, Faculty of Natural Resources and Marine Sciences, Tarbiat Modares University, Tehran 15119-43943, Iran; kourosh.ahmadi@modares.ac.ir
5 Department of Biological Sciences, Global Ecology and ARC Centre of Excellence for Australian Biodiversity and Heritage, College of Science and Engineering, Flinders University, GPO Box 2100, Adelaide, SA 5001, Australia; farzin.shabani@flinders.edu.au
6 Department of Biological Sciences, Macquarie University, Sydney, NSW 2109, Australia
* Correspondence: bahareh.kalantar@riken.jp

Received: 9 October 2020; Accepted: 5 November 2020; Published: 10 November 2020

Abstract: This study predicts forest fire susceptibility in Chaloos Rood watershed in Iran using three machine learning (ML) models—multivariate adaptive regression splines (MARS), support vector machine (SVM), and boosted regression tree (BRT). The study utilizes 14 set of fire predictors derived from vegetation indices, climatic variables, environmental factors, and topographical features. To assess the suitability of the models and estimating the variance and bias of estimation, the training dataset obtained from the Natural Resources Directorate of Mazandaran province was subjected to resampling using cross validation (CV), bootstrap, and optimism bootstrap techniques. Using variance inflation factor (VIF), weight indicating the strength of the spatial relationship of the predictors to fire occurrence was assigned to each contributing variable. Subsequently, the models were trained and validated using the receiver operating characteristics (ROC) area under the curve (AUC) curve. Results of the model validation based on the resampling techniques (none, 5- and 10-fold, bootstrap and optimism bootstrap) produced AUC values of 0.78, 0.88, 0.90, 0.86 and 0.83 for the MARS model; 0.82, 0.82, 0.89, 0.87, 0.84 for the SVM and 0.87, 0.90, 0.90, 0.90, 0.91 for the BRT model. Across the individual model, the 10-fold CV performed best in MARS and SVM with AUC values of 0.90 and 0.89. Overall, the BRT outperformed the other models in all ramification with highest AUC value of 0.91 using optimism bootstrap resampling algorithm. Generally, the resampling process enhanced the prediction performance of all the models.

Keywords: machine learning; remote sensing; computational intelligence; bootstrapping; cross validation (CV)

1. Introduction

Over the last century, uncontrollable forest fires caused severe long-term destruction to wildlife, property and the environment, including forest and agricultural land across Asia, Australia, Africa and

the Americas [1]. Although climate change (i.e., increase of temperature, reduction in precipitation) play a role in the increase in fire occurrences, the impact of human activities should not be disregarded. For example, in areas where lightning rarely occurs or lightning is not concurrent with dry conditions, human is to be blamed for the source of ignitions [2]. The frequency and risk of forest fires has continued to intensify in response to global warming and climate change and its attendant effects such as extreme temperature, decrease in the amount of rainfall and longer dry season in many regions [3]. Forest fire affects soil characteristics, contributes to the global carbon emission and climate change [4–6]. Most importantly, fire constitutes a major disturbance in forest ecological balance, alteration in the forest successional rates and, reduction of forest biomass [4,7].

Forests are important natural resources that perform vital economic and ecological functions, including provision of goods and livelihoods, safeguards biodiversity, protect soils from degradation and erosion, regulate water flow, and regulate climate by trapping carbon that could have been added to greenhouse gases [8]. Therefore, effective forest fire control is essential for productive use of forest resources, protection of the environment, and maintaining wide-reaching ecological balances [9]. Since the late 1970s, satellite-based remote sensing (RS) data have been extensively used to both detect active forest fires and map burned areas by exploiting thermal contrast between burning fire and the background and by assessing the effects of fire on vegetation reflectance [10,11]. RS data, such as MODIS [11–13], Landsat [10,14], and lately, the European Space Agency (ESA) Sentinel-2 satellites [15–17], allow accurately estimating the extent of fire-affected areas and the burn severity at different scales (local, regional and global) taking advantage of their high-quality temporal, spatial, and spectral resolutions.

The first step to preventing or mitigating forest fire is to improve accuracy of the detection of the exact location and extents of potential fire to optimize response time. Today, availability of multi-sensor data and advance computational intelligence, including artificial intelligence (AI) and machine leaning (ML) algorithms have improved the accuracy of forest fire prediction [18]. Several studies have attempted mapping forest fire susceptibility to improve the detection and response time to potential fire outbreak by combining several data layers such as from RS, topographical features and climatic factors. The primary objective of achieving forest fire prediction map with extremely high level of reliability resulted in the application of several ML methods in geographic information system (GIS) environment (see Jain et al. [18] and the references therein). These methods include artificial neural network (ANN) and its family [19,20], random forest (RF) [21], support vector machines (SVM) [21], logistic regression (LR) [22], decision tree (DT) [23], dynamic bayesian network [24], multi-layer perception neural networks (MLP) [24], and multivariate adaptive regression splines (MARS) [25], and many others comprehensively discussed in [18]. These methods have been identified with good result but incapable of producing the much-expected degree of accuracy at regional scale [26].

There are several groups of effective resampling techniques such as cross validation (CV), nonparametric bootstrapping, Jackknife resampling which are useful for validating models using random subsets [27]. Single iteration model validation using CV is the most common resampling techniques [28]. CV is a more refined resampling for splitting sampled fire inventory dataset into test and validation to minimize variability in model's extrapolation which makes it a reliable means of evaluating ML models especially with a limited sampled dataset [29]. In the work of Ghorbanzadeh et al. [11], the authors applied ANN, multi-criteria decision making (GIS-MCDM) to map forest fire susceptibility in Amol County, the Mazandaran province, Iran using 16 relevant environmental variables and 34 historical forest fire polygons. The forest fire sampling polygons was randomly split into 10-folds of about 1742 pixels and the result yielded area under the curve (AUC) value of 0.801. One of the advantages of CV method is that it resolves the adverse effect of randomness in data layers.

In a related study, Wijayanto et al. [30] proposed the application of adaptive neuro-fuzzy inference system (ANFIS) utilizing forest fires hotspot data to develop classification models for hotspots occurrence in Central Kalimantan. They used three k-fold CV resampling technique (i.e., Fold 1 = 60/40, Fold 2 = 40/60, and Fold 3 = 20/60–20) to select the best combination of training and testing datasets.

The results showed the 2-fold producing the best accuracy of 99.99% with the lowest training error of 0.003. Dodangeh et al. [27] applied multi-time resampling methods namely, random subsampling, bootstrapping combined with ML models (generalized additive model (GAM), boosted regression tree (BTR), MARS) for flood susceptibility mapping. The results showed that the employing the resampling methods improved the performance of ML models. Especially, the bootstrapping approach is performed better than random subsampling algorithm in terms of performance assessment.

Several studies have investigated forest fire susceptibility but no reference in literature known to the authors have reported comparing different resampling algorithms to optimize partitioning of inventory data for model training and validations. Most studies adopt the traditional splitting train-test data set (such as 70/30 percent or 65/35 percent ratio) for model training and validation. Therefore, this study optimizes three well-known ML algorithms–MARS, SVM, and BRT–by testing four resampling methods CV (5-fold and 10-fold) and bootstrapping (bootstrap and optimism bootstrap) to predicts areas that are extremely susceptible to forest fire in Chaloos Rood watershed, Iran. In summary the contribution of this study is twofold. First, four resampling techniques are used on inventory datasets to determine the best combination of training and testing datasets. This paper presents the first attempt to apply optimism bootstrap resampling technique for forest fire susceptibility prediction. Second, three ML algorithms were considered on 14 forest fire conditioning factors to predict forest fire susceptibility in Chaloos Rood watershed, Iran.

2. Study Area

The Chaloos Rood watershed is part of the large Chaloos catchment and one of the seven sub-basins of the Caspian basin. This watershed is located within geographical coordinates 50°58′E to 51°40′E longitude and 36°08′N to 36°36′N latitude, with approximate basin area of about 1634 km^2 (Figure 1) with altitude ranging between −26 m and 4256 m above mean sea level (msl). The Chaloos Rood watershed in the west leads to the Kourkoursar catchment in the south to the Karaj catchment, and in the north to the Caspian Sea. In terms of climate, the Caspian basin belongs to the cold semi-humid and cold humid climatic zones; although, cold and semi-arid climates have also been identified in some lowlands [31]. In the study area also, annual rainfall ranges between 288.3 mm and 1538.2 mm. The study area covered mostly by range (713.71 km^2), forest (671.72 km^2), and agriculture (143.4 km^2).

An increase in temperature and waste compaction caused a fire occurrence on 1 July 2016 in the study area. Later (17 November 2016), another fire incident burnt around 12 hectares of the forest in the studied area, possibly due to human activities (i.e. hunting). On 24 March 2018, an additional 5-hectares of the forests were burnt by fire to the extent of leaf carcasses, and between 30 and 40 rotten and broken trees, however, the environmental conditions such as high temperature along with hot winds made controlling the incident difficult. Recently (22 June 2019), this area had lost one hectare of the vegetation cover by another fire occurrence, possibly caused by human factors (visitors) together with dryness of the forage (natural fuel load). Overall, a combination of anthropogenic activities (lightning, intentional land clearing, and hunting) are identified as the main reasons for the incidences in the studied area.

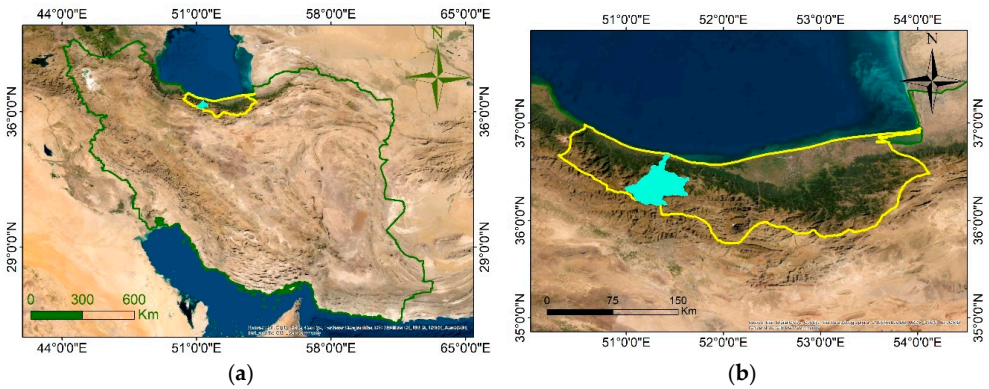


Figure 1. *Cont.*

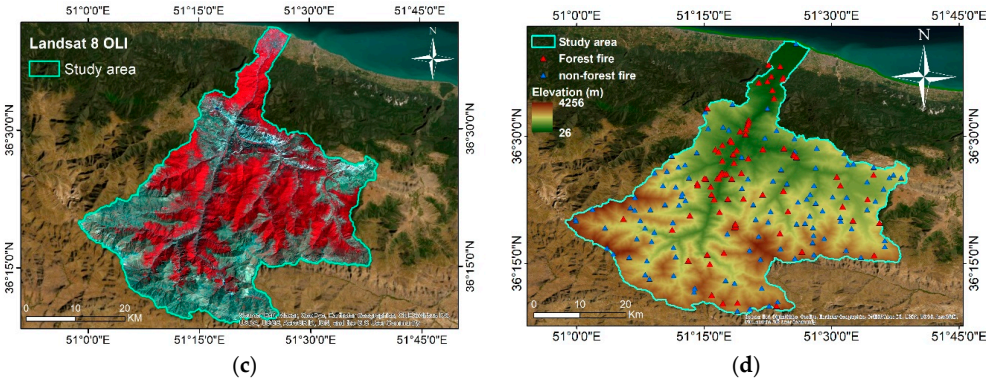


Figure 1. The study area describing: (**a**) map of Iran; (**b**) location of study area; (**c**) Landsat 8 satellite imagery of study area (bands 5, 4, 3); (**d**) The Chaloos catchment including digital elevation model (DEM), forest fire and non-forest fire inventories.

3. Methodology

3.1. Overview

In this study, the process of predicting susceptibility to fire event in Chaloos Rood watershed involves a number of biophysical and geomorphological factors derived from remotely sensed data, ASTER digital elevation model (DEM) (Figure 1d) and Landsat 8 OLI (Figure 1c). First, we obtained sampling dataset comprising of 109 fire inventory points taken from the fire incident archive of the General Directorate Administration of Natural Resources of Mazandaran province and additional 109 non-fire points for the purpose of modeling. This was followed by generation of 14 fire conditioning factors (independent variables) from the remotely sensed and climatic data. After attaching the raster values of the fire conditioning factors into the corresponding sampling points, it was subjected to multi-collinearity tests for quality assurance. Thereafter, the sampling points were divided into 70 and 30 percent ratio for model training and testing, respectively. Then, the training data was infused into the four resampling methods and the result subsequently used in MARS, SVM and BRT models for training and generation of forest fire susceptibility maps. Finally, the models were validated using 30 percent sample dataset as testing and the performance evaluated using the sensitivity, specificity, negative predictive values (NPV), positive predictive values (PPV) and AUC metrics. Figure 2 describes the flow of data processing and analysis steps.

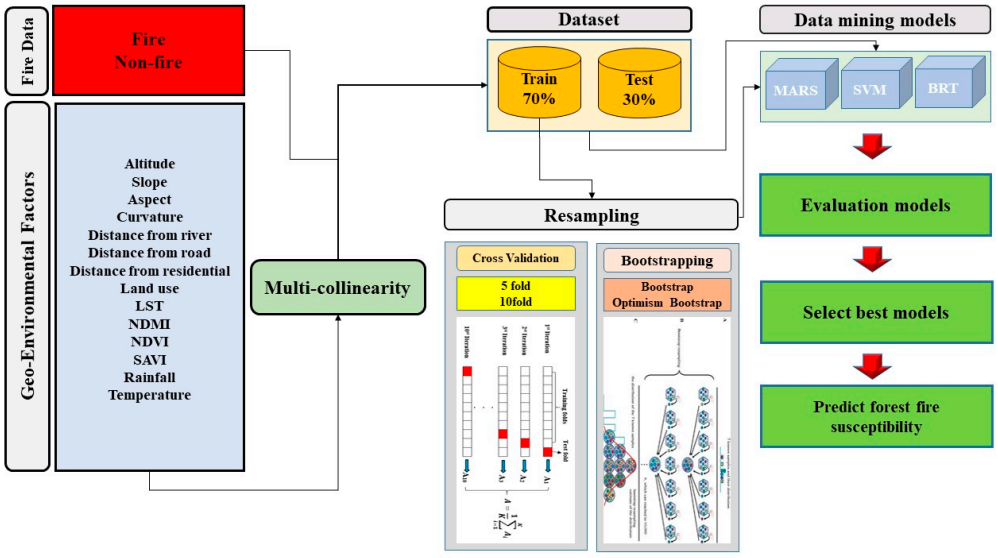


Figure 2. Framework of the proposed methodology.

3.2. Data Preparation

To map the vulnerability of the study area to forest fire, historical data recorded from previous forest fire incidents and assemblage of forest fire predictors [26,32] were prepared. Within the study area, the General Directorate Administration of Natural Resources of Mazandaran province identified a total number of 109 fire incidents. In addition to the fire points, 109 randomly sampled points were generated outside the fire zones such that average distance between points is >30 m to ensure no two adjacent points are selected within the same pixel. Sufficient sampling points were created randomly and distributed to represent a balance of fire and non-fire occurrences. The sampling points were subsequently split into two (training and testing datasets) for model training and validation [1,12]. In forest fire modeling, there is no general consensus on the partitioning ratio for the sampling dataset, the choice varies in the literature usually depending on the quantity and quality of the sample data. However, in this study, splitting of the sample data into two was objectively achieved through the process of randomly selection which resulted into 70/30 percent ratio for building and testing of the model [21].

To derive photosynthetic, non-photosynthetic, climatic and geomorphologic indices, we employed Landsat 8 OLI satellite imagery, 30 m resolution ASTER DEM, meteorological data and other vector GIS data layers (road, rivers and residential area). Landsat 8 satellite imagery collected in July 21, 2019 with low cloud cover (<5%) was provided free of charge by the United States Geological Survey (USGS) through its data depository (http://earthexplorer.usgs.gov and http://glovis.usgs.gov). The image came geometrically corrected; however, in ENVI software environment, the imagery was preprocessed to represent standardized surface reflectance by correcting for atmospheric effects using quick atmospheric correction (QUAC) technique [33] which determines atmospheric correction parameters from the spectra of the image scene without the need for any other external information. In addition, it is capable of producing accurate pixels that reasonably represent the image object spectra [33].

Several studies have mapped or predicted forest fire basically using satellite-based approach [10,18]. However, limitations arise with respect to the accuracy of the prediction where the topographical factors such as effect of shadow cast by steep slope and terrain ruggedness and environmental conditions such as accessibility to or human activities are not considered [13]. Therefore, this study systematically incorporates 14 geo-environmental factors (Figure 3) to predict fire susceptibility in the study area: four from topographical data (altitude, aspect, slope and curvature); three from vector data layers (distance from residential area, river and road); six from Landsat 8 imagery (land surface temperature (LST), normalized difference moisture index (NDMI), normalized difference vegetation index (NDVI), soil adjusted vegetation index (SAVI), land use and land cover—(LULC)); and finally, two climatic indices—air temperature and rainfall.

The effects of topographic complexity on forest fire have been widely reported [4,26]. Thus, the altitude, slope, aspect and curvature (Figure 3a–d) indices were derived from the 30 m resolution ASTER DEM. Canopy structure and forest species are greatly influenced by topographical variations in the landscape. Experience has shown that the possibility of wide spread of fire is higher at high altitude due to the influence of wind action [34]. Similarly, the rate at which fires burn is faster on a steeped slope in the direction facing the eastern aspect which is much more influenced by the incoming solar radiation [35]. Slope aspect has effect on the micro-climate in terms of exposure to solar illumination and wind. For instance, while the eastern and western facing slopes have direct sunlight (depending on the time of the day), the northern and southern aspects have enriched moisture contents that support healthy vegetation and canopy undergrowth. Curvature, which characterizes concavity or convexity of the topography [35], is useful for measuring and predicting soil water content, controls overall water and accumulation processes of the downhill flow, and rainfall response. Curvature is a physiographic variable that controls the distribution of vegetation in response to erosive and depositional processes of the topography.

Natural forces such as lightning, high atmospheric temperatures and dryness with low humidity are favorable for forest fire initiation but most often, forest fires are originated through human activities,

including agricultural activities, hunting, or when naked flame comes into contact with inflammable forest biomass [36]. The more accessible the forest is to the publics the more prone it is to bush burning. We consider road and river important accessibility factors that contribute to fire ignition. The closer the road is to the forest, the higher the risk of fire. On the other hand, vegetation closer to the river hold more moisture that make them healthy with enriched greenness which incidentally slows down fire rage. Using Euclidean distance, distance from river and road (Figure 3e,f) were generated.

LULC of a place contributes to fire at varying degrees depending on the composition of the cover type and human interaction. For instance, it not unusual to expect greater risk of fire arising from agricultural land close to forest or orchard with undergrowth and dry trunks as fuel. So, the Landsat imagery was classified into six LULC classes: agriculture, forest, orchard, range, urban and water (Figure 3g) using supervised classification Maximum Likelihood classifier algorithm [7]. We used the Landsat image in 2019 because the majority of fire points happened during last three years. As the study area is a part of Hyrcanian forest, when we evaluated the initial Landsat images, we figured out that during last three years the changes in forest cover was insignificant due to the effective and sustainable management and conservation of natural resource systems. So, a single image was selected for extracting the vegetation indices and LULC mapping. Using the LULC, another data layer, distance from residential area, was created as an indicative fire factor (Figure 3h).

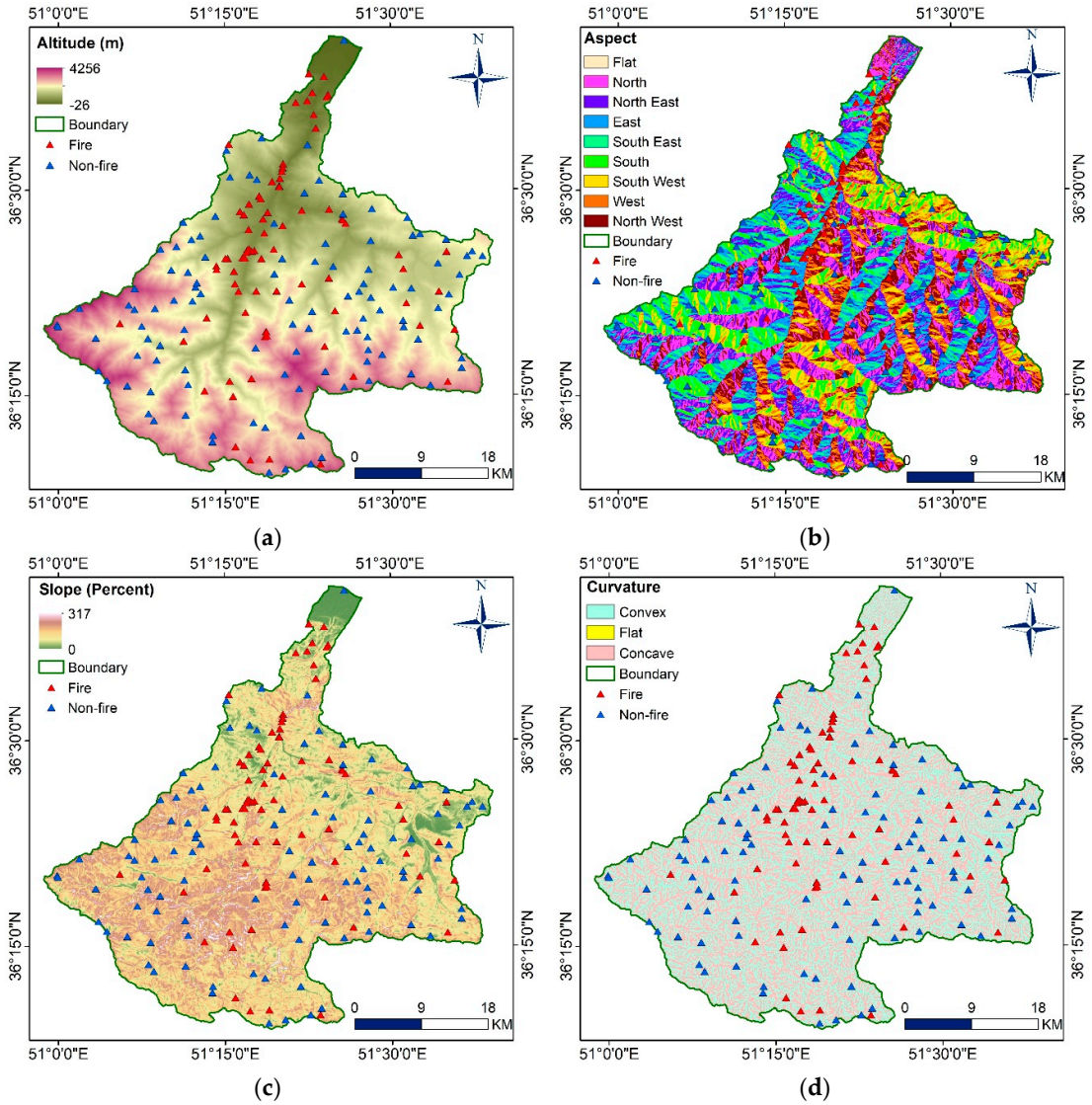


Figure 3. *Cont.*

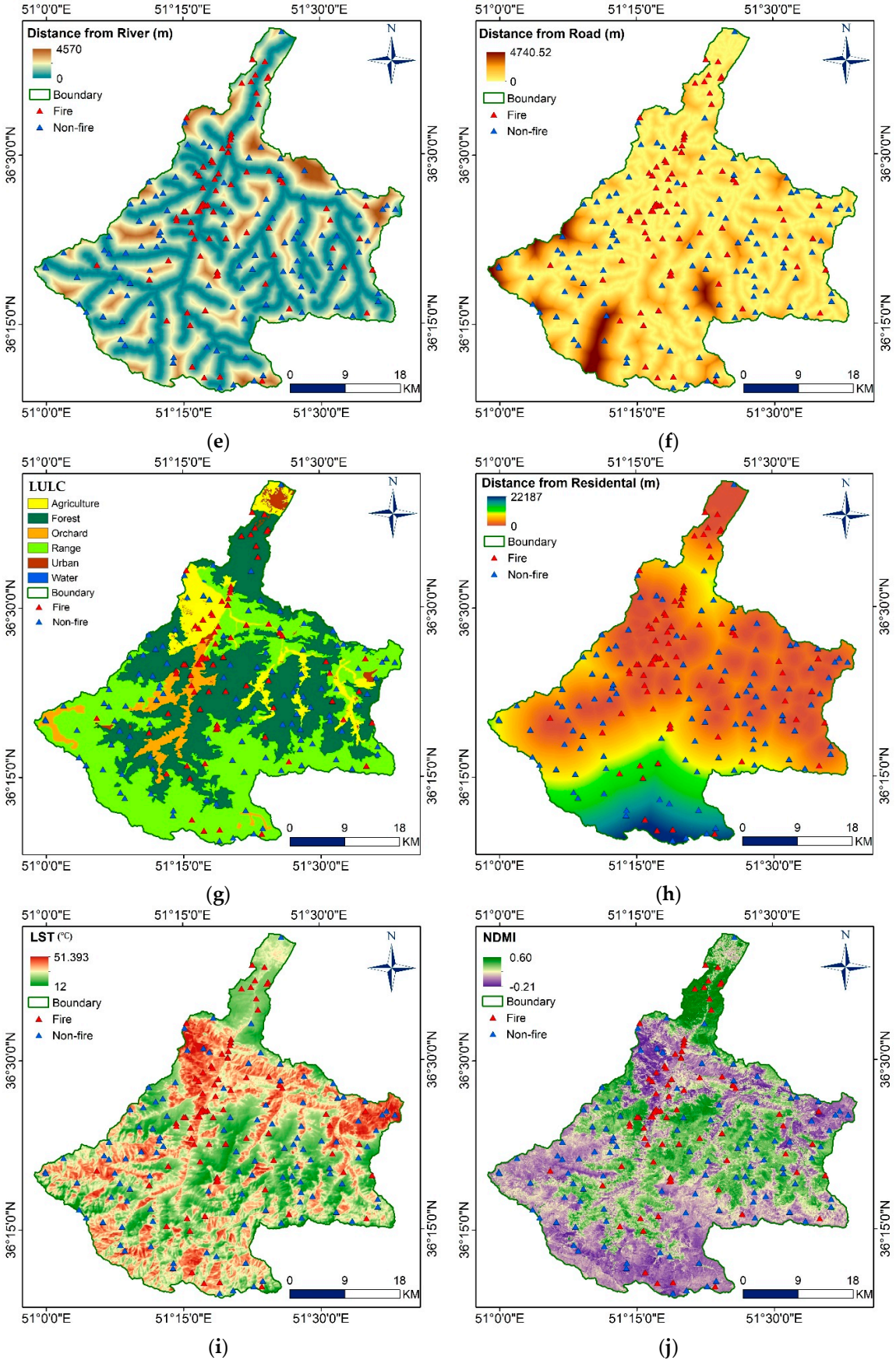


Figure 3. *Cont.*

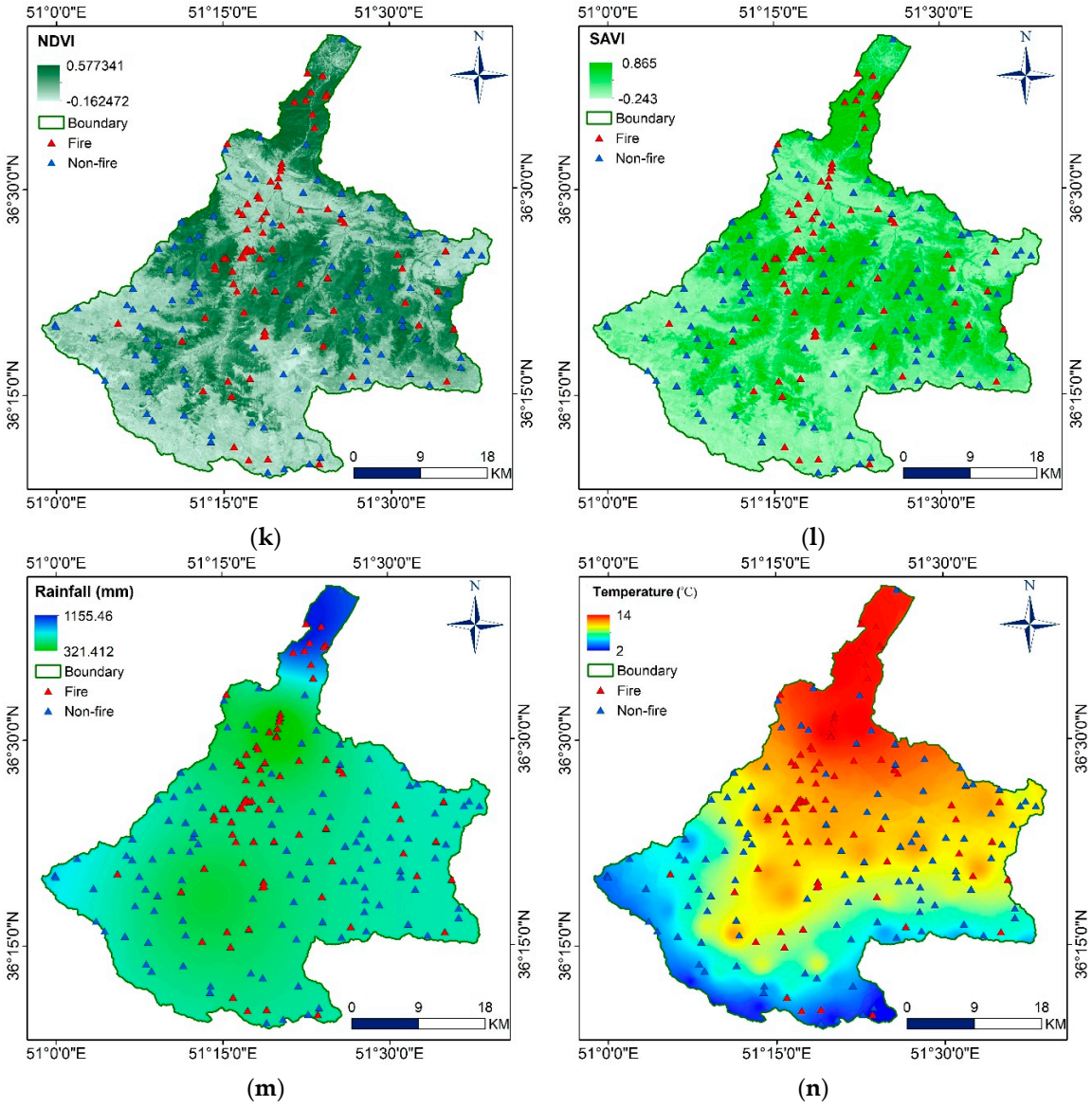


Figure 3. Forest fire conditioning factors: (**a**) altitude, (**b**) aspect, (**c**) slope, (**d**) curvature, (**e**) distance from river, (**f**) distance from road, (**g**) distance from residential, (**h**) LULC, (**i**) LST, (j) NDMI, (**k**) NDVI, (**l**) SAVI, (**m**) rainfall, (**n**) temperature.

Also, forest fire is directly influenced by the vegetation structure usually assessed through the measurement of vegetation canopy using NDVI. Because every index has its limitations [26], we added two other well-known vegetation indexes, SAVI and NDMI, to the NDVI for more accurate analysis of the vegetation composition in the study area. NDVI is certainly the most widely used index for vegetation analysis [17]. NDVI describes the state of health of vegetation computed as the ratio between the difference and the sum of the reflected radiations in the near-infrared (NIR) and the red bands of the electromagnetic spectrum (Equation (1)). Important vegetation parameters, such as live green vegetation, photosynthetic activity of the plant, percent of vegetation ground cover, the amount of biomass and the leaf area and surface water can be obtained from NDVI.

Reflectance of light in the red and NIR spectral affect NDVI values in low vegetation cover (<40%) areas where the surface is exposed [16]. Because of this sensitivity of NDVI to the effects of soil and atmosphere, Huete [37] developed the SAVI (Equation (2)) to minimize soil brightness influences by including a soil adjustment factor L to the NDVI formula so as to correct for soil noise effects (soil variability, color and moisture). Although SAVI accounts for variations in soils, but unlike the

NDVI, it is highly sensitive to atmospheric variations and less sensitive to changes in vegetation amount and cover of vegetation greenness. Compared to NDVI and SAVI which deals with vegetation vigor, the NDMI (Equation (3)) determines vegetation water stress level (vegetation water content) evaluated from the output of the ratio between the difference and the sum of the reflected energy in the NIR and short-wave infrared (SWIR) region of the electromagnetic spectrum [38]. The percentage of recorded energy in the NIR, red, and the SWIR bands of the Landsat 8 OLI imageries provided quantitative parameters of the vegetation health and water stress as indicators of canopy cover structure in forest fire vegetation index maps (Figure 3j–l).

$$\mathrm{NDVI} = \frac{(\mathrm{NIR} - \mathrm{RED})}{(\mathrm{NIR} + \mathrm{RED})} \tag{1}$$

$$\mathrm{SAVI} = \left(\frac{(\mathrm{NIR} - \mathrm{RED})}{\mathrm{NIR} + \mathrm{RED} + \mathrm{L}}\right) * (1 + \mathrm{L}) \tag{2}$$

$$\mathrm{NDMI} = \frac{(\mathrm{NIR} - \mathrm{SWIR})}{(\mathrm{NIR} + \mathrm{SWIR})} \tag{3}$$

where NIR, RED, SWIR and L are near infrared, red, short-wave infrared and soil adjustment factor, respectively. Note that for SAVI, in very high vegetation regions, the value of L = 0 (i.e., SAVI = NDVI); L = 1 in no green vegetation area whereas L = 0.5 is suitable for most conditions and usually the default value. For all the three indexes, the values range from −1 to 1 where lower value indicate lower amount/cover of vegetation greenness or water stress [38].

The two climatic variables, annual rainfall and air temperature (Figure 3m,n) were generated based on 25-year long record of eight rain gauge and meteorological stations employing inverse distance weighted (IDW) interpolation method [38] to produce raster data layers of the respective indices in GIS software environment. Increase in temperature increases the predisposition for fire rage, especially with accompanying speedy dry wind. On the other hand, precipitation increases the soil moisture and supports vegetation greenness that decreases the fire and spread of forest fire. In addition to the climate variables, the LST was derived from the Landsat 8 thermal bands (bands 10 and 11) to obtain the surface heat variation across the area. Table 1 shows the source and impact of each conditioning factors.

3.3. Factor Analysis

The presence of intercorrelated variables, called multi-collinearity, can considerably reduce the accuracy of models produced using bivariate and multivariate methods. Multi-collinearity exists where a particular predictor variable has equal prediction potential of other variables with high level of accuracy in a multiple regression analysis [32]. In this study, the fire conditioning factors were subjected to multi-collinearity test to determine whether or not there is occurrence of collinearity using the quantitative measures of the tolerance and variance inflation factor (VIF) Equation (5). VIF is the measure of the disagreement in a model with multiple relationships by the variance of a model with a single term alone [34]. The values provide indication to the criticality of interrelationship by evaluating the degree of disagreement of an estimated regression coefficient amplified due to collinearity in an ordinary least square regression analysis.

The process requires calculating k different VIFs (one for each X_i) as presented in Equations (4) and (5) below. Consider the linear model with k independent variables; first, an ordinary least square regression with X_i as a function of all the other explanatory variables in Equation (4) is run.

For example, if $i = 1$, the linear regression equation would be:

$$X_i = \alpha_0 + \alpha_0 X_2 + \alpha_3 X_3 + \ldots + \alpha_k X_k + e \tag{4}$$

where α_0 is a constant and e is the error term. Then, the VIF factor for $\hat{\alpha}_i$ is calculated as Equation (5):

$$VIF_i = \frac{1}{1 - R_1^2} \tag{5}$$

where R_1^2 is the coefficient of determination of the regression equation in Equation (4), where X_i is on the left-hand side, and all other predictors are at the other side. The result allows decision on whether or not to exclude a variable(s) from the modelling process based on the VIF value. Generally, the baseline is that if VIF > 10, multicollinearity is considered to be high. However, the cutoff threshold of 5 is commonly used [24].

Table 1. Spatial dataset and data sources.

Conditioning Factor	Source	Impact
Altitude	ASTER DEM	Controls the microclimate in terms of vegetation distribution, composition and flammability.
Aspect	ASTER DEM	Influences the local climate of the slope with respect to solar insolation, wind moisture content, etc. The hill side facing away from the direct sunshine usually retain more moisture supporting vegetation greenness and vigor.
Slope	ASTER DEM	Slope regulates vegetation distribution and composition, with high impact on the direction in which the fire rage and the speed at which it spreads, particularly at steep slope.
Curvature	ASTER DEM	Curvature indicates convergence or divergence of water in the landscape simultaneously with respect to downhill flow interpreted as either negative, zero or positive curvature. Negative curvature represents concave flow channel, zero curvature shows flat surfaces while positive curvature depicts convex flow waterway.
Distance from river	GIS data	Vegetation close to rivers tend to be more greenish and healthier. Providing fuel for wildfire. However, moisture content in the vegetation could insulate inflammability.
Distance from road	GIS data	Roads provide accessibility to forests and consequently forest fire initiation through human activities.
NDVI	Landsat 8 OLI	It measures vegetation surface cover and density which indicates availability of fuel for fire spreading.
SAVI	Landsat 8 OLI	It measures vegetation amount, vigor and cover of greenness as indicators of flammability or otherwise of the area.
NDMI	Landsat 8 OLI	Moisture content of stressed vegetation fuel forest fire. NDMI measures the stress level and consequently degree of flammability.
LULC	Landsat 8 OLI	Describes the various use the land is put into and the activities involved, including fuel types and level of exposure to fire.
Distance from residential	LULC	The further away forest is to residential areas the lesser its vulnerability to fire occurrence.
LST	Landsat 8 OLI	Shows surface heat variation and its contribution to the spread of fire
Temperature	Meteorological data	Influences atmospheric air mass which controls relative humidity, air mass and soil moisture content.
Rainfall	Meteorological data	Increases the soil moisture and supports vegetation greenness that decreases the rage and spread of forest fire.

3.4. Resampling: Cross Validation (CV), Bootstrap and Optimism Bootstrap

As mentioned earlier, the validation process involved distributing the inventory data into 70% building and 30% for model evaluation. Out of the 218 inventory location points, 152 location points representing 70% was used for training the models and the other 30% (66 points) used for validation. To assess the adequacy of a statistical model and estimating the variance and the bias of estimation, the training dataset was subjected to resampling process using CV, bootstrap, and optimism bootstrap.

Resampling is the process of taking samples from a training data set repeatedly and refitting a model to get more information about it.

CV is a resampling method commonly employed in classification problems to assess the suitability of a statistical model. In principle, CV randomly splits the sample data into model building and test dataset for assessment of the models' performance accuracy for prediction. This approach eliminates bias in sampled data and prevents under/overfitting of the model during optimization [39]. In this study, 5- and 10-fold CV implementation scheme were used. Figure 4 presents CV implementation schema.

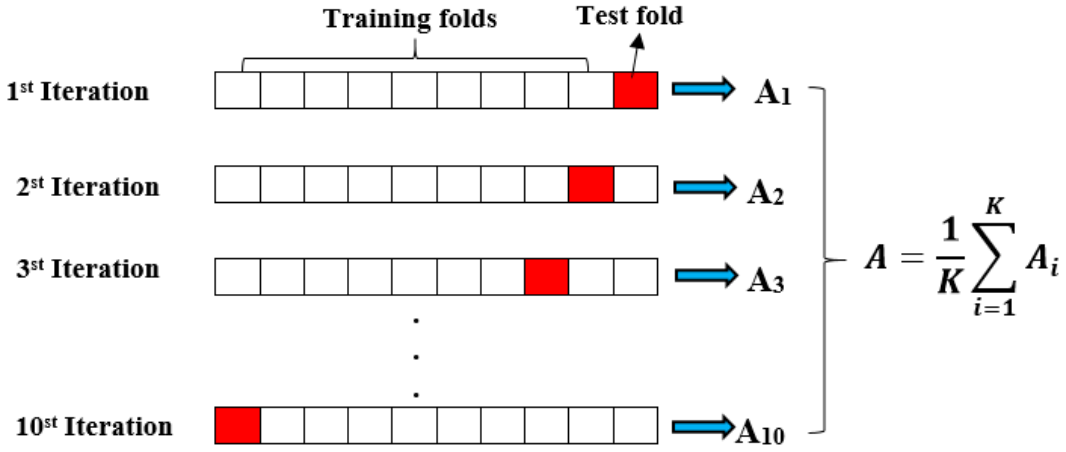


Figure 4. Schematic representation of cross validation (CV) implementation.

Bootstrap, a nonparametric resampling method developed in the mid-1970s by Efron, is another resampling technique that has gained popularity as a flexible, statistically efficient resampling method that makes it possible to estimate statistical parameters such as standard errors, *p*-values, confidence intervals (CI) [39]. The ability to arrive at unbiased inferential decision by sampling the observed data a specified number of times, with replacement, distinguished bootstrap from other resampling methods. More information on the principle and mathematical bases of bootstrap can be found in [39,40]. The legend advanced this process by using his own bootstraps to get out of sea. His legendry discovery resulted to the name "bootstrapping", a technique for estimating the distribution of an estimator where sampling is drawn from the original sample with replacement [41]. Bootstrapping is similar to the process of selecting population sample from a universal dataset. Since only a sample of the population is considered, it does not truly provide complete representation of the reality. Bootstrap is based on probability theory where samples are drawn with replacement from the original sample to create a probabilistic element [39,41]; nonetheless, the sample maintains its original sample size to ensure that the precision of estimates obtained in each bootstrap sample remains invariant.

Bootstrapping is an important tool for measuring the optimism of prediction models [39]. This is usually done by repeatedly fitting a model in bootstrap samples using a simple bootstrap modifier and the performance assessed by comparing the result with the initial sample. Average performance improvement of the bootstrap methods on the original sample are useful for predicting potential forest fire more accurately. For better accuracy, bootstrap is employed to assess optimism by determining variation in performance of the original sample and that of the bootstrapped sample and using the estimate to adjust the performance of the bootstrap sample, called optimism-bootstrap.

3.5. Model Implementation

Since early 1990s, AI has been applied in wildfire studies using neural networks and expert systems [18,42–45]. Ever since then, the field has grown rapidly with the wide adoption of several ML algorithms in forest fire prediction for wildfire disaster preparedness, mitigation and

assessment [14]. This development has brought about a natural evolution in the science of forest fire management to a data-centric approach which leverage on empirical and statistical models; motivated by the availability of sufficient datasets obtainable from different sensors providing high-quality spatiotemporal data [14,46]. In addition to the development in the computational efficiency of modern computer systems, there has been a growing interest in the use of ML in forest fire prediction in recent years. The use of computational intelligence in mapping has grown over the years [43]. This has resulted to the development of dozens of suitable approaches to approximate the decision boundary in binary sampled dataset. As mentioned earlier, in the current study, MARS, BRT, and SVM models are explored.

3.5.1. Multivariate Adaptive Regression Splines (MARS)

Inspired by the classification and regression techniques, Friedman [47] proposed the MARS prediction algorithm by simplifying splines methods [25]. Based on literature sources, MARS has demonstrated effectiveness in determining a classification model capable of producing a good prediction accuracy and at the same time prevents overfitting [14]. In principle, MARS automatically partitions the input datasets into sub-groups, and the data in each sub-space fitted using a linear function and smoothing spline regression to produce an adaptive global model. Moreover, the model has the capability to identify association in the input data effectively and to differentiate the underlining variables that influences the model. Details on the principles of MARS algorithm can be found in the work of Jaafari et al. [48], Jain et al. [18], Tien et al. [25].

3.5.2. Support Vector Machine (SVM)

SVM is a popular statistical ML algorithm with a supervised learning binary classifier first proposed by Cortes and Vapnik [1,4]. Over time, particularly with increase in sensor development and increasing data availability, SVM has been used successfully in many earth observations related real-world problems, including fire prediction [7,21]. Empirically, SVM maximizes the margin between the data points to be separated to find an optimal separating hyperplane. This implies that only support vectors are important to maximizing the margin with only a linear classifier, while other training examples are ignorable. Compared to other ML algorithm, SVM is capable of solving classification and regression problem and can handle large feature space [49,50]. In addition, it offers flexibility in choosing a similarity function and minimizes overfitting. Readers can refer to the works of Brown et al. [7], Ghorbanzade et al. [11], Jain et al. [18] and Tehrany et al. [23] for detail information on SVM model.

3.5.3. Boosted Regression Tree (BRT)

The BRT model is tree decision technique developed by fusing several simple classical regression tree models into a single model. BRT combines the potentials of classical regression trees and boosting algorithms to improve predictive performance [39]. The regression tree employs a piece-wise linear estimate of a regression function built by iteratively splitting the data into sub sample space; while, concurrently, the boosting algorithm increases predictive performance of the regression trees by averaging and merging the results of several contending models which are fitted repeatedly to a subset of the training data [39] the BRT model has several advantages [24], including the ability to handle different types of predictor variables and automatically addressing interaction effects between them, and fitting complex nonparametric relationships. Furthermore, the BRT model accommodates missing data and does not require prior data transformation, and ultimately it decreases overfitting. More information can be obtained from from Chernick [39], Pourghasemi et al. [24], and Shabani et al. [22].

3.6. Validation

The fire inventory dataset randomly split into 70% and 30% provided the basis for training and validation process, respectively. While the first subset (70%) was used for training the models, the second

part (30%) was employed to validate the models using the (receiver operating characteristics) ROC-AUC accuracy assessment parameter. Furthermore, consideration of the performance evaluation of each model relative to CV, bootstrap and optimism-bootstrap resampling techniques were determined, similarly assessing their superiority by the ROC-AUC illustrated both using their success and prediction rate results supported with the sensitivity, specificity, NPV, and the PPV measures. The training and validation process of the model building phase, with and without optimization, provides quantitative parameters to quantify the model performance of the models using ROC-AUC curve metrics; sensitivity, specificity, NPV, PPV. Each of these parameters are evaluated with respect to the accuracy obtained with the training and testing subset. The graph of the sensitivity-specificity plots provides visual and statistical measure of how well the models and the different resampling algorithms are able to predict forest fire in the study area.

4. Results

Assessment of the independent variables for multi-collinearity provided insight of their respective importance and position for optimal model building. The degree of association between two or more independent variables is measured by considering the duo of VIF and tolerance. By standard, an individual variable with VIF exceeding 5 and tolerance less than 0.1 is considered inappropriate for inclusion into the model building [51,52]. From the result, the maximum and minimum VIF and tolerance are (4.2, 1.1) and (0.9, 0.2), respectively (Table 2). It is observed that the VIF values of NDMI and NDVI variables are closer to the critical margin than any other variable, similarly their respective tolerance values. Perhaps, this arises because both measures related vegetative quantities in dataset (i.e., the amount of water content present, vegetation health and greenness). High soil moisture content means that the vegetation will have sufficient water to support photosynthesis resulting to healthy and greener leaves [25]. In this study, none of the predictive variables exceeds the critical threshold of the evaluating pair of VIF and tolerance.

Table 2. Multi-collinearity assessment result.

Row	Variables	VIF	Tolerance
1	Altitude	3.08	0.32
2	Slope	1.51	0.66
3	Aspect	1.10	0.91
4	Curvature	1.19	0.84
5	Distance from river	1.84	0.54
6	Distance from road	1.54	0.65
7	Distance from residential	2.12	0.57
8	LULC	1.90	0.53
9	LST	3.26	0.31
10	NDMI	3.95	0.25
11	NDVI	4.23	0.24
12	SAVI	3.06	0.33
13	Rainfall	1.41	0.71
14	Temperature	2.48	0.40

In this study we investigate the performance of the 5-fold, 10-fold, bootstrap and optimism-bootstrap resampling techniques on the three ML algorithms (MARS, SVM and BRT). Overall, the BRT model yielded the best result across all the resampling methods. Table 3 presents detailed result of the performance evaluation for the models using the optima combination of training and testing percentage ratio from the historical forest fire sample dataset. Across the models, the success rate of the test dataset for the non-resampling model output is generally lower than those produced from 5/10-fold CV and bootstrapping.

Table 3. Predictive capability of forest fire models using train and test dataset.

Models	Resampling	Stage	Evaluate Parameters				
			Sensitivity	**Specificity**	**NPV**	**PPV**	**AUC**
MARS	non	Train	0.72	0.75	0.73	0.71	0.83
		Test	0.65	0.74	0.70	0.69	0.78
	5 fold CV	Train	0.84	0.86	0.88	0.83	0.92
		Test	0.77	0.77	o.79	0.75	0.88
	10 fold CV	Train	0.88	0.93	0.92	0.91	0.94
		Test	0.74	0.86	0.79	0.82	0.90
	bootstrap	Train	0.85	0.89	0.86	0.84	0.91
		Test	0.71	0.77	0.75	0.73	0.86
	optimism bootstrap	Train	0.89	0.93	0.93	0.91	0.93
		Test	0.77	0.80	0.80	0.77	0.83
SVM	non	Train	0.78	0.86	0.88	0.91	0.92
		Test	0.65	0.83	0.75	0.77	0.82
	5 fold CV	Train	0.79	0.86	0.87	0.90	0.92
		Test	0.65	0.83	0.75	0.77	0.82
	10 fold CV	Train	0.93	0.92	0.96	0.94	0.97
		Test	0.84	0.83	0.85	0.81	0.89
	bootstrap	Train	0.86	0.92	0.91	0.89	0.95
		Test	0.77	0.80	0.80	0.77	0.87
	optimism bootstrap	Train	0.87	0.82	0.81	0.86	0.92
		Test	0.84	0.74	0.74	0.84	0.84
BRT	non	Train	0.79	0.84	0.79	0.84	0.91
		Test	0.74	0.83	0.78	0.79	0.87
	5-fold CV	Train	0.94	0.95	0.97	0.96	0.98
		Test	0.87	0.88	0.89	0.87	0.90
	10-fold CV	Train	0.95	0.95	0.98	0.97	0.98
		Test	0.84	0.83	0.85	0.S81	0.90
	bootstrap	Train	0.96	0.94	0.98	0.93	0.98
		Test	0.87	0.80	0.87	0.79	0.90
	optimism bootstrap	Train	0.98	0.97	0.99	0.98	0.99
		Test	0.87	0.83	0.89	0.82	0.91

The performance evaluation process has shown that the resampling method have significant influence on the overall accuracy of the resulting models. For example, the 10-fold CV step performed best in the MARS model with AUC value of 0.90 compared to non-resampling (0.78), 5-fold CV (0.88), bootstrap (0.86) and optimism bootstrap (0.83), respectively. Similarly, for the SVM model, the 10-fold CV indicated the best with AUC value of 0.89 as against the AUC value of 0.82 for both of non-resampling and 5-fold CV, bootstrap and optimism bootstrap of AUC values of 0.87 and 0.84, respectively. The BRT model has shown superiority in performance with all the resampling methods. Ultimately, the BRT model proven to have the best prediction accuracy at all stages of assessment and sampling techniques. For instance, the BRT non-sampling model outperforms its counterparts in the MARS and SVM models with AUC value of 0.87. Likewise, the 5-fold CV, the 10-fold CV and the bootstrap resampling methods each of which produced AUC value of 0.90 across board while the optimism bootstrap indicated the overall best accuracy with AUC value of 0.91. The BRT model produced the best result across all the tested methods with particular attention to 5-fold, 10-fold, bootstrap and optimism bootstrap resampling methods which produced AUC values of 0.90, 0.90, 0.90, and 0.91, respectively. The forest fire map was generated from the fire-related indicators considered

in this study (see Table 1). The modeling process exploits the interaction between these indicators (independent variables) based on the sampled dataset to establish relationship that accurately produce the forest fire susceptibility maps with the MARS, SVM, and BRT model outputs (10-fold, and optimism bootstrap resampling methods—Figure 5).

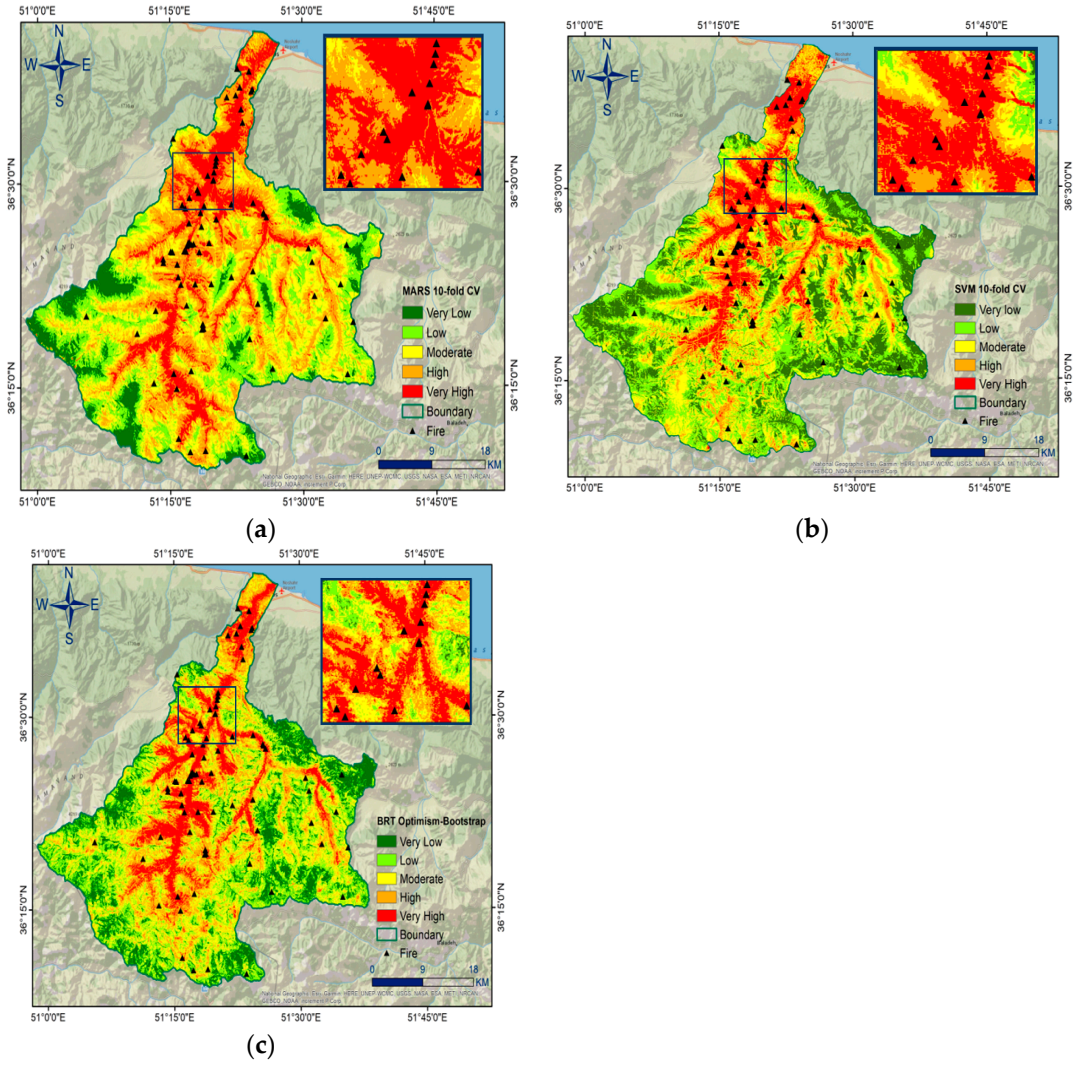


Figure 5. Forest fire susceptibility maps produced with (**a**) MARS 10-fold CV (**b**) SVM 10-fold CV and (**c**) BRT optimism bootstrap.

In each map, the degree of susceptibility is categorized into five classes using natural break method: very high, high, moderate, low, and very low, as described in [25]. The maps exhibit similar pattern of susceptibility in their respective grouping. For example, the very high and high susceptibility classes are prominent north of the study area which trends north-south, occupying the central zone (Figure 5a–c). Nonetheless, from the output maps, it is observed that the very high and high fire susceptibility classes of the BRT using optimism bootstrap resampling is more concentrated at the central region (Figure 5c). Conversely, the high and very high classes are more spread out in the 10-fold MARS and 10-fold SVM model generated maps (Figure 5a,b). Again, the low to very low susceptible classes are majorly found around the east, south and west border, though at varying spatial extent with 10-fold MARS and 10-fold SVM model but less prominent in the optimism Bootstrap BRT result. This phenomenon resulted in spatial variation in the respective susceptibility class across the models. Table 4 provides the quantitative analysis of susceptibility class coverage across the models.

Table 4. Forest fire susceptibility classes' area.

Susceptibility Class	Models					
	10-Fold CV MARS		10-Fold CV SVM		Optimism Bootstrap BRT	
	Area (Km^2)	Area (%)	Area (Km^2)	Area (%)	Area (Km^2)	Area (%)
Very Low	132.01	8.09	301.96	18.51	241.44	14.80
Low	320.71	19.66	447.21	27.41	425.78	26.1
Moderate	479.41	29.39	387.32	23.74	426.26	26.13
High	468.11	28.69	284.25	17.42	326.53	20.02
Very High	231.1	14.17	210.6	12.91	211.33	12.95

Based on the 10-fold MARS map, the very high, high and moderate forest fire susceptibility classes cover 14.17%, 28.69%, and 29.39% approximately 231.1, 468.11 and 479.41 square kilometers of the study area, respectively. The low and very low classes take up 19.66% (320.71 km^2) and 8.09% (132.01 km^2). Similarly, the 10-fold resampling SVM model output map predicts the very high, high and moderate classes to cover 12.91% (210.6 km^2), 17.42% (284.25 km^2) and 23.74% (387.32 km^2) of the study area. Whereas the low and very low classes are forecasted to cover 27.41% (447.21 km^2) and 18.51% (301.96 km^2) of the area under investigation, respectively. For the optimism bootstrap BRT, the very high, high and moderate classes take up 12.95% (211.33 km^2), 20.02% (326.53 km^2) and 26.13% (426.26 km^2) land mass, in that order. The same model predicts low and very low classes to cover 26.10% and 14.80%, approximately 425.78 km^2 and 241.44 km^2, respectively. The area coverage of the fire susceptibility classes of 10-fold CV MARS model varied slight from those of 10-fold CV SVM and optimism bootstrap BRT counterparts, which give relatively similar results. Aggregation of the 10-fold CV MARS area coverage of the very high, high and moderate susceptibility classes covers 72.25% of the study area compared to 54.07% and 57.10% with the 10-fold CV SVM and optimism bootstrap, respectively.

Further analysis of the accuracy assessment statistical results, in addition to the AUC, is the plots of the ROC curve expressing the sensitivity/specificity (Figure 6). Comparatively, sensitivity shows how good each technique is at determining portions are truly susceptible forest fire (true positive rate) while the specificity indicates the ability of the models to correctly identify non-fire susceptible areas (true negative rate). Sensitivity and specificity values range from 0 to 1; the closer their values to 1 indicate how good the model is at determining the probability of susceptibility fire, or otherwise.

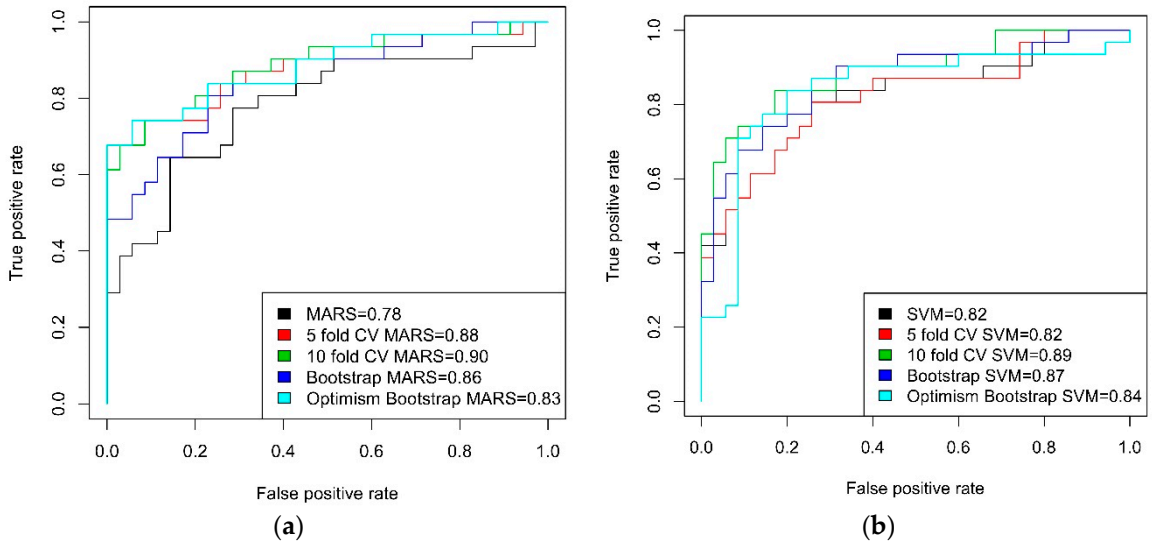
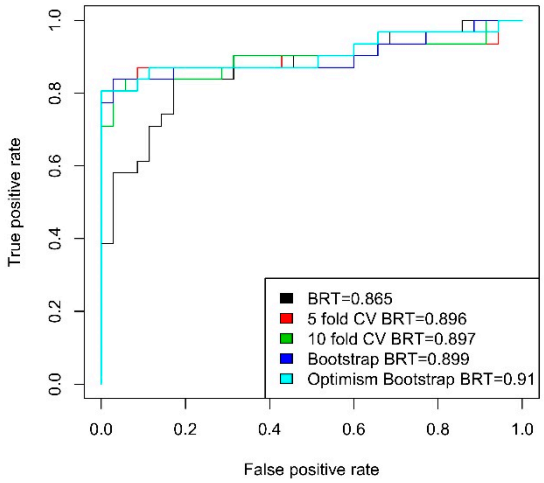


Figure 6. *Cont.*



(**c**)

Figure 6. ROC curve analysis of (**a**) MARS, (**b**) SVM and (**c**) BRT for forest fire susceptibility models using the test validation dataset.

For the MARS model, the sensitivity and specificity values of the resampling methods range from 0.65 to 0.77 and 0.74 to 0.86, respectively. Similarly, for the SVM model, the sensitivity/specificity values produced 0.65 to 0.84 and 0.74 to 0.83 while the BRT model yielded 0.74 to 0.87 and 0.80 to 0.88, respectively. Basically, in the three models, the 5- and 10-fold CV, bootstrap and optimism bootstrap resampling methods improved the prediction accuracy. With AUC, sensitivity and specificity values close to 1, it is evidenced, statistically, that the models reliably predict forest fire in the study area.

Further model evaluation was made by producing forest fire density graphs, as shown in Figure 7. The number of forest fires in each susceptibility class was calculated and plotted in a two-dimensional chart, where its horizontal axis showed the susceptibility class, and the number of forest fires was presented in the vertical axis. The results showed that the BRT optimism bootstrap model predicted the highest forest fires percentage (58.71%) in the very high class. In addition, the SVM 10-fold CV and MARS 10-fold CV models predicted 57.79% and 43.11% of forest fires in the very high susceptibility class.

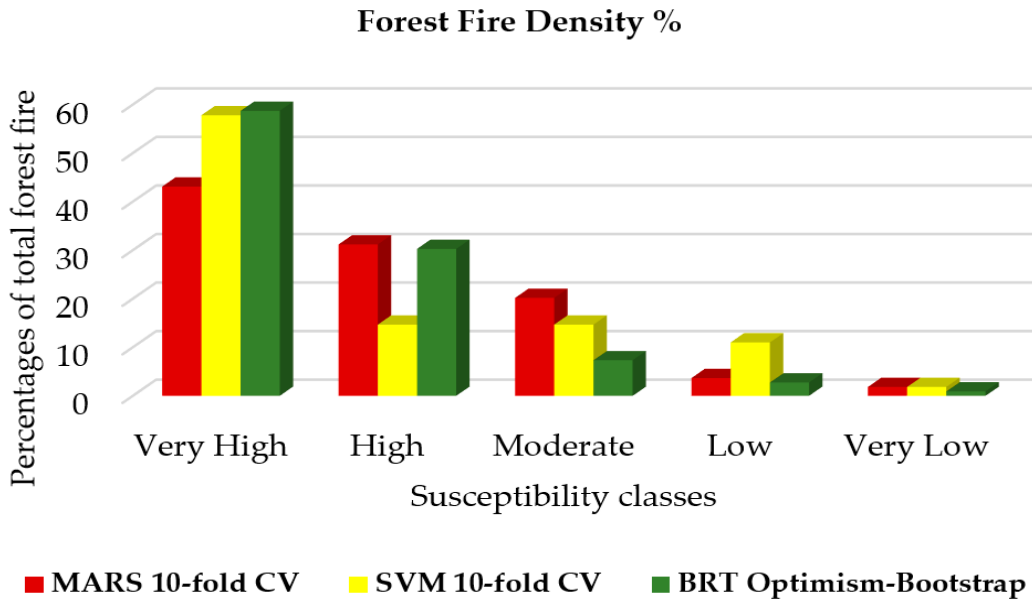


Figure 7. Forest fire density graph.

It is unscientific to assume that the predictor variables contribute to the model building process equally. Therefore, the relative importance of the variables (Table 5) was assessed by ranking using the optimism bootstrap BRT technique. The percentage importance ranges from 0.7% to 16.8% with SAVI at the bottom while altitude ranks first. Next to altitude in order of importance are distance from river, rainfall and distance from residential area with 13.0%, 10.8% and 9.0%, respectively. This is followed by curvature (8.5%), NDMI (7.0%), temperature (6.8%), aspect (6.4%), distance from road (6.0%), and LST (5.9%). The factors constituting the least influential variables are NDVI (3.8%), slope (3.4%), LULC (1.9%) and SAVI (0.7%).

Table 5. Percentage importance of predictor variables.

No	Variables	Value	% Importance
1	Altitude	33.77	16.8
2	Distance from river	26.25	13.0
3	Rainfall	21.71	10.8
4	Distance from residential	18.09	9.0
5	Curvature	17.19	8.5
6	NDMI	14.07	7.0
7	Temperature	13.76	6.8
8	Aspect	12.87	6.4
9	Distance from road	12.06	6.0
10	LST	11.94	5.9
11	NDVI	7.56	3.8
12	Slope	6.78	3.4
13	LULC	3.81	1.9
14	SAVI	1.43	0.7

The response curve of the first four most important variables (altitude, distance from river, rainfall, distance from residential areas) using lattice-based partial dependence plots (PDPs) from the optimism bootstrap BRT is presented in Figure 8. In Figure 8a, it is observed that areas with height less than 1000 m have high susceptibility to fire and with increasing altitude the degree of susceptibility decreases. Conversely, the level of susceptibility decreases with increase in height above 2000 m but between 3000 to 4000 m susceptibility level is fixed. From Figure 8b, areas close to the river (less than 1000 m) reveal high susceptibility to forest fire. Annual rainfall changes on fire susceptibility indicate that areas with 400 to 500 mm of rainfall are more susceptible to fire. As the amount of rainfall increases from 500 mm, the susceptibility decreases and this trend continues until the rainfall amounts 600 mm. Examination of the changes in the distance from residential areas shows that areas with distance less than 5000 m from residential areas have higher susceptibility to fire but decrease with increasing distance from residential areas to a distance of 10,000 m and continues until between 15,000 and 20,000 m where sensitivity to changes remain constant and unchanged.

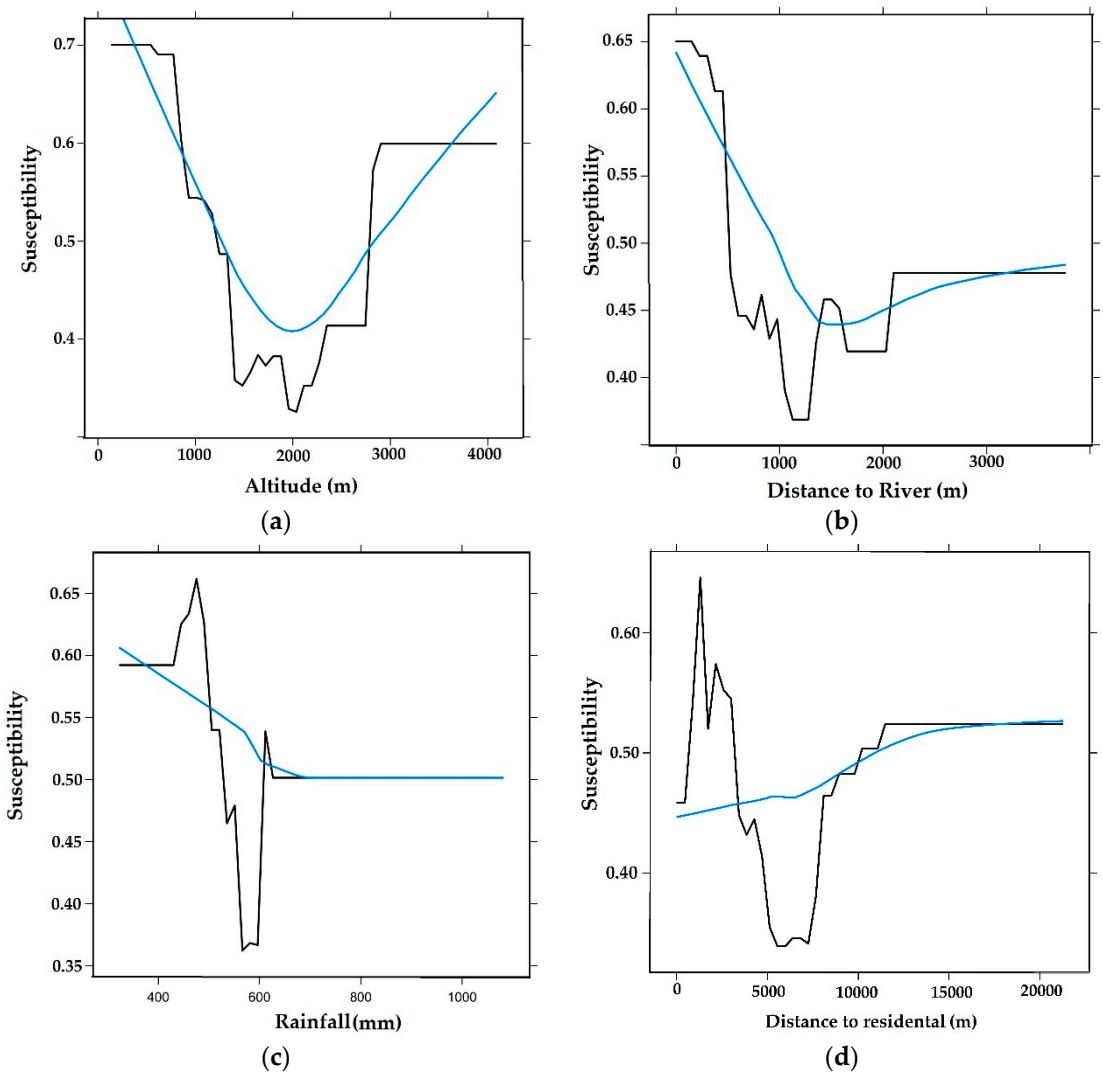


Figure 8. Response curve of variable importance: (**a**) altitude, (**b**) distance from river, (**c**) rainfall and (**d**) distance from residential using the lattice-based PDP.5.

5. Discussion

Determining the best method to predict wildfire is challenging due to associated spatial heterogeneity of the conditioning factors. Multicollinearity reduces the accuracy of estimate quantities by weakens the significance of a regression model. So, the first step to ensure reliable forest fire susceptibility map from the models was to evaluate possible association among each of the 14 input fire conditioning factors considered. Since there is no serious linear association between any of the explanatory data based on the estimated values of the VIF and tolerance, all the 14 variables were included in the model building.

Among the various ML algorithms available, those of MARS, SVM, and the BRT have been used to predicate the level and spatial variability of forest fire susceptibility in Chaloos Rood watershed. To understand this, we selected the best map (Figure 5) among the resampling methods in each ML algorithm on the basis of the statistical performance evaluation of the testing samples (Table 3). So, each map was classified into five classes (Very high, High, Moderate, Low and Very Low) of susceptibility for analysis. The results showed that the forest fire susceptible map of the 10-fold CV MARS model varies slight from those of 10-fold CV SVM and optimism bootstrap BRT in terms of spatial distribution of the susceptibility classes. It is observed that the moderate fire risk zones cover almost equal spatial extent across the three models (between 23.7% and 29.4%) but the upper and

lower risk categories are different. The very high and high classes in 10-fold CV MARS model occupies ~47% of the study area compared to 30% and 33% in the 10-fold CV SVM and optimism bootstrap BRT models. Conversely, low and very low risk categories cover ~28% in the 10-fold CV MARS model while the same bands occupy ~46% and ~41% for the other two models.

Our approach has shown the robustness of optimizing the process of selecting the appropriate testing and validation percentage ratio by employing different resampling methods. The performance evaluation process has shown this through analysis of the accuracy of the MARS, SVM and BRT models through the AUC values. AUC has a standard scale, measured from 0–1, which indicates the degree of accuracy; value < 0.6 is interpreted to have low accuracy while those between 0.6–0.7, 0.7–0.8, 0.8–0.9, and >0.9 are interpreted to mean that the evaluation has moderate, good, very good, and excellent accuracy, respectively. In this study, the performance evaluation yielded AUC values of 0.89, 0.90 and 0.91 for the 10-fold CV MARS, 10-fold CV SVM and optimism bootstrap BRT models, correspondingly–all of which fall in the very good and excellent accuracy performance classification. Our findings were in agreement with previous works that investigated the same models. As an example, Shabani et al. [22] discovered that BRT is more effective in terms of prediction ability with an AUC value of 0.94 compared to LR. Pourghasemi et al. [24] indicated that BRT (AUC training = 88.90% and AUC testing = 88.2%) was more effective model compare to mixture discriminant analysis (MDA) and general linear model (GLM).

Sensitivity of the fire conditioning factors was analyzed as a function of percentage contribution to the model building. The variable with minimum impact has 0.7% contribution while the variable with the maximum influence constitutes 16.8% (Table 5). Top on the list are altitude, distance from river, rainfall and distance from residential which contributed 16.8%, 13.0%, 10.8% and 9.0%, respectively. This is consistent with several previous studies, i.e., [20,24,36,53]. For example, Bui et al. [20] found that distance to residence area is one of the most important variables with predictive ability of 0.281 in forest fire susceptibility. The finding is reasonable because distance from residence area is related to anthropogenic factor that is the main cause of forest fire. In other study done by Guo et al. [53] indicated the importance of elevation as underlying factor of fire occurrence. More recently, Eskandari et al. [36] concluded that the mean annual rainfall had the highest relative importance in fire spatial occurrence. Pourghasemi et al. [24] emphasized on the importance of annual mean rainfall, and elevation, while the distance to river and roads were not highly remarkable for forest fire events which it demonstrates the site's dependency of the explanatory variables. This is obviously not a coincidence; all the other variables are directly linked to topography, precipitation and the anthropogenic factors. For instance, the amount of vegetation cover, canopy structure and moisture content are controlled by altitude and the amount of rainfall [34,48] which in turn determine derivatives such as NDSM, SAVI, and NDMI. Similarly, the altitude of a place defines the slope, aspect, curvature, river channel and, to a great extent, LULC. Graphical illustration demonstrates the implication of these most influential variables in the response curve (Figure 8). For example, at lower altitude, the risk of fire ignition is higher, and decreases with increasing altitude up to ~3000m above msl beyond which the risk curve flattens. Similarly, forest areas close to the river 1 km have higher risk of forest fire initiation than those further away. This phenomenon is also observed in the relationship between forest fire and residential area; forest within 5 km radius from residential areas are more prone to fire outbreak and the risk reduces with increase in distance. But, unlike the distance to river and residential area, increasing rainfall counters this trend as increase in the amount of precipitation lessen the chance of forest fire. This study has proven that application of ML algorithms (MARS, SVM and BRT) allows predicting forest fire in agreement with similar studies in similar environment, see [11,12,32,42,48,54]. However, the application of resampling methods to optimize training/testing ratio from the sampled data produced fire susceptibility map with better accuracy.

6. Conclusions

The paper examined the application of MARS, SVM and BRT ML methods to predict forest fire proneness in the Chaloos Rood watershed, Iran. A set of vegetation indices, climatic variables, environmental factors, and topographical/physiographic elements were integrated with the forest fire inventory location points to train and validate the models. Experimenting with different optimization techniques enhanced the models' performances yielding accuracy of prediction at AUC of 0.90, 0.89 and 0.91, for the MARS, SVM and BRT, respectively. Interestingly, the 10-fold CV produced the best accuracy for the MARS and SVM models whereas the optimism bootstrap produced the overall best accuracy with the BRT model. This study reveals that, altitude, distance from river, rainfall, distance from residential, and curvature contributed effectively to forest fire prediction. While all the models produced good results in mapping the spatial variability of potential forest fire, the BRT model produced a more distinct pattern of coverage and class distribution from the MARS and SVM models. Eve though there are several instances of using ML techniques including MARS and SVM separately for producing fire risk map, we add BRT, a rarely used ML method in this study. The novelty of our study is using both data-driven ML algorithms and optimizing them through different resampling methods to produce high-quality forest fire vulnerability map. The reasons for incorporating different data layers is to provide a broader perspective of social, environmental, climate and anthropogenic forest fire triggering agents. Our results emphasize the importance of eliminating bias in dataset during the modeling process using resampling techniques so as to improve the reliability of the resulting forest fire risk maps.

Author Contributions: S.J. and K.A. acquired the data; B.K., and S.J. conceptualized and performed the analysis; B.K. and M.O.I. wrote the manuscript, discussion and analyzed the data; N.U. supervised including the funding acquisition; B.K. and F.S. provided technical sights, as well as edited, restructured, and professionally optimized the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: The APC is supported by the RIKEN Centre for Advanced Intelligence Project (AIP), Tokyo, Japan.

Acknowledgments: The authors would like to thank the RIKEN AIP, Japan for providing all facilities during the research.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bui, D.T.; Le, K.T.T.; Nguyen, V.C.; Le, H.D.; Revhaug, I. Tropical forest fire susceptibility mapping at the Cat Ba National Park area, Hai Phong City, Vietnam, using GIS-based Kernel logistic regression. *Remote Sens.* **2016**, *8*, 347. [CrossRef]
2. Balch, J.K.; Bradley, B.A.; Abatzoglou, J.T.; Chelsea Nagy, R.; Fusco, E.J.; Mahood, A.L. Human-started wildfires expand the fire niche across the United States. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 2946–2951. [CrossRef] [PubMed]
3. Randerson, J.T.; Liu, H.; Flanner, M.G.; Chambers, S.D.; Jin, Y.; Hess, P.G.; Pfister, G.; Mack, M.C.; Treseder, K.K.; Welp, L.R.; et al. The impact of boreal forest fire on climate warming. *Science* **2006**, *314*, 1130–1133. [CrossRef] [PubMed]
4. Ireland, G.; Petropoulos, G.P. Exploring the relationships between post- fire vegetation regeneration dynamics, topography and burn severity: A case study from the Montane Cordillera Ecozones of Western Canada. *Appl. Geogr.* **2015**, *56*, 232–248. [CrossRef]
5. Nölte, A.; Meilby, H.; Yousefpour, R. Multi-purpose forest management in the tropics: Incorporating values of carbon, biodiversity and timber in managing Tectona grandis (teak) plantations in Costa Rica. *For. Ecol. Manag.* **2018**, *422*, 345–357. [CrossRef]
6. Lamb, D.; Erskine, P.D.; Parrotta, J.A. Restoration of degraded tropical forest landscapes. *Science* **2005**, *310*, 1628–1632. [CrossRef]
7. Brown, A.R.; Petropoulos, G.P.; Ferentinos, K.P. Appraisal of the Sentinel-1 & 2 use in a large-scale wildfire assessment: A case study from Portugal's fires of 2017. *Appl. Geogr.* **2018**, *100*, 78–89.
8. Bruinsma, J. Towards sustainable forestry. In *World Agriculture: Towards 2015/2030: An FAO Perspective*; Diouf, J., Ed.; Earthscan Publications Ltd.: London, UK, 2003.

9. Pricope, N.G.; Binford, M.W. A spatio-temporal analysis of fire recurrence and extent for semi-arid savanna ecosystems in southern Africa using moderate-resolution satellite imagery. *J. Environ. Manag.* **2012**, *100*, 72–85. [CrossRef]
10. Key, C.H.; Benson, N.C. Landscape assessment: Remote sensing of severity, the normalized burn ratio and ground measure of severity, the composite burn index. In *FIREMON: Fire Effects Monitoring and Inventory System*; General Technical Report; RMRS-GTR-164-CD; LA1-LA51; U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station: Fort Collins, CO, USA, 2005; pp. 305–325.
11. Ghorbanzadeh, O.; Valizadeh Kamran, K.; Blaschke, T.; Aryal, J.; Naboureh, A.; Einali, J.; Bian, J. Spatial Prediction of Wildfire Susceptibility Using Field Survey GPS Data and Machine Learning Approaches. *Fire* **2019**, *2*, 43. [CrossRef]
12. Pourtaghi, Z.S.; Pourghasemi, H.R.; Aretano, R.; Semeraro, T. Investigation of general indicators influencing on forest fire and its susceptibility modeling using different data mining techniques. *Ecol. Indic.* **2016**, *64*, 72–84. [CrossRef]
13. Tshering, K.; Thinley, P.; Shafapour Tehrany, M.; Thinley, U.; Shabani, F. A Comparison of the qualitative analytic hierarchy process and the quantitative frequency ratio techniques in predicting forest fire-prone areas in Bhutan using GIS. *Forecasting* **2020**, *2*, 36–58. [CrossRef]
14. Syifa, M.; Panahi, M.; Lee, C.W. Mapping of post-wildfire burned area using a hybrid algorithm and satellite data: The case of the camp fire wildfire in California, USA. *Remote Sens.* **2020**, *12*, 623. [CrossRef]
15. Lang, N.; Schindler, K.; Wegner, J.D. Country-wide high-resolution vegetation height mapping with Sentinel-2. *Remote Sens. Environ.* **2019**, *233*, 111347. [CrossRef]
16. Roteta, E.; Bastarrika, A.; Padilla, M.; Storm, T.; Chuvieco, E. Development of a Sentinel-2 burned area algorithm: Generation of a small fi re database for sub-Saharan Africa. *Remote Sens. Environ.* **2019**, *222*, 1–17. [CrossRef]
17. Navarro, G.; Caballero, I.; Silva, G.; Parra, P.C.; Vázquez, Á.; Caldeira, R. Evaluation of forest fire on Madeira Island using Sentinel-2A MSI imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2017**, *58*, 97–106. [CrossRef]
18. Jain, P.; Coogan, S.C.P.; Subramanian, S.G.; Crowley, M.; Taylor, S.; Flannigan, M.D. A review of machine learning applications in wildfire science and management. *arXiv* **2020**, arXiv:2003.00646. [CrossRef]
19. Liang, H.A.O.; Zhang, M.; Wang, H. A neural network model for wildfire scale prediction using meteorological factors. *IEEE Access* **2020**, *7*, 176746–176755. [CrossRef]
20. Tien, D.; Bui, Q.; Nguyen, Q.; Pradhan, B. A hybrid artificial intelligence approach using GIS-based neural-fuzzy inference system and particle swarm optimization for forest fire susceptibility modeling at a tropical area. *Agric. For. Meteorol.* **2017**, *233*, 32–44. [CrossRef]
21. Gigović, L.; Pourghasemi, H.R.; Drobnjak, S.; Bai, S. Testing a new ensemble model based on SVM and random forest in forest fire susceptibility assessment and its mapping in Serbia's Tara National Park. *Forests* **2019**, *10*, 408. [CrossRef]
22. Shabani, S.; Reza, H.; Blaschke, T. Forest stand susceptibility mapping during harvesting using logistic regression and boosted regression tree machine learning models. *Glob. Ecol. Conserv.* **2020**, *22*, e00974. [CrossRef]
23. Tehrany, M.S.; Jones, S.; Shabani, F.; Martínez-Álvarez, F.; Tien Bui, D. A novel ensemble modeling approach for the spatial prediction of tropical forest fire susceptibility using LogitBoost machine learning classifier and multi-source geospatial data. *Theor. Appl. Climatol.* **2019**, *137*, 637–653. [CrossRef]
24. Pourghasemi, H.R.; Gayen, A.; Lasaponara, R.; Tiefenbacher, J.P. Application of learning vector quantization and different machine learning techniques to assessing forest fire influence factors and spatial modelling. *Environ. Res.* **2020**, *184*, 109321. [CrossRef] [PubMed]
25. Tien, D.; Hoang, N.; Samui, P. Spatial pattern analysis and prediction of forest fire using new machine learning approach of multivariate adaptive regression splines and differential flower pollination optimization: A case study at Lao Cai province (Vietnam). *J. Environ. Manag.* **2019**, *237*, 476–487. [CrossRef] [PubMed]
26. Gibson, R.; Danaher, T.; Hehir, W.; Collins, L. A remote sensing approach to mapping fire severity in south-eastern Australia using Sentinel 2 and random forest. *Remote Sens. Environ.* **2020**, *240*, 111702. [CrossRef]
27. Dodangeh, E.; Choubin, B.; Eigdir, A.N.; Nabipour, N.; Panahi, M.; Shamshirband, S.; Mosavi, A. Integrated machine learning methods with resampling algorithms for flood susceptibility prediction. *Sci. Total Environ.* **2020**, *705*, 135983. [CrossRef]

28. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the Appears in the International Joint Conference on Artificial Intelligence, Montreal, QC, Canada, 20–25 August 1995; Volume 14, pp. 1137–1145.
29. Kalantar, B.; Ueda, N.; Saeidi, V.; Ahmadi, K.; Halin, A.A.; Shabani, F. Landslide susceptibility mapping: Machine and ensemble learning based on remote sensing big data. *Remote Sens.* **2020**, *12*, 1737. [CrossRef]
30. Kane, S.N.; Mishra, A.; Dutta, A.K. Preface: International conference on recent trends in physics (ICRTP 2016). *J. Phys. Conf. Ser.* **2016**, *755*. [CrossRef]
31. Pourtaghi, Z.S.; Pourghasemi, H.R.; Rossi, M. Forest fire susceptibility mapping in the Minudasht forests, Golestan province, Iran. *Environ. Earth Sci.* **2015**, *73*, 1515–1533. [CrossRef]
32. Jaafari, A.; Gholami, D.M.; Zenner, E.K. A Bayesian modeling of wildfire probability in the Zagros Mountains, Iran. *Ecol. Inform.* **2017**, *39*, 32–44. [CrossRef]
33. Markham, B.; Barsi, J.; Kvaran, G.; Ong, L.; Kaita, E.; Biggar, S.; Czapla-Myers, J.; Mishra, N.; Helder, D. Landsat-8 operational land imager radiometric calibration and stability. *Remote Sens.* **2014**, *6*, 12275–12308. [CrossRef]
34. Hong, H.; Jaafari, A.; Zenner, E.K. Predicting spatial patterns of wildfire susceptibility in the Huichang County, China: An integrated model to analysis of landscape indicators. *Ecol. Indic.* **2019**, *101*, 878–891. [CrossRef]
35. Fernández-Moya, J.; Alvarado, A.; Forsythe, W.; Ramírez, L.; Algeet-Abarquero, N.; Marchamalo-Sacristán, M. Soil erosion under teak (*Tectona grandis* L.f.) plantations: General patterns, assumptions and controversies. *Catena* **2014**, *123*, 236–242. [CrossRef]
36. Eskandari, S.; Miesel, J.R.; Pourghasemi, H.R. The temporal and spatial relationships between climatic parameters and fire occurrence in northeastern Iran. *Ecol. Indic.* **2020**, *118*, 106720. [CrossRef]
37. Huete, A.R. A soil-adjusted vegetation index (SAVI). *Remote Sens. Environ.* **1988**, *25*, 295–309. [CrossRef]
38. Mukti, A.; Prasetyo, L.B.; Rushayati, S.B. Mapping of fire vulnerability in Alas Purwo National Park. *Procedia Environ. Sci.* **2016**, *33*, 290–304. [CrossRef]
39. Chernick, M.R. Resampling methods. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2012**, *2*, 255–262. [CrossRef]
40. Beasley, W.H.; Rodgers, J.L. Re-Sampling Methods. In *The SAGE Handbook of Quantitative Methods in Psychology*, 1st ed.; SAGE Publications Ltd.: Newbury Park, CA, USA, 2009; pp. 362–386.
41. Steyerberg, E.W. Overfitting and Optimism in Prediction Models. In *Clinical Prediction Models, Statistics for Biology and Health*; Springer: New York, NY, USA, 2019; pp. 95–112. ISBN 9783030163990.
42. Jaafari, A.; Zenner, E.K.; Panahi, M.; Shahabi, H. Hybrid artificial intelligence models based on a neuro-fuzzy system and metaheuristic optimization algorithms for spatial prediction of wildfire probability. *Agric. For. Meteorol.* **2019**, *266–267*, 198–207. [CrossRef]
43. Pourghasemi, H.R.; Kariminejad, N.; Amiri, M.; Edalat, M.; Zarafshar, M.; Blaschke, T.; Cerda, A. Assessing and mapping multi- hazard risk susceptibility using a machine learning technique. *Sci. Rep.* **2020**, *10*, 1–11. [CrossRef]
44. Sakr, G.E.; Elhajj, I.H.; Mitri, G.; Wejinya, U.C. Artificial intelligence for forest fire prediction. In Proceedings of the IEEE/ASME International Conference Advanced Intelligent Mechatronics, AIM, Montreal, QC, Canada, 6–9 July 2010; pp. 1311–1316.
45. Stula, M.; Krstinic, D.; Seric, L. Intelligent forest fire monitoring system. *Inf. Syst. Front.* **2012**, *14*, 725–739. [CrossRef]
46. Kato, A.; Thau, D.; Hudak, A.T.; Meigs, G.W.; Moskal, L.M. Quantifying fire trends in boreal forests with Landsat time series and self-organized criticality. *Remote Sens. Environ.* **2020**, *237*, 111525. [CrossRef]
47. Friedman, J.H. Multivariate adaptive regression splines. *Ann. Stat.* **1991**, *19*, 1–141. [CrossRef]
48. Jaafari, A.; Mafi-Gholami, D.; Thai Pham, B.; Tien Bui, D. Wildfire probability mapping: Bivariate vs. multivariate statistics. *Remote Sens.* **2019**, *11*, 618. [CrossRef]
49. Roy, D.P.; Huang, H.; Boschetti, L.; Giglio, L.; Yan, L.; Zhang, H.H.; Li, Z. Landsat-8 and Sentinel-2 burned area mapping—A combined sensor multi-temporal change detection approach. *Remote Sens. Environ.* **2019**, *231*, 111254. [CrossRef]
50. Tien, D.; Le, V.H.; Hoang, N. Ecological informatics GIS-based spatial prediction of tropical forest fi re danger using a new hybrid machine learning method. *Ecol. Inform.* **2018**, *48*, 104–116. [CrossRef]

51. Kalantar, B.; Ueda, N.; Lay, U.S.; Al-Najjar, H.A.H.; Halin, A.A. Conditioning factors determination for landslide susceptibility mapping using support vector machine learning. In Proceedings of the International Geoscience and Remote Sensing Symposium (IGARSS), Yokohama, Japan, 28 July–2 August 2019.
52. Al-Najjar, H.A.H.; Kalantar, B.; Pradhan, B.; Saeidi, V. Conditioning factor determination for mapping and prediction of landslide susceptibility using machine learning algorithms. In *Earth Resources and Environmental Remote Sensing/GIS Applications X*; International Society for Optics and Photonics: Bellingham, WA, USA, 2019; Volume 19. [CrossRef]
53. Guo, F.; Selvalakshmi, S.; Lin, F.; Wang, G.; Wang, W.; Su, Z.; Liu, A. Geospatial information on geographical and human factors improved anthropogenic fire occurrence modeling in the Chinese boreal forest. *Can. J. For. Res.* **2016**, *46*, 582–594. [CrossRef]
54. Ghorbanzadeh, O.; Blaschke, T.; Gholamnia, K.; Aryal, J. Forest fire susceptibility and risk mapping using social/infrastructural vulnerability and environmental variables. *Fire* **2019**, *2*, 50. [CrossRef]