PIXEL BASED LANDSLIDE IDENTIFICATION USING LANDSAT 8 AND GEE

P. Singh[1], V. Maurya[1], R. Dwivedi[2*]

[1]GIS Cell, MNNIT Allahabad, Prayagraj, India- (pawansingh1610, vipinmaurya2010)@gmail.com
[2]GIS Cell, MNNIT Allahabad, Prayagraj, India- ramjid@mnnit.ac.in

KEYWORDS: Machine Learning, Landslide, GEE, Landsat, Rudraprayag, BHUKOSH

ABSTRACT:

Landslide is one of the most common natural disasters triggered mainly due to heavy rainfall, cloud burst, earthquake, volcanic eruptions, unorganized constructions of roads, and deforestation. In India, field surveying is the most common method used to identify potential landslide regions and update the landslide inventories maintained by the Geological Survey of India, but it is very time-consuming, costly, and inefficient. Alternatively, advanced remote sensing technologies in landslide analysis allow rapid and easy data acquisitions and help to improve the traditional method of landslide detection capabilities. Supervised Machine learning algorithms, for example, Support Vector Machine (SVM), are challenging to conventional techniques by predicting disasters with astounding accuracy. In this research work, we have utilized open-source datasets (Landsat 8 multi-band images and JAXA ALOS DSM) and Google Earth Engine (GEE) to identify landslides in Rudraprayag using machine learning techniques. Rudraprayag is a district of Uttarakhand state in India, which has always been the center of attention of geological studies due to its higher density of landslide-prone zones. For the training and validation purpose, labeled landslide locations obtained from landslide inventory (prepared by the Geological Survey of India) and layers such as NDVI, NDWI, and slope (generated from JAXA ALOS DSM and Landsat 8 satellite multi-band imagery) were used. The landslide identification has been performed using SVM, Classification and Regression Trees (CART), Minimum Distance, Random forest (RF), and Naïve Bayes techniques, in which SVM and RF outperformed all other techniques by achieving an 87.5% true positive rate (TPR).

1. INTRODUCTION

It has been observed that natural disasters are increasing year by year due to the effect of global climate change and the rapid human settlement (Bamisaiye, 2019). Landslide, one of the hazardous geological events, is the downslope movement of rock mass and debris. It has no one particular reason for its occurrence. It can happen due to multiple reasons like Heavy rainfall, cloud burst, earthquake, improper human settlement or unorganized constructions (Haigh et al., 2012). A landslide can alter the natural surroundings causing a change of land cover. However, the most adverse effect is the loss of lives and livelihood in that region, and sometimes blockage of the roads causing a delay in transportation and emergency medical services (Pardeshi et al., 2013).

Field surveying is the standard method to identify and update landslide inventories. However, it is very costly, time-consuming, inefficient, and ineffective, which may delay the update of landslide inventories by half-decade or more. Previously, the landslide inventory of our study area (Rudraprayag) was updated in 2016 which is not encouraged when this area is regularly facing cases of instabilities and landslides. Traditional methods could be improved by using advanced remote sensing techniques. For example, landslides could be identified in an improved way if satellite imageries train machine learning models (supervised machine learning) with various thematic layers obtained from various sources (DEM derivatives and morphological, lithological, etc.).

In supervised machine learning, a model is trained using the classified (or labeled) datasets. Initially, the whole dataset is divided into two categories, training data and validation data. Training data is used to train the machine learning model for a particular purpose and identify the data class. Further, the validation data is used to check the model's performance, deciding whether the model performs well. If the model is not performing well, the model is again trained by changing parameters.

There have been good number of attempts to identify landslides using multiple approaches (Ghorbanzadeh et al., 2019; Semlali et al. 2019, Mohan et al., 2020; Devara et al., 2021). Goetz et al. (2015) presented a comparison of logistic regression (LR), support vector machine (SVM), general additive models (GAM), bundling classification, random forest (RF), and weight of evidence techniques for landslide susceptibility modeling and concluded best results obtained from RF and bundling classification methods. Mahrooghy et al. (2015) conducted a study for detection of landslides using SVM, maximum likelihood (ML) and back-propagation neural network (BPNN), and superiority of the RF method is observed in comparison to other techniques. Wang et al. (2020) exploited five machine learning algorithms(Logistic regression, Support vector machine, Random forest, Discrete Adaboost, LogitBoost, Gentle Adaboost) and deep learning methods (CNN-6 and DCNN-11) on landslide databases (recent, relict, and joint) and evaluated their robustness and potential in identification of landslides. Ji et al. (2020) used a convolution neural network (CNN) and used multiple models such as VHH-13, ResNet-50, ResNet-101, Inception-v3, DenseNet, and various others, which were an extension of these modules. We have also observed that most of the papers declaring high accuracy in landslide identification have used high spatial resolution (Wang et al., 2020; Ji et al., 2020).

Research communities are already dedicated to landslide identification in various regions, but the potential of open-

*Corresponding author

source data is not very well explored in the Indian subcontinent. In most research studies of landslide identification, high-resolution satellite imagery (resolution better than 5m /pixel) is utilized, and the use of open-source satellite imagery such as Landsat 8 available on the web with a low resolution of 30m/pixel is very limited.

In this proposed paper, we have explored and compared multiple supervised machine learning algorithms, SVM, CART, Minimum Distance, Random forest, and Naïve Bayes, to evaluate the potential of the open-source multi-band satellite imagery in the identification of landslides.

2. STUDY AREA

Rudraprayag district is one of the worst-affected districts in India, which lies in lesser and greater Himalaya and consists of many faults, for example, MCT (Main crustal thrust). This area represents rugged and immature topography characterized by moderate to steep slopes that are intervened by narrow valleys. Mandakini is a significant stream of the study area which meets Alaknanda in Rudraprayag city, and it is North south-oriented basin-shaped between higher and lesser Himalayas. As a result of heavy rainfall (higher than average) during the monsoon season, many landslides are observed mainly on the Mandakini and Alaknanda banks. Study area ranges from $30^0$10'36"N to $30^0$48'50"N in latitude and $78^0$48'46"E to $79^0$21'45"E in longitude, covering the Rudraprayag and its neighboring districts of Uttarakhand state in India (Figure 1).

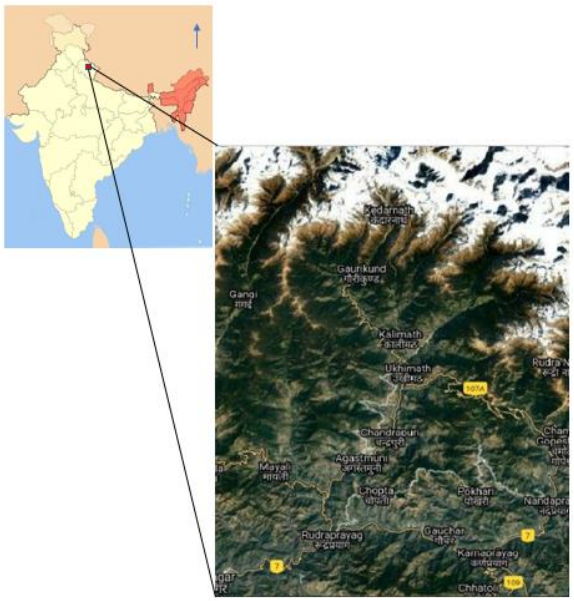

Figure 1. Google Earth Image of Study Area

The datasets used in this paper are JAXA ALOS World 3D - 30m (AW3D30) and Landsat 8 multi-band satellite images of the study area. AW3D30 is a global digital surface model (DSM) dataset with a horizontal resolution of approximately 30 meters (1 arcsec mesh) which is generated using the DSM dataset (5-meter mesh version) of the World 3D Topographic Data (Tadano et al., 2014). Landsat 8 carries two sensors, the OLI and the thermal infrared sensor (TIRS). The OLI collects image data for the nine shortwave spectral bands over a 185 km swath with 30 m spatial resolution for all the bands except for a 15 m panchromatic band (Acharya et al., 2015). The TIRS collects image data for two thermal bands with a 100 m resolution over a 185 km swath. In 11 bands of Landsat 8, red, green, and blue sensors are numbered as 4, 3 and 2 respectively, so we get a true-color image when we combine them. In 11 available bands, only bands of very short wavelengths (band 1-4 and 8) sense visible light, the remaining of the spectrum could not be seen. Band 5 measures the near-infrared, or NIR, vital for vegetation analysis because it is reflected by healthy plants. Band 6 and 7 cover different slices of the short wave infrared, or SWIR. These bands are beneficial for observing the humidity level of earth and geology: rock and soil that look similar in other bands, but they could be differentiated due to solid contrast in SWIR. Using these data, we calculated the Slope (Figure 2), and from landslide inventory data, annotation was done for training and validation purposes. Landsat 8 images are of the interval April 2015 - October 2015 has been utilized to acoomplish the research objectives. Specifically, this time interval was chosen because the data in the inventory are updated to the same duration, and during this time, a large portion of the landmass is exposed after melting of the snow. We have divided the whole dataset in the 75% and 25% ratio for the training and validation purposes respectively.

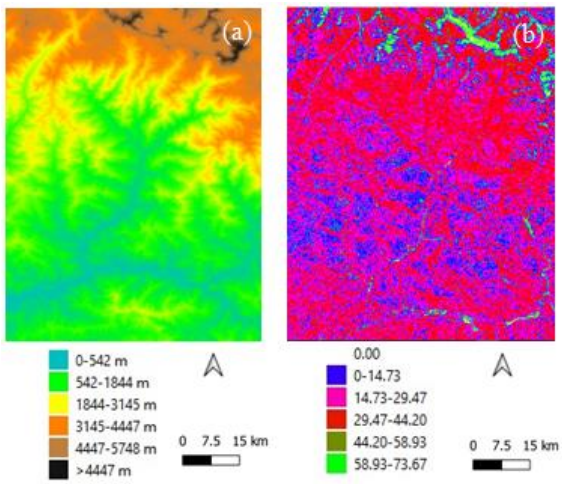

Figure 2. (a) Digital Elevation Model and (b) and slope of the study area

From the image collection of multi-band satellite imagery, we calculated different indices such as normalized difference vegetation index (NDVI) and normalized difference water index (NDWI) to be used as parameters in the Machine learning Algorithm (Figure 3). NDVI is the mathematical ratio used to identify dense vegetation canopy, snowfields, and bare land and shown in equation (1). Here, RED and NIR stand for the spectral reflectance measurements acquired in the red (visible) and near-infrared regions. NDWI is one of the indices related to water content which is computed by formulae shown in equation (2). SWIR and NIR stand for the spectral reflectance measurements acquired in the short wave infrared region and near-infrared regions. Using NIR and SWIR, we can monitor the changes in the water content of vegetation.

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

$$NDWI = \frac{NIR - SWIR}{NIR + SWIR} \quad (2)$$

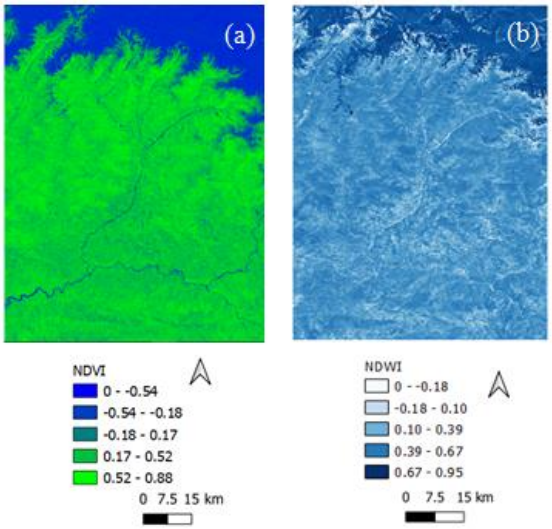

Figure 3. (a) NDVI and (b) NDWI

GSI contributes a significant role in geology and geomorphology-related research in India and has been designate as the nodal agency for natural hazards, i.e. landslides, earthquakes etc. Therefore, GSI is accountable for landslide inventory preparation, susceptibility mapping, and hazard assessments. Under the national landslide susceptibility mapping (NLSM) program, GSI aimed to create a dynamic national landslide susceptibility geodatabase for India and prepared GIS-based seamless landslide susceptibility maps in India on a 1:50,000 scale. Further, GSI created a comprehensive repository on GIS-based landslide inventory and all data related to the geoscientific field that could be visualized, downloaded, and digitized using an online gateway called BHUKOSH. Landslide information of the study area has been collected from the BHUKOSH portal, which is point-shaped type vector data and provides the location and status of landslides (Figure 4).

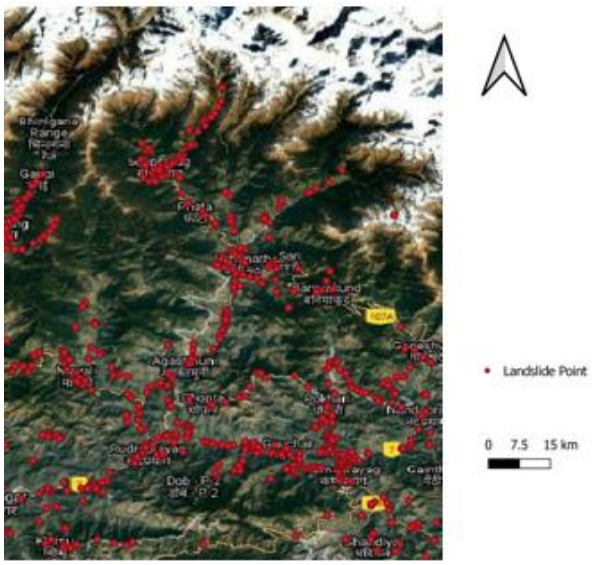

Figure 4. Location of Landslides (red points) obtained from BHUKOSH portal

To prepare the landslides dataset, few cares have been taken. First, in the historical landslide inventory data collected from the field survey, the Landslide was initially located over satellite images. Second, while delineating the landslide boundary, special care was taken to not to include outside landcover. Furthermore, few landslides which cannot be recognized from the satellite imagery were removed as they can only be identified from field investigation, and if we use them, they could confuse models.

3. METHODOLOGY

3.1 Google Earth Engine (GEE)

Earth Engine consists of a multi-petabyte analysis-ready data catalog with a high-performance, intrinsically parallel computing service. GEE has a catalog of publicly available satellite imagery and spatial datasets, including observations from various satellites in optical and non-optical wavelengths, environmental variables, weather, climate forecasts, land cover, topographic and socio-economic datasets. It also provides the capabilities of processing various analyses that help detect changes, finding differences on the earth's surface. It also provides internet accessible application programming interface (API) and a web-based interactive development environment (IDE) that enables rapid prototyping and visualization of results (Gorelick et al., 2017). Such facilities also help other applications to use its services. Users can just signup to access the earth engine homepage and access its user interface, user's guide, tutorials, examples, references, and educational curricula. Users can access data available in the earth engine catalog and their private data using a library of operators provided by earth engine API. Earth Engine includes satellite imagery from Landsat, sentinel-1, sentinel-2, climate forecast, and various other environmental, geophysical, and socio-economic datasets.

3.2 Advanced Machine learning algorithms

Multiple machine learning algorithms use different approaches for the classifications. For example, SVM finds a hyper-plane that creates a boundary between classes SVM along with linear classification, SVMs can efficiently perform a non-linear classification using kernel tricks and kernel functions (Girosi et al., 1995; Smola et al., 1998). Classification and Regression Trees (CART) models, also referred to as "decision trees," are obtained by partitioning the data space and fitting a simple prediction model within each partition which is done recursively. CART provides a foundation for the algorithm like Random forest. The representation used for the CART is a binary tree, and predictions are made by traversing the binary tree. A tree is learned using a greedy algorithm on training data and stop criteria define how much tree learns (Loh, 2011), naïve Bayes uses the probabilistic approach, which means it predicts based on the probability of an object, it assumes specific feature is independent of the occurrence of the other features it depends on the principle of Baye's theorem and minimum distance uses a close distance approach and classify unknown data by minimizing the distance between the data and the class in multi-feature space. Random forest is the tree-based method with reliable prediction performance by combining many decision trees to yield a single consensus prediction. The main feature of random forest is that it cannot consider most of the available features at each split of the tree (Friedman et al., 2001).

Initially, landslide inventory and Landsat-8 imagery, and ALOS DSM are used to create data for training and validation for the four machine learning models. The distribution percentage of the training and the validation distribution made 75-25 percent, respectively. The proposed methodology is shown in Figure 5.

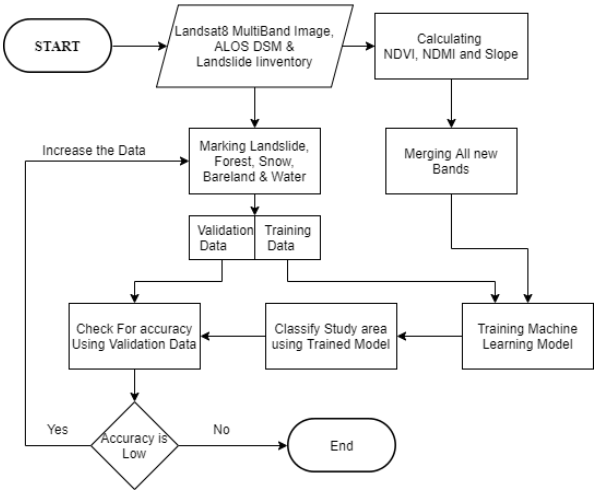

Figure 5. Proposed workflow of the methodology

4. RESULTS AND DISCUSSION

Initially, we import the required dataset, including Landsat 8 satellite imagery, DSM, and the landslides point obtained from the landslide inventory. Then, the landslide points are used to label landslides in Landsat-8 imagery by marking point features, and for the catalog purpose, the labeled landslides are also annotated by polygons. Further, using google earth engine, satellite images of the study area aare filter according to the requitred time interval and cloud cover over the study area. Using these acquired data, we calculated the parameters such as slope and indices such as NDVI and NDWI. For the classification process, we divided the landcover of the study area into five categories: forest, water, snow, bare land, and Landslide, and we marked the data for each of the five categories for training and for validation.

After calculating the parameters and plotting them over a scatter plot, we visualize the behavior of the parameter as per the landcover type to decide which machine learning algorithm can show up the better result in the classification process. After plotting the data on the scatter plot, we observed that the Landslide's properties are distinguishable to landcover such as snow, river, and vegetation, but it was found similar to the few of the bare land properties at few locations. In all landcover, only snow was found directly differentiable from other landcover types. Linear classification was observed to classify few categories very well but not the landslides. Therefore, we extended the classification of the landcover types to utilize non-linear methods to achieve expected results. We used multiple algorithms for the classification process like decision tree, naive Bayes, SVM, random forest, and minimum distance. Every algorithm mentioned has different methods based on the dataset's behaviors and the parameter (NDVI, NDWI, DEM, and slope) provided to train data. In our data, there is a non-linear relationship between the predictors and the outcomes. In that case, we consider enlarging the feature space using a function of the predictors, such as quadratic and cubic terms, to address this non-linearity. We could address the possible non-linear boundaries between classes similarly by enlarging a feature space using quadratic, cubic, and even high-order polynomial function of the predictors.

A total of 525 landslide pixels, 400 pixels for training, and 125 pixels are selected for validation purposes. The landside identification has been performed using SVM, CART, Minimum Distance, Random forest, and Naïve Bayes techniques, and obtained individual results are shown in Figure 6. In classification, SVM and RF outperform all other techniques by interpreting true positive rate (TPR), false-negative rate (FNR), positive predictive value (PPV), and false detection rate (FDR) collectively as shown in Table 1. From Figure 6, we can see a high frequency of landslides is observed near the river, which is confirmed from the landslide inventory.

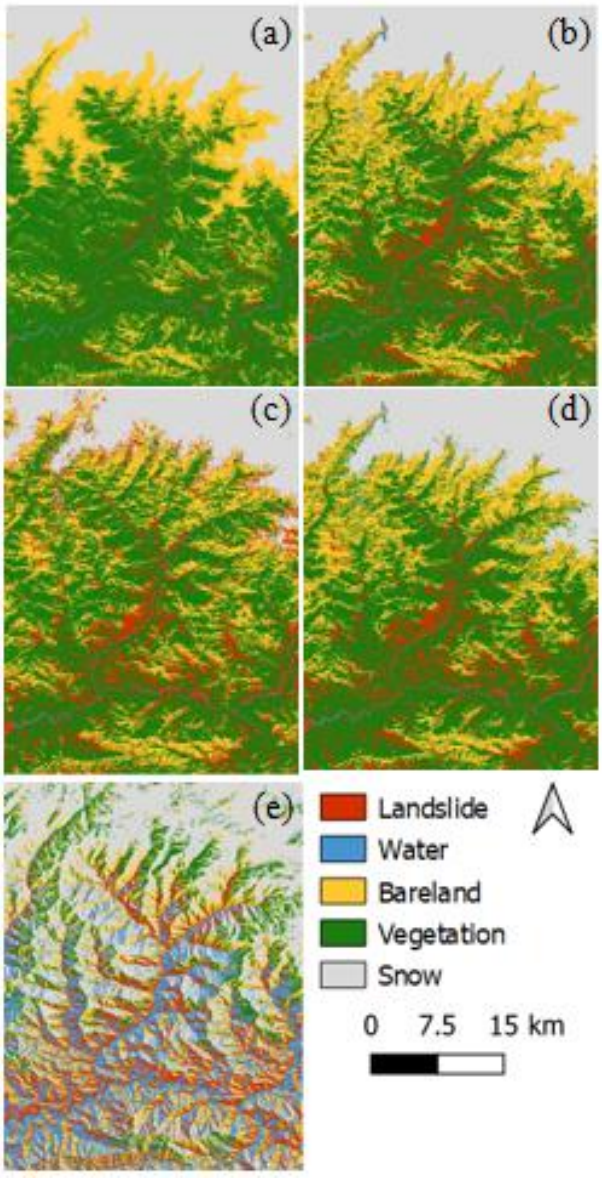

Figure 6. Classification of the study area by (a) SVM, (b) CART, (c) Minimum Distance, (d) Random Forest, (e) Naïve Bayes

To visualize the efficacy of the result, two Landslides, one from Mansoona (Ukhimath district) and another in Banadhar are shown in Figure 7 and Figure 8 with their Google Earth images. The Landslide is well identified in the images using the SVM algorithm, and the landslide area is marked as red in the image. We can also see the area at the river bank is also highlighted as red even though it is not a landslide. However, due to the

similarity in properties, a lot of riverbank having deteriorated bank is considered in the landslide category, the same issue arises in the bare land category few locations of bare land is considered landslides or vice versa. The landslides close to river are identified with higher frequency, it showa the effect of river erosion in our study area.

S.no	Method	TPR	FNR	PPV	FDR
1.	SVM	87.5	12.5	87.5	12.5
2.	CART	56.3	43.8	75	25
3.	Minimum Distance	81.2	18.8	92.9	7.1
4.	Random Forest	87.5	12.5	66.7	33.3
5.	Naïve Bayes	68.8	31.2	100	0

Table 1. Confusion matrix of the landslide identification using different Machine learning Algorithm

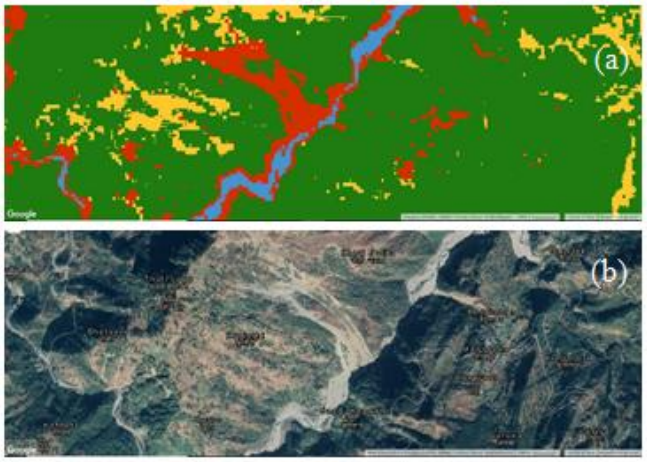

Figure 7. Location near Ukhimath identified as Landslide

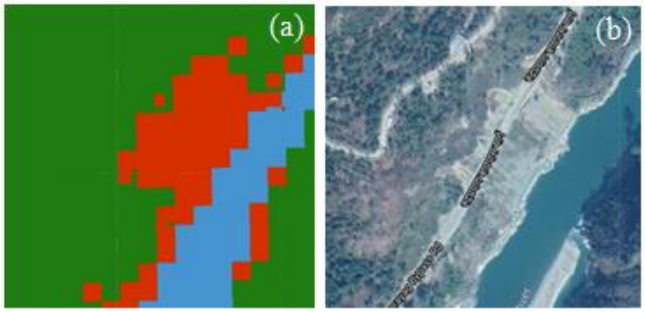

Figure 8. Location near Banadhar identified as Landslide

There are three limitations that we have identified in this research work. First is the unavailability of records of dumping zones created during road construction work. The characteristic of dumping zones is quite similar to landslides, which is problematic for landslide classification techniques. If a record is available about locations of dumping zones, these sites could be excluded from the classification process. The second limitation is related to the dataset, and we can only provide pixels of landslide for the training purpose to the model, which shows the properties of landslides. Other unmapped pixels could not be included in training or validation. The third limitation is the unavailability of a set of instructions (SOI) to select suitable indices out of all available indices for landslide identification. In this research work, we used NDVI and NDWI, but hundreds of other indices can be used but for the Landslide. For example, we are more focused on the SWIR band, but soil and rock properties could not be derived adequately due to low spectral resolution of the Landsat imagery.

5. CONCLUSIONS

The presented research work explores coarse-resolution satellite dataset's capability in identifying landslides by using various machine learning methods in the Rudraprayag region of the Uttarakhand state in India. This study demonstrated the significant potential of the BHUVAN landslide inventory, Landsat 8, and GEE. The Machine learning models, SVM and RF, achieved 87.5% TPR, indicating good accuracy of the model. We can achieve better results by overcoming the mentioned limitations. This paper shows the applicability of machine learning methods in classification problems. For the future, we can use deep learning methods to take our study to the next level for better understanding and to get output performance from the classification model.

ACKNOWLEDGEMENTS

The authors are grateful to GEE for providing free access to the geospatial data and an environment of working in a parallel processing environment. We also want to acknowledge the BHUVAN platform provided by the Indian Space Research Organization (ISRO). We also acknowledge our colleagues for their valuable assistance during field investigation.

REFERENCES

Acharya, T. D., & Yang, I. (2015). Exploring Landsat 8. *International Journal of IT, Engineering and Applied Sciences Research (IJIEASR)*, *4*(4), 4-10.

Bamisaiye, O. A. (2019). Landslide in parts of southwestern Nigeria. *SN Applied Science.* https://doi.org/10.1007/s42452-019-0757-0

Devara, M., Tiwari, A., & Dwivedi, R. (2021). Landslide susceptibility mapping using MT-InSAR and AHP enabled GIS-based multi-criteria decision analysis. *Geomatics, Natural Hazards and Risk*, *12*(1), 675-693. https://doi.org/10.1080/19475705.2021.1887939

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1, No. 10). New York: Springer series in statistics.

Ghorbanzadeh O, Blaschke T, Gholamnia K, Meena SR, Tiede D, Aryal J. Evaluation of Different Machine Learning Methods and Deep-Learning Convolutional Neural Networks for Landslide Detection. *Remote Sensing*. 2019; 11(2):196. https://doi.org/10.3390/rs11020196

Girosi, F., Jones, M., & Poggio, T. (1995). Regularization theory and neural networks architectures. *Neural computation*, *7*(2), 219-269. https://doi.org/10.1162/neco.1995.7.2.219

Goetz, J. N., Brenning, A., Petschko, H., & Leopold, P. (2015). Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. *Computers & geosciences*, *81*, 1-11. https://doi.org/10.1016/j.cageo.2015.04.007

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote sensing of Environment*, *202*, 18-27. https://doi.org/10.1016/j.rse.2017.06.031

Haigh, M., & Rawat, J. S. (2012). Landslide disasters: Seeking causes–A case study from Uttarakhand, India. In *Management of Mountain Watersheds* (pp. 218-253). Springer, Dordrecht. https://doi.org/10.1007/978-94007-2476-1_18

Ji, S., Yu, D., Shen, C., Li, W., & Xu, Q. (2020). Landslide detection from an open satellite imagery and digital elevation model dataset using attention boosted convolutional neural networks. *Landslides*, *17*(6), 1337-1352. https://doi.org/10.1007/s10346-020-01353-2

Kumar, R., & Anbalagan, R. (2013, July). Pixel based terrain analysis for landslide hazard zonation, a case study of Tehri reservoir region, Uttarakhand, India. In *2013 IEEE International Geoscience and Remote Sensing Symposium-IGARSS* (pp. 2868-2871). IEEE. https://doi.org/10.1109/IGARSS.2013.6723423

Loh, W. Y. (2011). Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, *1*(1), 14-23. https://doi.org/10.1002/widm.8

Mahrooghy, M., Aanstoos, J. V., Nobrega, R. A., Hasan, K., Prasad, S., & Younan, N. H. (2015). A machine learning framework for detecting landslides on earthen levees using spaceborne SAR imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *8*(8), 3791-3801. https://doi.org/10.1109/JSTARS.2015.2427337

Meghanadh, D., Tiwari, A., & Dwivedi, R. (2019, October). Multicriteria analysis for landslide inventory mapping using PS-InSAR. In *2019 IEEE Recent Advances in Geoscience and Remote Sensing: Technologies, Standards and Applications (TENGARSS)* (pp. 30-33). IEEE. https://doi.org/10.1109/TENGARSS48957.2019.8976054

Mohan, A., Singh, A. K., Kumar, B., & Dwivedi, R. (2020). Review on remote sensing methods for landslide detection using machine and deep learning. *Transactions on Emerging Telecommunications Technologies*, e3998. https://doi.org/10.1002/ett.3998

Pardeshi, S. D., Autade, S. E., & Pardeshi, S. S. (2013). Landslide hazard assessment: recent trends and techniques. *SpringerPlus*, *2*(1), 1-11. https://doi.org/10.1186/2193-1801-2-523

Semlali, I., Ouadif, L., & Bahi, L. (2019). Landslide susceptibility mapping using the analytical hierarchy process and GIS. *Current Science*, *116*(5), 773. http://dx.doi.org/10.18520/cs/v116/i5/773-779

Wang, H., Zhang, L., Yin, K., Luo, H., & Li, J. (2021). Landslide identification using machine learning. *Geoscience Frontiers*, *12*(1), 351-364.