land

MDPI

Article

Factors Affecting Landslide Susceptibility Mapping: Assessing the Influence of Different Machine Learning Approaches, Sampling Strategies and Data Splitting

Minu Treesa Abraham [1], Neelima Satyam [1], Revuri Lokesh [1], Biswajeet Pradhan [2,3,*] and Abdullah Alamri [4]

[1] Department of Civil Engineering, Indian Institute of Technology Indore, Indore 453552, India; phd1901204011@iiti.ac.in (M.T.A.); neelima.satyam@iiti.ac.in (N.S.); ce170004028@iiti.ac.in (R.L.)
[2] Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney P.O. Box 123, Australia
[3] Earth Observation Center, Institute of Climate Change, Universiti Kebangsaan Malaysia, Bangi 43600, Malaysia
[4] Department of Geology & Geophysics, College of Science, King Saud University, P.O. Box 2455, Riyadh 11451, Saudi Arabia; amsamri@ksu.edu.sa
* Correspondence: Biswajeet.Pradhan@uts.edu.au or Biswajeet24@gmail.com

Abstract: Data driven methods are widely used for the development of Landslide Susceptibility Mapping (LSM). The results of these methods are sensitive to different factors, such as the quality of input data, choice of algorithm, sampling strategies, and data splitting ratios. In this study, five different Machine Learning (ML) algorithms are used for LSM for the Wayanad district in Kerala, India, using two different sampling strategies and nine different train to test ratios in cross validation. The results show that Random Forest (RF), K Nearest Neighbors (KNN), and Support Vector Machine (SVM) algorithms provide better results than Naïve Bayes (NB) and Logistic Regression (LR) for the study area. NB and LR algorithms are less sensitive to the sampling strategy and data splitting, while the performance of the other three algorithms is considerably influenced by the sampling strategy. From the results, both the choice of algorithm and sampling strategy are critical in obtaining the best suited landslide susceptibility map for a region. The accuracies of KNN, RF, and SVM algorithms have increased by 10.51%, 10.02%, and 4.98% with the use of polygon landslide inventory data, while for NB and LR algorithms, the performance was slightly reduced with the use of polygon data. Thus, the sampling strategy and data splitting ratio are less consequential with NB and algorithms, while more data points provide better results for KNN, RF, and SVM algorithms.

Keywords: landslide; susceptibility; machine learning; GIS; Kerala

Citation: Abraham, M.T.; Satyam, N.; Lokesh, R.; Pradhan, B.; Alamri, A. Factors Affecting Landslide Susceptibility Mapping: Assessing the Influence of Different Machine Learning Approaches, Sampling Strategies and Data Splitting. *Land* **2021**, *10*, 989. https://doi.org/10.3390/land10090989

Academic Editors: Enrico Miccadei, Cristiano Carabella and Giorgio Paglia

Received: 26 August 2021
Accepted: 17 September 2021
Published: 19 September 2021

1. Introduction

Catastrophic landslides in mountainous terrains interact with human environment and cause adverse impacts on lives and properties [1]. Aids for managing the risk due to landslides is a topic of which several decades of research has been devoted [2,3]. Mapping the spatial distribution of landslide hazard is one of the most-adopted strategies for risk management, as the landslide susceptibility maps can be used by the government for strategic planning and development [4]. With the recent advancements in Machine Learning (ML) techniques and computational facilities, Landslide Susceptibility Mapping (LSM) have become much easier.

Data driven methods are extensively used for LSM, and the earlier statistical methods using Geographical Information System (GIS)-based approaches are now being replaced by advanced ML algorithms. Different ML algorithms are being widely used for this purpose [5], and the literature shows that no single ML algorithm can be said to be the best for LSM. The choice of an ML algorithm for a particular region is subjected to the scientific goals and objectives of the LSM [5]. Five different algorithms are considered in this study,

viz., Naïve Bayes (NB), Logistic Regression (LR), K Nearest Neighbors (KNN), Random Forest (RF), and Support Vector Machines (SVM). All the algorithms are popular in LSM, but the best suited model for each scenario has to be decided by a quantitative comparison of the model performances. The data used for training and testing of the ML algorithm should be prepared with utmost care, as the quality of data is the key parameter which decides the performance of any ML model. The data includes the landslide inventory and the Landslide Conditioning Factors (LCF). The LCFs are selected considering the topographical and meteo-geological conditions of the study area, and most conditioning factors are often derived from Digital Elevation Models (DEM), satellite data, and existing regional maps. In most cases, landslide inventories are obtained based on satellite images and field investigations [6].

Even though the quality of LCFs are found to be satisfactory and with good resolution DEMs available from satellite-based missions such as TanDEM-X and ALOS, the landslide inventories are often incomplete [7]. The quality of the landslide inventory is subjected to the positional accuracy and sampling strategy. In many studies, the inventories are prepared using points representing landslide crowns. The training and testing data for LSM are prepared using the data from all LCFs extracted using the landslide points. Hence, the positional accuracy of the inventory significantly affects the dataset used for testing and training. When the region is affected by shallow landslides only, the Crown Point provides a satisfactory representation of the landslide-affected area. However, when a region is affected by long runout landslide events, such as debris flows and avalanches, the runout zones cannot be represented using single point information [7], and the events can cause adverse effects in downslope areas [1,8]. The LCFs of the initiation zones and runout zones are entirely different, and a model which is trained using only the initiation zones will ignore the runout zones that may be affected by landslides [9,10]; however, in most studies, landslides are represented using point data, due to the limitations in data availability [11]. Hence, in this study, both point data (single point at the crown of landslide) and polygon data (cluster of points covering the area affected by landslides) are used for LSM. Each point in the cluster represents a cell in the landslide body and is used for LSM. The difference between both the approaches is that the point data considers only the crown area, while the polygon data considers the whole area affected by landslides, including the crown and the runout zone.

The resampling technique of cross validation is a recent advancement in ML, applied to test cases with limited data samples [5]. k-fold cross validation techniques are being widely used for LSM applications, in which the data is split into k parts and are internally resampled such that k−1 parts are used for training and 1 part of testing at each stage of sampling. Even though the method is being widely used for the purpose of validation, there are no guidelines for the number of k to be chosen for an analysis, and, in most studies, the value is chosen as 5 or 10 arbitrarily [12]. The number of k decides the ratio of train to test data, which can affect the performance of the ML model. Hence, in this study, the value of k is also varied from 2 to 10 in order to find the optimum value of k for each algorithm.

To test the objectives, the Wayanad district in Kerala, India, was selected as the test site. The district has suffered from a number of landslides after the incessant rains that occurred during monsoon seasons of 2018, and the landslide inventory data of 2018 was used for LSM.

2. Study Area

The Wayanad district is in the southern part of India (Figure 1), which belongs to Western Ghats, the most prominent orographic feature of the peninsular India. This district is highly prone to landslides [13,14] and has a total area of 2130 km^2, of which 40% is covered by forests. The topography falls mostly in plateau region sloping towards east, for this hilly district is located at the southern tip of Deccan plateau. A major share of the district contributes to the east-flowing river Kabani and its tributaries (Figure 1). The

natural drainage system is constituted by a number of streams, rivulets, and small springs, and the district landscape with flood plains and ridges is formed by this drainage system. Many debris flows that have occurred in the district have runout distances of a few hundred meters, and the longest one ranges up to 3 km. All these slides have contributed to the process of landscape evolution in the district, and minor order streams are originated along the debris flow paths. Thus, the development of drainage paths and watersheds are highly related to the occurrence of landslides, especially debris flows in the region. The flood plains are formed by alluvial deposits with a thickness of more than 10 m. The northwest, southwest, and western parts of the region are formed by higher elevation hill ranges, with steep slopes and a rugged topography. Most of the forest areas are also along these hilly regions. The continuous erosion, transportation, and deposition of the rocks have resulted in the formation of valleys in between the hill ranges. The long runout debris flows that are common in the region also contribute to this process of landscape evolution. Geologically, the district is composed of a peninsular gneissic complex, charnockite group, Wayanad group, and the migmatite complex [15]. Bands of the Wayanad group are found in the northern part of the district, while the rocks of south and southeast are formed by the charnockite group [15]. The northcentral part is composed of a peninsular gneissic complex and the southcentral part is of the migmatite complex.

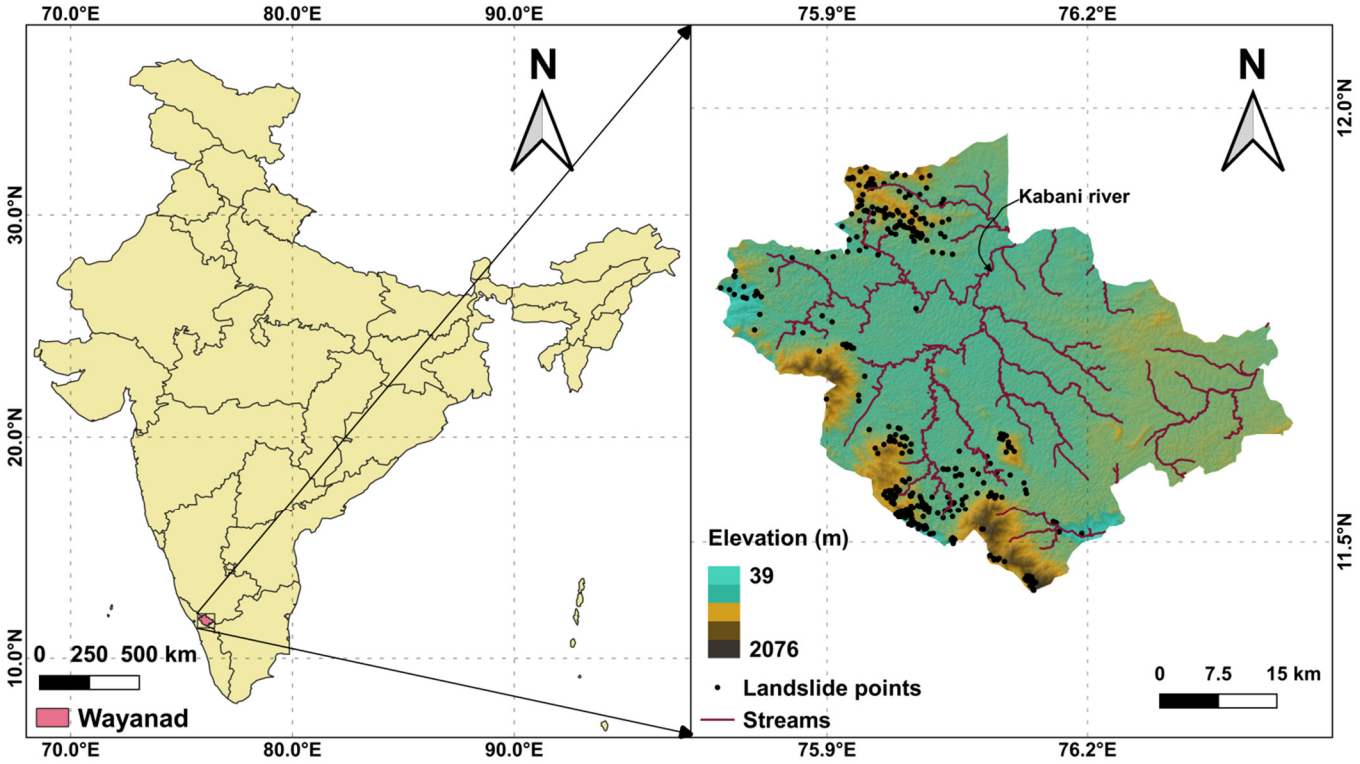

Figure 1. Location map of Wayanad.

A major share of the district is covered by reddish-brown lateritic soil with higher fine content. The forest zones are covered by forest soil with rich organic content, and the riverbanks are formed by thick alluvial deposits. The larger regolith thickness often leads to the bed erosion and bulking of landslides, which increases the landslide volume and destruction potential [16].

The district is highly affected by geohazards such as landslides and floods due to its topographic and geomorphological conditions. The highly dissected hills and valleys along the west, northwest, and southwest parts of the district are highly prone to landslides. During August of 2018, the district was affected by a number of landslides due to torrential rains [17]. A total of 388 landslides (Figure 1) were mapped within the district using

government reports and pre- and post-event satellite images from Google Earth, and have been verified using a recently published dataset [18]. The inventory data were prepared separately for LSM, and are different from the dataset used for previous studies conducted by the authors [13] in which they derived the rainfall thresholds for the region. For deriving rainfall thresholds, multiple landslides occurring on the same day were considered to be a single landslide event, and approximate locations were used, as the focus was on the day of occurrence of the landslide event. However, the inventory data of LSM needs to be accurate, and the spatial distribution of landslides is more important than the time of occurrence of landslides. Hence, the high resolution satellite images available from Google Earth were utilized to prepare a separate landslide inventory database of 388 landslides which occurred in 2018 alone. The district faced major setbacks during the disaster and the catastrophic landslides repeated in the years 2019 and 2020 as well. The increasing frequency of landslides in the districts calls for an updated landslide susceptibility map using data-driven approaches.

3. Methodology

This study aims at evaluating the uncertainties in LSM using ML by adopting different ML algorithms, sampling strategies, and train to test ratios. The first step was the preparation of the dataset, starting from the landslide inventory. The data has to be preprocessed before using it for training and testing. Five different ML approaches were used in this study for comparison.

3.1. Machine Learning Algorithms

Data-based methods are often used to solve real-world problems when the knowledge of the theoretical part is limited and the data is of a large size [19]. Being a non-linear problem, ML models are highly suitable for LSM. The algorithms can learn the association between the occurrence or non-occurrence of landslides and the LCFs using the landslide and no landslide points used for training. Five different ML algorithms are considered in this study, which are explained as follows:

3.1.1. Naïve Bayes

The name of the NB algorithm is formed by two words, 'Naïve' and 'Bayes'. While the latter word stands for the Bayes (named after Thomas Bayes) theorem, which is used for calculating the conditional probability of the occurrence of landslides, in NB, the first term stands for the assumption that the algorithm naively considers all parameters to be independent of each other. The use of simple Bayes' theorem helps the model to have good mathematical control and the results can be achieved fast by using an NB algorithm [20]. The equation for calculating conditional probability of occurrence of landslide (L), subject to the occurrence of conditioning factors C $(C_1$ *to* $C_n)$ is given in the following equation:

$$P\ (L|\ C_1,\ C_2,\ \ldots\ldots C_n) = \frac{P(L) \times P(C_1,\ C_2,\ \ldots C_n|L)}{P(C_1,\ C_2,\ \ldots C_n)} \quad (1)$$

The advantage of an NB algorithm is its simplicity and lower calculation time. The model does not require any hyper parameter tuning and can be easily implemented on any dataset. The major limitation is its assumption of independent parameters. The assumption does not hold true for most of the real-world problems and hence the algorithm may not provide reliable results when the parameters are highly dependent on each other. The algorithm has been used in LSM for more than a decade [21].

3.1.2. Logistic Regression

An LR algorithm is formed from regression analyses, deriving a linear relationship amongst the LCFs by using coefficients [22]. This algorithm, which is derived from statistics, produces a regression output in the form of a mathematical function, and can calculate the probability of the occurrence of landslides. The sigmoid function or logistic function,

which is used in this algorithm, is where the name of LR originates. The sigmoid function in 'S' shape is a core part of LR, which sets an asymptote, based on the positive or negative values of x. For positive values of x, an asymptote is set to $y = 1$, and for negative values of x, asymptote is set for $y = 0$.

The algorithm is easy to implement and does not require any hyper parameter tuning. The model finds its application in LSM due to this simplicity and its usage of probability to predict the solution. A non-linear relationship is established with the landslide and non-landslide points and LCFs and finds a fitting function. The probability of the occurrence of landslides $P(L)$ is calculated by LR as follows:

$$P\,(L) = \frac{e^x}{1 + e^x} \tag{2}$$

where, x is a linear fitting function, using the LCFs, given by:

$$z = a_0 + a_1C_1 + a_2C_2 + \cdots + a_nC_n \tag{3}$$

where, a_0 is the intercept, $a_1, a_2 \ldots\ a_n$ are the regression coefficients, and $C_1,\ C_2, \ldots C_n$ are the LCFs. For dependent variables in binary form and large input data with minimum duplicates and minimum multi collinearity, the algorithm can produce satisfactory results in LSM [23].

3.1.3. K-Nearest Neighbors

The classification of a data point using a KNN algorithm is carried out by using the properties of the neighboring data points [20]. It is a more efficient form of the ball tree concept [24], which can be applied to larger dimensions. The algorithm is widely used in LSM applications [25] and the probability of a data point to be allocated in any class is determined by the classification of its nearest neighbors [26]. The data point takes the classification in which the maximum number of its neighbors is classified. The number of K shall be decided by tuning process for better results.

KNN is classified as a non-parametric model, as the computation process does not depend upon the distributions of the dataset. This is another advantage while using KNN for LSM applications where the number of parameters is more and the data seldom fits to standard distributions. For a set of unclassified points, the algorithm calculates the distance from each point to find K closest neighbors. The classification of these neighbors are then used for voting, and the classification with the maximum votes is assigned to the unclassified data point.

3.1.4. Random Forest

As the name indicates, RF is a combination of many Decision Trees (DT) and the concept was developed in 1995 [27]. Each DT has nodes and branches. The decisions are made at nodes and the classification continues on a particular branch based on the decision. The decisions are continued by considering all LCFs, and each DT assigns a class for the object. RF then considers the class predicted by all DTs and assigns a class for the object based on voting. Each DT is a subset of the whole dataset, and is independently sampled by bootstrapping. The randomness of selection at each node is the major advantage of RF model, which often results in highly accurate predictions, making it suitable for LSM [21,28–30].

The use of splitting at nodes, bootstrapping, and several number trees reduces overfitting in RF by increasing randomness. The model can be fine-tuned by varying the depth of trees, number of trees to be combined, and the number of features considered at each node.

3.1.5. Support Vector Machines

The SVM algorithm classifies a data point using a hyperplane in a multidimensional space, first proposed by Vapnik and Lerner [31–33]. The hyperplanes are boundaries

that decide the classification of an object. The number of LCFs used for the analysis determines the dimensions of the hyperplane. For each dataset, multiple hyperplanes are possible, which can classify the points into different classes. Hence, the SVM algorithm should choose a hyperplane which can maximize the distance between the data points of both classes using statistical learning theory [31,34]. The distance is maximized in order to accommodate the future data points. The data points which are located near to the hyperplane determine the orientation and position of the hyperplane, and these data points are called support vectors.

The SVM algorithm classifies the objects by using different kernel functions, and the choice of kernel function is critical in the results produced by the algorithm. The algorithm is widely used for LSM applications [29,35] and has been in practice since the 2000s [34].

3.2. Data Collection and Sampling Strategies

The landslide inventory map for the study was prepared manually after interpreting satellite images before and after the event. A total of 388 landslides which occurred in 2018 were identified within the boundary of Wayanad. The 2018 disaster was chosen for the study as the district was widely affected by this particular event. The locations where historical landslides were reported were affected, and many new landslides were also reported. Two datasets were prepared from the landslide data collected (Figure 2). In the first approach, the landslide was represented by a point in the crown area and, in the second method, the shape of landslide was demarcated using pre- and post-satellite images; the polygon was marked as inventory data. The district was highly affected by long runout debris flows, as 309 events out of the total 388 were classified as debris flow events. Among the remaining events, 68 were shallow landslides and 11 were rock falls or rockslides. The 388 landslides were represented by 388 cells in the first sampling strategy (Figure 2a) and 9431 cells using the second strategy (Figure 2b). The developed landslide susceptibility map thus provides the probability of occurrence of any of these landslide typologies in the region, and it is not specific for any single landslide typology. The debris flows have very long runout distances [16], and even the locations which are a few kilometers away from the crown points, with entirely different LCFs, were also affected. Hence, using point data for the training and testing of the model might ignore the probability of the occurrence of hazards in the runout zones. To avoid this issue, polygon inventory data was also used in the analysis. The polygon data represents all the cells affected by landslides, unlike the single point used in the first approach. However, the polygon does not differentiate between the crown area and the runout zone. The objective is to train the ML model to predict the probability of the occurrence of a landslide in each cell, and the focus of this manuscript is to compare the probabilities predicted by different approaches. The methodology does not differentiate between crown and landslide body, and checks only if the cell is affected by landslide or not.

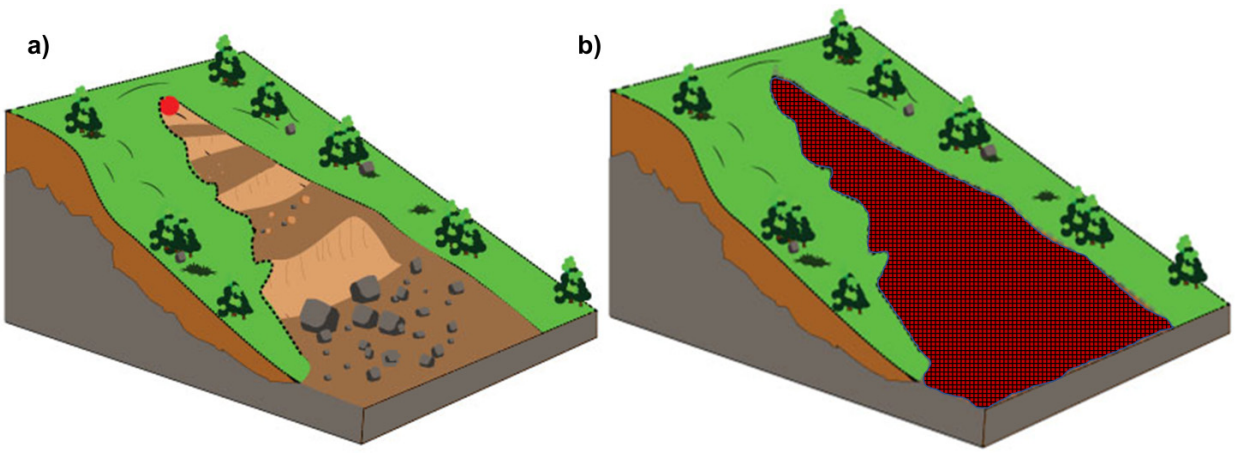

Figure 2. Different sampling strategies adopted in this study: (**a**) point data, and (**b**) polygon data.

The DEM for the study was collected from an Advanced Land Observing Satellite–Phased Array type L-band Synthetic Aperture Radar (ALOS PALSAR) [36], with a resolution of 12.5 m. All the other layers were also prepared in the same resolution as the DEM, and all GIS operations were carried out using QGIS version 3.10. The LCFs, such as slope, aspect, Stream Power Index (SPI), and Topographic Wetness Index (TWI) were derived from the DEM. The first LCF used in this study was elevation, which was directly obtained from the DEM. Slope angle is another significant factor which is critical in triggering the landslides. Slope is defined as the ratio of the vertical to horizontal distance between two points, expressed in terms of the tangent angle in degrees. The value of slope may vary between 0 and 90 degrees. The orientation of the sloping face is expressed using the direction and is termed as aspect. From previous studies, it was found that the value of aspect is critical when landslides occur after the formation of tension cracks in clay [37] and hence it is considered as an LCF. The value of aspect ranges from 0 to 360 degrees and it is classified into 9 categories based on the orientation.

The drainage map for the district was also prepared using the DEM. The locations of the streams were then verified using satellite images and were used for calculating the distance from the streams' layer, which is considered as an LCF. The observation from the inventory data was that many of the long runout debris flows occurred near the streams in the locality. The DEM was also used to create the flow accumulation map and the SPI and TWI layers were developed using the values of flow accumulation. Both SPI and TWI are significant in the process of the initiation of landslides, as SPI represents the power of a flowing water source to erode the material. As the values of the SPI ranges over multiple orders, the natural logarithm of SPI was used for calculation. TWI indicates the wetness of the location, which quantifies the topographic control on different hydrological processes.

The Normalized Difference Vegetation Index (NDVI) is considered to be an important LCF, as it indicates the amount of greenness of a location [38]. When the NDVI values are higher, it represents the presence of vegetation [39,40] and can be correlated with the canopy cover [23]. Thus, the NDVI values are maximum for forest regions and minimum for water bodies and non-vegetated surfaces. Most landslides have occurred within the forest region itself, and the long runout debris flows have originated in the forest area. The net cropped area is 1129.76 km^2, and a major share of cropped area is being used for perennial crops such as coffee, arecanut, and coconut [41]. The cash crops such as coffee and tea and spices such as cardamom are widely cultivated along the hill slopes, while the other crops are cultivated in flatter areas. The NDVI value was calculated using two bands of the electromagnetic spectrum, the Near Infra-Red (NIR) and Red (R) bands [42]. For Landsat 8 images, Band 5 represents NIR and Band 4 represents R. Hence, for this study, the NDVI values were calculated from Landsat 8 images captured in December 2017 and January 2018. As a major share of the cultivated areas is dedicated to perennial crops, the collected images can also satisfactorily represent the conditions at the time of landslides. From the collected images, NDVI is derived using the following formula:

$$NDVI = \frac{(Band\ 5 - Band\ 4)}{Band\ 5 + Band\ 4} \quad (4)$$

The rainfall data for the Wayanad district was collected from the Indian Meteorological Department (IMD) [43]. The data from four different rain gauge stations from 2010 to 2018 were interpolated using inverse distance weighted method of interpolation to get the average annual rainfall values across the district.

The geology, geomorphology, road network, and lineaments of the district were collected from maps published by the Geological Survey of India (GSI). The lineaments and roads were first rasterized and then used to develop the distance rasters, which were used as LCFs. The geology and geomorphology layers were classified and rasterized. The geology was classified into 7 groups, such as migmatite complex, charnockite, younger intrusive, basic intrusive, wayanad group, acid intrusive and peninsular gneissic complex (Figure 3). Geomorphologically, the region was classified into four categories: the highly

dissected hills and valleys, moderately dissected hills and valleys, low dissected hills and valleys, and pediment complex. The collected layers are shown in Figure 3. The layers were then further processed to prepare the database for LSM.

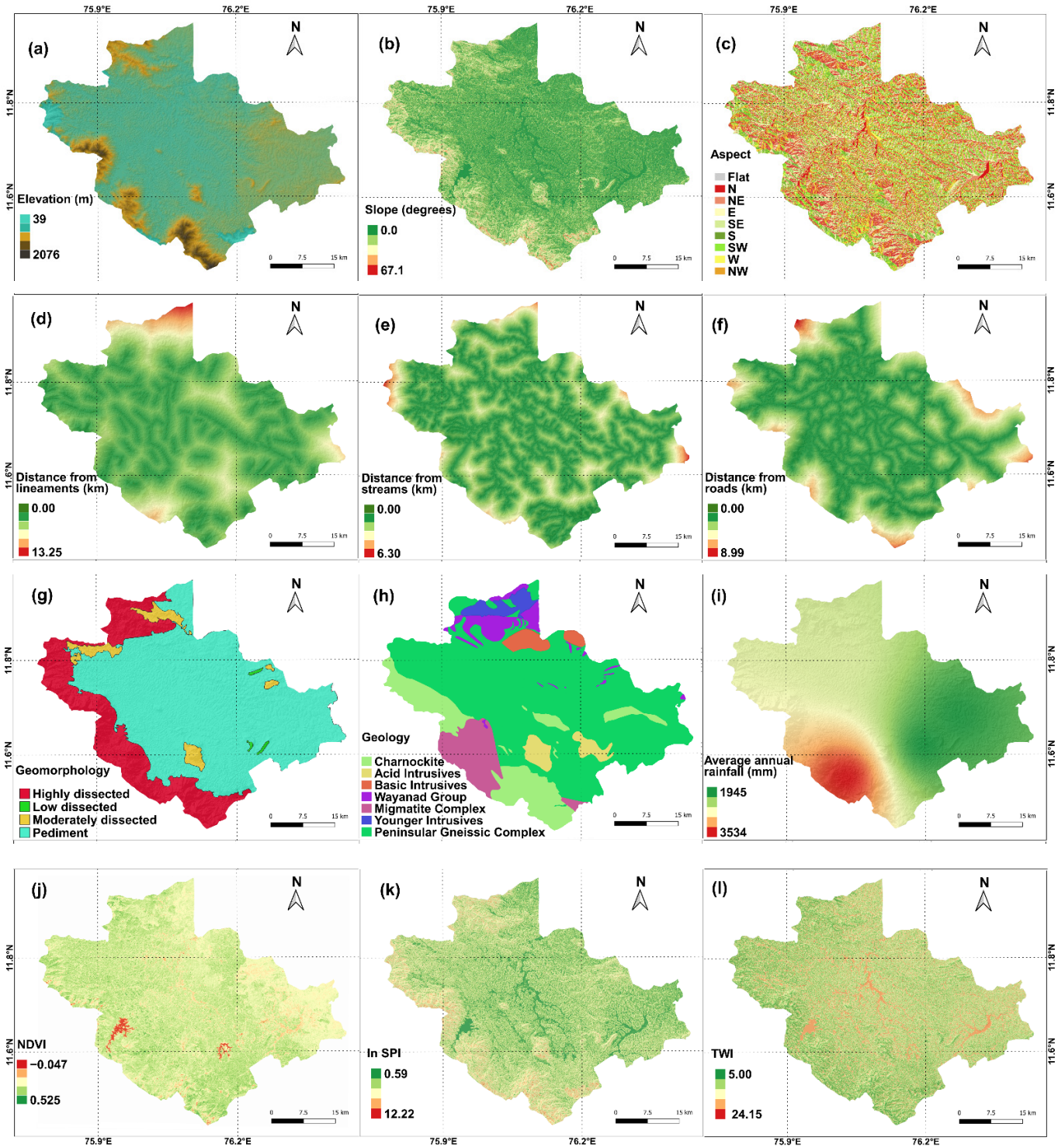

Figure 3. Different LCFs used for LSM: (**a**) elevation, (**b**) slope, (**c**) aspect, (**d**) distance from lineaments, (**e**) distance from streams, (**f**) distance from roads, (**g**) geomorphology, (**h**) geology, (**i**) rainfall, (**j**) NDVI, (**k**) ln SPII, and (**l**) TWI.

The processing of different LCFs is depicted in detail in Figure 4. The processing is different for raster and vector layers. The vector layers are first rasterized and then

converted to XYZ format. For roads, streams, and lineaments, the distance from each feature is first calculated, and the distance rasters were used as LCF (Figure 3).

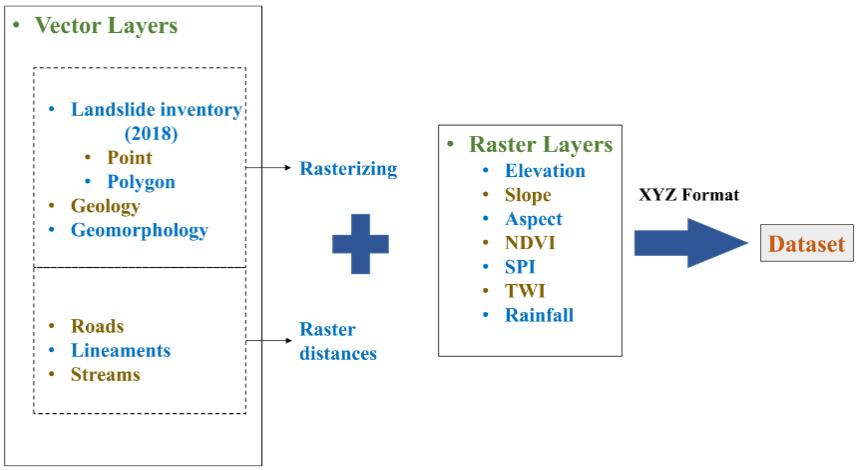

Figure 4. Schematic representation of dataset preparation from different spatial layers collected.

After preparing the landslide inventory data, an equal number of no landslide points were also prepared for the purpose of training and testing for both the sampling methods (point and polygon data). The landslide cells are represented using 1 and no landslide cells using 0 in the dataset. The data from all LCFs were then extracted for the landslide and no landslide points to develop the training and testing dataset. The derived model was later applied to the whole dataset to develop the landslide susceptibility map for the study area.

3.3. K-Fold Cross Validation and Data Splitting

Validation techniques are used to evaluate the performance of ML models. When the dataset is limited, cross validation techniques are often adopted to overcome the limitations associated with the size of the dataset. For k-fold cross validation, the value of k is the only input required, and the dataset is then divided into k different subsets or folds (Figure 5). Among the k-folds, k−1 folds are used for training the model and the last fold is used for testing. The process is repeated k−1 times so that each subset in the dataset is considered for testing.

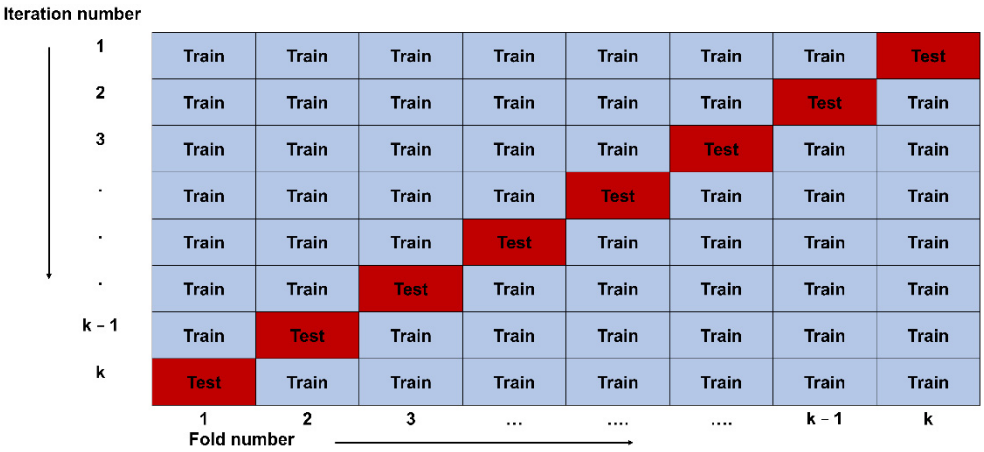

Figure 5. k-fold cross validation represented graphically.

The number of k decides the ratio of train to test ratio of validation and, in most studies, the value of k is randomly is chosen as 5 (train to test ratio 80:20) or 10 (train to test ratio 90:10) [44]. However, detailed studies on performance of cross validation suggest that repeated cross validation should be carried out to determine the optimum value of k [12].

3.4. Quantitative Comparison

The Receiver Operating Characteristic (ROC) curve approach is used for the quantitative comparison of different models. The curve is a plot between the False Positive Rate (FPR) on the x axis and the True Positive Rate (TPR) on the y axis. These parameters are calculated using a conventional confusion matrix where true positives are correctly predicted landslide points, true negatives are correctly predicted, no landslide points, false positives are incorrectly predicted, no landslide points and false negatives are landslide points missed by the model. From these four values, TPR and FPR are calculated as follows:

$$TPR = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{5}$$

$$FPR = \frac{False\ Positives}{False\ Positives + True\ Negatives} \tag{6}$$

The plot with maximum Area Under the Curve (AUC) had the best performance. The landslide susceptibility maps were then prepared using the probabilities predicted by the derived ML models. The model predicts the probability of the occurrence of landslides in each cell, varying from 0 to 1. Based on the probability, the district is categorized into five [45–47] (0.0 to 0.2, 0.2 to 0.4, 0.4 to 0.6, 0.6 to 0.8, and 0.8 to 1.0) and the corresponding susceptibility classes are defined as very low, low, medium, high, and very high. The classification based on equal interval was chosen over the other approaches such as natural break and quantiles, as this study focuses on the comparison of probabilities predicted by different approaches. By using equal interval, the susceptibility classes predicted by each approach can be compared directly to evaluate the agreement or disagreement between the predicted probability values. In other approaches, relative values predicted by each model are used separately for defining the classes and hence the comparison of predicted probabilities is difficult. The statistical attributes such as accuracy and AUC do not provide insights into the agreement and disagreement between the different landslide susceptibility maps prepared. Hence, another parameter, called the Empirical Information Entropy (EIE), or H index, is used to evaluate the agreement between different maps. H index can be calculated as:

$$H = -\sum_{I=1}^{n} P(i)\log(P(i)) \tag{7}$$

where, $P(i)$ is the likelihood of the susceptibility class (very low, low, etc.) i, which is numbered from 1 to 5 in this study (1 is very low and 5 is very high), and n is the number of classes (5 in this case). When all the maps agree with each other, the value of H is zero and as the value increases; the disagreement also increases.

The value of the H-index can be used as an indication to quantify the mutual agreement between the landslide susceptibility maps considered [4]. When two landslide susceptibility maps are compared, there are two outcomes. When both the outcomes are same, the probability of occurrence of one susceptibility class becomes 1 and that of all the other classes are zero. Hence, the H-index becomes zero. In cases where both the outcomes are different, the probability of occurrence of two susceptibility classes is 0.5 and that of remaining classes are zero. The H-index value is the absolute value of twice the product of 0.5 and $\log(0.5)$; i.e., 0.30. When five landslide susceptibility maps are compared, the possible combinations of outcomes and H index values are given in Table 1 below. The number of landslide susceptibility maps predicting each class is interchangeable along the row, and all combinations result in the same value of H index.

From Table 1, it is clear that, as the value of H-index increases, the entropy increases [48], i.e., the disagreement between landslide susceptibility maps increases [4]. Hence, the value can be used to quantify the agreement amongst the results. If more landslide susceptibility maps predict the same class for a cell, the predicted results can be considered to be highly reliable.

Table 1. Possible H index values while comparing the landslide susceptibility maps produced using five algorithms.

Number of Landslide Susceptibility Maps Predicting Each Class					H-Index
Class 1	Class 2	Class 3	Class 4	Class 5	
5	0	0	0	0	0.00
4	1	0	0	0	0.22
3	2	0	0	0	0.29
3	1	1	0	0	0.41
2	2	1	0	0	0.46
2	1	1	1	0	0.58
1	1	1	1	1	0.70

The numbers in rows three to nine can be interchanged among the first five columns. The resulting H-index will remain the same.

4. Results

The performance of the test dataset was first evaluated using the ROC approach to find out the model with best performance. The analysis was carried out with the values of k ranging from 2 to 10 for algorithms, using both a point and polygon dataset, and the ROC curves are plotted in Figure 6.

The minimum and maximum accuracy of the model with NB algorithm and point data are 82.70% and 83.30%, respectively, and the corresponding AUC values are nearly the same, i.e., 0.903 and 0.904. The accuracy values remained the same, while the AUC values reduced when the polygon data is used with the NB algorithm. The trend is nearly the same for the LR algorithm as well. The AUC values are slightly better than NB, with the maximum value of 0.920 with point data. The pattern is different for the other three algorithms, and the performance is significantly improved with polygon data in all the three cases. With the point data, the maximum accuracy values are 84.71%, 88.12%, and 86.63% for KNN, RF, and SVM, respectively, while the maximum AUC values are 0.911, 0.954, and 0.930. With the use of polygon data, the maximum accuracy of KNN increased up to 95.22%, while that of RF became 98.14% and the same for SVM became 91.61%. The AUC values also increased up to 0.981, 0.993, and 0.963 for KNN, RF, and SVM, respectively. Another important observation is that the performance of SVM is better than KNN while using point data, with a difference of 1.92% in accuracy, albeit when polygon data is used. KNN performed better than SVM, with a difference of 3.61% accuracy (Table 2). In both the cases, the RF model outperforms the other models with the highest values of accuracy and AUC.

From Figure 6, it can be observed that the AUC values of KNN, RF and SVM have improved significantly by using polygon inventory data, while the variation is minimum in the case of NB and LR. Moreover, the effect of varying the value of k in k-fold cross validation is insignificant while using polygon data for NB, LR, and SVM algorithms, while, in the case of KNN and RF, variation in the number of folds can result in a variation of approximately 2% accuracy with polygon data. Even though the variation is not significant, the best performance of all models was obtained at k = 8, using point data. A summary of quantitative comparison is provided in Table 2, with the k values corresponding to minimum and maximum performances in the brackets.

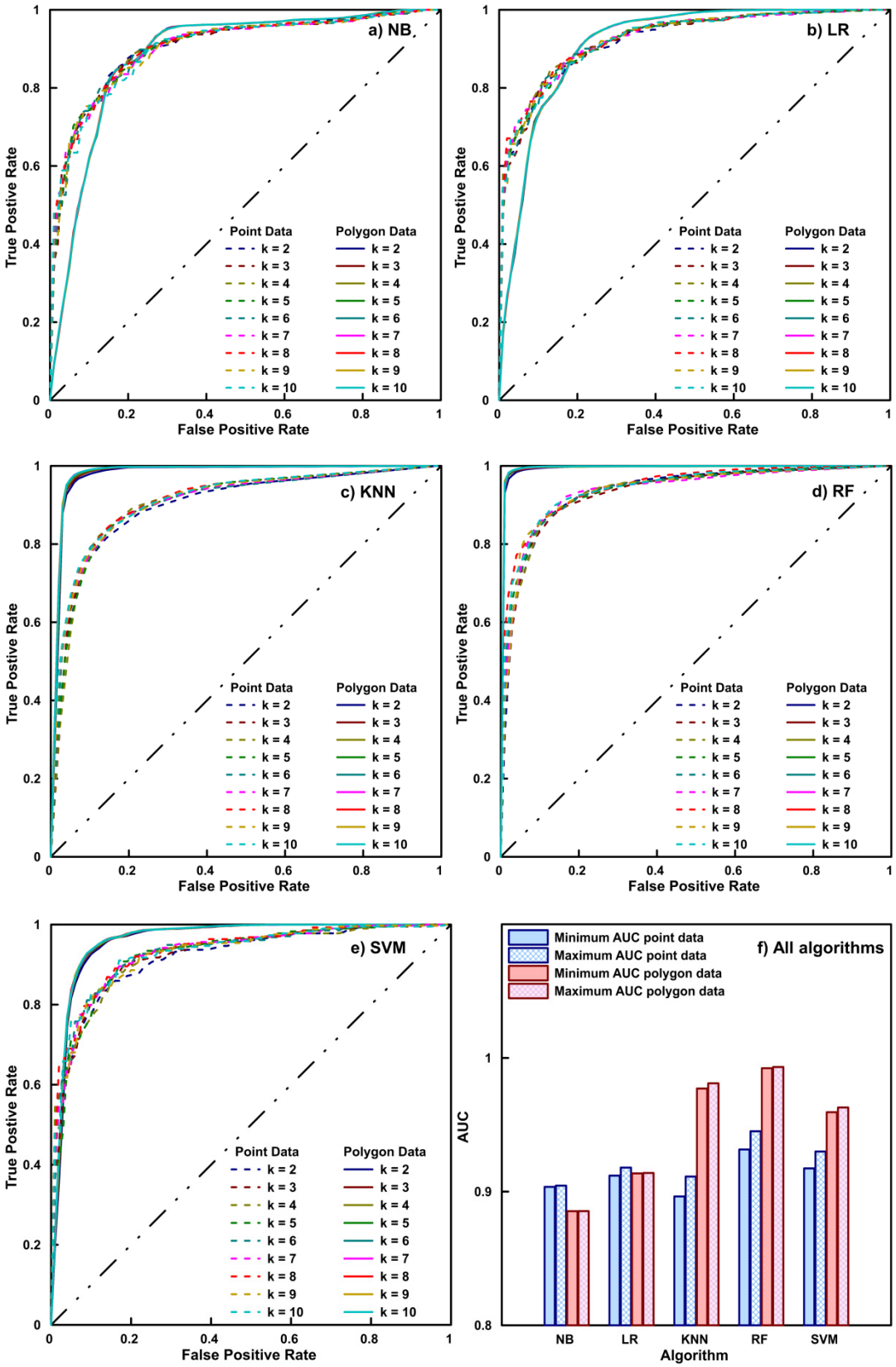

Figure 6. ROC curves, AUC, and accuracy of different models: (**a**) Naïve Bayes, (**b**) Logistic Regression, (**c**) K Nearest Neighbors, (**d**) Random Forest, (**e**) SVM, and (**f**) comparison of AUC of all five algorithms.

Table 2. Quantitative comparison of different algorithms, sampling strategies and data splitting using accuracy and AUC values.

Algorithm	NB	LR	KNN	RF	SVM
Point Data					
Min Accuracy (%) (k)	82.70 (3)	86.67 (3)	83.00 (2)	86.20 (3)	84.80 (2)
Max Accuracy (%) (k)	83.30 (8)	87.41 (8)	84.71 (8)	88.12 (8)	86.63 (8)
Min AUC (k)	0.903 (3)	0.912 (3)	0.896 (2)	0.932 (3)	0.917 (2)
Max AUC (k)	0.904 (8)	0.920 (8)	0.911 (8)	0.954 (8)	0.930 (8)
Polygon Data					
Min Accuracy (%) (k)	83.32 (2)	83.44 (2)	93.23 (2)	96.13 (2)	91.00 (2)
Max Accuracy (%) (k)	83.34 (6)	83.45 (5)	95.22 (8)	98.14 (9)	91.61 (9)
Min AUC (k)	0.885 (2)	0.914 (2)	0.977 (2)	0.992 (2)	0.959 (2)
Max AUC (k)	0.885 (6)	0.914 (5)	0.981 (8)	0.993 (9)	0.963 (9)

From the comparison of statistical performance obtained as per Figure 6 and Table 2, it can be observed that the RF algorithm with polygon inventory data is performing better than all other models. The performance of KNN and RF are comparable while using polygon data and the scores of RF and SVM are comparable while using point data. Still, the best suited model cannot be selected on the basis of statistical scores only. The choice needs a detailed understanding of the distribution of susceptibility classes and a detailed evaluation based on practical perspectives. The purpose of landslide susceptibility maps is to help the planners and authorities in making strategic decisions for future development. Hence, it is important to provide clear information about the susceptibility classes. Based on the value of probability of the occurrence of landslides, the district is divided into five susceptibility classes: very low, low, medium, high, and very high. The statistical attributes provide the prediction performance on the test data only [49]. From a practical perspective, a landslide susceptibility map with an acceptable performance should classify all the landslides correctly within the very high, high, or medium classes. At the same time, the model cannot be too conservative, which may restrict the developmental activities within a larger area. The landslide susceptibility maps prepared using both point and polygon data using each algorithm with the best performing model are evaluated in detail along with the H-index map for a better understanding of spatial agreement.

The number of pixels in each category and the number of landslides that occurred in each class are also important concerns. By using a reliable landslide susceptibility map, the landslides should occur within medium, high, and very high susceptible zones. The landslides which occur outside these zones are missed events, which should be considered with utmost care. Any model with an increased number of missed alerts fails to predict the possible occurrence of landslides.

The landslide susceptibility maps prepared using NB algorithm classifies 15.07% of the total area in the very high category with point data and 18.29% with polygon data (Figure 7). It can also be understood from Figure 7 that, among the 388 landslides considered, 72.64% occurred in very high classified areas, itself with point data, and the percentage increased to 80.64% using polygon data. Exactly 74.49% of the total area is classified as very low using point data and 73.27% using polygon data. The performance of the model is slightly reduced while using polygon data due to the increased number of false alarms within the increased percentage area covered by very high and high category. Considering the mutual agreement between the predictions made by both sampling strategies, 86.42% of the total predictions are in perfect agreement with each other (Figure 7c), while the classification of susceptibility predicted by both methods are different in the remaining area.

The LR algorithm classifies 6.90% of the total area as very high, 9.04% as high, 10.55% as medium, 22.21% as low, and 51.30% as very low susceptible classes using point data

(Figure 8). The number of landslides that occurred in the very high classified locations are reduced to 58.60% when compared with NB, but, at the same time, the number of landslides that occurred in the very low category was also reduced to 6.78%, which in turn slightly improved the performance of LR. While using polygon data, LR algorithm classifies 8.63% of the total area as very high, 7.82% as high, 9.03% as medium, 14.58% as low, and 59.95% as very low. Even though the missed alarms are reduced by this case, the increased number of false alarms resulted in a marginal decrease in accuracy and the AUC values. For 72% of the total area, the susceptibility class predicted using both point data and polygon data perfectly agreed with each other, with an H-index of 0.

From the AUC values (Figure 6), it is evident that the performance of KNN is comparable with NB and LR algorithms while using point data, but it has increased significantly while using polygon data. The reason for this is the drop in the areas classified into very high, high, and medium classes to 3.48%, 3.27%, and 3.44% while using polygon data when compared to 7.15%, 7.60%, and 7.82% while using point data (Figure 9). This reduction has resulted in a considerable reduction of false alarms and in the improvement of accuracy and AUC values. The variation is also reflected in the H-index map, as only 68.70% of the total area agrees with the prediction made using different sampling methods.

Similar to KNN, RF also shows a significant improvement in performance while using polygon data when compared to the point data. The reason is also very similar, as the percentage of very high, high, and medium classified points are reduced while using the polygon data. With the use of point data, 7.86% of the total area was classified under the very high category, which comprises 61.26% of the total landslide occurrences (Figure 10). However, with polygon data, 97.90% of the total landslides are happening within the 1.06% of the total area, which are classified into the very high category. The number of missed events is also reduced by using polygon data as only 0.13% and 0.06% of landslides occurring in the low and very low classified areas, respectively. The mutual agreement between the landslide susceptibility maps produced by point and polygon data is also the least in case of RF algorithm, as 71.20% of the total area has been classified into different categories by using different sampling strategies.

Similar to NB and LR, SVM also shows an increase in percentage of area classified into the very high category with the use of polygon data when compared with the landslide susceptibility map prepared using point data (Figure 11). However, the percentage increase in this category does not result in false alarms, as in the case of NB and LR, as most pixels classified as high and medium categories using point data were classified as in the very high category while using polygon data. Thus, the true positives have increased, and false negatives have been reduced by using polygon data, which in turn resulted in an increase in performance using polygon data. For 75.28% of the area, the categorization is same when using both point and polygon data, as depicted by the H-index plot.

While comparing the performance of different models, RF provides better performance by using both point and polygon data. Moreover, while using polygon data, the performance of KNN and RF are comparable and, while using point data, the performance of SVM and RF are comparable. Apart from statistical comparison, a better understanding of the pixel-wise distribution of susceptibility classes and mutual agreement between the landslide susceptibility maps can help in deciding the best suited landslide susceptibility map for a region.

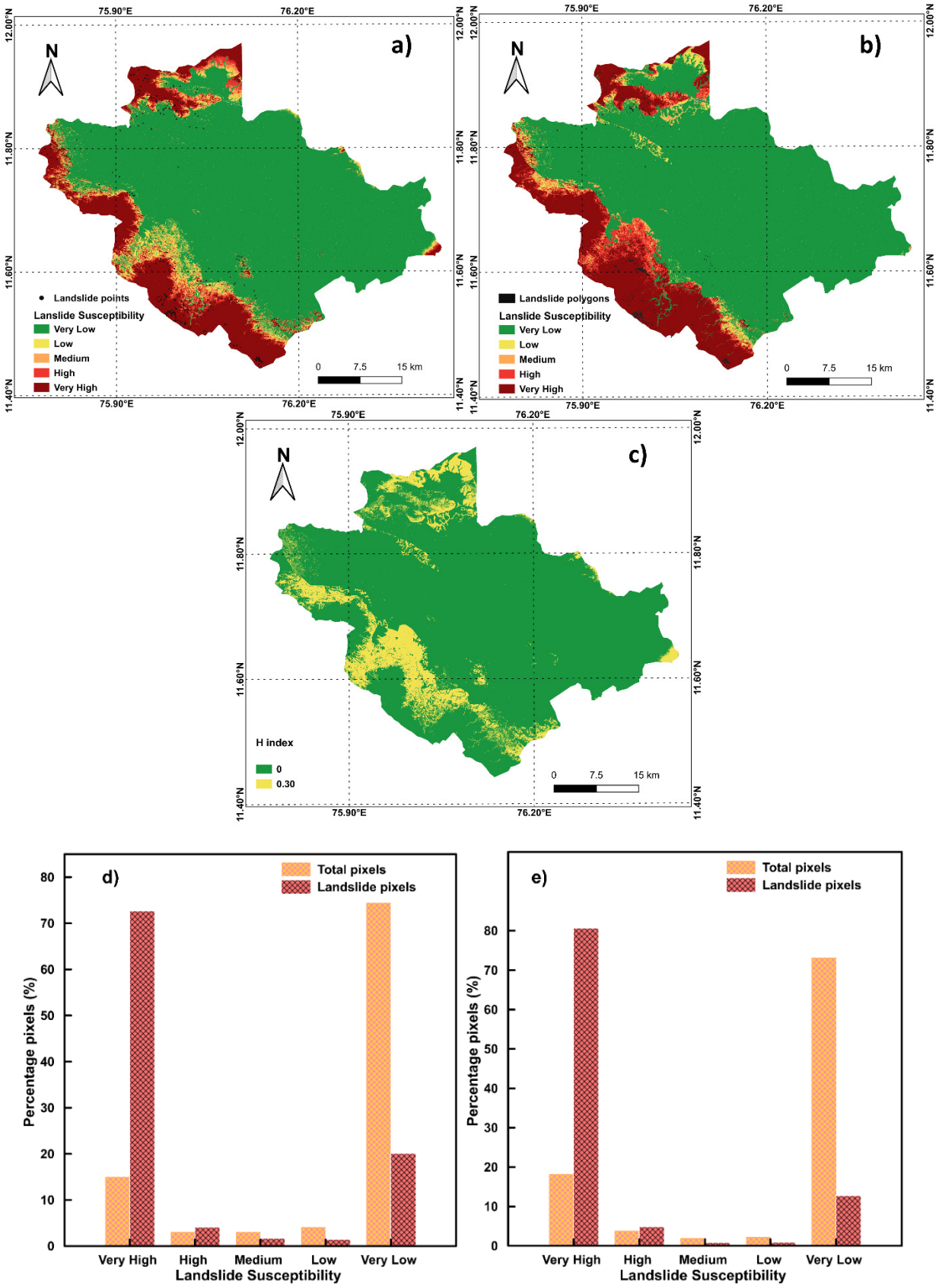

Figure 7. Details of landslide susceptibility maps prepared using NB algorithm: (**a**) using point data, (**b**) using polygon data, (**c**) H-index plot, (**d**) percentage distribution of using point data, and (**e**) percentage distribution of pixels using polygon data.

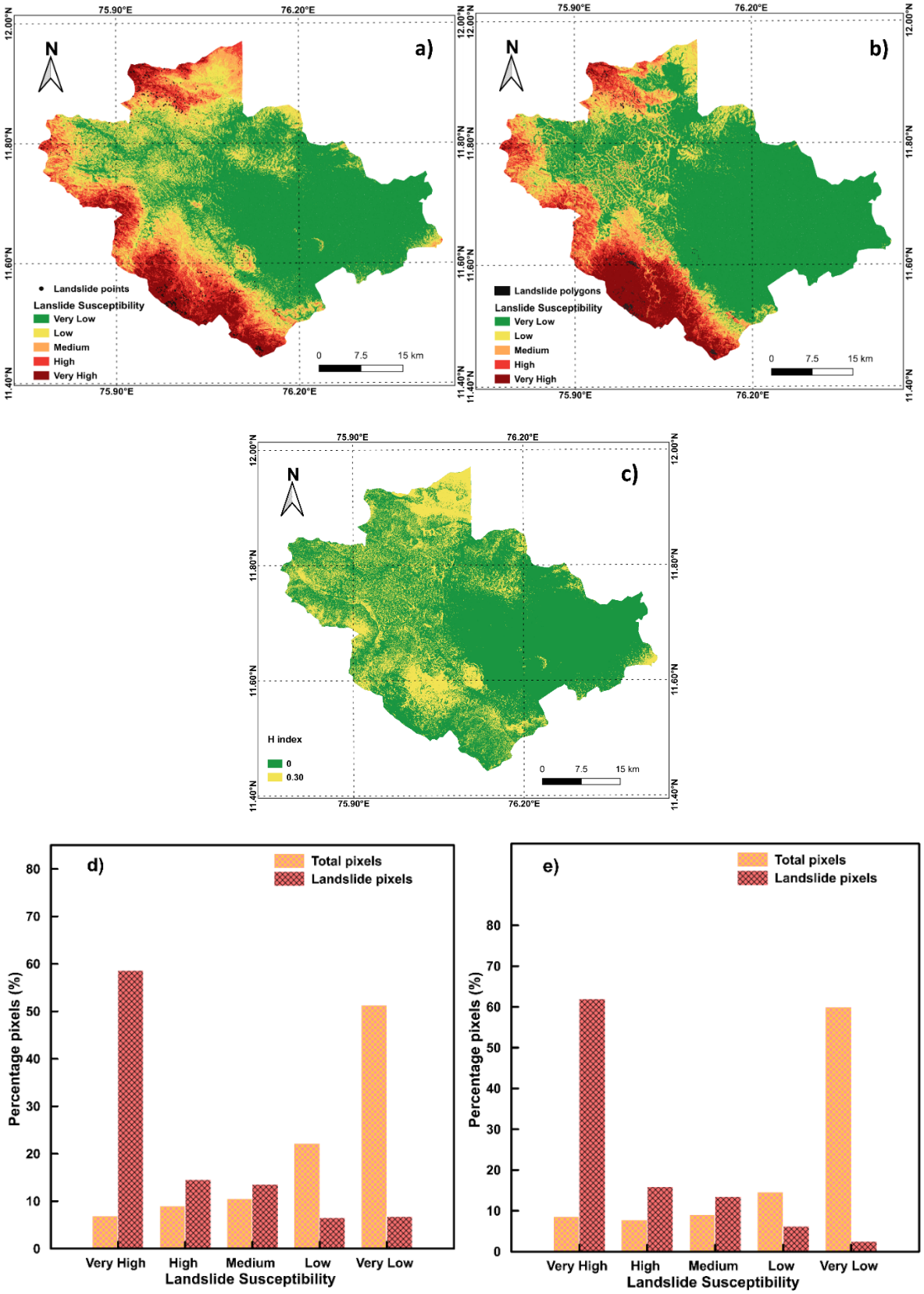

Figure 8. Details of landslide susceptibility maps prepared using LR algorithm: (**a**) using point data, (**b**) using polygon data, (**c**) H-index plot, (**d**) percentage distribution of pixels using point data, and (**e**) percentage distribution of pixels using polygon data.

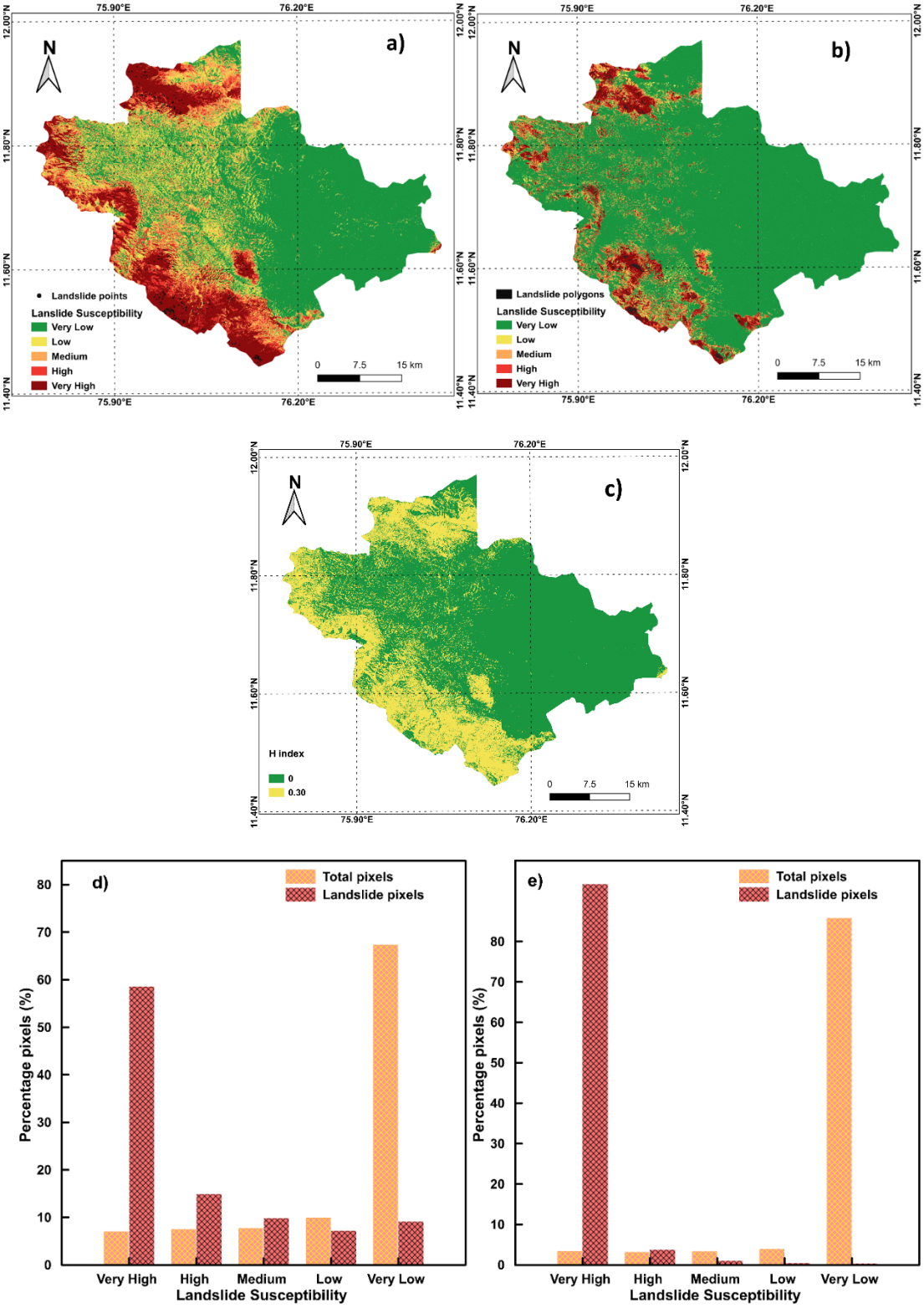

Figure 9. Details of landslide susceptibility maps prepared using KNN algorithm: (**a**) using point data, (**b**) using polygon data, (**c**) H-index plot, (**d**) percentage distribution of pixels using point data, and (**e**) percentage distribution of pixels using polygon data.

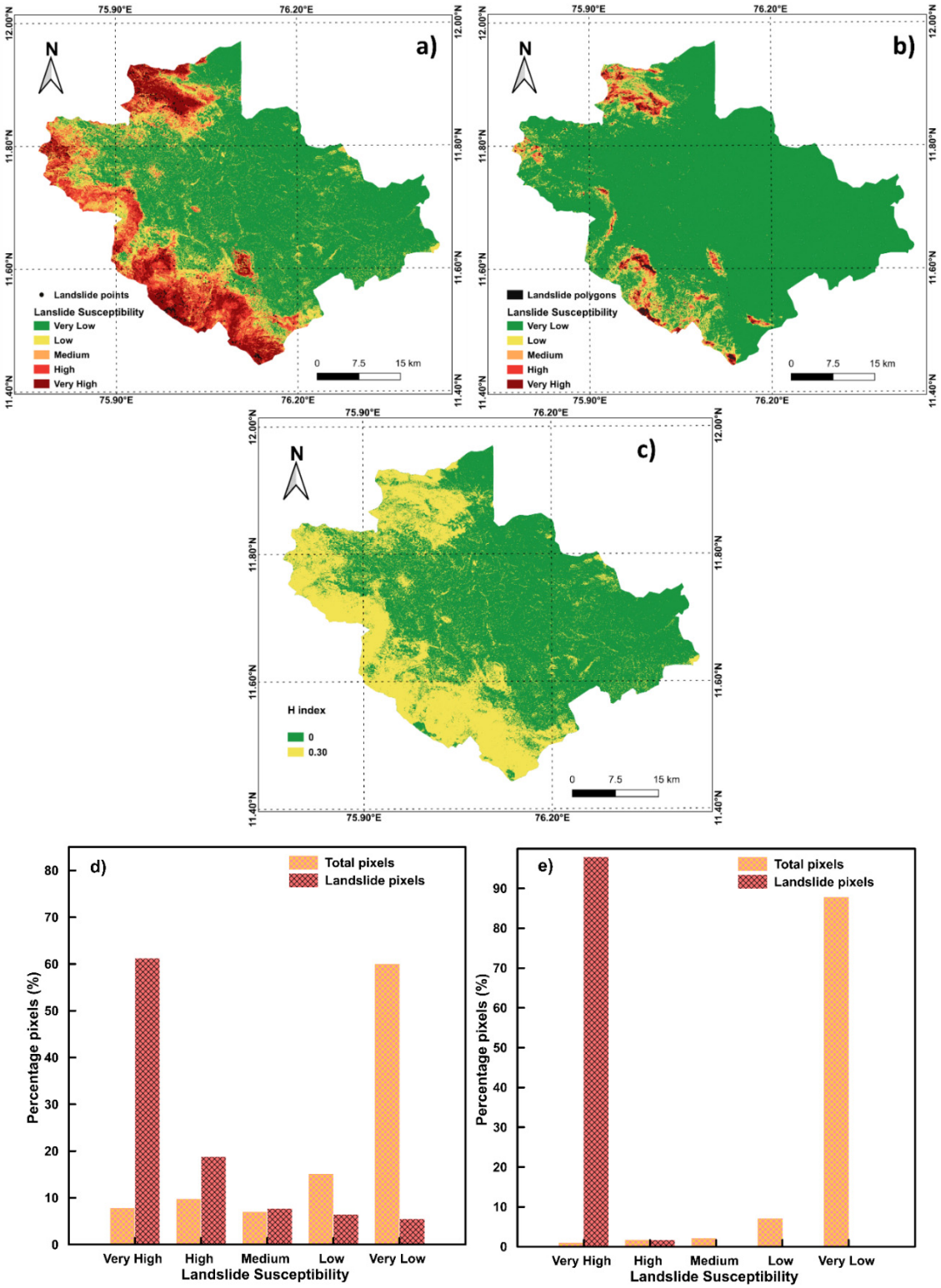

Figure 10. Details of landslide susceptibility map prepared using RF algorithm: (**a**) using point data, (**b**) using polygon data, (**c**) H-index plot, (**d**) percentage distribution of pixels using point data, and (**e**) percentage distribution of pixels using polygon data.

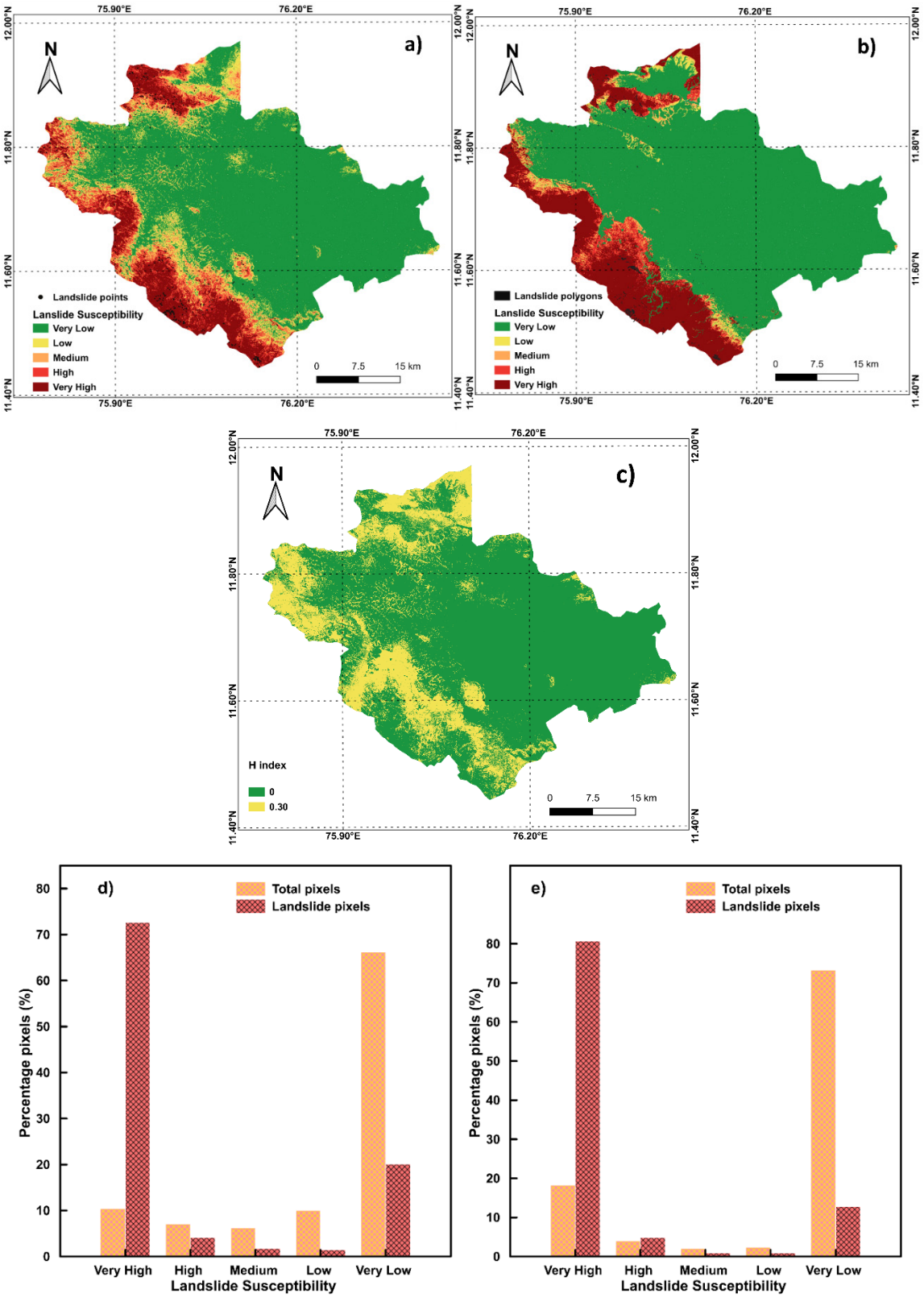

Figure 11. Details of landslide susceptibility maps prepared using SVM algorithm: (**a**) using point data, (**b**) using polygon data, (**c**) H-index plot, (**d**) percentage distribution of pixels using point data, and (**e**) percentage distribution of pixels using polygon data.

5. Discussion

From the obtained results (Table 2), it is evident that the choice of algorithm and sampling strategies can affect the prediction performance of a landslide susceptibility map significantly. The effect of data splitting is crucial for only RF, KNN, and SVM algorithms while using the point data for sampling. The landslide susceptibility maps and H-index plots provide more insights into the effects of different sampling strategies in the performance of different algorithms. From the H-index maps and AUC values, it is evident that the sampling strategy is least effective in the case of NB and most effective in the case of RF.

Figure 12 shows the H-index plots prepared to understand the mutual agreement between different algorithms using the same sampling strategy. It can be observed that, in the case of low susceptible area, the different algorithms are in good agreement with each other, and the LR algorithm classifies the least area in the very low category, which is 51.30% of the total area. While using point data, all algorithms agree in the classification of 47.56% of the total area and all algorithms differ in the case of 0.17% of the total area.

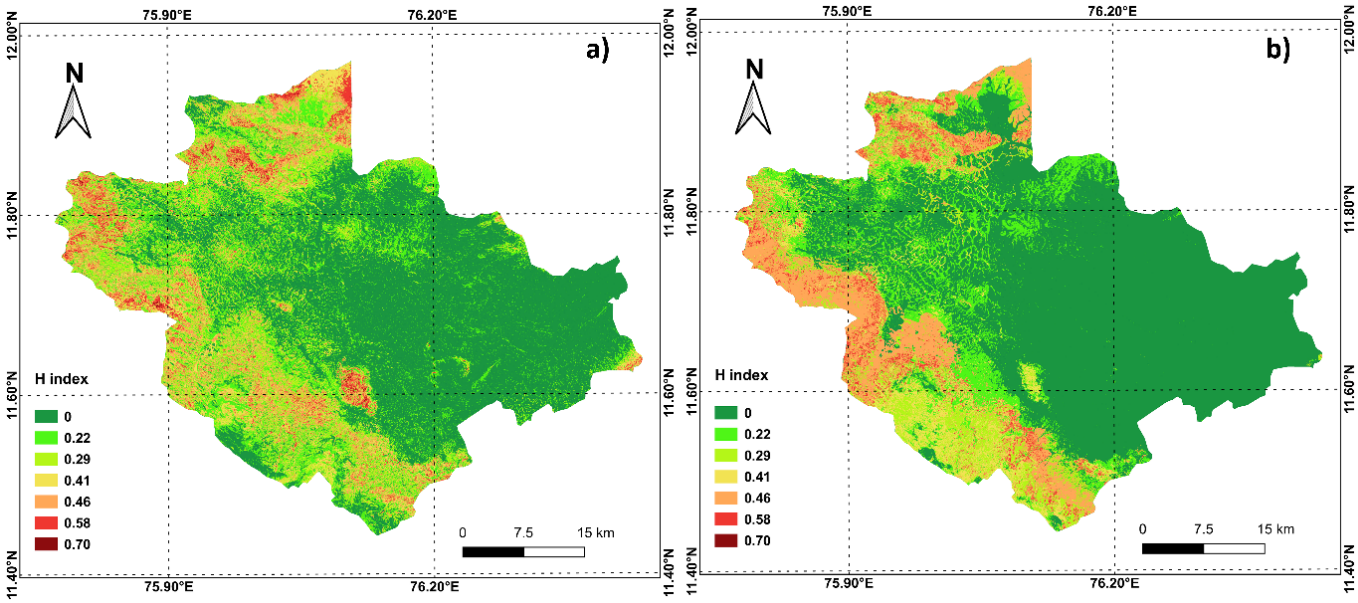

Figure 12. H-index maps plotted using all five algorithms with: (**a**) point data, and (**b**) polygon data.

The percentage distribution of each value of H-index is provided in Table 3 below. While using polygon data, the mutual agreement between algorithms is improved, with perfect agreement in 58.06% of the total area. In no pixels, the classification of all algorithms is entirely different and at least two algorithms agree with the predicted classification. As can be observed from Figure 12b and Table 3, there are no pixels with a H-index value of 0.70 when polygon data is used.

For NB and LR algorithms, the performance is reduced when a greater number of data points in the polygon dataset is used. This is a result of increased correlation between the LCFs with more data points, which violates the basic assumption of independent variables in both the cases. The use of linear fitting function in the case of LR also results in a slight decrease in the accuracy and AUC values with the increased number of data points. However, the advantage of using these algorithms is the reduced computational time involved, as they do not require any hyper parameter tuning.

Table 3. Percentage distribution of H-index values in the total area, using different sampling strategies: comparison between all algorithms.

H Index	Point Data	Polygon Data
	Percentage Pixels (%)	
0.00	47.56	58.06
0.22	19.31	13.25
0.29	13.04	7.47
0.41	8.93	6.96
0.46	7.52	11.48
0.58	3.47	2.77
0.70	0.17	0.00

In the case of KNN, SVM, and RF, the ratio of the train to test dataset can also result in a performance variation while using point data. The performance of these algorithms is significantly increased with the use of polygon data. The improvement in performance can be attributed to the improved size of data used for training the model. All three models demand a long time for the fine-tuning process. The models are highly sensitive to the parameters, train to test ratio, and the size of the dataset [5]. All the three models are widely used for LSM and, hence, if computational facilities are available, the train to test ratio should also be varied to produce the best results from these algorithms.

Even though the performance is comparable with KNN and RF, a higher number of landslides in the very low category make the landslide susceptibility maps made using SVM unsuitable for practical applications. This is an important aspect to be considered. From Figure 11, it is evident that the model using polygon data with an AUC of 0.963 is classifying 13% of the landslides in the very low susceptible zone. This is visible in the landslide susceptibility maps in Figure 11b. The performance can be further improved by using different data sampling approaches and ensemble algorithms and neural networks. In the case of RF, even though the results are statistically better from a practical perspective, the very high, high, and medium classes are bounded by the polygon data used for training and the model is too optimistic, which does not leave room for possible landslides in the surrounding areas in the future. The same issue is observed with the landslide susceptibility map prepared with the KNN algorithm using polygon data. Even though these three algorithms (KNN, SVM, and RF) are having the highest statistical attributes, they cannot be considered to be the best suited for the landslide susceptibility map, due to the limited part of the study area classified into very high, high, and medium classes. The landslide susceptibility map must be conservative, which considers the possible occurrence of landslides in areas other than the ones used for training and testing, and, at the same time, should not classify the safe zones as landslide-susceptible regions. The landslide susceptibility map produced using the RF algorithm with point data is an optimum solution with good statistical performance (AUC = 0.952 and accuracy = 88.12%) and practical applications. It classifies 7.87% of the total area into the very high category and 9.79%, 7.09%, 15.17%, and 60.08% into the high, medium, low, and very low categories, while the best performing model is developed using RF with polygon dataset, with an accuracy of 97.30% and an AUC of 0.993.

From the results, it can be inferred that both the choice of algorithm and sampling strategy can influence the prediction performance of LSM, but the choice of the landslide susceptibility map should not be based on the statistical performance only.

6. Conclusions

The influence of the choice of the ML algorithm, sampling strategies, and data splitting for LSM is evaluated in detail using a case study from the Wayanad district in Kerala.

12 LCFs were used to develop different models using five different ML algorithms (NB, LR, KNN, RF, and SVM), two sampling strategies (point data and polygon data), and different values of k in k-fold cross validation. The results show that data splitting is least effective among the considered parameters. The performance of NB and LR are unaffected by the variation of k values, but the performance of KNN, RF, and SVM are slightly varied by k values, with the best performance at k = 8 in all cases using point data.

The performance of NB and LR did not improve with the use of a large dataset with polygon inventory. The inter dependency of parameters is a critical factor affecting the performance of these algorithms while, in the case of KNN, RF, and SVM, the performance is significantly improved with the use of polygon data. By comparing the H index values, it was observed that the landslide susceptibility maps perfectly agreed with each other in the case of 47.56% of the total area while using point data and 58.06% while using polygon data.

The results produced by KNN and RF using the polygon dataset have a very good statistical performance with very high values for accuracy and AUC. The best performing model developed using an RF algorithm and polygon dataset has an accuracy of 97.30% and an AUC of 0.99.

Author Contributions: Conceptualization, M.T.A., N.S. and B.P.; methodology, M.T.A. and N.S.; data curation, M.T.A. and R.L.; writing—original draft preparation, M.T.A.; writing—review and editing, B.P. and A.A.; supervision, N.S. and B.P. All authors have read and agreed to the published version of the manuscript.

Funding: The study is supported by the Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), University of Technology Sydney. This research was also supported by Researchers Supporting Project number RSP-2021/14, King Saud University, Riyadh, Saudi Arabia.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The publicly archived datasets used for the analysis are cited in the manuscript. The analysis has been carried out at Indian Institute of Technology Indore, and the derived data can be provided upon request to the corresponding author (Biswajeet.Pradhan@uts.edu.au).

Acknowledgments: The authors express their sincere gratitude to Geological Survey of India, Kerala SU, District Soil Conservation Office Wayanad and Kerala State Disaster Management Authority (KSDMA) for their support throughout the study. Authors would like to thank three anonymous reviewers for their critical reviews which helped to improve the quality of the manuscript.

Conflicts of Interest: Authors declare no conflict of interest.

References

1. Froude, M.J.; Petley, D.N. Global fatal landslide occurrence from 2004 to 2016. *Nat. Hazards Earth Syst. Sci.* **2018**, *18*, 2161–2181. [CrossRef]
2. Dou, J.; Yunus, A.P.; Merghadi, A.; Shirzadi, A.; Nguyen, H.; Hussain, Y.; Avtar, R.; Chen, Y.; Pham, B.T.; Yamagishi, H. Different sampling strategies for predicting landslide susceptibilities are deemed less consequential with deep learning. *Sci. Total Environ.* **2020**, *720*, 137320. [CrossRef]
3. Abraham, M.T.; Satyam, N.; Rosi, A.; Pradhan, B.; Segoni, S. Usage of antecedent soil moisture for improving the performance of rainfall thresholds for landslide early warning. *Catena* **2021**, *200*, 105147. [CrossRef]
4. Pradhan, B.; Sameen, M.I. *Effects of the Spatial Resolution of Digital Elevation Models and Their Products on Landslide Susceptibility Mapping*, 1st ed.; Pradhan, B., Ed.; Springer: Cham, Switzerland, 2017; ISBN 978-3-319-55341-2.
5. Merghadi, A.; Yunus, A.P.; Dou, J.; Whiteley, J.; Pham, B.T.; Bui, D.T.; Avtar, R.; Abderrahmane, B. Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance. *Earth Sci. Rev.* **2020**, *207*, 103225. [CrossRef]
6. Reichenbach, P.; Rossi, M.; Malamud, B.D.; Mihir, M.; Guzzetti, F. A review of statistically-based landslide susceptibility models. *Earth Sci. Rev.* **2018**, *180*, 60–91. [CrossRef]
7. Korup, O.; Stolle, A. Landslide prediction from machine learning. *Geol. Today* **2014**, *30*, 26–33. [CrossRef]
8. Li, Y.; Liu, X.; Han, Z.; Dou, J. Spatial proximity-based geographically weighted regression model for landslide susceptibility assessment: A case study of Qingchuan area, China. *Appl. Sci.* **2020**, *10*, 1107. [CrossRef]

9. Simon, N.; Crozier, M.; de Roiste, M.; Rafek, A.G. Point based assessment: Selecting the best way to represent landslide polygon as point frequency in landslide investigation. *Electron J. Geotech. Eng.* **2013**, *18*, 775–784.
10. Süzen, M.L.; Doyuran, V. Data driven bivariate landslide susceptibility assessment using geographical information systems: A method and application to Asarsuyu catchment, Turkey. *Eng. Geol.* **2004**, *71*, 303–321. [CrossRef]
11. Tien Bui, D.; Shirzadi, A.; Shahabi, H.; Geertsema, M.; Omidvar, E.; Clague, J.; Thai Pham, B.; Dou, J.; Talebpour Asl, D.; Bin Ahmad, B.; et al. New ensemble models for shallow landslide susceptibility modeling in a semi-arid watershed. *Forests* **2019**, *10*, 743. [CrossRef]
12. Rodríguez, J.D.; Pérez, A.; Lozano, J.A. Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 569–575. [CrossRef]
13. Abraham, M.T.; Satyam, N.; Rosi, A.; Pradhan, B.; Segoni, S. The selection of rain gauges and rainfall parameters in estimating intensity-duration thresholds for landslide occurrence: Case study from Wayanad (India). *Water* **2020**, *12*, 1000. [CrossRef]
14. Abraham, M.T.; Satyam, N.; Rosi, A. Empirical rainfall thresholds for occurrence of landslides in Wayanad, India. *EGU Gen. Assem.* **2020**, 5194. [CrossRef]
15. Department of Mining and Geology Kerala. *District Survey Report of Minor. Minerals*; Department of Mining and Geology Kerala: Thiruvananthapuram, India, 2016.
16. Abraham, M.T.; Satyam, N.; Reddy, S.K.P.; Pradhan, B. Runout modeling and calibration of friction parameters of Kurichermala debris flow, India. *Landslides* **2021**, *18*, 737–754. [CrossRef]
17. United Nations Development Programme. *Kerala Post Disaster Needs Assessment Floods and Landslides-August 2018*; United Nations Development Programme: Thiruvananthapuram, India, 2018.
18. Hao, L.; van Westen, C.; Martha, T.R.; Jaiswal, P.; McAdoo, B.G. Constructing a complete landslide inventory dataset for the 2018 monsoon disaster in Kerala, India, for land use change analysis. *Earth Syst. Sci. Data* **2020**, *12*, 2899–2918. [CrossRef]
19. Dou, J.; Yunus, A.P.; Bui, D.T.; Merghadi, A.; Sahana, M.; Zhu, Z.; Chen, C.W.; Khosravi, K.; Yang, Y.; Pham, B.T. Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic Island, Japan. *Sci. Total Environ.* **2019**, *662*, 332–346. [CrossRef] [PubMed]
20. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
21. Miner, A.; Vamplew, P.; Windle, D.J.; Flentje, P.; Warner, P. A Comparative study of various data mining techniques as applied to the modeling of landslide susceptibility on the Bellarine Peninsula, Victoria, Australia. In Proceedings of the 11th IAEG Congress of the International Association of Engineering Geology and the Environment, Auckland, New Zealand, 5–10 September 2010; pp. 1327–1336.
22. Cabrera, A.F. Logistic regression analysis in higher education: An applied perspective. *High. Educ. Handb. Theory Res.* **1994**, *10*, 225–256.
23. Huang, X.; Wu, W.; Shen, T.; Xie, L.; Qin, Y.; Peng, S.; Zhou, X.; Fu, X.; Li, J.; Zhang, Z.; et al. Estimating forest canopy cover by multiscale remote sensing in northeast Jiangxi, China. *Land* **2021**, *10*, 433. [CrossRef]
24. Omohundro, S.M. *Five Balltree Construction Algorithms*; Tech. Rep. TR-89-063; International Computer Science Institute (ICSI): Berkeley, CA, USA, 1947.
25. Marjanovic, M.; Bajat, B.; Kovacevic, M. Landslide susceptibility assessment with machine learning algorithms. In Proceedings of the 2009 International Conference on Intelligent Networking and Collaborative Systems, IEEE, Barcelona, Spain, 4–6 November 2009; pp. 273–278.
26. Bröcker, J.; Smith, L.A. Increasing the reliability of reliability diagrams. *Weather Forecast.* **2007**, *22*, 651–661. [CrossRef]
27. Ho, T.K. Random decision forests. In Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, Montreal, QC, Canada, 14–16 August 1995; Volume 1, pp. 278–282.
28. Chen, W.; Sun, Z.; Zhao, X.; Lei, X.; Shirzadi, A.; Shahabi, H. Performance evaluation and comparison of bivariate statistical-based artificial intelligence algorithms for spatial prediction of landslides. *ISPRS Int. J. Geo Inf.* **2020**, *9*, 696. [CrossRef]
29. Zhou, X.; Wu, W.; Lin, Z.; Zhang, G.; Chen, R.; Song, Y.; Wang, Z.; Lang, T.; Qin, Y.; Ou, P.; et al. Zonation of landslide susceptibility in Ruijin, Jiangxi, China. *Int. J. Environ. Res. Public Health* **2021**, *18*, 5906. [CrossRef] [PubMed]
30. Zhang, Y.; Wu, W.; Qin, Y.; Lin, Z.; Zhang, G.; Chen, R.; Song, Y.; Lang, T.; Zhou, X.; Huangfu, W.; et al. Mapping landslide hazard risk using random forest algorithm in Guixi, Jiangxi, China. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 695. [CrossRef]
31. Cortes, C.; Vapnik, V. Suppport vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
32. Vapnik, V.N. *The Nature of Statistical Learning Theory*; Springer: New York, NY, USA, 1995.
33. Vapnik, V.; Lerner, A.Y. Recognition of patterns with help of generalized portraits. *Avtomat. Telemekh* **1963**, *24*, 774–780.
34. Yao, X.; Dai, F.C. Support vector machine modeling of landslide susceptibility using a GIS: A case study. *IAEG2006* **2006**, *793*, 1–12.
35. Gao, R.; Wang, C.; Liang, Z.; Han, S.; Li, B. A research on susceptibility mapping of multiple geological hazards in Yanzi river basin, China. *ISPRS Int. J. Geo Inf.* **2021**, *10*, 218. [CrossRef]
36. Alaska Satellite Facility Distributed Active Archive Center (ASF DAAC) Dataset: ASF DAAC 2015, ALOS PALSAR Radiometric Terrain Corrected high res; Includes Material© JAXA/METI 2007. Available online: https://asf.alaska.edu/data-sets/derived-data-sets/alos-palsar-rtc/alos-palsar-radiometric-terrain-correction/ (accessed on 13 December 2020).

37. Capitani, M.; Ribolini, A.; Bini, M. The slope aspect: A predisposing factor for landsliding? *Comptes Rendus Geosci.* **2013**, *345*, 427–438. [CrossRef]
38. Zhang, T.; Han, L.; Zhang, H.; Zhao, Y.; Li, X.; Zhao, L. GIS-based landslide susceptibility mapping using hybrid integration approaches of fractal dimension with index of entropy and support vector machine. *J. Mt. Sci.* **2019**, *16*, 1275–1288. [CrossRef]
39. Achour, Y.; Pourghasemi, H.R. How do machine learning techniques help in increasing accuracy of landslide susceptibility maps? *Geosci. Front.* **2020**, *11*, 871–883. [CrossRef]
40. Ray, R.L.; Jacobs, J.M.; Cosh, M.H. Landslide susceptibility mapping using downscaled AMSR-E soil moisture: A case study from Cleveland Corral, California, US. *Remote Sens. Environ.* **2010**, *114*, 2624–2636. [CrossRef]
41. Department of Economics and Statistics Government of Kerala Official website of Department of Economics & Statistics, Government of Kerala. Available online: http://www.ecostat.kerala.gov.in/index.php/agri-state-wyd (accessed on 5 September 2021).
42. Fiorucci, F.; Ardizzone, F.; Mondini, A.C.; Viero, A.; Guzzetti, F. Visual interpretation of stereoscopic NDVI satellite images to map rainfall-induced landslides. *Landslides* **2019**, *16*, 165–174. [CrossRef]
43. India Meteorological Department (IMD) Data Supply Portal. Available online: http://dsp.imdpune.gov.in/ (accessed on 3 May 2019).
44. Sun, D.; Xu, J.; Wen, H.; Wang, Y. An optimized random forest model and its generalization ability in landslide susceptibility mapping: Application in two areas of Three Gorges Reservoir, China. *J. Earth Sci.* **2020**, *31*, 1068–1086. [CrossRef]
45. Ou, P.; Wu, W.; Qin, Y.; Zhou, X.; Huangfu, W.; Zhang, Y.; Xie, L.; Huang, X.; Fu, X.; Li, J.; et al. Assessment of landslide hazard in Jiangxi using geo-information technology. *Front. Earth Sci.* **2021**, *9*, 178. [CrossRef]
46. Chalkias, C.; Ferentinou, M.; Polykretis, C. GIS-based landslide susceptibility mapping on the Peloponnese Peninsula, Greece. *Geosciences* **2014**, *4*, 176–190. [CrossRef]
47. El-Fengour, M.; El Motaki, H.; El Bouzidi, A. Landslides susceptibility modelling using multivariate logistic regression model in the Sahla Watershed in northern Morocco. *Soc. Nat.* **2021**, *33*. [CrossRef]
48. Sharma, S.; Mahajan, A.K. Information value based landslide susceptibility zonation of Dharamshala region, northwestern Himalaya, India. *Spat. Inf. Res.* **2019**, *27*, 553–564. [CrossRef]
49. Frattini, P.; Crosta, G.; Carrara, A. Techniques for evaluating the performance of landslide susceptibility models. *Eng. Geol.* **2010**, *111*, 62–72. [CrossRef]