Indian Journal of Science and Technology, *Vol 11(12), DOI: 10.17485/ijst/2018/v11i12/99745*, March 2018
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Machine Learning Methods of Kernel Logistic Regression and Classification and Regression Trees for Landslide Susceptibility Assessment at Part of Himalayan Area, India

Binh Thai Pham[1*] and Indra Prakash[2]

[1]Department of Geotechnical Engineering, University of Transport Technology, 54 Trieu Khuc, Thanh Xuan, Ha Noi, Viet Nam; binhpt@utt.edu.vn
[2]Department of Science & Technology, Bhaskarcharya Institute for Space Applications and Geo-Informatics (BISAG), Government of Gujarat, Gandhinagar - 382421, Gujarat, India ; indra52prakash@gmail.com

Abstract

Objectives: To evaluate performance of machine learning methods for assessment of landslide susceptibility at Himalayan area, India. **Methods/Statistical analysis:** Machine learning methods namely Kernel Logistic Regression (KLR) and Classification and Regression Trees (CART) were applied and compared in this study. Landslide affecting parameters and 930 historical landslides were used for generating datasets. Receiver Operating Characteristic (ROC) curve and Statistical analysis methods were used for validation and comparison. **Findings:** Result analysis shows that both the KLR and CART models perform well for landslide susceptibility assessment but the KLR model (AUC = 0.894) outperforms the CART model (AUC = 0.842). Thus, both these methods can be considered as promising machine learning techniques for landslide susceptibility assessment; however, the KLR is better than the CART. **Application/Improvements:** Results of this study would be useful for susceptibility assessment and landslide hazard management in landslide prone areas.

Keywords: Classification and Regression Trees (CART), Kernel Logistic Regression (KLR), Landslides, Machine Learning

1. Introduction

Landslides were about 4.89% of the geo-environmental hazards all over the world during the period 1990 to[1]. Landslide studies are receiving global attention not only because of increasing awareness of socio-economic harmful impacts but also from increasing pressure of urbanization on the mountain regions[2]. Nowadays, due to increase unplanned urbanization, increased regional precipitation as a result of climate change, and continued deforestation, landslide problems are enhancing, which seems to be more challenging in the future[3,4].

Landslide susceptibility mapping is an important task for proper land use planning and environmental management[1,5]. Based on mapped landslide high or very high susceptible areas, governmental agencies could make proper decisions to combat and prevent landslide occurrences which can help in reduction of losses caused by landslides[6,7]. Machine learning methods are used mostly in recent decades for landslide susceptibility mapping. Machine learning algorithms namely Support Vector Machines (SVM)[8–10], Artificial Neural Networks (ANN)[11–13], Logistic Regression (LR)[8,14–16] are at present most popular for assessment of landslide susceptibility. Additionally,

*Author for correspondence

the KLR and CART can be applied for landslide spatial prediction[17,18].

The KLR is known as a robust method for classification in noisy, complex problems, and resulting good performance in many studies[19–21]. Even so, the application of KLR is still limited for spatial prediction of landslides[22] stated that the KLR has better performance than artificial neural networks, and logistic model tree, it is also indicating as an encouraging method for landslide susceptibility assessment that could be applied also for other landslide affected areas.

The CART was applied efficiently in other fields such as medical science[23,24], agriculture[25]. However, for landslide prediction, the CART has been rarely applied[17] applied the CART for mapping landslide susceptibility, and stated that the CART has the highest accuracy compared with Maximum Entropy, Multiple Adaptive Regression Splines, and LR. In another study[18] stated that the CART is a promising method for landslide susceptibility mapping.

In this study, the main objective is to evaluate and compare the performance of the KLR and CART methods for assessment of landslide susceptibility at part of Himalayan area, India. Statistical analysis methods and the ROC curve were used for validation and comparison. ArcGIS 10.2 and Weka 3.7.12 were used for data analysis and modeling.

2. Machine Learning Methods

2.1 Kernel Logistic Regression (KLR)

The KLR is a common probabilistic non-linear form of logistic regression classification method[26], which estimates the class-posterior probabilities through a log-linear combination of kernel functions using the penalized maximum likelihood method to learn their parameters[27].

Let $\{(t_1,z_1),(t_2,z_2),\ldots,(t_n,z_n)\}$ to be set as a training dataset whereas $t \in \mathbb{R}^n$ are landslide affecting parameters and $z \in \{1,-1\}$ are output variables (non-landslide and landslide). Based on the posterior probability for any x to be assigned class y, the KLR-based classification function is expressed as:

$$\begin{cases} p\left(z=1 \mid f(t)\right) = 1/\left[1+\exp\left(w^T f(t)\right)\right] \\ p\left(z=-1 \mid f(t)\right) = \exp\left(w^T f(t)\right)/\left[1+\exp\left(w^T f(t)\right)\right] \end{cases} \quad (1)$$

During classification process, the regularized optimization problem is carried out by below expression:

$$\min \frac{\lambda}{2}\left|f(t)\right|^2 + \frac{1}{n}\sum_{i=1}^{n}\ln\left(1+e^{y_i f(t_i)}\right) \quad (2)$$

Kernel functions can be used in KLR including linear, polynomial, and radial basis function[21]. In this study, the Radial Basis Function (RBF) was selected to train the KLR as it is considered as a common kernel function[19], the RBF kernel is expressed as below:

$$\mathrm{K}(\mathrm{t},\mathrm{t_i}) = \exp\left\{-\left\|\mathrm{t}\text{-}\mathrm{t_i}\right\|_2^2/\sigma^2\right\}, \quad (3)$$

σ^2 is the squared bandwidth

2.2 Classification and Regression Tree (CART)

The CART is a statistic approach based on tree-building algorithms to classify or predict problems[28]. This method was first proposed by[29]. It is different compared to conventional tree-building methods (J48 or C45) in selection of important variables from a set of predicted variables as it is based on the performance of outcomes for classification[30]. One noticeable advantage of CART is that it can handle small datasets and be scalable to large problems[31] as it is a non-parametric procedure for predicting output variables with input variables.

The CART analysis can be carried out in four main steps: (a) tree building, (b) tree building stop, (c) tree pruning, and (d) optimal tree selection[30].

Tree building: This step is stated with a root node, and then the CART checks all possible splitting variables to find the best possible variable for splitting root node into two child nodes.

Tree building Stop: Node splitting is repeated for each child node until one of three following conditions occurs (i) each of the child nodes has only one observation and (ii) observations inside each child node have same distribution of input variables[30].

Tree Pruning: The "cost-complexity" method is used to simple trees by the cutting of important nodes[30]. When complexity parameter is increased, simpler and simpler trees are created due to more and more nodes are pruned away[30].

Optimal Tree Selection: The purpose of this step is to find the maximal tree that fits the learning dataset with highest accuracy compared to other trees. It is based on finding the correct complexity parameter which the information in training dataset is suitable but not overfitting[30].

3. Study Area

The study area is located at tri junction of Rudrapryag, Tehri Garhwal and Pauri Garhwal districts, Uttarakhand, Himalaya, India between Longitudes: 78°37'40'E to 79°00'50"E and Latitudes: 30°23'15"N to 30°03'58"N, covering an area of about 1325.47 km^2 (Figure 1). Temperature ranges from sub-zero to 45°C. Relative humidity varies from 25% to 85%. Heavy rainfall usually happens during monsoon season (June to September), and annual average rainfall varies from 200 mm to 1000 mm.

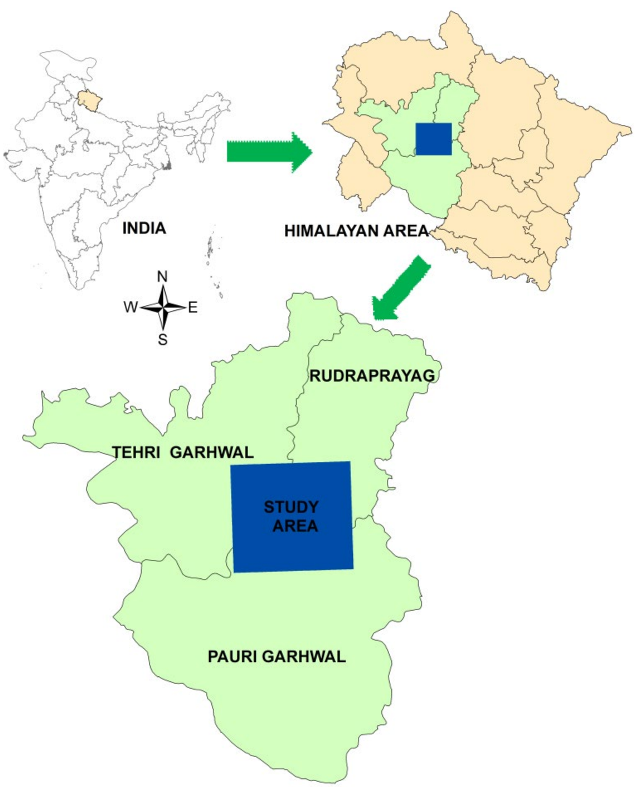


Figure 1. Location of the study area.

Topography of the area is hilly, elevation ranges from 450 m to 2738 m with mean elevation of about 1373 m. Slopes of the hills are relatively steep, up to 70°. Geologically, the area is occupied by Jaunsar Group of Rocks mainly phyllite and quartzite. Tectonically, the area is highly disturbed with folding and faulting[32]. Loamy soil occupies major part of the area. Sand, silt and gravel are present in the valleys. The area is covered by scrub land, non-forest (cultivated land and built up area), forest (dense and open), and deforested area.

4. Spatial Database

Landslide inventory map was first built with 930 historical landslides identified from Google Earth images with the help of Google Earth pro 7.0. These landslides were validated by comparing with field reports (Figure 2). Out of these, 730 landslide locations are classified as translational type, 130 landslide locations are classified as debris flows, and 70 landslide locations are rotational type.

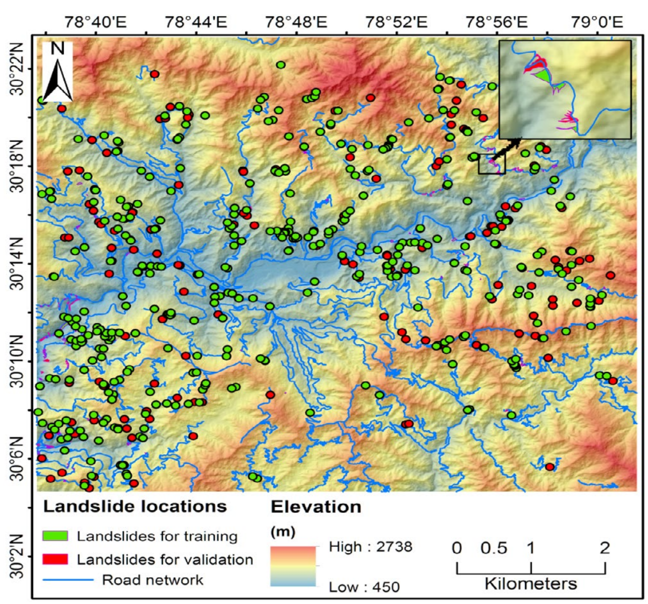


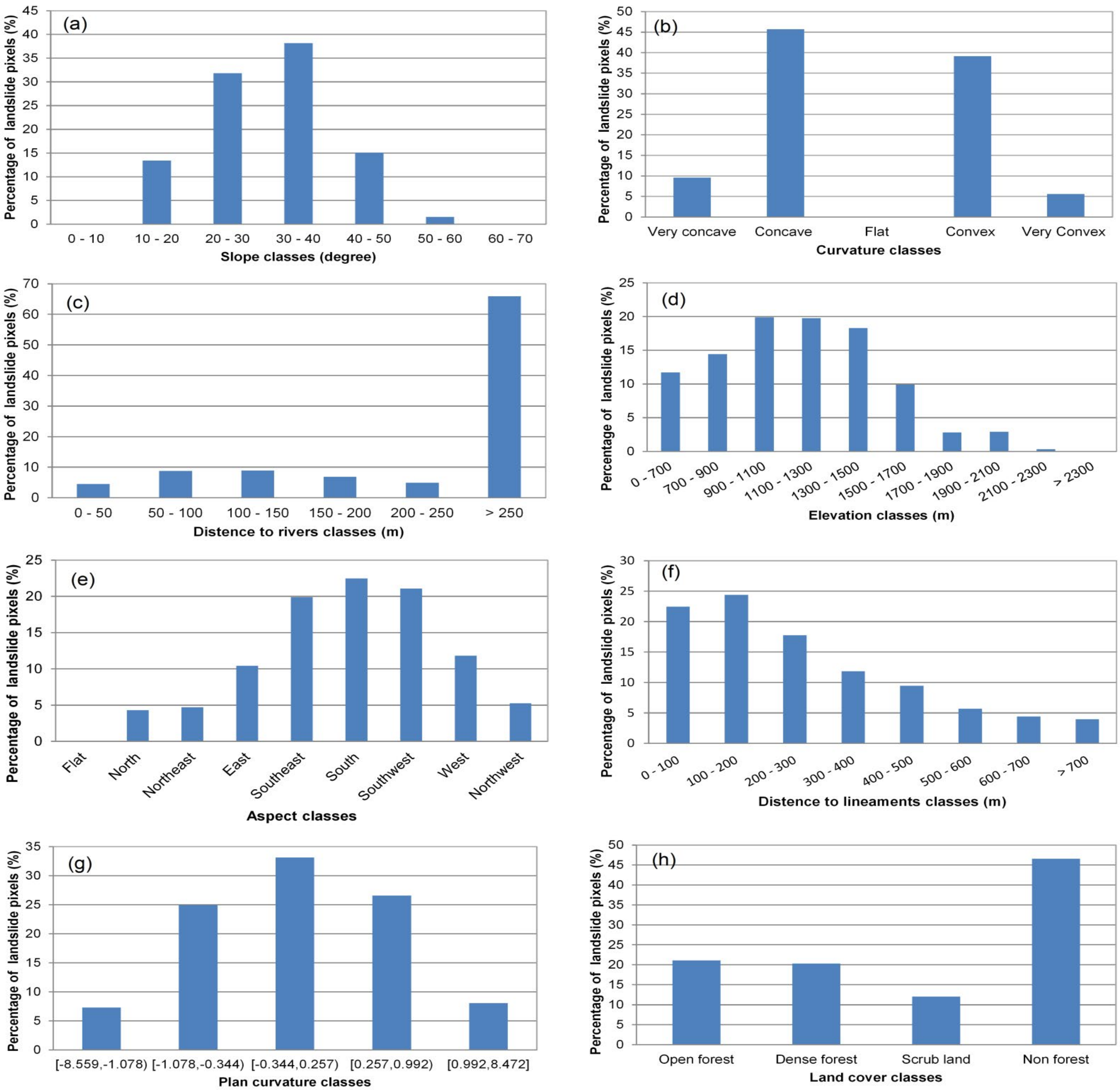
Figure 2. Landslide locations and elevation map.

In addition, a total of fifteen landslide affecting parameters (distance to roads, slope angle, road density, curvature, elevation, distance to lineaments, slope aspect, lineament density, profile curvature, river density, soil type, plan curvature, distance to rivers, land cover, and rainfall) were selected for landslide susceptibility mapping in the present study. Maps of these parameters were extracted from Aster Global DEM, Landsat images, available thematic maps, meteorological maps using ArcGIS software. These maps were constructed with different classes of landslide influencing parameters (Figures 3, 4 and 5)[32]. Frequency analysis of different classes of affecting parameters was done for the development of susceptibility model (Figure 3).

5. Results and Discussion

5.1 Model Construction and Landslide Susceptibility Mapping

Models namely KLR and CART were constructed for assessment of landslide susceptibility at the study area using training dataset which was generated from 651



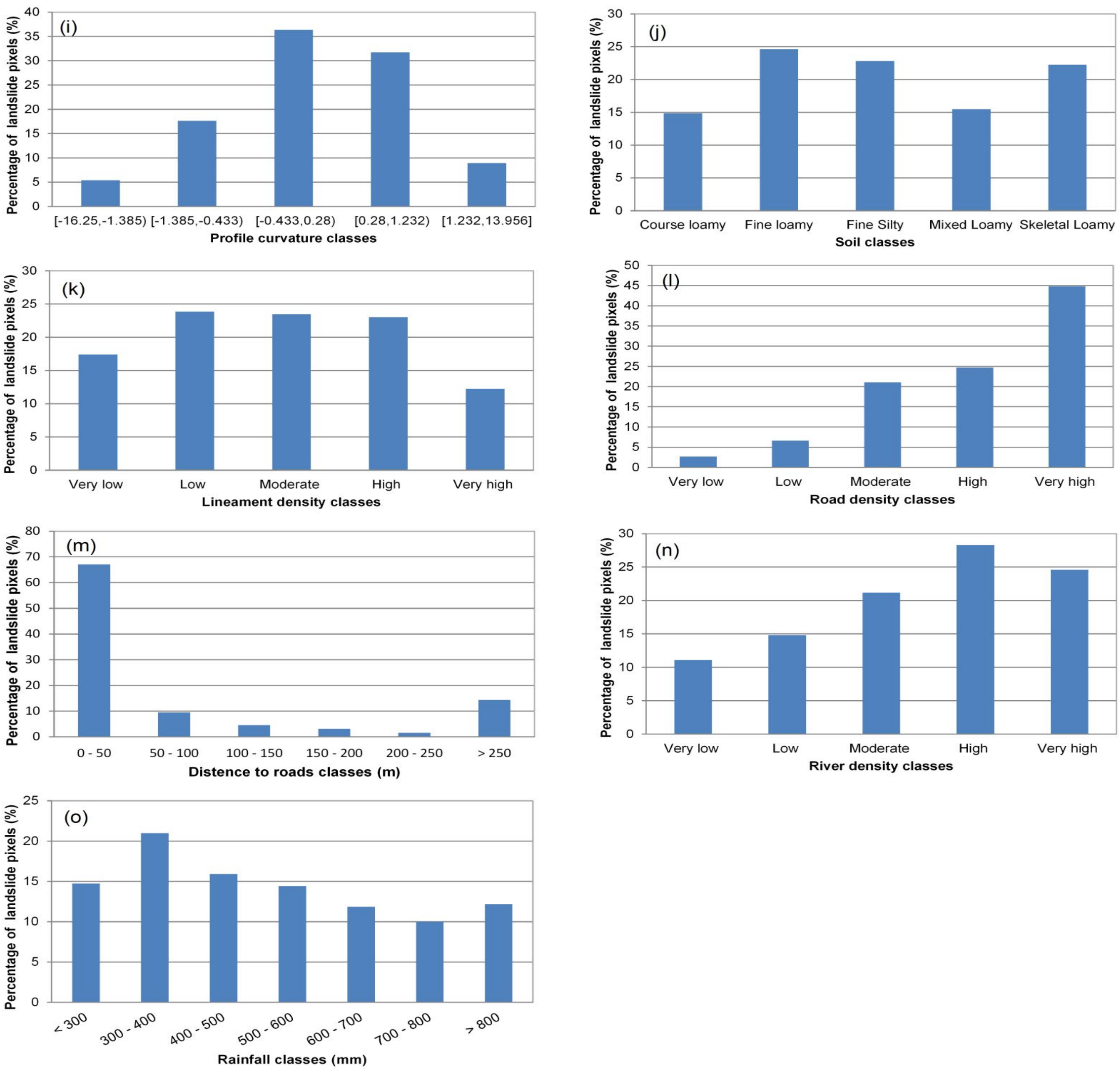


Figure 3. Frequency analysis of landslides on thematic maps: (a) slope angle, (b) curvature, (c) distance to rivers, (d) elevation, (e) slope aspect, (f) distance to lineaments, (g) plan curvature, (h) land cover, (i) profile curvature, (j) soil, (k) lineament density, (l) road density, (m) distance to roads, (n) river density, and (o) rainfall.

landslides and 651 non-landslides in conjunction with landslide affecting parameters. Thereafter, landslide susceptibility maps were constructed using the results from training the KLR and CART models (Figures 6 and 7). Classes namely very high, moderate, low, and very low on the landslide susceptibility maps were classified using geometrical intervals method[32,33].

Landslide Density (LD) was also calculated to validate the reliability of landslide susceptibility maps (Table 1). It can be seen that landslide susceptibility maps have good

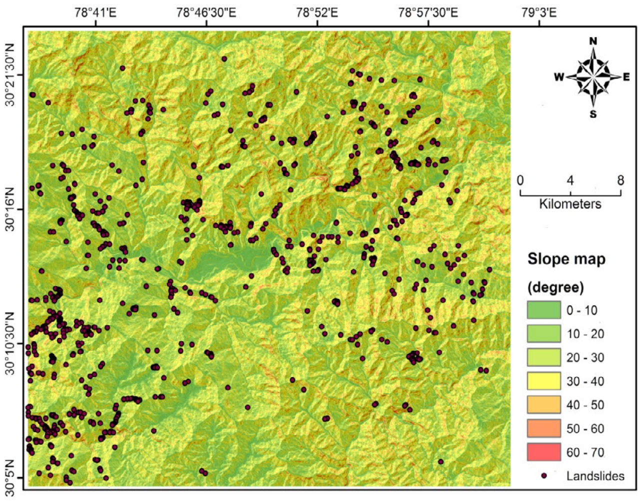


Figure 4. Slope angle map.

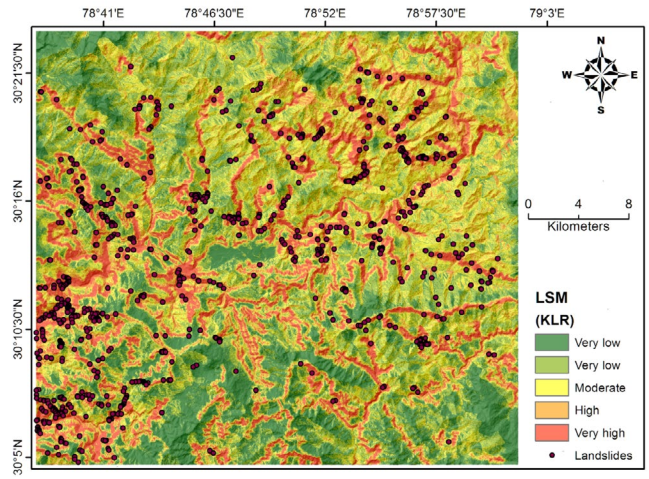


Figure 6. Landslide Susceptibility Map (LSM) using KLR method.

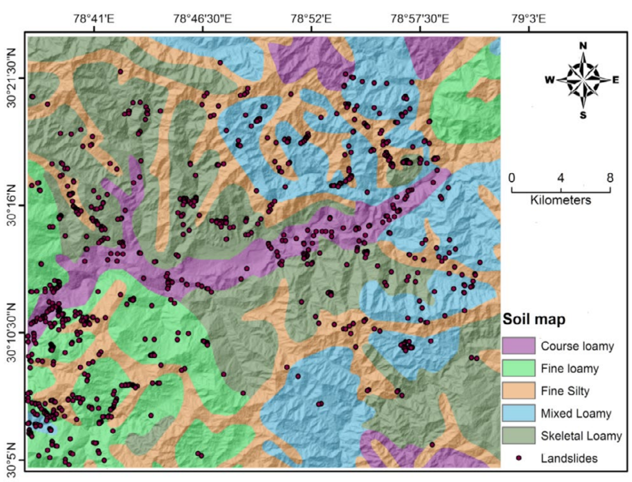


Figure 5. Soil map.

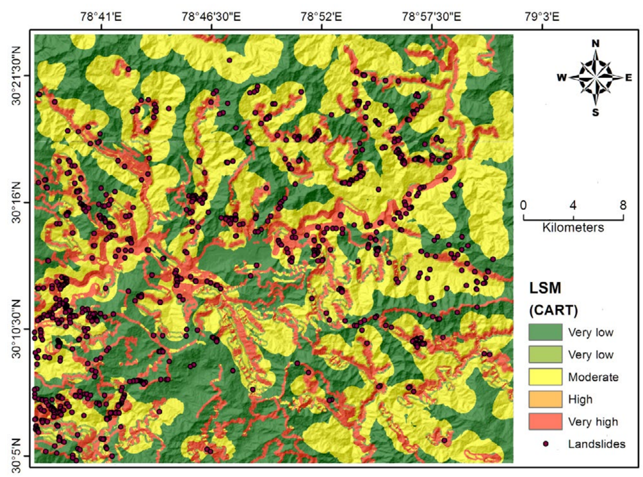


Figure 7. Landslide Susceptibility Map (LSM) using CART method.

Table 1. Landslide density on the susceptibility maps using the KLR and CART models

No	Classes	KLR			CART		
		% class pixels	%Landslides pixels	LD	% class pixels	%Landslides pixels	LD
		17	1	0.03	35.20	4.84	0.14
2	Low	25	2.58	0.1	2.71	0.43	0.16
3	moderate	24	6.99	0.29	39.13	11.51	0.29
4	High	17	9.46	0.55	3.23	1.08	0.33
5	Very high	16	80.43	4.97	19.74	82.15	4.16

performance as the LD values are highest in high and very high class.

5.2 Evaluation and Comparison of Machine Learning Landslide Models

Machine learning landslide models of KLR and CART were evaluated and compared using the testing dataset which was generated from 270 landslides and 270 non-landslides in conjunction with landslide affecting parameters. Statistical analyzing methods and ROC curve were applied to validate the models[8].

In the present study, statistical indexes namely Root Mean Squared Error (RMSE), kappa, sensitivity, accuracy, and specificity were used to validate the KLR and CART models. Detail description of these indexes is shown in[34,8]. Results are shown in Tables 2 and 3. Result analysis show that both the KLR and CART models have good performance in the present study as the values of sensitivity, specificity, accuracy are very high (81.31% - 83.80%) for both training and testing datasets, the value of kappa is relatively high (0.6364 - 0.6676), and the value of RMSE is relatively low (0.3409 - 0.3803).However, the KLR model outperforms the CART model for landslide spatial prediction as the values of sensitivity, specificity, accuracy, kappa of the KLR model is higher than those of the CART model, and the value of RMSE of the KLR model is lower than those of the CART model (Tables 2 and 3).

Table 2. Performance of the KLR and CART using training dataset

No	Parameter	KLR	CART
1	RMSE	0.3409	0.3775
2	kappa	0.6676	0.6484
3	Sensitivity (%)	82.97	81.85
4	Specificity (%)	83.80	83.01
5	Accuracy (%)	83.38	82.42

Table 3. Performance of the KLR and CART using testing dataset

No	Parameter	KLR	CART
1	RMSE	0.3597	0.3803
2	kappa	0.6395	0.6364
3	Sensitivity (%)	82.68	81.79
4	Specificity (%)	81.31	81.85
5	Accuracy (%)	81.98	81.82

Moreover, performance of the KLR and CART models was also validated using the ROC curve analysis[6]. The AUC (area under the ROC Curve) is then utilized to validate the models[35]. The AUC value of 1 indicates perfection of the models. Higher AUC values show better models[36]. Results are shown in Figure 8 and Figure 9. Result analysis shows that both the KLR and CART models perform well for landslide spatial prediction as the values of AUC range from 0.842 to 0.919 for both training and testing datasets. However, the KLR model has better performance than the CART model as the AUC value of the KLR model is higher 7.4% for training dataset, and 5.2% for testing dataset than those of the CART model.

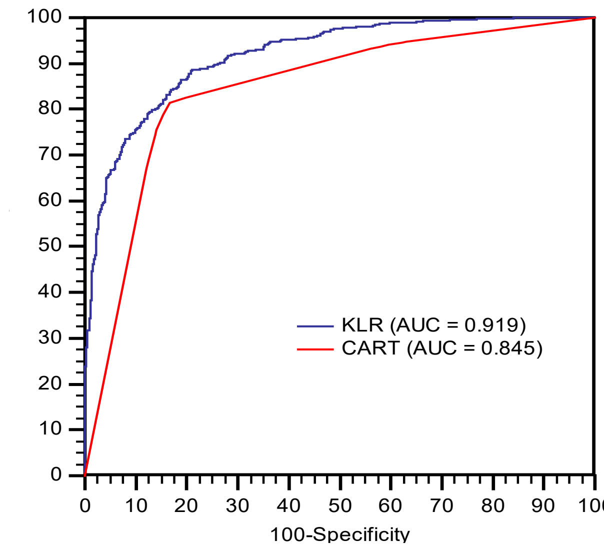


Figure 8. The ROC curves ofthe KLR and CART models using training dataset.

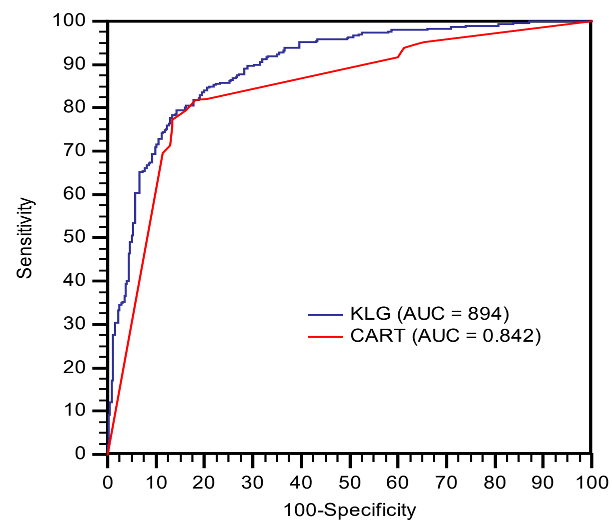


Figure 9. The ROC curves of the KLR and CART models using testing dataset.

Result analysis of both evaluation methods shows that both the KLR and CART models have good performance for landslide spatial prediction in the present study but the KLR model outperforms the CART model. Reason of these results is because the KLR method has many advantages which can improve its performance compared to the CART method such as (i) the KLR can offer a natural estimation of the probability[37], (ii) It has simplicity as well as ability to explore the contribution of neighbors to the classification functions[38], and (iii) the KLR has the advantages of both the logistic regression and kernel algorithms which are known very efficient for landslide prediction[22].

For the CART model, it is inherently non-parametric technique which helps in handling highly skewed or multi-modal numerical data[29]. It is also able to search all possible variables to identify "splitting" variables which helps in dealing with missing variables[31]. Additionally, there is no assumption about distribution of predictor variable's value to be set during training process which can eliminate processing time for determining whether variables are normally distributed or unclassified[30]. However, CART still has a disadvantage of independent assumption of parameters which is not really true for landslide susceptibility assessment[9,39].

6. Conclusions

Machine learning methods are more effective compared with conventional methods for assessment of landslide susceptibility. In the present study, well known KLR and CART methods, which were widely applied to solve classification problems in other fields, were applied for assessment of landslide susceptibility and predictive capability of these methods was evaluated. The ROC curve and statistical analysis methods were selected to validate and compare performance of the models.

The result analysis shows that both the KLR and CART models have good predictive capability for assessment of landslide susceptibility but the KLR (AUC = 0.894) outperforms the CART (AUC = 0.842). Thus, the KLR and CART indicate as promising methods assessment of for landslide susceptibility but the KLR is better than the CART. Therefore, both these models can be used for landslide hazard assessment and management also in other landslide prone areas.

7. Acknowledgement

Authors are thankful to the Director, Bhaskarcharya Institute for Space Applications and Geo-Informatics (BISAG), Department of Science & Technology, Government of Gujarat, Gandhinagar, Gujarat, India for providing facilities to carry out this research work.

8. References

1. Kanungo D, Arora M, Sarkar S, Gupta R. A comparative study of conventional, ANN black box, fuzzy and combined neural and fuzzy weighting procedures for landslide susceptibility zonation in Darjeeling Himalayas. Engineering Geology. 2006; 85(3):347–66. crossref
2. Aleotti P, Chowdhury R. Landslide hazard assessment: summary review and new perspectives. Bulletin of Engineering Geology and the Environment. 1999; 58(1):21–44. crossref
3. Ercanoglu M, Gokceoglu C, Van Asch TW. Landslide susceptibility zoning north of Yenice (NW Turkey) by multivariate statistical techniques. Natural Hazards. 2004; 32(1):1–23. crossref
4. Schuster RL. Socioeconomic significance of landslides. Landslides: Investigation and mitigation. Washington (DC): National Academy Press Transportation Research Board Special Report. 1996; 247:12–35.
5. Pham BT, Tien Bui D, Indra P, Dholakia MB. Landslide susceptibility assessment at a part of Uttarakhand Himalaya, India using GIS – based statistical approach of frequency ratio method. International Journal of Engineering Research & Technology. 2015; 4(11):338–44.
6. Bui DT, Pham BT, Nguyen QP, Hoang ND. Spatial prediction of rainfall-induced shallow landslides using hybrid integration approach of Least-Squares Support Vector Machines and differential evolution optimization: A case study in Central Vietnam. International Journal of Digital Earth. 2016; 9(11):1–21.
7. Pham BT, Tien Bui D, Pourghasemi HR, Indra P, Dholakia MB. Landslide susceptibility assesssment in the Uttarakhand area (India) using GIS: a comparison study of prediction capability of naïve bayes, multilayer perceptron neural networks, and functional trees methods. Theoretical and Applied Climatology. 2015; 122(3–4):1–19.
8. Pham BT, Pradhan B, Tien Bui D, Prakash I, Dholakia MB. A comparative study of different machine learning methods for landslide susceptibility assessment: A case study of Uttarakhand area (India). Environmental Modelling & Software. 2016; 84:240–50. crossref
9. Pham BT, Bui DT, Prakash I, Dholakia M. Evaluation of predictive ability of support vector machines and naive

Bayes trees methods for spatial prediction of landslides in Uttarakhand state (India) using GIS. Journal of Geomatics. 2016; 10(1):71–9.

10. Yao X, Tham LG, Dai FC. Landslide susceptibility mapping based on Support Vector Machine: A case study on natural slopes of Hong Kong, China. Geomorphology. 2008; 101(4):572–82. crossref
11. Chauhan S, Sharma M, Arora M, Gupta N. Landslide susceptibility zonation through ratings derived from artificial neural network. International Journal of Applied Earth Observation and Geoinformation. 2010; 12(5):340–50. crossref
12. Conforti M, Pascale S, Robustelli G, Sdao F. Evaluation of prediction capability of the artificial neural networks for mapping landslide susceptibility in the Turbolo River catchment (northern Calabria, Italy). Catena. 2014; 113:236–50. crossref
13. Pradhan B, Buchroithner MF. Comparison and validation of landslide susceptibility maps using an artificial neural network model for three test areas in Malaysia. Environmental & Engineering Geoscience. 2010; 16(2):107–26. crossref
14. Conoscenti C, Angileri S, Cappadonia C, Rotigliano E, Agnesi V, Märker M. Gully erosion susceptibility assessment by means of GIS-based logistic regression: A case of Sicily (Italy). Geomorphology. 2014; 204:399–411. crossref
15. García-Rodríguez MJ, Malpica JA, Benito B, Díaz M. Susceptibility assessment of earthquake-triggered landslides in El Salvador using logistic regression. Geomorphology. 2008; 95(3–4):172–91. crossref
16. Yesilnacar E, Topal T. Landslide susceptibility mapping: A comparison of logistic regression and neural networks methods in a medium scale study, Hendek region (Turkey). Engineering Geology. 2005; 79(3–4):251–66. crossref
17. Felicísimo ÁM, Cuartero A, Remondo J, Quirós E. Mapping landslide susceptibility with logistic regression, multiple adaptive regression splines, classification and regression trees, and maximum entropy methods: A comparative study. Landslides. 2013; 10(2):175–89. crossref
18. Nefeslioglu HA, Sezer E, Gokceoglu C, Bozkir A, Duman T. Assessment of landslide susceptibility by decision trees in the metropolitan area of Istanbul, Turkey. Mathematical Problems in Engineering. 2010; 2010:1–15.
19. Wu B, Huang B, Fung T. Projection of land use change patterns using kernel logistic regression. Photogrammetric Engineering & Remote Sensing. 2009; 75(8):971–9. crossref
20. Rahayu S, Purnami S, Embong A. Applying Kernel Logistic Regression in data mining to classify credit risk. IEEE, International Symposium on Information Technology; 2008. p. 1–6. crossref
21. Karsmakers P, Pelckmans K, Suykens JA, hamme HV. Fixed-size kernel logistic regression for phoneme classification. Interspeech; 2007. p. 78–81.
22. Bui DT, Tuan TA, Klempe H, Pradhan B, Revhaug I. Spatial prediction models for shallow landslide hazards: A comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. Landslides; 2015. p. 1–18.
23. Knable M, Barci B, Bartko J, Webster M, Torrey E. Molecular abnormalities in the major psychiatric illnesses: Classification and Regression Tree (CRT) analysis of post-mortem prefrontal markers. Molecular Psychiatry. 2002; 7(4):392–404. crossref. PMid:11986983
24. Royston P, Altman DG. Risk stratification for in-hospital mortality in acutely decompensated heart failure. Jama. 2005; 293(20):2467–8. crossref. crossref. PMid:15914743
25. Tittonell P, Shepherd KD, Vanlauwe B, Giller KE. Unravelling the effects of soil and crop management on maize productivity in smallholder agricultural systems of western Kenya - An application of classification and regression tree analysis. Agriculture, Ecosystems & Environment. 2008; 123(1):137–50. crossref
26. Sugiyama M, Hachiya H, Yamada M, Simm J, Nam H. Least-squares probabilistic classifier: A computationally efficient alternative to kernel logistic regression. Proceedings of International Workshop on Statistical Machine Learning for Speech Processing (IWSML2012). Kyoto, Japan; 2012. p. 1–10. PMCid:PMC3816674
27. Minka TP. A comparison of numerical optimizers for logistic regression. Unpublished draft; 2003. p. 1–18.
28. Razi MA, Athappilly K. A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. Expert Systems with Applications. 2005; 29(1):65–74. crossref
29. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. CRC press; 1984.
30. Lewis RJ. An introduction to classification and regression tree (CART) analysis. Annual Meeting of the Society for Academic Emergency Medicine in San Francisco. California; 2000. p. 1–14.
31. Markham IS, Mathieu RG, Wray BA. Kanban setting through artificial intelligence: A comparative study of artificial neural networks and decision trees. Integrated Manufacturing Systems. 2000; 11(4):239–46. crossref
32. About the Geometrical Interval classification method [Internet]. [cited 2007 Oct 18]. Available from: crossref.
33. Pham BT, Tien Bui D, Prakash I, Dholakia MB. Rotation forest fuzzy rule-based classifier ensemble for spatial prediction of landslides using GIS. Natural Hazards. 2016; 83(1):1–31. crossref

34. Bennett ND, Croke BF, Guariso G, Guillaume JH, Hamilton SH, Jakeman AJ, Marsili-Libelli S, Newham LT, Norton JP, Perrin C. Characterising performance of environmental models. Environmental Modelling & Software. 2013; 40:1–20. crossref
35. Dou J, Bui DT, Yunus AP, Jia K, Song X, Revhaug I, Xia H, Zhu Z. Optimization of causative factors for landslide susceptibility evaluation using remote sensing and GIS data in parts of Niigata, Japan. PloS one. 2015; 10(7):133–262. crossref. PMid:26214691 PMCid:PMC4516333
36. Tsangaratos P, Ilia I. Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size. Catena. 2016; 145:164–79. crossref
37. Zhu J, Hastie T. Kernel logistic regression and the import vector machine. Advances in neural information processing systems; 2001. p. 1081–8.
38. Lee H, Tu Z, Deng M, Sun F, Chen T. Diffusion kernel-based logistic regression models for protein function prediction. Omics: A journal of integrative biology. 2006; 10(1):40–55.
39. Pham BT, Tien Bui D, Dholakia MB, Prakash I, Pham HV. A comparative study of least square support vector machines and multiclass alternating decision trees for spatial prediction of rainfall-induced landslides in a tropical cyclones area. Geotechnical and Geological Engineering. 2016; 34(1):1–18. crossref