Land subsidence assessment using geospatial (UAV, DInSAR), artificial intelligence and GPR techniques in Coal mining regions of East India

File Number : PDF/2023/002337

Submitted By : Dr. Pankaj Prasad

[**S**ERB **Q**ualified **U**nique **I**dentification **D**ocument: SQUID-1993-PP-5965]

Submission Date : 10-Aug-2023

PROPOSAL DETAILS

(PDF/2023/002337)

Principal Investigator	Mentor & Host Institution
Dr. Pankaj Prasad ppankaj.earthscience@gmail.com (RemoteSensing) **Contact No** : +919883329432 **Date of Birth** : 26-Feb-1993 **Name of Father/Spouse :** PradipPrasad	**Jeganathan Chockalingam** jeganathanc@bitmesra.ac.in Professor(Remote Sensing) **Birla Institute of Technology, Mesra** Mesra, ranchi, Ranchi, Jharkhand-835215 **Contact No.** : +917763859236 **Registrar Email** : registrar@bitmesra.ac.in **No. of PHD Scholars** : 8 **No. Post-Doctoral Fellow** : 0

Details of Post Doctorate

Ph.D. (Earth Science)[Degree Awarded on : 18-Jul-2022]

Environmental Geomorphology of Central West Coast of India: an Integrated Approach Using Geospatial, Machine Learning, and Ground Penetrating Radar Techniques

Research Supervisor/Guide & Institution :

C. Jeganathan

Birla Institute of Technology, Mesra

Brief details of Thesis work :

Earth's surface dynamics and changing environment constitute one of the most compelling and fascinating areas of study. Environmental geomorphology, a comparatively new field among the earth science disciplines, deals with the practical application of geomorphology. It helps explore resources and finds out appropriate solutions for environmental problems using geomorphic principles. In line with the essence of the discipline, the main objectives of this research are i) to prepare the geomorphological resource maps (coastal groundwater, land cover/land use, wetland) and geomorphological hazard maps (coastal flood, shoreline erosion) in the central west coast of India using Geospatial, Machine learning, and Ground Penetrating Radar (GPR) techniques. ii) to reconstruct the Holocene paleo-shoreline and associated regional relative sea-level highstands. To fulfill these objectives, several remote sensing data including optical and microwave have been applied. This huge satellite dataset has been processed using different machine learning models. The sub-surface information has been gathered by GPR method and age of sediment sample has been determined by Optically Stimulated Luminescence (OSL) dating method.
The application of the machine learning models have been successful for coastal hazard and resource prediction mapping. This study is the first of its kind to assess and compare the total five satellite sensors including optical and SAR data using five machine learning algorithms for coastal land cover and land use mapping. Novel stacking ensemble model (Random Forest (RF)-Support Vector Machine (SVM)-Multivariate Adaptive Regression Splines (MARS)) has been proposed for probabilistic wetland mapping. The study has introduced adabag model as a base classifier with five other ensemble classifiers including RF, Logitboost (LB), Nearest Shrunken Centroids (NSC), K-Nearest Neighbour (KNN), and Boosted Regression Tree (BRT) for coastal flood susceptibility mapping with very high precision. In India, this is the pioneering attempt for reconstruction of the Holocene regional relative sea-level and associated paleo-shoreline using GPR sub-surface image. In nutshell, outputs of this work can be utilized as supporting database for coastal planning, sustainable development and the entire methodology of this study can be opted for carrying out future investigation in the research region as well as in alike coastal environment. This study makes itself distinct with this modest novel attempt of integrating Geospatial, Machine learning, and GPR techniques in a single platform.

Technical Details :

Research Area : Earth & Atmospheric Sciences (Earth & Atmospheric Sciences)

Project Summary :

Land subsidence resulting from coal mining activities poses significant environmental, social, and economic challenges in the regions of East India. To effectively mitigate the impacts of land subsidence, a comprehensive understanding of its dynamics is essential. Therefore, it is important to apply appropriate methods for effective land subsidence monitoring. As a result of remarkable technological advancement, the method of spatial data generation in the form of maps has shifted from traditional topographical surveys to satellite remote sensing. Remote sensing is considered as one of the most important tools for land subsidence mapping and monitoring because of its extensive and repetitive coverage of the earth's surface. Optical and microwave remote sensing data along with Unmanned Aerial Vehicles (UAVs) and Differential Interferometric Synthetic Aperture Radar (DInSAR) techniques, furnishes complementary information, hence land subsidence mapping tasks can take advantage of the fusion of both data types resulting in an increase of mapping precision. In addition, high resolution ground penetrating radar (GPR) data helps to understand stratigraphy of the regions. The bulk dataset processing in a vast area is an arduous task by traditional classification methods. In this decade, approaches through the machine and deep learning models have gradually increased for different studies. However, machine learning and deep learning approaches have not been applied to land subsidence in this region so far. This study makes itself distinct with this modest novel attempt of geospatial, machine learning, deep learning, and GPR methods in a single platform for land subsidence study. Considering the above-mentioned research gaps, the main aim of the project is to employ a multi-modal approach that integrates UAV, DInSAR, Artificial Intelligence (AI), and GPR to assess and analyze land subsidence in coal mining regions of East India. Additionally, detailed investigation of sedimentary architecture using GPR survey will be carried out in the selected vulnerable sites. The output of the research will be of immense help to policymakers for sustainable development, environmental planning and management. Moreover, the proposed methodology of this work can be opted in other regions of the world facing such land subsidence.

Objectives :

• To monitor and quantify land subsidence in coal mining regions of East India using UAV-based imagery and DInSAR techniques.
• To establish the robust novel machine and deep learning predictive models for land subsidence assessment.
• To validate the results obtained from UAV, DInSAR, and AI using ground-truth data collected through GPR surveys.
• To provide actionable recommendations for sustainable mining practices and land subsidence mitigation strategies.

Keywords :

East India, Land Subsidence, Unmanned Aerial Vehicles, Synthetic Aperture Radar, Artificial Intelligence, Ground Penetrating Radar

Expected Output and Outcome of the proposal :

The land subsidence map as an important tool for decision support system will be of immense help to policy makers for sustainable development, environmental planning and management, and hazard mitigation. This approach can be opted in other regions of the world suffering from such land subsidence. The major expected outcomes of this work are:
• Demarcation of the most vulnerable area under land subsidence regions.
• Novel machine learning and deep learning method with higher precision can be established for land subsidence assesment.
• Scientific interpretation of sub-surface radar images of land subsidence affected regions.
• Recommendations for sustainable mining practices and subsidence mitigation strategies to minimize environmental and societal impacts.

Reference Details :

S.No	Reference Details
1	**Dr. Victor Joseph Loveson** **Former Chief Scientist** **Geological Oceanography** **CSIR National Institute of Oceanography** **Dona Paula-403004, Goa[+8806827575]** vjloveson@gmail.com
2	**Prof. Onkar Singh Chauhan** **Former Chief Scientist** **Geological Oceanography** **CSIR National Institute of Oceanography** **Dona Paula-403004, Goa[+9422439151]** chauhanonkar@gmail.com

Methodology:

The first and key step for the land subsidence assessment is to collect the inventory data of land subsidence. These datasets will be assembled from the field survey, previous documents, high resolution UAV and synthetic aperture radar (SAR) images. At the next level, various factors (topographical, geological, environmental, and hydrological) will be considered for land subsidence mapping. After that, the entire datasets will be classified as training and validation applying k-fold cross-correlation method to avoid the bias. These training and validation data are intended for model establishment and testing purposes, respectively. At last, the results of the predictive models will be validated from the GPR survey.

Application of geospatial technology (UAV, DInSAR) for land subsidence inventory and conditioning factors

The inventory of the land subsidence will be collected from optical and SAR datasets. High-resolution UAV imagery will be collected periodically to monitor surface deformations in the study area and DInSAR techniques will be applied to satellite radar data to generate displacement maps and quantify subsidence rates. Advanced InSAR analysis will be performed to distinguish between natural and mining-induced subsidence. All the conditioning variables of the hazards will be prepared employing different tools in GIS platform.

Application of machine and deep learnings in subsidence mapping

Machine learning and deep learning algorithms will be employed to establish relationships between mining activities and subsidence patterns. A number of machine learning methods such as random forest (RF), boosted regression tree (BRT), support vector machine (SVM), logitboost (LB), etc. and deep learning namely convolutional neural networks (CNN), recurrent neural networks (RNN), radial basis function networks (RBFNs), deep belief networks (DBNs) etc. will be applied to compare the precision of the algorithms. Apart from this, based on the nature of data and applied models, new models are supposed to be developed for land subsidence study. For the model validation, various statistical criteria including area under the receiver operating characteristics curve (AUROC), accuracy, specificity, sensitivity, root mean square error (RMSE), mean absolute error (MAE), kappa coefficient will be employed.

Application of ground penetrating radar technique for detailed sub-surface information

GPR surveys will be conducted in the selected land subsidence regions of the study area to verify subsurface structural changes associated with land subsidence. Data from GPR surveys will be correlated with UAV and DInSAR results to validate the accuracy of the multi-modal

analysis. GPR antennas of 100, 200 and 400 MHz (central frequency) will be used to acquire the sub-surface information. The attenuation and velocity spectra models will be applied to estimate velocity using cross-correlation method. In this method, inversion of reflection amplitudes are employed to measure the reflection coefficients for calculating velocity of internal layers.

Tentative Time Schedule of project activity

Activity Details	First Year				Second Year			
Months	1-3	4-6	7-9	10-12	13-15	16-18	19-21	22-24
Literature Review								
Primary and secondary data collection								
Data processing and establishment of the novel model								
Preparation and evaluation of land subsidence map								
Report and Manuscript preparation								

Bibliography

Chatterjee, R.S., Thapa, S., Singh, K.B., Varunakumar, G. and Raju, E.V.R., 2015. Detecting, mapping and monitoring of land subsidence in Jharia Coalfield, Jharkhand, India by spaceborne differential interferometric SAR, GPS and precision levelling techniques. Journal of Earth System Science, 124, 1359-1376. https://doi.org/10.1007/s12040-015-0606-5

Park, S. and Choi, Y., 2020. Applications of unmanned aerial vehicles in mining from exploration to reclamation: A review. Minerals, 10(8), 663. https://doi.org/10.3390/min10080663

Karanam, V., Motagh, M., Garg, S. and Jain, K., 2021. Multi-sensor remote sensing analysis of coal fire induced land subsidence in Jharia Coalfields, Jharkhand, India. International Journal of Applied Earth Observation and Geoinformation, 102, 102439. https://doi.org/10.1016/j.jag.2021.102439

Zhang, Y., Lian, X., Ge, L., Liu, X., Du, Z., Yang, W., Wu, Y., Hu, H. and Cai, Y., 2022. Surface Subsidence Monitoring Induced by Underground Coal Mining by Combining DInSAR and UAV Photogrammetry. Remote Sensing, 14(19), 4711. https://doi.org/10.3390/rs14194711

Wu, Z., Xia, T., Nie, J. and Cui, F., 2020. The shallow strata structure and soil water content in a coal mining subsidence area detected by GPR and borehole data. Environmental Earth Sciences, 79, 1-13. https://doi.org/10.1007/s12665-020-09178-x

PROFORMA FOR BIO-DATA (to be uploaded)

1. Name and full correspondence address: **Dr. Pankaj Prasad**
Department of Remote Sensing,
Birla Institute of Technology, Mesra
Ranchi-835215, Jharkhand

2. Email(s) and contact number(s): **ppankaj.earthscience@gmail.com**
9883329432

3. Institution: **Birla Institute of Technology, Mesra**

4. Date of Birth; **26/02/1993**

5. Gender (M/F/T): **M**

6. Category Gen/SC/ST/OBC: **Gen**

7. Whether differently abled (Yes/No); **No**

8. Academic Qualification (Undergraduate Onwards)

	Degree	Year	Subject	University/Institution	% of marks
1.	**BSc**	**2013**	**Geography (Hons)**	**University of Calcutta**	**55**
2.	**MSc**	**2015**	**Geography with specialization in Remote Sensing & GIS**	**Banaras Hindu University**	**72.3**
3.	**PhD**	**2022**	**Earth Science**	**Goa University and CSIR-NIO**	**64.5**

9. Ph.D thesis title**: Environmental Geomorphology of Central West Coast of India: an Integrated Approach Using Geospatial, Machine Learning, and Ground Penetrating Radar Techniques**
Guide's Name: **Dr Victor Joseph Loveson**
Institute/Organization/University: **Goa University and CSIR- National Institute of Oceanography, Goa**
Year of Award: **2022**

10. Work experience (in chronological order).

S.No.	Positions held	Name of theInstitute	From	To	Pay Scale
1	**Junior Research fellow**	**CSIR-National Institute of Oceanography**	**16/08/2016**	**15/08/2018**	**25000**
2	**Senior Research fellow**	**CSIR-National Institute of Oceanography**	**16/08/2018**	**15/08/2021**	**28000**
3	**Senior Project Associate**	**CSIR-National Institute of Oceanography**	**26/11/2016**	**25/05/2018**	**42000**
4	**Project Associate**	**Birla Institute of Technology**	**31/03/2023**	**10/08/2023**	**31000**

11. Professional Recognition/ Award/ Prize/ Certificate, Fellowship received by the applicant.

S.No	Name of Award	Awarding Agency	Year
1	**Junior Research fellow**	**University Grant Commission**	**2015**
2	**Best Research Paper Award**	**International Conference on Biofouling and Ballast Water Management (ICBAB)**	**2021**

12. Publications *(List of papers published in SCI Journals, in year wise descending order).*

S. No.	Author(s)	Title	Name of Journal	Volume	Page	Year
1	**Prasad, P.,** Loveson, V.J., Mondal, S., and Chandra, P	Multi-resource potentiality and multi-hazard susceptibility assessments of the central west coast of India applying machine learning and geospatial techniques	**Environmental Earth Sciences,. (IF=3.11)**	82	226	2023
2	Thorat, B.R., **Prasad, P.** and Ram, A.,	Heavy metal accumulation in a moderately polluted Ulhas estuary, Western India.	**Regional Studies in Marine Science, (IF=2.16)**	60	102818	2023
3	Kulimushi, L.C., Bashagaluke, J.B., **Prasad, P., et al**	Soil erosion susceptibility mapping using ensemble machine learning models: A case study of upper Congo river sub-basin.	**Catena, (IF=6.5)**	222	106858	2023
4	**Prasad, P.,** Loveson, V.J., Chandra, P. andKotha, M	Evaluation and Comparison of the Landsat-8, Sentinel-1, Sentinel-2, LISS III, andLISS IV sensors in land cover/land use studies	**Ecological Informatics (IF3.14)**	68	101522	2022
5	**Prasad, P.,** Loveson, V.J., Das,B., and Kotha, M	Novel Ensemble Machine Learning Models in FloodSusceptibility Mapping	**Geocarto International (IF 4.89)**	37	4571-4593	2022
6	**Prasad, P.,** Loveson, V.J., Das,S., and Chandra, P 2021	Artificial intelligence approaches for spatial prediction of landslides in Western Ghats of India	**Environmental Earth Sciences (IF 2.78)**	80	720	2021

7	Nigam, R., Luis, A.J., **Prasad, P.,** Kuttikar, S., Yadav, R., Vaz, E. and Kotha, M.,	Spatio-temporal assessment of COVID-19 lockdown impact on beach litter status and composition in Goa, India	**Marine Pollution Bulletin(IF 5.55)**	174	113293	2021
8	Shivakrishna, A., Ramteke, K.K., Kesavan, S., **Prasad, P.,** Naidu, B.C., Dhanya, M. and Abidi, Z.J.,	Monitoring of current land use pattern of Ramsar designated Kolleru Wetland, India using geospatial technologies	**Journal of Environmental Biology (0.78)**	42	1106-1111	2021
9	**Prasad, P.,** Loveson, V.J., Kotha, M. and Yadav, R.,	Application of machine learning techniques in groundwater potential mapping along the west coast of India	**GIScience & Remote Sensing (IF 6.22)**	57	735-752	2020
10	**Prasad, P.,** and V JLoveson	Signature of buried channels as deduced fromsubsurface GPR survey atSouthwest coast of Tamil Nadu, India	**Arabian Journal of Geoscienc es(IF 1.82)**	13	1-12	2020

13. Detail of patents.

S.No	Patent Title	Name of Applicant(s)	Patent No.	Award Date	Agency/Country	Status

14. Books/Reports/Chapters/General articles etc.

S.No	Title	Author's Name	Publisher	Year of Publication

15. Any other Information (maximum 500 words)
Apart from this, I was member of several projects

- Present Ongoing Project is " **Geomorphological Sand Mining of the Rivers of Goa**, India"
- Working in the project "**Submarine Groundwater Discharge of India**" sponsored by theMinistry of earth Sciences, March 2018 to till date.
- Worked in the project "**Scientific study on beach sand and budgeting from Valliyar toMelmidalam**" sponsored by the Indian Rare Earth Limited, April 2018-July 2017.
- Worked in the project " **Geo-archaeological investigation to assess the existence and reconstruction of an ancient port at Gopakapattanam, Goa: scientific and cultural aspects**" sponsored by the Ministry of Earth Sciences, Nov 2017- Sep 2019.

Conferences

Prasad, P., Loveson, V.J., and Kotha, M (2019). A comparative assessment of statistical and machine learning models for flash flood susceptibility mapping at Sindhudurg coast, Maharashtra, India. International Symposium on 'Advances in Coastal Research with special reference to Indo Pacific, Chennai, India.

Prasad, P., and V J Loveson (2020). Landslide susceptibility mapping using GIS based machine learning algorithms in the north Western Ghats of India. 4th Disaster Risk and Vulnerability Conference, Kerala, India.

Prasad, P., Loveson, V.J., Kotha, M. and Yadav, R., (2020). Application of GIS based graph neural networks in groundwater potential mapping. International Conference on Graph Labeling and Applications, Goa, India.

Reviewer of Peer review Journals

- Environmental Earth Sciences (Springer)
- Journal of Flood Management (Willey)
- Geomatics, Natural Hazards and Risk (Taylor & Francis)
- Geocarto International (Taylor & Francis)
- Stochastic Environmental Research and Risk Assessment (Springer)
- Geosystems and Geoenvironment (Elsevier)
- Environmental Monitoring and Assessment (Springer)
- Indian Society of Remote Sensing (Springer)
- Acta Geophysica (Springer)

Training

- Three months Certificate Course on "**Remote Sensing, GIS & GNSS**" by Indian Institute of Remote Sensing, ISRO.

- Participated in two week day Certificate Course on **"Microwave Remote Sensing and application"**organised by National Remote Sensing Centre (ISRO) Hyderabad, India,17-28April, 2017.

- Participated in 2 months Certificate course on **"Application in Remote Sensing"** at Indian Institute of Remote Sensing, Dehradun organised by **Indian Space Research Organisation.**

- Participated in **"INQUA Meet"**, jointly organized by the **"INQUA and Indian Quaternary Group"**,

- Participated in a 5 day workshop on **"High Performence Computing"**, jointly organized by the **Centre for Development of advanced computing and Goa University**, 28 Jan-1st Feb, 2019

Declaration:

All the information provided above are true to my conscience.

Curriculum Vitae

Dr. Jeganathan Chockalingam
Date of Birth: 15th March 1971
Place of Birth: Virudhunagar, Tamil Nadu, India

e-mail Address: **jegan_iirs@yahoo.com**

ORCID: 0000-0002-2375-7677

Years of Experience (Jan, 1993 to Jun, 2023):
30+ years
(3.5 years in RRSSC-D, ISRO + 12 years in IIRS, ISRO + 3 years in University of Southampton, UK + from 1st Sep, 2011 onwards in BIT, Mesra)

Areas of Research:
Remote Sensing based vegetation and environmental Analysis, Modelling Space-Time vegetation dynamics, Land Surface Phenology, Vegetation and Climate, Landscape Metrics and Modelling, Land Cover Change Modelling, Geostatistics, Downscaling of Satellite derived vegetation data, Multi-criteria decision modelling, Fuzzy Logic and Software Development.

Educational Qualifications :
1991 **B.Sc. Physics**, Madurai Kamaraj University, Tamil Nadu, India (91%)
1995 **M.Sc. Physics**, Madurai Kamaraj University, Tamil Nadu, India (82%)
2003 **M.Sc. Geoinformatics**, International Institute for Geoinformation Science and Earth Observation (ITC), Enschede, The Netherlands. (86% Distinction, *Klaas Jan Beek Awardee*).
2007 **Ph.D.(Forest Geoinformatics),** Forest Research Institute (FRI) University, Dehradun, India.

Peer Reviewed Papers Published: 60+ ; Books: 4 ; Book Chapters: 13
Number of Phds Guided: Completed: 5, Submitted: 2 & Ongoing: 8
Number of M.Tech. & MSc students supervised: 50+
Number of Sponsored Projects Handled & Grant: 25+ ; ~Rs. 14.5Crores
Number of Invited Lectures: 30+

Work Experience:

1993 – 1996	Scientific Assistant 'B', Regional Remote Sensing Service Centre (RRSSC), Indian Space Research Organisation (ISRO), Dehradun, India
1996 (Apr – Sep)	Scientific Assistant 'C', RRSSC,ISRO, Dehradun, India
1996-1997	Scientist/Engineer 'SB', Indian Institute of Remote Sensing (IIRS), National Remote Sensing Centre (NRSC, ISRO), Dehradun, India
1998 – 2000	Scientist/Engineer 'SC', IIRS, NRSC, ISRO, Dehradun, India
2001-2005	Scientist/Engineer 'SD', IIRS, NRSC, ISRO, Dehradun, India
2005-2008	Scientist/Engineer 'SE', IIRS,NRSC, ISRO, Dehradun, India
2008 (June) – 2011 (July)	Senior Researcher in Spatial Analysis, School of Geography, University of Southampton, Southampton, England, UK
1st Sep 2011 onwards -	Professor, Department of Remote Sensing, Birla Institute of Technology (BIT) University, Mesra, Ranchi, Jharkhand
9th May, 2007 – 31st Dec, 2020	Head, Department of Remote Sensing, Birla Institute of Technology (BIT) University, Mesra, Ranchi, Jharkhand
1st Jan 2021 onwards	Dean (Research, Innovation and Entrepreneurship), BIT, Mesra

Research Papers Published **(Peer-Reviewed, SCI papers, Last 5 years only)**

1. Alex Praveen, **Jeganathan, C.** and Mondal, S. (2023). Mapping Annual Cropping Pattern from Time-Series MODIS EVI Using Parameter-Tuned Random Forest Classifier. *Journal of Indian Society of Remote Sensing*, https://doi.org/10.1007/s12524-023-01676-2.
2. Mallika Bhuyan, Beependra Singh, Swayam Vid and **Jeganathan, C.** (2023). Analysing the spatio-temporal patterns of vegetation dynamics and their responses to climatic parameters in Meghalaya from 2001 to 2020. *Environmental Monitoring and Assessment*, https://doi.org/10.1007/s10661-022-10685-6.
3. Sujit M. Ghosh, Behera, M.D., Kumar, S., Das,P., Prakash,A.J., Bhaskaran,P.K., Roy,P.S., Barik, S.K., Jeganathan, C., Srivastava, P.K., and Behera, S.K. (2022). Predicting the Forest Canopy Height from LiDAR and Multi-Sensor Data Using Machine Learning over India. *Remote Sensing*, 14(23). https://doi.org/10.3390/rs14235968.
4. Swadhina Koley and **Jeganathan, C.** (2022). Evaluating the climatic and socio-economic influences on the agricultural drought vulnerability in Jharkhand. *Environmental Monitoring and Assessment*, https://doi.org/10.1007/s10661-022-10557-z.
5. Farzana Shaheen, Nayama Scariah, Mili Ghosh, A.P. Krishna, **Jeganathan, C**. and Nick M. Hoekzema (2022). Shadow method retrievals of the atmospheric optical depth above Gale crater on Mars using HRSC images. *Icarus*, 388(3-4):115229. DOI: 10.1016/j.icarus.2022.115229.
6. Saptarshi Mondal and **Jeganathan, C.** (2022). Effect of scale, landscape heterogeneity and terrain complexity on agriculture mapping accuracy from time-series NDVI in the Western-Himalaya region. *Landscape Ecology*, https://doi.org/10.1007/s10980-022-01533-6.
7. Atkinson, P.M., Stein, A., and **Jeganathan, C.** (2022). Spatial sampling, data models, spatial scale and ontologies: Interpreting spatial statistics and machine learning applied to satellite optical remote sensing. *Spatial Statistics*, 100646. (Aug, 2022)
8. Swadhina Koley and **Jeganathan, C.** (2022). Sentinel 1 and Sentinel 2 for cropland mapping with special emphasis on the usability of textural and vegetation indices. *Advances in Space Research*, 69(4), 1768-1785. (Feb, 2022)
9. Niraj Priyadarshi, V.M. Chowdary, K. Chandrasekar, **Jeganathan C**., Soumya Bandyopadhyay, Y.K. Srivastava, D. Dutta , Neeti Neeti, and Chandra Shekhar Jha (2021). Multi-resolution analysis based data mining approach to assess vegetation dynamics in Jharkhand using time series MODIS products. *Geocarto International*, https://doi.org/10.1080/10106049.2021.2024610. (Dec, 2021)
10. Beependra Singh, **Jeganathan C**, Rathore VS, Behera MD, Singh CP, Roy PS, Atkinson PM (2021). Resilience of the Central Indian Forest Ecosystem to Rainfall Variability in the Context of a Changing Climate. *Remote Sensing*. 2021; 13(21):4474. https://doi.org/10.3390/rs13214474 (Nov, 2021)
11. Beepedra Singh, **Jeganathan, C.** and Rathore, V.S. (2020). Improved NDVI based proxy leaf-fall indicator to assess rainfall sensitivity of deciduousness in the central Indian forests through remote sensing, *Nature-Scientific Reports*, 10:17638.(Oct,2020)
12. Swadhina Koley and **Jeganathan C.** (2020). Estimation and evaluation of high spatial resolution surface soil moisture using multi-sensor multi-resolution approach. *Geoderma - The Global Journal of Soil Science* (Elsevier), 378, 114618. (Nov, 2020).
13. Niraj Priyadarshia, Soumya Bandyopadhyaya, V.M. Chowdary, K. Chandrasekar, C., **Jeganathan C.**, Uday Raj, and Chandra Shekhar Jha (2020). Segmentation-based approach for trend analysis and structural breaks in rainfall time series (1851–2006) over India, *Journal of Hydrological Sciences*, 65(9), 1583-1595. (Taylor & Francis) https://doi.org/10.1080/02626667.2020.1761022. (May, 2020).

14. Shawky Mansour and **Jeganathan, C.** (2020). Diagnostically counting Palm Date Trees in Al-Ahhssa Governorate of Saudi Arabia: An integrated GIS and remote sensing processing of IKONOS imagery. *Spatial Information Research*, (Springer), DOI: 10.1007/s41324-020-00318-w. (Jan, 2020)
15. Dash, J., Behera, M.D., **Jeganathan, C**., Jha, C.S., Sharma, S., Lucas, R., et al. (2020). India's contribution to mitigating the impacts of climate change through vegetation management. *Tropical Ecology*, 61, 168-171.
16. Suman Sinha, Shiv Mohan, A.K. Das, L.K. Sharma, **C. Jeganathan**, A. Santra, S. Santra Mitra and M.S. Nathawat (2019). Multi-sensor approach integrating optical and multi-frequency synthetic aperture radar for carbon stock estimation over a tropical deciduous forest in India. *Carbon Management*, DOI: 10.1080/17583004.2019.1686931. (Nov, 2019)
17. Harshit Rajan and **Jeganathan, C.** (2019). Understanding Spatio-temporal Pattern of Grassland Phenology in the western Indian Himalayan State. *Journal of Indian Society of Remote Sensing*, Springer, 47(7). 1137-1151. DOI:10.1007/s12524-019-00976-w. (Jul, 2019)
18. Suman Sinha, A. Santra, A.K. Das, L.K. Sharma, Shiv Mohan, M.S. Nathawat, S. Santra Mitra and **C. Jeganathan** (2019). Regression-based integrated Bi-Sensor SAR Data Model to Estimate Forest Carbon Stock. *Journal of Indian Society of Remote Sensing*, Springer, 47(7), 1599-1608. https://doi.org/10.1007/s12524-019-01004-7. (May, 2019)
19. Suman Sinha, A. Santra, A. K. Das,L. K. Sharma, Shiv Mohan,M. S. Nathawat, S. S. Mitra & **C. Jeganathan** (2019). Accounting tropical forest carbon stock with synergistic use of space-borne ALOS PALSAR and COSMO-Skymed SAR sensors. *Tropical Ecology*, DOI 10.1007/s42965-019-00011-6. (Apr, 2019)
20. Margaret Johnson, Petrutza C. Caragea, Wendy Meiring**, C. Jeganathan** and Peter M. Atkinson (2018). Bayesian Dynamic Linear Model for estimation of phenological events from Remote sensing data. *Journal of Agricultural, Biological and Environmental Statistics*. https://doi.org/10.1007/s13253-018-00338-y. (Mar, 2019)
21. Saptarshi Mondal and **Jeganathan, C.** (2018). Evaluating the performance of multi-class and single-class classification approaches for mountain agriculture extraction using time series NDVI. *Journal of Indian Society of Remote Sensing*, 46(12), 2045-2055. (Dec, 2018)
22. Niraj Priyadarshi, V.M. Chowdary, Iswar Chandra Das, **Jeganathan Chockalingam**, Y.K. Srivastava, G Srinivasa Rao, Uday Raj & Chandra Shekhar Jha (2018). Wavelet and non-parametric statistical based approach for long term land cover trend analysis using time series EVI data. *Geocarto International*. DOI: 10.1080/10106049.2018.1520925 (Sep, 2018)
23. Suman Sinha, Santra, A., Sharma, L.K., **Jeganathan, C**., Nathawat, M.S., Das, A.K. and Shiv Mohan (2018). Multi-polarized Radarsat-2 satellite sensor in assessing forest vigor from above ground biomass. *Journal of Forestry Research*, DOI 10.1007/s11676-017-0511-7, 29(4), 1139-1145. (July, 2018)
24. Saptarshi Mondal and **Jeganathan, C.** (2018). Extracting Mountain Agriculture from Time-Series MODIS NDVI using Dynamic Time Warping Technique. *International Journal of Remote Sensing*, 39(11), 3679-3704. (SCI) (Feb, 2018)

Awards, Recognition & Fellowship:

Letter of Appreciation from Director-General, Forest Survey of India (2023)
INSA Teachers Award (2020) – Indian National Science Academy Award
Samaj Bandhu Award – Education (2019) from Prantik - Care the Earth
India-UK Water Centre Travel Grant (2018)
Indian National Geospatial Award (2015)
Erasmus Mundus EU Fellowship (2015)
Adjunct Faculty (Centre for Space Science and Technology Education - CSSTE-AP, Affiliated to UN, Dehradun, India, 1996 to 2008)
Visiting Scholar (School of Geography, University of Southampton, till 2014)
Visiting Scholar (School of RS & GIS, University of Peking, China, 2012)
Visiting Scholar (School of Geographical Information Science, University of Nanjing, China, 2012)
Visiting Scholar (School of Geography, University of California Santa Barbara, 2010)
Visiting Scholar (NASA Centre of Excellence, School of Geography, Boston University, 2009)
Klaas Jan Beek Award for best Research Thesis of ITC, Netherlands (2003)
IIRS-ITC Masters Fellowship (Oct, 2001 to Mar, 2003)
Japanese Doctoral Fellowship (Awarded, but did not avail) (2000)
Gold Medalist (B.Sc. Physics, VHNSN College, Madurai Kamaraj University)

Professional activities and Memberships

Life Member: Indian Society of Remote Sensing (ISRS)
Life Member: Indian Society of Agriculture Information Technology (INSAIT)
Life Member: Indian Society of Geomatics (ISG)
Life Member: Deccan Geographers Association
Life Member: Indian Planetary Association
Associate Editor, Journal of Indian Society of Remote Sensing (Springer)
Guest Editor – Special Issue in *Remote Sensing* Journal
Guest Editor – Special Issue in *Science of Remote Sensing* Journal (Elsevier)
National Expert under ISWT programme, IIT Kharagpur (2014)
Expert Panel Member, Scientific Review Committee, IEEE/IGARSS
Editorial Board Member of International Journal of Remote Sensing Applications
Member, Executive Council, Indian Society of Agriculture Information Technology (INSAIT)
Member, Network for Conserving Central India (NCCI)

Involvement in the Institutional Committee

President, Institute Innovation Council
Chairman, Research Policy & Product Development
Chairman, Seed Money Allottment Committee
Chairman, Institute Report Preparations for Convocation
Member, Administrative Committee on Academic Programmes

Jury Member, Institute Annual Athletic-Meet Committee
Member, Institute Academic Council
Member, Institute Foundation Day Committee
Member, Institute Convocation Ceremony Committee
Member, Department Board of Studies
Member, Department Policy Committee

Invited Lectures/Presentations/Conference Papers:

Delivered 30+ invited lectures at various universities, institutions, societies under ATAL FDPs, NNRMS training programmes, and published 40+ conference papers.

Referees:

1. **Dr. George Joseph** **E-Mail**: *georgejoseph1938@hotmail.com*
   Padma Bhusan
   Ex-Director (CSSTE-AP & SAC)
   Honorary Distinguished Professor (ISRO)
   Space Application Centre, Ahmedabad, India

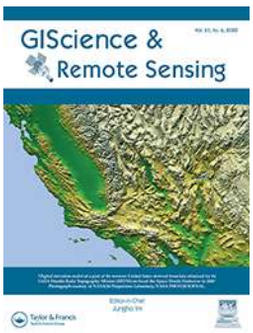
2. **Dr. Peter M. Atkinson** **E-Mail**: *pma@lancaster.ac.uk*
   Dean
   Faculty of Science and Technology
   University of Lancaster
   Lancaster
   United Kingdom
   Tel:+44-23-80595000

3. **Dr. P.L.N. Raju** **E-Mail:** *raju.pln@assam.gov.in; rajupln@gmail.com*
   Special Secretary, Department of Science and Technology
   Assam Government, Guwahati.
   &
   Ex-Director
   North-Eastern Space Application Centre (NESAC), Shillong, Meghalaya
   Tel: +91-9411768991

Declaration:

All the information provided above are true to my conscience. For any clarification/details I may be contacted.

Place: BIT, Mesra **(Dr. C. Jeganathan)**
August, 2023



GIScience & Remote Sensing

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tgrs20

Application of machine learning techniques in groundwater potential mapping along the west coast of India

Pankaj Prasad, Victor Joseph Loveson, Mahender Kotha & Ramanand Yadav

To cite this article: Pankaj Prasad, Victor Joseph Loveson, Mahender Kotha & Ramanand Yadav (2020) Application of machine learning techniques in groundwater potential mapping along the west coast of India, GIScience & Remote Sensing, 57:6, 735-752, DOI: 10.1080/15481603.2020.1794104

To link to this article: https://doi.org/10.1080/15481603.2020.1794104

GISCIENCE & REMOTE SENSING
2020, VOL. 57, NO. 6, 735–752
https://doi.org/10.1080/15481603.2020.1794104

Taylor & Francis
Taylor & Francis Group

Application of machine learning techniques in groundwater potential mapping along the west coast of India

Pankaj Prasad [a,b], Victor Joseph Loveson[a,c], Mahender Kotha[b] and Ramanand Yadav[a,c]

[a]Geological Oceanography Division, CSIR- National Institute of Oceanography, Dona Paula, Goa, India; [b]School of Earth, Ocean and Atmospheric Sciences, Goa University, Taleigao, Goa, India; [c]Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, India

ABSTRACT

Groundwater potential mapping (GWPM) in the coastal zone is crucial for the planning and development of society and the environment. The current study is aimed to map the groundwater potential zones of Sindhudurg coastal stretch on the west coast of India, using three machine learning models: random forest (RF), boosted regression tree (BRT), and the ensemble of RF and support vector machine (SVM). In order to achieve the objective, 15 groundwater influencing factors including elevation, slope, aspect, slope length (LS), profile curvature, plan curvature, topographical wetness index (TWI), distance from streams, distance from lineaments, lithology, geomorphology, soil, land use, normalized difference vegetation index (NDVI), and rainfall were considered for inter-thematic correlations and overlaid with spring and well occurrences in a spatial database. A total of 165 spring and well locations were identified, which had been divided into two classes: training and validation, at the ratio of 70:30, respectively. The RF, BRT, and RF-SVM ensemble models have been applied to delineate the groundwater potential zones and categorized into five classes, namely very high, high, moderate, low, and very low. RF, BRT, and ensemble model results showed that 33.3%, 35.6%, and 36.8% of the research area had a very high groundwater potential zone. These models were validated with area under the receiver operating characteristics (AUROC) curve. The accuracy of RF (94%) and hybrid model (93.4%) was more efficient than BRT (89.8%) model. In order to further evaluate and validate, four different sites were subsequently chosen, and we obtained similar results, ensuring the validity of the applied models. Additionally, ground-penetrating radar (GPR) technique was applied to predict the groundwater table and validated by measured wells. The mean difference between measured and GPR predicted groundwater table was 14 cm, which reflected the importance of GPR to guide the location of new wells in the study region. The outcomes of the study will help the decision-makers, government agencies, and private sectors for sustainable planning of groundwater in the area. Overall, the present study provides a comprehensive high-precision machine learning and GPR-based groundwater potential mapping.

ARTICLE HISTORY
Received 2 December 2019
Accepted 6 July 2020

KEYWORDS
Groundwater potential; GIS; machine learning; ensemble model; GPR

1. Introduction

Groundwater is the most valuable resource on our planet. It reflects the socio-economic condition and development of an area (Naghibi, Pourghasemi, and Dixon 2016). Since the last century, groundwater has been in high demand for domestic, agricultural, and industrial purposes worldwide (Mogaji, Lim, and Abdullah 2014; Chen et al. 2019). Extensive groundwater extraction has led to a continuous drop of water table (Naghibi, Pourghasemi, and Dixon 2016; Das 2019). Therefore, Groundwater management is necessary for sustainable use of water resources. Groundwater availability and its movement depend on topographical, hydrological, ecological, geological, and atmospheric factors (Oh et al. 2011; Golkarian et al. 2018). As groundwater is a hidden natural resource, the demarcation of groundwater potential zones is essential for planning, management, and sustainable development of an area.

Many studies have been done by researchers on GWPM using different methods. Earlier groundwater mapping was based on field surveys, which was more expensive and time-consuming (Ganapuram et al. 2009; Chen et al. 2018; Das 2019). In the present time, remote sensing (RS), geographic information system (GIS), statistical, and geophysical techniques have been applied to map the groundwater potentiality of a large area in time and cost-effective manner.

In earlier studies, GIS modeling is very successfully applied to identify the groundwater prospect region with a high prediction rate (Das 2019; Chen et al. 2019). The combined use of RS and GIS techniques has been employed in different research for GWPM (Prasad et al. 2008; Oh et al. 2011; Magesh, Chandrasekar, and Soundranayagam 2012; Naghibi, Pourghasemi, and Dixon 2016; Murasingh, Jha, and Adamala 2018; Lee, Hong, and Jung 2018; Golkarian et al. 2018; Chen et al.

CONTACT Pankaj Prasad ppankaj@nio.org
Supplemental data for this article can be accessed here.

2018; Das 2019). Besides, several statistical techniques have been employed along with RS and GIS techniques for GWPM such as frequency ratio (Guru et al., 2017; Oh et al. 2011; Pourtaghi and Pourghasemi 2014; Naghibi et al. 2015; Das 2019), weights of evidence (Lee, Kim, and Oh 2012; Tahmassebipoor et al. 2016; Chen et al. 2018), and logistic regression (Nampak, Pradhan, and Manap 2014; Chen et al. 2018).

In this decade, approaches through machine learning models have continuously been increased for groundwater mapping, such as random forest (Naghibi and Pourghasemi 2015; Rahmati, Pourghasemi, and Melesse 2016; Naghibi, Pourghasemi, and Dixon 2016; Golkarian et al. 2018; Naghibi et al. 2019; Arabameri et al. 2019), boosted regression tree (Naghibi et al. 2015; Naghibi, Pourghasemi, and Dixon 2016; Naghibi et al. 2019; Kordestani et al. 2019), support vector machine (Naghibi, Ahmadi, and Daneshi 2017b; Lee, Hong, and Jung 2018; Naghibi, Pourghasemi, and Abbaspour 2018), C5.0 (Duan et al. 2016; Golkarian et al. 2018), classification and regression tree (Naghibi et al. 2015; Naghibi, Pourghasemi, and Dixon 2016), and artificial neural network (Lee, Hong, and Jung 2018). Such models are also applied in numerous fields, namely landslide susceptibility mapping (Pourghasemi, Moradi, and Aghda 2013; Youssef et al. 2016; Kim et al. 2018), gully erosion susceptibility mapping (Arabameri et al. 2019; Gayen et al. 2019), ecological study (Lek and Guegan 1999; Recknagel 2001; Elith, Leathwick, and Hastie 2008; Crisci, Ghattas, and Perera 2012), flood susceptibility mapping (Tehrany, Pradhan, and Jebur 2014; Tehrany et al. 2015; Khosravi et al. 2018; Shafizadeh-Moghadam et al. 2018), and land use land cover change detection (Friedl, Brodley, and Strahler 1999; Gislason, Benediktsson, and Sveinsson 2006; Rodriguez-Galiano and Chica-Rivas 2014). The prediction rate of machine learning models is very high compared to statistical models as reported in different studies (Naghibi and Pourghasemi 2015; Chen et al. 2018).

The aforementioned studies have used the advantages of a single model. However, the ensemble model is a combination of statistical and machine learning techniques (Naghibi et al. 2017a; Kordestani et al. 2019). Recently, Naghibi et al. (2017a), Kordestani et al. (2019), and Naghibi et al. (2019) have used ensemble models for GWPM with satisfactory accuracy. The present work was applied the ensemble model of RF and SVM for GWPM.

In the previous studies, the researchers used different models, but their applicability was restricted to a specific study region. The objective of the present research is to map the groundwater prospect zones along the Sindhudurg coastal stretch using machine learning techniques and evaluate the results in different regions for the validity of the models. Besides, GPR technology was introduced to identify the groundwater table. Groundwater potential maps of the study region can be helpful for better planning and management of groundwater resources.

2. Study area

Sindhudurg coast stretches from 15°43′11.43″N to 16° 33′45.63″N latitude and 73°18′36.53″E to 73°55′50.07″E longitude covering an area around 3177 sq km along the west coast of India. The study region is bounded in the north and south by rivers and in the east and west by Western Ghats and shoreline of the west coast of India (Figure 1a). With these natural boundaries, the area under study represents a typical coastal environment as it is a transition zone between land and sea.

The geology of the study area shows formation from the Archean to the Recent age and the lithology of which mainly consists of granite, basalt, lateritic, and alluvial deposits. The Deccan basalts occupy more than 30% of the study area and the aquifers are mostly associated with fractures and joints. However, the aquifers associated with laterites are substantiated with porous nature. The elevation of the study area varies from 0 to 450 m above mean sea level from seashore to landward dissected hill ranges. From the geomorphological perspective, the study area is reclassified into six distinct regions viz., denudational origin-pediment pediplain complex, coastal origin-younger coastal plain, structural origin medium dissected plateau, denudational origin moderately dissected plateau, structural origin-low dissected plateau, and others. Karli and Gad, the main river networks of the study area, originate from the Western Ghats and debouch into the Arabian Sea. Climatologically, the research area is sub-tropical with a minimum temperature of 15°C to a maximum temperature of 40°C having three seasons (rainy, winter, and summer) throughout a year. The annual average rainfall of the area is around 3000 mm, and the maximum rainfall occurs during the southwest monsoon season (June–September). Monsoon rainfall profoundly influences the groundwater in different parts of the study region. According to the groundwater surveys and development agency and central groundwater board reports, the groundwater level ranges between 0.20 and 21 m/bgl. During the field visit, it was observed that groundwater is the primary source for drinking and irrigation extracted through dug and pumping well. Hence, the groundwater management is a vital concern for the study region.

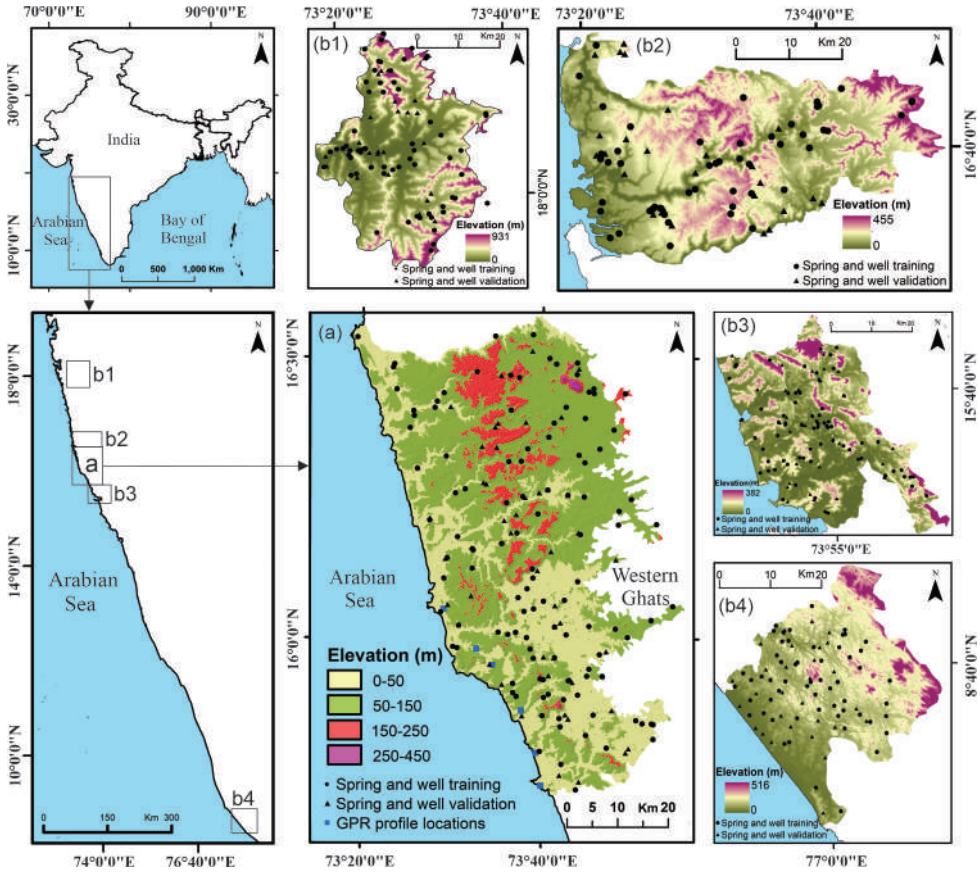


Figure 1. Location maps of the study area (a), and secondary regions (b1, b2, b3, b4).

Apart from this primary study region, four secondary regions were selected, shown in Figure 1 (b1, b2, b3, and b4), for further assessment of the applied models. The geographical details of these areas are given in Table 1.

3. Materials and methods

The methodology applied in the current research is presented in Figure 2, which involves the following stages: Firstly, different thematic layers and inventory map of spring and well were prepared and transformed into the spatial database. Secondly, the groundwater potential maps were produced using machine learning models. Finally, the accuracy of the models was examined by applying the AUROC curve.

3.1. Spring and well inventory map

Many researchers have used the locations of spring, well, and quant as inventory for groundwater potential mapping. In the present research, both spring and well points were considered for GWPM. The inventory map of the study region contains 165 spring and well points, identified from numerous sources (Table 2), and field observation. Random partition algorithm was deployed to separate the spring and well points for training and validation purposes, where 116 (70%) points were

Table 1. Geographical description of selected sites for cross-validation.

	Areal extension			
Regions	Latitude	Longitude	No. of springs and wells	Area (sq. km)
South-east Raigad (b1)	17°50′57.75″– 18°19′4.39″	73°17′54.76″– 73°40′9.03″	87	1077
South Ratnagiri (b2)	16°29′58.34″– 16°49′44.78″	73°18′0.97″– 73°50′57.34″	76	1221
North Goa (b3)	15°24′22.51″– 15°47′52.15″	73°41′19.26″– 74°7′29.49″	99	933
South Kerala (b4)	8°23′21.62″–8°50′13.49″	76°48′3.46″–77°11′53.69″	87	1036

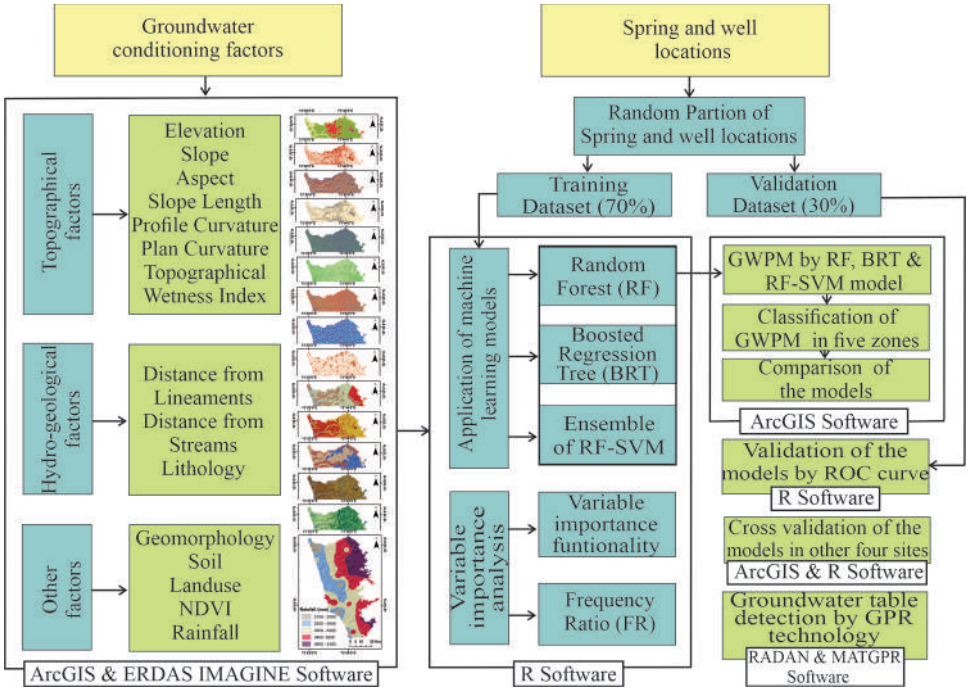


Figure 2. Flowchart of the present study for groundwater potential mapping.

preferred for training and the remaining 49 (30%) for validation of dataset. In the same way, the inventory maps of other regions were prepared.

3.2. Groundwater conditioning factors

It is essential to select the effective parameters for preparing the groundwater prospect map of an area. Based on the previous studies (Naghibi, Pourghasemi, and Dixon 2016; Rahmati, Pourghasemi, and Melesse 2016; Golkarian et al. 2018; Chen et al. 2018) and field examination, 15 groundwater conditioning factors (thematic maps) viz. elevation, slope, aspect, LS, profile curvature, plan curvature, TWI, distance from the streams, distance from the lineaments, lithology, geomorphology, soil, land use, NDVI, and rainfall (Figure 3a–o) were taken into account for GWPM in the study area. These thematic maps were prepared using the ArcGIS 10 software from several data (Table 2). Each thematic layer was resampled into a uniform grid size of 30 × 30 m, and the grid of the research area was prepared by 2201 columns and 3159 rows (3,530,426 pixels; 3177 km^2). Similarly, thematic maps of the selected secondary regions were generated.

SRTM (30 m resolution) digital elevation model (DEM) was used to produce the following topographic factors. Elevation, one of the potential indicators of groundwater, plays a vital role for GWPM (Oh et al. 2011; Naghibi, Pourghasemi, and Dixon 2016; Patra, Mishra, and Mahapatra 2018; Chen et al. 2019). The elevation map was created from the DEM. The slope is considered as the most relevant topographic variable for groundwater potentiality (Naghibi and Pourghasemi 2015; Lee, Hong, and Jung 2018). The slope map was produced in the ArcGIS environment, and the slope values were grouped into six classes using the natural break method. Aspect is one of the important controlling factors for the GWPM. It defines the direction of the slope, which is exposed to sunlight, winds, lineament, and rainfall (Goudie 2013; Chen et al. 2018). The aspect map was used to correlate the groundwater availability at the different directions of the slope. Slope length (LS) defines the length (L) and steepness (S) of the topography that influences the amount of groundwater storage. LS is calculated with the following equation (Moore and Burch 1986).

$$LS = (fa \times cellsize/22.13)^{0.4} \times (sin\theta/0.0896)^{1.3} \quad (1)$$

where *fa* refers to flow accumulation and θ represents the slope in degrees.

Curvature affects the surface and subsurface hydrology (Regmi et al. 2015). Profile curvature is parallel to the maximum slope in a particular direction. The negative value of profile curvature indicates the water flow decelerated in the surface, whereas the positive value indicates the water flow accelerated on the surface, and zero indicates the surface is linear. In another side, plan curvature defines the maximum slope in a perpendicular

Table 2. Details of database used in groundwater potential mapping.

Data layers	Source of data	Scale	Time period
Spring and well locations	Topographical Maps http://www.surveyofindia.gov.in/	1:50,000 and 1:25,000	1967
	Groundwater Surveys and Development Agency https://gsda.maharashtra.gov.in		1990–2018
	Central Groundwater Board http://cgwb.gov.in/		1990–2018
	Field Survey		
Groundwater level	Groundwater Surveys and Development Agency	–	2007–2016
	Central Groundwater Board		
Rainfall	Department of Agriculture Maharashtra state http://maharain.gov.in/	–	2013–2018
	India Meteorological Department https://mausam.imd.gov.in/		
Geology	Geological Survey of India https://www.gsi.gov.in	1:250,000	2001
Geomorphology	National Remote Sensing Centre https://bhuvan.nrsc.gov.in	1:50,000	2005–2006
	SRTM		
	Landsat8 OLI		
	Field Survey		
Soil map	National Bureau of Soil Survey and Land Use Planning https://www.nbsslup.in/	1:500,000	1996
Digital elevation model (DEM)	SRTM https://earthexplorer.usgs.gov/	1 arc second,	23 September 2014
Satellite image	Landsat8 OLI https://earthexplorer.usgs.gov/	30 m spatial resolution	19 October 2016, 17 December 2016.

direction. It describes the convergence and divergence of water flow in the earth's surface. Negative values represent the concave slope of the surface, which causes the confluence of water flow. In contrast, positive values indicate the convex slope of the surface that determines the divergence of water flow in the region (ESRI 2016). TWI expresses the effect of topography on the location, which is related to soil conditions of the area. TWI is calculated as follows (Moore, Grayson, and Ladson 1991).

$$TWI = In(fa/tan\beta) \quad (2)$$

Here fa is the flow accumulation, and β is the slope angle at the point.

Distance from the stream is inversely related to the identification of groundwater prospects in an area. Lineaments are hydro-geologically very meaningful to control the groundwater movement and storage (Magesh, Chandrasekar, and Soundranayagam 2012). It is in a linear or curvilinear pattern on the earth's surface and identified from the satellite imagery, DEM, and field survey (Pradhan and Lee 2010; Magesh, Chandrasekar, and Soundranayagam 2012; Goudie 2013; Rahmati et al. 2015). For this study, lineaments were extracted from the superimposed shaded relief maps at an interval of 45° azimuth angle. High lineament density indicates more groundwater potentiality in the area (Magesh, Chandrasekar, and Soundranayagam 2012). Euclidean distance method was adopted to examine the relationship between inventory and distance from streams and lineaments. In the study area, the age of the rock traced from the Archean to Recent, which controls the groundwater storage. Based on lithofacies and geological ages, the lithology of the study area was reclassified into six broad classes: Group 1: Archean schist and gneisses (Granite gneiss, quartzite, meta-gabbro, amphibole schist) (Ask), Group 2: Dharwar supergroup (Meta graywacke, metabasalt, granite) (Dsg), Group 3: Kalladgi supergroup (Sedimentary quartzite, shale) (Ksg), Group 4: Sahyadri supergroup (Unclassified flows, Aa flow, Mega cryst flow) (Ssg), Group 5: Laterite (Lat), Group 6: Alluvium (Alv). Geomorphology is another predisposing factor for predicting the potential of groundwater. In this study, the geomorphic unit was categorized into six major groups: 1. Denudational origin-pediment pediplain complex (DoPPc), 2. Coastal origin-younger coastal plain (CoYCp), 3. Structural origin moderately dissected plateau (SoMDp), 4. Denudational origin moderately dissected plateau (DoMDp), 5. Structural origin-low dissected plateau (SoLDp), 6. Others (Oth), using the topographical maps, and Landsat 8 OLI image. Soil is considered one of the most important indicators of the surface and sub-surface runoff, recharge, and infiltration processes (Mogaji, Lim, and Abdullah 2014; Rahmati, Pourghasemi, and Melesse 2016). Soil map from the National Bureau of Soil Survey and Land use planning (NBSSLUP) was used to reclassify the soil into four categories: Ultic Typic Haplustalfs, Lithic Ustorthents, Ultic Haplustalfs, and Typic Ustropepts. Land use and land-use changes influence the groundwater storage and aquifer yield (Ibrahim-Bathis and Ahmed 2016; Guru,

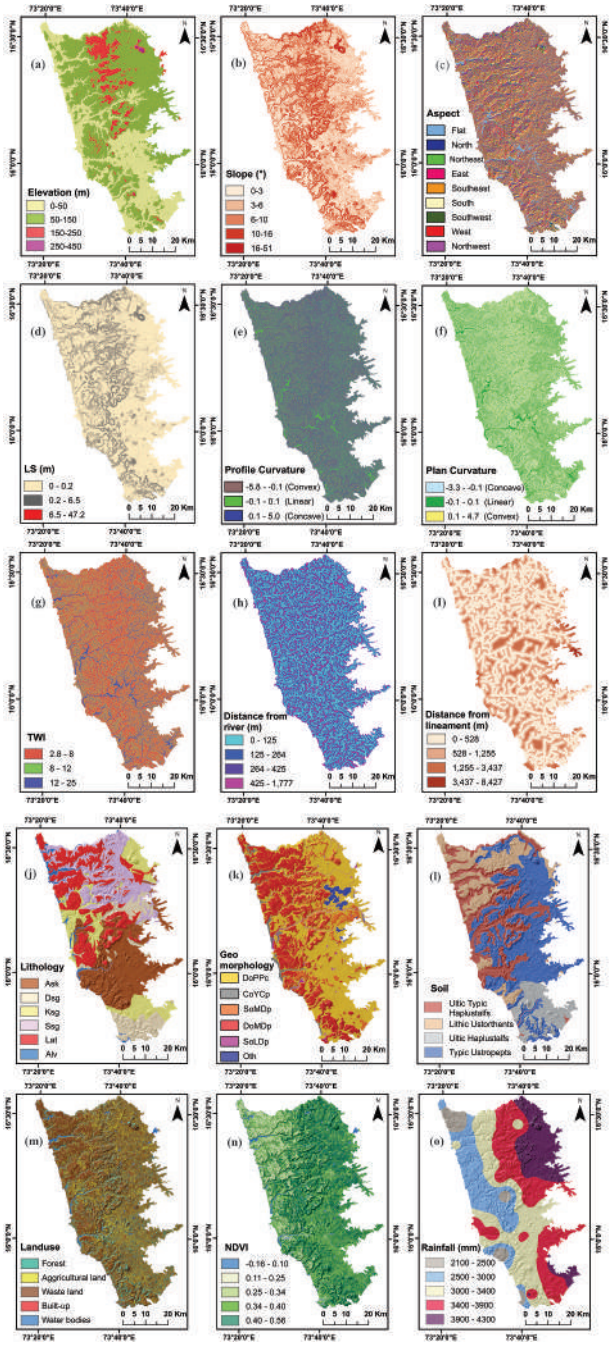


Figure 3. Thematic layers of the groundwater affecting factors (a) elevation, (b) slope, (c) aspect, (d) LS, (e) profile curvature, (f) plan curvature, (g) TWI, (h) distance from rivers, (i) distance from lineaments, (j) lithology, (k) geomorphology, (l) soil, (m) land use, (n) NDVI, (o) rainfall.

Seshan, and Bera 2017; Chen et al. 2018). Unsupervised classification technique was employed to produce the land use map from the Landsat 8 OLI image. The study area was classified into five major land use classes: forest, agricultural land, wasteland, built-up, and waterbodies. The accuracy of the land use classification was calculated as 86% using the Kappa index. NDVI provides the health of vegetation, and it is usually used to relate the vegetation density and groundwater potentiality (Pourghasemi, Moradi, and Aghda 2013; Chen et al. 2018). The range of the NDVI value lying from −1 to 1, with higher NDVI value indicates the healthy vegetation and vice versa. Based on the Landsat 8 OLI image, the NDVI map was created using the following equation:

$$NDVI = (NIR - R)/(NIR + R) \quad (3)$$

Here near-infrared (NIR) and red (R) bands represent the spectral reflectance measurement of these bands.

The distribution, duration, and intensity of rainfall are significant affecting factors for infiltration, runoff, and recharge conditions (Magesh, Chandrasekar, and Soundranayagam 2012). The rainfall map was prepared from the rain gauge data of Maharashtra government using the inverse distance weighted method and classified into five groups.

3.3 Methods

In this research, three machine learning models (RF, BRT, and ensemble of RF-SVM) were employed for GWPM. The relationship between different groundwater conditioning factors and inventory locations was calculated by the frequency ratio method. On the other side, the importance of these groundwater effective factors was measured by "variable importance" function in R software. The raster values of 15 factors of each spring and well location were imported to R software; then, the models were applied using different packages in the R environment. To improve the classification accuracy and avoid the biasness, a 10-fold cross-validation method (with five repetitions) was used in the models. The final output values of the models were transformed into a spatial dataset for GWPM using the ArcGIS software. At last, models were validated by the AUROC curve and also examined in different regions of the west coast of India. In addition, GPR technology was used to measure the groundwater table by RADAN and MATGPR software.

3.3.1 Application of frequency ratio (FR) and variable importance function

FR is a bivariate statistical technique to explain the prospect of occurrence of a certain attribute (Bonham-Carter 1994; Oh et al. 2011; Manap et al. 2014; Naghibi, Pourghasemi, and Dixon 2016; Guru, Seshan, and Bera 2017). It is defined by the relationships between dependent variables (spring and well location) and independent variables (groundwater conditioning factors) (Guru, Seshan, and Bera 2017; Das 2019). In this context, the FR model was used to show the quantitative relationship

between the inventory and each sub-class of groundwater variables. FR is calculated as;

$$FR = (P_s/T_s)/(P_a/T_a) \quad (4)$$

where P_s is the number of spring and well under each sub-class of the groundwater effective factors, T_s denotes the total spring and well of the study area, P_a is the number of pixels of each sub-class of the conditioning parameter, and T_a is the total pixels of the study area. FR is the ratio of spring and well occurrences to the total area of each class of the affecting factors. So the FR value 1 indicates the average of the model. If the value is more than 1, it considers the high groundwater prospect and less than 1 value indicates the low groundwater potential in each sub-class of the parameters (Lee and Pradhan 2007; Pradhan and Lee 2010; Oh et al. 2011; Manap et al. 2014; Nampak, Pradhan, and Manap 2014).

"Variable importance" function of the RF model was used for measuring the effectiveness of each factor. It is a generic method for calculating the feature importance by trained methods (Kuhn et al. 2018). Each factor was evaluated individually using a filter approach. The importance values of the features range from 0 to 100. Higher the importance value, greater is the influence of the factor, and vice versa. The details of variable importance function are given in Kuhn et al. (2018).

3.3.2 Application of random forest (RF)

Random forest is an ensemble machine learning technique for both classification and regression tasks (Breiman 2001; Youssef et al. 2016; Naghibi, Ahmadi, and Daneshi 2017b; Kim et al. 2018). For classification, RF uses the resampling technique by randomly changing the predictive variables to increase the diversity in each tree (Youssef et al. 2016; Naghibi, Ahmadi, and Daneshi 2017b). This method consists of multiple decision trees and merges them to explain the spatial relationship between controlling variables of groundwater and inventory of spring and well (Kim et al. 2018). The decision tree is generated by bootstrap samples and leaves few samples for validation to test the accuracy of the decision tree. The mean-squared error of each decision tree with their OOB samples (E_{oob}) is used to calculate the learning error. E_{oob} is expressed as:

$$E_{OOB} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \Delta_i)^2 \quad (5)$$

Here n denotes the total number of OOB samples; y_i is the observed output, and $\hat{y}_i$ is the model output. The main advantages of this approach are; i) it can handle large datasets with high dimensionality as well as avoid the over-fitting of the datasets; ii) this does not need any assumption regarding the explanatory variables and response variables, and iii) it does not require any prior data to transformation and rescaling.

3.3.3 Application of boosted regression tree (BRT)

BRT model is advanced and different from classical regression methods. It uses statistical and machine learning techniques to enhance the performance of a single model by fitting many models and combining them for prediction (Schapire 2003; Elith, Leathwick, and Hastie 2008; Naghibi, Pourghasemi, and Dixon 2016). It is a combination of two algorithms, namely boosting and regression tree (decision tree) (Elith, Leathwick, and Hastie 2008; Youssef et al. 2016; Kim et al. 2018). In the BRT model, the initial decision tree (DT) reduces the loss function. At each iteration, the main target was to decrease the root-mean-square error and the residuals. Then, the next DT is fit for prediction residuals of the first tree. In this stagewise process, the existing trees are unchanged as the model develops increasingly larger. At each step, the fitted value of each observation is reestimated to express the contribution of the recently added tree (Elith, Leathwick, and Hastie 2008; Naghibi, Pourghasemi, and Dixon 2016). DT was used for visualization and explicit decision-making. The advantages of the algorithm are that the predictive variable can be of any type (numeric, binary, and categorical, etc.), and the outcomes of the model are not affected by monotone transformations and different scales of measurement among predictors. It replaces the missing data in predictor variables using surrogates (Breiman 2001; Elith, Leathwick, and Hastie 2008; Youssef et al. 2016). Boosting is a technique to get higher accuracy from the predictive variables of regression trees. It is a sequential procedure to average many rough rules of thumb (Schapire 2003; Elith, Leathwick, and Hastie 2008).

By combining the algorithms, the BRT model establishes a binary tree with general classification and regression tree. The classified data split into two samples. Each sample defines the best point for the data partition, and it formulates the observed deviation and residuals at each partition (Kim et al. 2018). Finally, the model is capable of estimating the observed value.

In the BRT model, three parameters such as number of trees, shrinkage or learning rate, and interaction depth are required for tuning. Interaction depth defines the number of nodes in trees and the learning rate determines the importance of each tree in the built model. Based on these two parameters, the number of trees was decided for optimal prediction (Elith, Leathwick, and Hastie 2008; Naghibi, Pourghasemi, and Dixon 2016).

3.3.4 Application of ensemble of RF and support vector machine (SVM)

SVM model is another popular supervised machine learning technique that is based on the concept of structural risk minimization and statistical learning theory (Tehrany, Pradhan, and Jebur 2014; Mojaddadi et al. 2017; Naghibi, Pourghasemi, and Abbaspour 2018). The principle of the method is to separate the hyperplane formation from the dataset. The hyperplane is defined as the center of the maximum margin of separation (Marjanovic et al. 2011; Tehrany et al. 2015). On the basis of the hyperplane, the point was classified as +1 or −1. For the case of linear separable data, a separating hyperplane can be computed with the help of the following equation (Hong et al. 2017):

$$y_i(w.x_i + b) \geq 1 - \zeta_i \tag{6}$$

where w represents the coefficient vector which expresses the orientation of the hyperplane in the feature space, b is the offset of the hyperplane from the origin, and ζ_i defines the positive slack variables. The following optimization problem can be solved by defining an optimal hyperplane (Samui 2008).

$$Minimize \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j y_i a y_j (x_i x_j) \tag{7}$$

$$Subject\ to \sum_{i=1}^{n} a_i y_i = 0, 0 \leq a_i \leq c \tag{8}$$

Here a_i indicates the lag range multiplier and C denotes the penalty. The precision of the successful classification in SVM model depends on the selection of kernel type (Yao, Tham, and Dai 2008). In the present research, radial basis function (RBF) was applied because of its higher capability in interpolation, as reported by many researchers (Tehrany, Pradhan, and Jebur 2014; Tehrany et al. 2015; Gayen et al. 2019). RBF equation is calculated as follows:

$$RBF : K(x_i, x_j) = \exp\left(-yx_i - x_j^2\right) \tag{9}$$

where K (x_i,x_j) is the kernel function, y represents the RBF kernel function. The purpose of applying the algorithm was to reduce the error and model complexity (Naghibi, Pourghasemi, and Abbaspour 2018).

The ensemble model was developed by combining two or more than two different predictive models. In recent studies (Naghibi et al. 2019; Kordestani et al. 2019; Chen et al. 2019), the ensemble model was applied to improve the results. In this work, RF and SVM models were ensembled in R software using the weighted average method. For the implementation of these machine learning models, R statistical software 3.6.1 version was used with the help of different packages.

3.4 Validation of groundwater potential maps

Validation is a fundamental step in modeling for the scientific significance of the research (Naghibi, Pourghasemi, and Dixon 2016; Chen et al. 2019). In this research, the area under the receiver operating characteristic (ROC) curve was opted for evaluation of the models. ROC is a graphical plot, which determines the performance of the models in a diagnostic test (Egan 1975; Golkarian et al. 2018). The curve plots the true-positive rate (sensitivity) on Y-axis and false-positive rate (1 – specificity) on X-axis (Youssef et al. 2016; Golkarian et al. 2018). Model prediction for occurrence and non-occurrence of springs and wells was evaluated using the area under the ROC curve. The area under the curve (AUC) represents the value between 0 and 1, and the higher value represents the better performance of the model (Youssef et al. 2016; Naghibi, Pourghasemi, and Dixon 2016; Golkarian et al. 2018; Chen et al. 2018, 2019). To extract the generalized findings, it is necessary to apply the models in different regions. For this purpose, four different areas along the west coast of India were chosen (b1, b2, b3, and b4). The b1 and b4 regions show slightly different characteristics from the study area in terms of topography, lithology, and climatic conditions whereas b2 and b3 sites lying in close proximity to the study area appear to be similar with respect to aforesaid criteria. Additionally, six GPR profiles were selected to examine the groundwater level with the corresponding water level of wells. GPR is a noninvasive geophysical technique based on propagation and reflection of the transmitted electromagnetic waves (Annan 2003; Neal 2004; Billy et al. 2014). A SIR-4000 model of GSSI (Geophysical Survey System Inc.) with 200 MHz antenna was used to identify the water table depth. GPR data were processed in Radan 7 and MatGPR 3.2 (compatible with Matlab software) software by applying the time zero removal, filtering, background removal, and migration.

4 Results

4.1 Spatial relationship between groundwater conditioning factors and inventory of spring and well

It is necessary to know the relationship between inventory with effective factors for GWPM in an area. The results of the above relation are shown in Table 3 using the FR model. The FR value ranges from 0 to 1.78

Table 3. Spatial relationship between each groundwater effective factor and inventory of spring and well using frequency ratio (FR) model.

Parameters	Classes	% of total area (a)	% of inventory area (b)	Frequency ratio (b/a)
Elevation (m)	0–50	34.29	52.12	1.49
	50–150	56.36	43.64	0.78
	150–250	8.87	4.24	0.48
	250–450	0.48	0.00	0.00
Slope (degree)	0–3	31.78	30.30	0.95
	3–6	30.86	36.36	1.18
	6–10	18.01	18.79	1.04
	10–16	10.96	7.88	0.72
	16–51	8.39	7.27	0.87
Aspect	Flat	4.82	2.42	0.51
	North	11.04	7.88	0.71
	Northeast	10.71	12.73	1.19
	East	10.70	10.30	0.96
	Southeast	11.59	12.12	1.05
	South	12.37	9.09	0.73
	Southwest	13.22	14.55	1.10
	West	13.13	15.76	1.20
	Northwest	12.42	15.15	1.22
LS (m)	0–0.2	88.80	84.24	1.01
	0.2–6.5	16.17	15.76	0.97
	6.5–47.2	0.02	0.00	0.00
Profile curvature	Convex	29.31	20.61	0.70
	Linear	35.09	50.30	1.43
	Concave	35.60	29.09	0.82
Plan curvature	Concave	24.56	16.36	0.67
	Linear	45.71	64.24	1.41
	Convex	29.73	19.39	0.65
TWI	2.8–8	63.26	55.15	0.87
	8–12	22.69	29.09	1.29
	12–25	14.05	15.76	1.10
Distance from river (m)	0–125	38.13	45.45	1.19
	125–264	30.30	28.48	0.94
	264–425	22.80	21.82	0.94
	425–1777	8.77	4.24	0.51
Distance from lineament (m)	0–528	52.52	61.82	1.18
	528–1255	37.79	33.94	0.90
	1255–3437	9.19	4.24	0.46
	3437–8427	0.49	0.00	0.00
Lithology	Ask	27.40	37.58	1.37
	Dsg	11.30	5.45	0.48
	Ksg	14.44	23.64	1.53
	Ssg	20.15	18.79	0.93
	Lat	23.38	13.33	0.57
	Alv	3.33	1.21	0.53
Geomorphology	DoPPc	49.76	76.36	1.55
	CoYCp	2.25	3.03	1.35
	SoMDp	18.65	13.33	0.71
	DoMDp	25.96	5.45	0.21
	SoLDp	2.03	1.82	0.89
	Oth	1.35	0.00	0.00
Soil	Ultic Typic Haplustalfs	25.27	33.94	1.34
	Lithic Ustorthents	18.15	10.91	0.60
	Ultic Haplustalfs	10.92	19.39	1.78
	Typic Ustropepts	45.64	35.76	0.78
Land use	Forest	15.27	18.79	1.23
	Agricultural Land	26.23	39.39	1.50
	Wasteland	55.00	38.79	0.71
	Built-up	1.98	2.42	1.22
	Waterbodies	1.90	0.61	0.32
NDVI	−0.16–0.10	1.84	1.21	0.67
	0.10–0.25	9.06	1.82	0.20
	0.25–0.34	16.10	27.27	1.69
	0.34–0.40	34.23	40.00	1.17
	0.40–0.56	38.77	29.70	0.77
Rainfall(mm)	2100–2500	4.94	45.45	0.99
	2500–3000	19.98	27.88	0.73
	3000–3400	29.14	22.42	1.08
	3400–3900	28.53	4.24	0.98
	3900–4300	17.41	45.45	1.22

in different sub-classes of the groundwater conditioning variables. For elevation, FR value decreases with the higher altitude. In the case of slope, the class from 3°-6° is highly correlated with spring and well occurrences (FR = 1.18). The northwest slope direction with FR value of 1.22 has more spring and well locations compared to other directions of the slope. FR values of LS classes are decreasing with the increasing slope length. In the matter of profile and plan curvature, flat areas have the highest FR value of 1.43 and 1.41, respectively. In TWI class, the second class (8–12) has the maximum FR value (1.29). The FR values increase with decreasing distance from streams, indicating a high groundwater prospect. In relation to lineament, most of the springs and wells occur at 0 to 1.25 km distance from the lineaments. Regarding lithology, spring and well locations have mostly identified in the Kaladgi group of rock with FR value of 1.53. For geomorphology, FR of the denudational origin-pediment pediplain complex has a maximum value (1.55). In the soil groups, Ultic Haplustalfs soil has a strong correlation with spring and well occurrences with FR value 1.78. In relation to land use, agricultural land is the maximum FR value 1.50. In the case of NDVI, the highest FR value is 1.69 in the third class (0.25–-0.34). In the rainfall class, the highest rainfall (3900–4300 mm) shows the maximum FR value (1.22).

4.2 Groundwater potential models

Groundwater potential maps were prepared using the three machine learning algorithms, namely RF, BRT, and the hybrid of RF-SVM models (Figure 4). Based on Min-max normalization, the probability values of the models were normalized between 0 and 1 and subsequently classified into five zones: very low (0–0.20), low (0.20–0.45), moderate (0.45–0.70), high (0.70–0.85), and very high (0.85–1) with same class range for comparison among the models. The higher value represents a very good groundwater prospect of the area and vice versa.

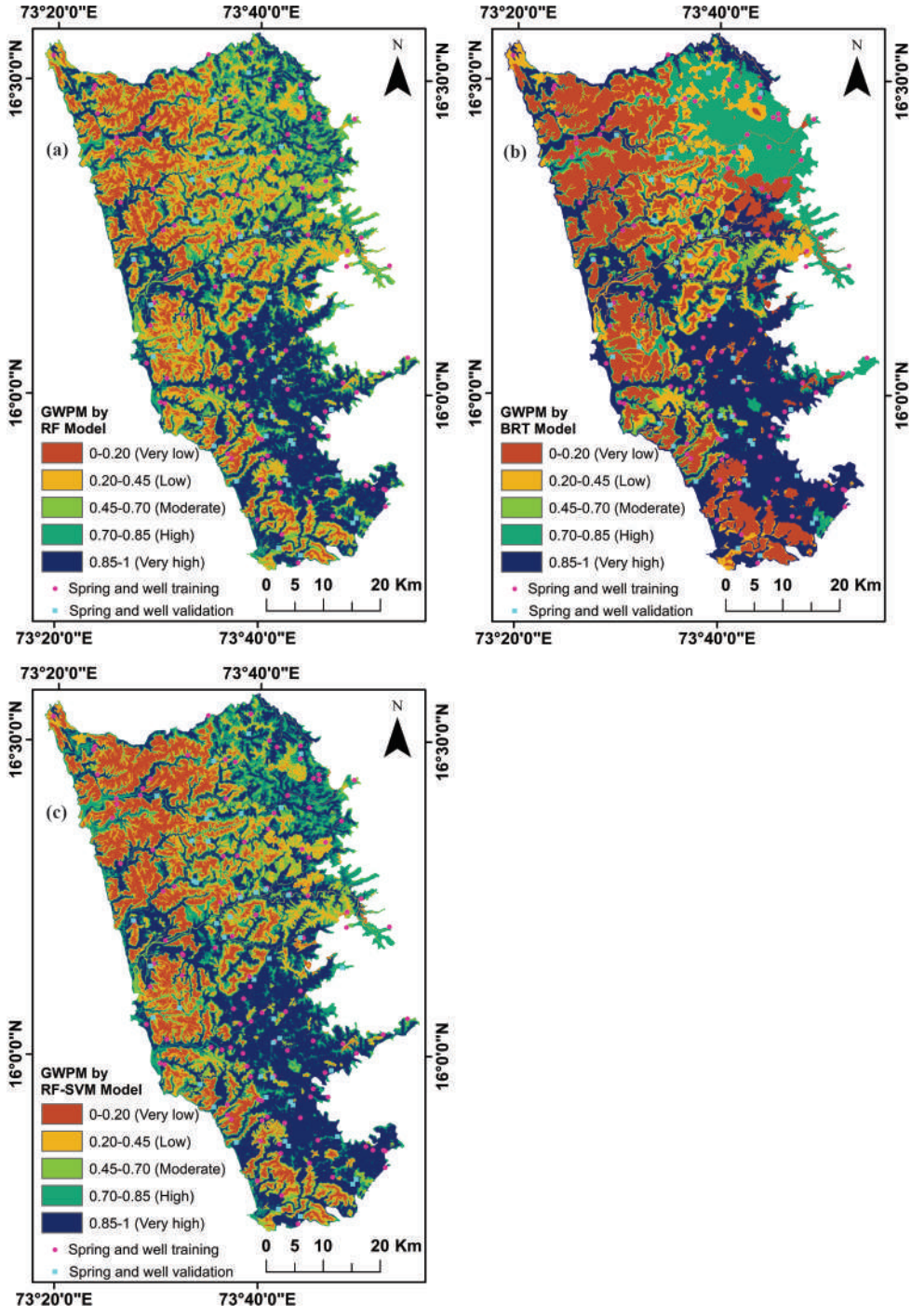


Figure 4. Groundwater potential maps derived from (a) RF, (b) BRT, and (c) RF-SVM models.

The model-wise spatial variation of the groundwater potentiality is shown in Figure 4 and Table 4. According to the RF model, the high and very high prospective groundwater zones covered 16.73% and 33.31% of the study area, respectively. Low and very low groundwater prospect zones comprise 17.98% and 12.48% of the total area, respectively. In the BRT model, it was found that 19.13% and 35.60% of the entire research area were of high and very high groundwater potentiality. Besides, low and very low groundwater potentiality covered 12.93% and 26.37% of the area. Based on the ensemble model result, the study region was classified as very high (36.85%), high (19.18%), moderate (13.14%), low (13.99), and very low (16.84%) groundwater prospect zones (Table 4).

Table 4. Area of different classes (%) of groundwater potential maps using RF, BRT, and RF-SVM ensemble models.

	Area in percentage		
Class	RF	BRT	RF-SVM
0–0.20 (very low)	12.48	26.37	16.84
0.20–0.45 (low)	17.98	12.93	13.99
0.45–0.70 (moderate)	19.50	5.97	13.14
0.70–0.85 (high)	16.73	19.13	19.18
0.85–1 (very high)	33.31	35.60	36.85
Total	100	100	100

4.3 Validation of machine learning models and GPR profiles

The validation of the model is crucial for the assessment of GWPM. In many studies, the ROC curve was used for quantitative validation of the models with a high prediction rate (Golkarian et al. 2018; Chen et al. 2018). In the present study, the validity of RF, BRT, and ensemble models was confirmed using the ROC curve. The AUC values ranging from lower (0) to higher (1) represent the worst to the best model prediction for GWPM. Based on the ROC result, the success rate of RF, hybrid of RF-SVM, and BRT models was measured as 94.0%, 93.4%, and 89.8%, respectively. The accuracy of the models from the selected sites was calculated. In the b1 area, the accuracy of RF, BRT, and RF-SVM was 82.2%, 72%, and 83.3%, respectively. The success rates of RF (94%, 96.5%), BRT (90.5%, 90.7%), and hybrid models (93.8%, 94.7%) were computed for the b2 and b3 regions, respectively. In the case of b4 region, the precision of RF, BRT, and RF-SVM was evaluated as 82.8%, 80.7%, and 82.5%, respectively.

The water table from the GPR profiles (a, b, c, e, and f) was identified at the depth of 3.6, 3.4, 4.9, 4.8, 4.3, and 4.1 m, respectively, and the nearest water level of wells was observed at 3.6, 3.4, 4.9, 5.1, 4.1, and 4.4 m, respectively (Table 5). Only in the case of profile "d," the water table could not be identified. From the five GPR profiles, the average and maximum difference between predicted and measured depths of groundwater were 14 and 30 cm, respectively. Overall, from six GPR profiles, five locations of water table were accurately determined, which means more than 80% accuracy achieved in detecting the groundwater table by GPR technology.

5 Discussion

The presence of spring and well at the particular segments of the study area indicates the potentiality of high groundwater yield (Oh et al. 2011; Naghibi et al. 2017a). However, to assess the groundwater prospect of the entire area, statistical and machine learning methods were used by many researchers with good results (Oh et al. 2011; Chen et al. 2019). The results of the present study are discussed as follows:

5.1 Important conditioning factors for GWPM

The comparative importance of the 15 influencing factors of groundwater was illustrated using "variable importance" function of RF model. In this context, geomorphology had the highest importance followed by elevation, NDVI, and distance from the stream, while soil was of lowest importance followed by slope length rainfall, and land use (Figure 5). Geomorphology is the most effective factor since the geomorphic features of different landforms of the study area control the groundwater potentiality to a maximum extent. Various landforms on the earth's surface are associated with a different kind of groundwater storage (Deepika, Avinash, and Jayappa 2013; Rajaveni, Brindha, and Elango 2017). Structural hill, residual hill, and linear ridge represent the low groundwater potential, whereas pediplain and valley fill have the high groundwater potential due to high infiltration and groundwater recharge (Rajaveni, Brindha, and Elango 2017; Berhanu and Hatiye 2020). Almost 50% of the study area is associated with a pediment-pediplain complex, which indicates the good groundwater potentiality. Elevation is another critical factor for GWPM, which was an agreement with the results of Naghibi and Pourghasemi (2015), Naghibi, Pourghasemi, and Dixon (2016), Rahmati, Pourghasemi, and Melesse (2016), Naghibi, Ahmadi, and Daneshi (2017b), and Naghibi, Pourghasemi, and Abbaspour (2018). The lower elevation of the study area has the highest potentiality of groundwater due to hydraulic gradient and presence of low water table. The NDVI, which was the highest contributing factor for GWPM, found in the study of

Table 5. Water table depths from GPR profiles and measured wells.

GPR profiles	Place name	Latitude (N)	Longitude (E)	Water level depth from the measured wells (m)	Water table depth from GPR profiles (m)	Deviation from well water level to GPR water table (m)
a	Dedoolwada	16.04812	73.47902	3.6	3.8	0.2
b	Betwa	15.98158	73.54611	3.4	3.4	0
c	Mopar	15.95280	73.57527	4.9	4.9	0
d	Daboli	15.87222	73.62916	5.1	Not detectable	-
e	Vengurla	15.86111	73.63194	4.1	4.3	0.2
f	Redi	15.73811	73.66579	4.4	4.1	0.3

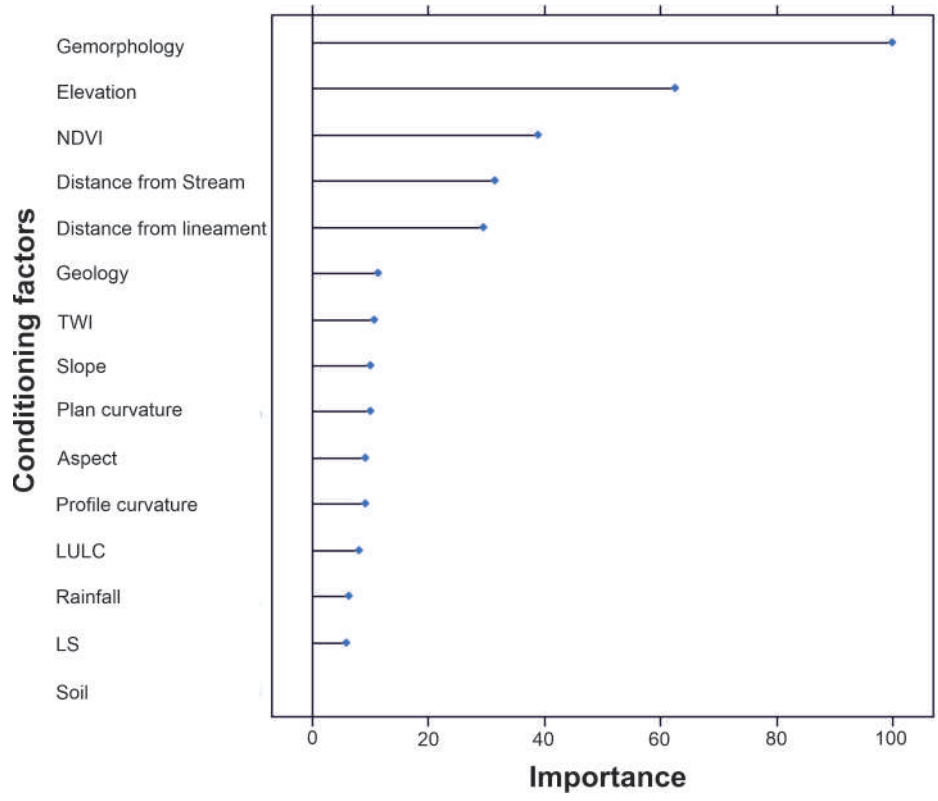


Figure 5. Variables importance in groundwater potential mapping.

Naghibi et al. (2017a) and Chen et al. (2019), is too incorporated in the present study. It is also considered as significant for land use classification, where forest class contains high FR value (1.23). Distance from the stream was a dominant factor for GWPM in the research of Kordestani et al. (2019), which reflects in the current study. The parallel drainage pattern indicates the presence of a fault system (Deffontaines and Chorowicz 1991) in the study area, which controls the groundwater movement and storage. The soil has the lowest controlling factor in GWPM because more than 30% of the area consists of hard rock, and which has negligible for primary porosity. Fractured rock, weathered basement, and depth of soil are favorable for the groundwater occurrence and movement (Prasad et al. 2008; Maiti et al. 2012; Das 2017). In the study area, the groundwater occurrence is mainly controlled by secondary porosity (weathered and fracture rocks) rather than the primary porosity of the soil.

On the other side, the FR model determines the importance of each sub-class of the groundwater affecting variables. The maximum values of the FR ratio were 1.78 (Ultic Haplustalfs), 1.69 (0.25–0.34), 1.55 (Denudational origin-pediment pediplain complex), and 1.53 (Kalladgi), for soil, NDVI, geomorphology, and geology factors, respectively. The importance of geomorphology and NDVI coincides with the result of the variable importance index. In the case of soil, the results of the variable importance index and FR were in contrast due to the consideration of a particular soil class that was conducive for high groundwater potentiality. The geological aspects (lithology, lineament) also influence the distribution and occurrence of groundwater (Berhanu and Hatiye 2020). The Kaladgi supergroup mainly consists of sedimentary rocks having more porosity in comparison to hard rocks. In the case of hard rock, groundwater occurrence is controlled by fault, boundaries between the different lithological units, weathered, and fracture zones. However, the properties of the study area and adopted methods have influenced the effective factors for GWPM.

5.2 Interpretation of the models and GPR profiles

In different studies (Naghibi, Ahmadi, and Daneshi 2017b; Golkarian et al. 2018) on GWPM, the RF model provides an excellent result. The results of the machine learning models from AUC exhibit that the RF is the best-fit model for the current study. The better performance of the RF model (AUC = 94%) may be because of the model consists of the multiple decision trees with no overfitting of the data. Besides, the model provides the interaction ability between effective factors and non-linearity (Catani et al. 2013). In recent years, ensemble

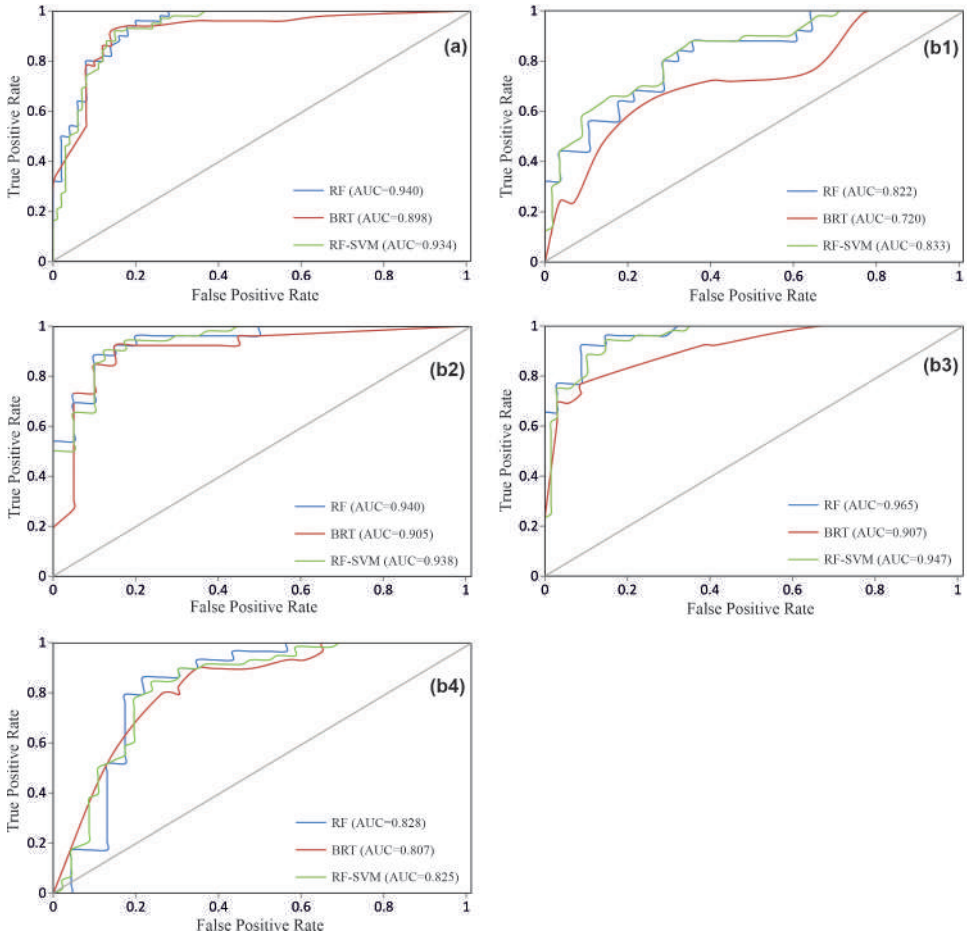


Figure 6. ROC curve of the models in study area (a), secondary areas (b1, b2, b3, b4).

models are increasingly used in groundwater potential mapping with very high accuracy (Chen et al. 2019; Kordestani et al. 2019; Naghibi et al. 2019). It was observed that the RF-SVM model had performed well with 93.40% accuracy from AUC (Figure 6). The advantages of RF and SVM models are the probable reason for high performance in the present work. The result of the BRT model has lesser accuracy compared to RF and ensemble models that may be due to the overfitting of the data. However, since the AUC values of the models are more than 0.7, the GWPM has been reliable for the study region (Naghibi, Pourghasemi, and Dixon 2016; Golkarian et al. 2018). The RF and RF-SVM models were successfully applied in the other selected parts of India. In the b1 and b4 areas, the RF and RF-SVM models performed better than the BRT model (Figure 6). On the other hand, the results from the b2 and b3 regions almost matched the result of the study area due to the homogenous hydro-geologic, topographic, and climatic properties. Moreover, the RF and hybrid models of RF-SVM are promising and sufficient to be advised as the method to prepare groundwater potential maps at the regional scale.

In many research works (Annan, Cosway, and Redman 1991; Nakashima, Zhou, and Sato 2001; Bano 2006; Mahmoudzadeh et al. 2012; Manu and Preko 2014), GPR was successfully used to identify the groundwater table for the better understanding of groundwater condition of an area. In this context, GPR technology was used to detect the groundwater table and advise the location for a new well in the research area. The strong radar reflection and amplitude variation from the groundwater table suggests the different dielectric contrast of the earth materials (Shih et al. 1986; Doolittle et al. 2006; Manu and Preko 2014). The results from GPR profiles revealed that radar reflection, amplitude variation, and high attenuation have prominent signatures of the water table. Groundwater table of the profile a, b, c, e, and f is precisely matched with the nearest water level data of the wells (Table 5). The water table from the GPR profile "d" was not identified due to the presence of different layers interpreted from the multiple radar reflections (Figure 7).

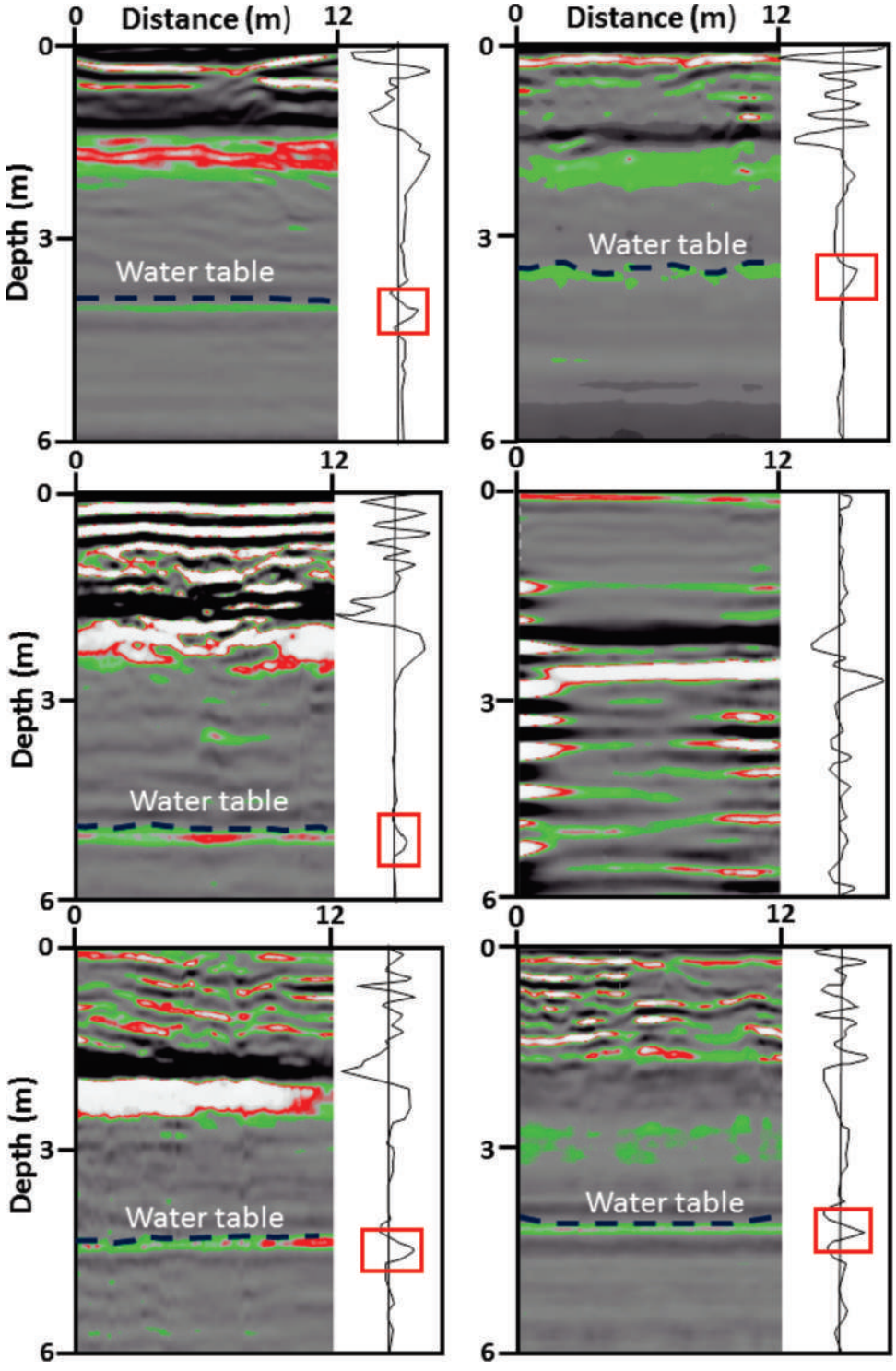


Figure 7. Water table depth from GPR profiles (a) 3.8 m (b) 3.4 m (c) 4.9 m (d) not detectable (e) 4.3 m and (f) 4.1 m.

5.3 Precision of the GWPM

A better model defines that the area of high and very high classes of the models is precisely matched with minimal variation (Naghibi, Ahmadi, and Daneshi 2017b). In this research, the outcomes of the five classes of groundwater potential area are consistent with the lowest percentage of variation for all the models. The validation results from the ROC suggest high preciseness of this model for GWPM. The cross-validation of the RF and RF-SVM models in all the selected sites has ensured the applicability of the

models with good precision. For water table detection, the GPR has produced an excellent result which in turn validates the groundwater mapping in the area of interest.

6. Conclusion

The increasing demand for groundwater made concern for groundwater potential mapping, especially on the west coast of India. In this research, machine learning algorithms have been applied to demarcate the groundwater potential zones in Sindhudurg coastal sector of the west coast of India. Based on literature review and field knowledge, 15 groundwater-related thematic layers were superimposed with inventory location in the GIS environment and integrated with RF, BRT, and ensemble of RF-SVM models. According to the results from RF, BRT, and RF-SVM, the very high groundwater potential zone occupies 33.31%, 35.60%, and, 36.85%, respectively, of the research area. The prediction of the models was validated with the AUROC curve. Based on AUROC curves, RF and RF-SVM models exhibited better performance than the BRT model for GWPM in the research area. Likewise, in the other four selected regions, RF and RF-SVM models proved to be superior in comparison to the BRT model. The most influencing factors of the groundwater prospect mapping were geomorphology, elevation, NDVI, distance from streams, and distance from lineament. The predicted depth of groundwater from GPR profiles and measured data during fieldwork on those wells are corroborating to each other, which helps to identify new potential well in the study region using GPR technology. The obtained results of the present study can be useful for government and private agencies in groundwater resource management, land use planning, and environmental protection in the study region. Furthermore, the methodology of this research can be adopted to study other coasts and watersheds with more or less similar hydro-geologic, topographic, and climatic properties.

Highlights

- RF and RF-SVM ensemble models performed very well with AUC>0.9.
- Evaluation of the models in four different regions with high-precision results.
- Geomorphology is the most important variable in groundwater potential mapping.
- GPR technique successfully measured the groundwater table.

Acknowledgements

We acknowledge the financial support from the Indian Rare Earth Limited (Project no. SSP-3232) and University Grant Commission (3160 NET-June 2015). The authors are thankful to the Director CSIR-NIO for support. The authors are also grateful to Prof. Jungho Im and anonymous reviewers for their critical comments and constructive suggestions to improve the manuscript. Field support from the survey team members KM Dubey, Lalit Arya, and Trilochan Kumar is thankfully acknowledged. The NIO contribution number is 6566.

Disclosure statement

The authors declare no conflict of interest.

Funding

This work was supported by the Indian Rare Earth Limited [SSP3232]; University Grants Commission [3160(NET-JUNE 2015)].

ORCID

Pankaj Prasad http://orcid.org/0000-0002-3118-2201

References

Annan, A. P. 2003. *Ground Penetrating Radar Principles, Procedures, and Applications*. Mississauga, ON, Canada: Sensors & Software .

Annan, A. P., S. W. Cosway, and J. D. Redman. 1991. "Water Table Detection with Ground-penetrating Radar." *Society of Exploration Geophysicists*. 494–496. doi:10.1190/1.1888793.

Arabameri, A., K. Rezaei, A. Cerda, L. Lombardo, and J. Rodrigo-Comino. 2019. "GIS-based Groundwater Potential Mapping in Shahroud Plain, Iran. A Comparison among Statistical (Bivariate and Multivariate), Data Mining and MCDM Approaches." *Science of the Total Environment*. 658:160–177. doi:10.1016/j.scitotenv.2018.12.115.

Bano, M. 2006. "Effects of the Transition Zone above a Water Table on the Reflection of GPR Waves." *Geophysical Research Letters*. 33. doi:10.1029/2006GL026158.

Berhanu, K. G., and S. D. Hatiye. 2020. "Identification of Groundwater Potential Zones Using Proxy Data: Case Study of Megech Watershed, Ethiopia." *Journal of Hydrology: Regional Studies*. 28:100676. doi:10.1016/j.ejrh.2020.100676.

Billy, J., N. Robin, C. J. Hein, R. Certain, and D. M. FitzGerald. 2014. "Internal Architecture of Mixed Sand-and-gravel Beach Ridges: Miquelon-Langlade Barrier, NW Atlantic." *Marine Geology*. 357:53–71. doi:10.1016/j.margeo.2014.07.011.

Bonham-Carter, G. F. 1994. *Geographic Information Systems for Geoscientists: Modelling with GIS*. UK: Elsevier.

Breiman, L. 2001. "Random Forests." *Machine Learning*. 45 (1):5–32.

Catani, F., D. Lagomarsino, S. Segoni, and V. Tofani. 2013. "Landslide Susceptibility Estimation by Random Forests Technique: Sensitivity and Scaling Issues." *Natural Hazards and Earth System Sciences*. 13(11):2815–2831.

Chen, W., M. Panahi, K. Khosravi, H. R. Pourghasemi, F. Rezaie, and D. Parvinnezhad. 2019. "Spatial Prediction of Groundwater Potentiality Using ANFIS Ensembled with

Teaching-learning-based and Biogeography-based Optimization." *Journal of Hydrology*. 572:435–448. doi:10.1016/j.jhydrol.2019.03.013.

Chen, W., H. Li, E. Hou, S. Wang, G. Wang, and T. Peng. 2018. "GIS-based Groundwater Potential Analysis Using Novel Ensemble Weights-of-evidence with Logistic Regression and Functional Tree Models." *Science of the Total Environment*. 634:853–867. doi:10.1016/j.scitotenv.2018.04.055.

Crisci, C., B. Ghattas, and G. Perera. 2012. "A Review of Supervised Machine Learning Algorithms and Their Applications to Ecological Data." *Ecological Modelling*. 240:113–122. doi:10.1016/j.ecolmodel.2012.03.001.

Das, S. 2017. "Delineation of Groundwater Potential Zone in Hard Rock Terrain in Gangajalghati Block, Bankura District, India Using Remote Sensing and GIS Techniques." *Modeling Earth Systems and Environment*. 3:1589–1599. doi:10.1007/s40808-017-0396-7.

Das, S. 2019. "Comparison among Influencing Factor, Frequency Ratio, and Analytical Hierarchy Process Techniques for Groundwater Potential Zonation in Vaitarna Basin, Maharashtra, India." *Groundwater for Sustainable Development*. 8:617–629.

Deepika, B., K. Avinash, and K. S. Jayappa. 2013. "Integration of Hydrological Factors and Demarcation of Groundwater Prospect Zones: Insights from Remote Sensing and GIS Techniques." *Environmental Earth Sciences*. 70(3):1319–1338.

Deffontaines, B., and J. Chorowicz. 1991. "Principles of Drainage Basin Analysis from Multisource Data: Application to the Structural Analysis of the Zaire Basin." *Tectonophysics*. 194 (3):237–263.

Doolittle, J. A., B. Jenkinson, D. Hopkins, M. Ulmer, and W. Tuttle. 2006. "Hydropedological Investigations with Ground-penetrating Radar (GPR): Estimating Water-table Depths and Local Ground-water Flow Pattern in Areas of Coarse-textured Soils." *Geoderma*. 131(3–4):317–329.

Duan, H., Z. Deng, F. Deng, and D. Wang. 2016. "Assessment of Groundwater Potential Based on Multicriteria Decision Making Model and Decision Tree Algorithms." *Mathematical Problems in Engineering*. doi:10.1155/2016/2064575.

Egan, J. P. 1975. *Signal Detection Theory and ROC-analysis*. New York: Academic Press.

Elith, J., J. R. Leathwick, and T. Hastie. 2008. "A Working Guide to Boosted Regression Trees." *Journal of Animal Ecology*. 77 (4):802–813.

ESRI. 2016. "ArcGIS for Desktop." http://desktop.arcgis.com

Friedl, M. A., C. E. Brodley, and A. H. Strahler. 1999. "Maximizing Land Cover Classification Accuracies Produced by Decision Trees at Continental to Global Scales." *IEEE Transactions on Geoscience and Remote Sensing*. 37(2):969–977.

Ganapuram, S., G. V. Kumar, I. M. Krishna, E. Kahya, and M. C. Demirel. 2009. "Mapping of Groundwater Potential Zones in the Musi Basin Using Remote Sensing Data and GIS." *Advances in Engineering Software*. 40(7):506–518.

Gayen, A., H. R. Pourghasemi, S. Saha, S. Keesstra, and S. Bai. 2019. "Gully Erosion Susceptibility Assessment and Management of Hazard-prone Areas in India Using Different Machine Learning Algorithms." *Science of the Total Environment*. 668:124–138. doi:10.1016/j.scitotenv.2019.02.436.

Gislason, P. O., J. A. Benediktsson, and J. R. Sveinsson. 2006. "Random Forests for Land Cover Classification." *Pattern Recognition Letters*. 27(4):294–300.

Golkarian, A., S. A. Naghibi, B. Kalantar, and B. Pradhan. 2018. "Groundwater Potential Mapping Using C5.0, Random Forest, and Multivariate Adaptive Regression Spline Models in GIS." *Environmental Monitoring and Assessment*. 190:149. doi:10.1007/s10661-6507-8.

Goudie, A. S. 2013. *Encyclopedia of Geomorphology*. London: Routledge.

Guru, B., K. Seshan, and S. Bera. 2017. "Frequency Ratio Model for Groundwater Potential Mapping and Its Sustainable Management in Cold Desert, India." *Journal of King Saud University-Science*. 29(3):333–347.

Hong, H., B. Pradhan, D. T. Bui, C. Xu, A. M. Youssef, and W. Chen. 2017. "Comparison of Four Kernel Functions Used in Support Vector Machines for Landslide Susceptibility Mapping: A Case Study at Suichuan Area (China)." *Geomatics, Natural Hazards and Risk*. 8(2):544–569.

Ibrahim-Bathis, K., and S. A. Ahmed. 2016. "Geospatial Technology for Delineating Groundwater Potential Zones in Doddahalla Watershed of Chitradurga District, India." *The Egyptian Journal of Remote Sensing and Space Science*. 19(2):223–234.

Khosravi, K., B. T. Pham, K. Chapi, A. Shirzadi, H. Shahabi, I. Revhaug, I. Prakash, and D. T. Bui. 2018. "A Comparative Assessment of Decision Trees Algorithms for Flash Flood Susceptibility Modeling at Haraz Watershed, Northern Iran." *Science of the Total Environment*. 627:744–755. doi:10.1016/j.scitotenv.2018.01.266.

Kim, J. C., S. Lee, H. S. Jung, and S. Lee. 2018. "Landslide Susceptibility Mapping Using Random Forest and Boosted Tree Models in Pyeong-Chang, Korea." *Geocarto International*. 33(9):1000–1015.

Kordestani, M. D., S. A. Naghibi, H. Hashemi, K. Ahmadi, B. Kalantar, and B. Pradhan. 2019. "Groundwater Potential Mapping Using a Novel Data-mining Ensemble Model." *Hydrogeology Journal*. 27(1):211–224.

Kuhn, M., J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, et al., 2018. "Package "Caret" Classification and Regression Training".

Lee, S., and B. Pradhan. 2007. "Landslide Hazard Mapping at Selangor, Malaysia Using Frequency Ratio and Logistic Regression Models." *Landslides*. 4(1):33–41.

Lee, S., S. M. Hong, and H. S. Jung. 2018. "GIS-based Groundwater Potential Mapping Using Artificial Neural Network and Support Vector Machine Models: The Case of Boryeong City in Korea." *Geocarto International*. 33 (8):847–861.

Lee, S., Y. S. Kim, and H. J. Oh. 2012. "Application of a Weights-of-evidence Method and GIS to Regional Groundwater Productivity Potential Mapping." *Journal of Environmental Management*. 96(1):91–105.

Lek, S., and J. F. Guegan. 1999. "Artificial Neural Networks as a Tool in Ecological Modelling, an Introduction." *Ecological Modelling*. 120(2–3):65–73.

Magesh, N. S., N. Chandrasekar, and J. P. Soundranayagam. 2012. "Delineation of Groundwater Potential Zones in Theni District, Tamil Nadu, Using Remote Sensing, GIS and MIF Techniques." *Geoscience Frontiers*. 3(2):189–196.

Mahmoudzadeh, M. R., A. P. Francés, M. Lubczynski, and S. Lambot. 2012. "Using Ground Penetrating Radar to Investigate the Water Table Depth in Weathered granites-Sardon Case Study, Spain." *Journal of Applied Geophysics*. 79:17–26.

Maiti, S., V. C. Erram, G. Gupta, and R. K. Tiwari. 2012. "ANN Based Inversion of DC Resistivity Data for Groundwater Exploration in Hard Rock Terrain of Western Maharashtra (India)." *Journal of Hydrology*. 464:294–308. doi:10.1016/j.jhydrol.2012.07.020.

Manap, M. A., H. Nampak, B. Pradhan, S. Lee, W. N. A. Sulaiman, and M. F. Ramli. 2014. "Application of Probabilistic-based Frequency Ratio Model in Groundwater Potential Mapping Using Remote Sensing Data and GIS." *Arabian Journal of Geosciences*. 7(2):711–724.

Manu, E., and K. Preko. 2014. "Estimation of Water Table Depths and Local Groundwater Flow Pattern Using the Ground Penetrating Radar." *International Journal of Scientific and Research Publications*. 4(8):1–12.

Marjanovic, M., M. Kovacevic, B. Bajat, and V. Vozenilek. 2011. "Landslide Susceptibility Assessment Using SVM Machine Learning Algorithm." *Engineering Geology*. 123(3):225–234.

Mogaji, K. A., H. S. Lim, and K. Abdullah. 2014. "Regional Prediction of Groundwater Potential Mapping in a Multifaceted Geology Terrain Using GIS-based Dempster-Shafer Model." *Arabian Journal of Geosciences*. 8(5):3235–3258.

Mojaddadi, H., B. Pradhan, H. Nampak, N. Ahmad, and A. H. B. Ghazali. 2017. "Ensemble Machine-learning-based Geospatial Approach for Flood Risk Assessment Using Multi-sensor Remote-sensing Data and GIS." *Geomatics, Natural Hazards and Risk*. 8(2):1080–1102.

Moore, I. D., and G. J. Burch. 1986. "Physical Basis of the Length-slope Factor in the Universal Soil Loss Equation 1." *Soil Science Society of America Journal*. 50(5):1294–1298.

Moore, I. D., R. B. Grayson, and A. R. Ladson. 1991. "Digital Terrain Modelling: A Review of Hydrological, Geomorphological, and Biological Applications." *Hydrological Processes*. 5(1):3–30.

Murasingh, S., R. Jha, and S. Adamala. 2018. "Geospatial Technique for Delineation of Groundwater Potential Zones in Mine and Dense Forest Area Using Weighted Index Overlay Technique." *Groundwater for Sustainable Development*. 7:387–399. doi:10.1016/j.gsd.2017.12.001.

Naghibi, S. A., D. D. Moghaddam, B. Kalantar, B. Pradhan, and O. Kisi. 2017a. "A Comparative Assessment of GIS-based Data Mining Models and a Novel Ensemble Model in Groundwater Well Potential Mapping." *Journal of Hydrology*. 548:471–483. doi:10.1016/j.jhydrol.2017.03.020.

Naghibi, S. A., and H. R. Pourghasemi. 2015. "A Comparative Assessment between Three Machine Learning Models and Their Performance Comparison by Bivariate and Multivariate Statistical Methods in Groundwater Potential Mapping." *Water Resources Management*. 29(14):5217–5236.

Naghibi, S. A., H. R. Pourghasemi, and B. Dixon. 2016. "GIS-based Groundwater Potential Mapping Using Boosted Regression Tree, Classification and Regression Tree, and Random Forest Machine Learning Models in Iran." *Environmental Monitoring and Assessment*. 188:44. doi:10.1007/s10661-015-5049-6.

Naghibi, S. A., H. R. Pourghasemi, and K. Abbaspour. 2018. "A Comparison between Ten Advanced and Soft Computing Models for Groundwater Qanat Potential Assessment in Iran Using R and GIS." *Theoretical and Applied Climatology*. 131:967–984. doi:10.1007/s00704-016-2022-4.

Naghibi, S. A., H. R. Pourghasemi, Z. S. Pourtaghi, and A. Rezaei. 2015. "Groundwater Qanat Potential Mapping Using Frequency Ratio and Shannon's Entropy Models in the Moghan Watershed, Iran." *Earth Science Informatics*. 8(1):171–186.

Naghibi, S. A., K. Ahmadi, and A. Daneshi. 2017b. "Application of Support Vector Machine, Random Forest, and Genetic Algorithm Optimized Random Forest Models in Groundwater Potential Mapping." *Water Resources Management*. 31(9):2761–2775.

Naghibi, S. A., M. Dolatkordestani, A. Rezaei, P. Amouzegari, M. T. Heravi, B. Kalantar, and B. Pradhan. 2019. "Application of Rotation Forest with Decision Trees as Base Classifier and a Novel Ensemble Model in Spatial Modeling of Groundwater Potential." *Environmental Monitoring and Assessment*. 191:248. doi:10.1007/s10661-019-7362-y.

Nakashima, Y., H. Zhou, and M. Sato. 2001. "Estimation of Groundwater Level by GPR in an Area with Multiple Ambiguous Reflections." *Journal of Applied Geophysics*. 47(3–4):241–249.

Nampak, H., B. Pradhan, and M. A. Manap. 2014. "Application of GIS Based Data Driven Evidential Belief Function Model to Predict Groundwater Potential Zonation." *Journal of Hydrology*. 513:283–300. doi:10.1016/j.jhydrol.2014.02.053.

Neal, A. 2004. "Ground-penetrating Radar and Its Use in Sedimentology: Principles, Problems and Progress." *Earth-Science Reviews*. 66(3–4):261–330.

Oh, H. J., Y. S. Kim, J. K. Choi, E. Park, and S. Lee. 2011. "GIS Mapping of Regional Probabilistic Groundwater Potential in the Area of Pohang City, Korea." *Journal of Hydrology*. 399(3–4):158–172.

Patra, S., P. Mishra, and S. C. Mahapatra. 2018. "Delineation of Groundwater Potential Zone for Sustainable Development: A Case Study from Ganga Alluvial Plain Covering Hooghly District of India Using Remote Sensing, Geographic Information System and Analytic Hierarchy Process." *Journal of Cleaner Production*. 172:2485–2502.

Pourghasemi, H. R., H. R. Moradi, and S. F. Aghda. 2013. "Landslide Susceptibility Mapping by Binary Logistic Regression, Analytical Hierarchy Process, and Statistical Index Models and Assessment of Their Performances." *Natural Hazards*. 69(1):749–779.

Pourtaghi, Z. S., and H. R. Pourghasemi. 2014. "GIS-based Groundwater Spring Potential Assessment and Mapping in the Birjand Township, Southern Khorasan Province, Iran." *Hydrogeology Journal*. 22(3):643–662.

Pradhan, B., and S. Lee. 2010. "Regional Landslide Susceptibility Analysis Using Back-propagation Neural Network Model at Cameron Highland, Malaysia." *Landslides*. 7(1):13–30.

Prasad, R. K., N. C. Mondal, P. Banerjee, M. V. Nandakumar, and V. S. Singh. 2008. "Deciphering Potential Groundwater Zone in Hard Rock through the Application of GIS." *Environmental Geology*. 55(3):467–475.

Rahmati, O., A. N. Samani, M. Mahdavi, H. R. Pourghasemi, and H. Zeinivand. 2015. "Groundwater Potential Mapping at Kurdistan Region of Iran Using Analytic Hierarchy Process and GIS." *Arabian Journal of Geosciences*. 8(9):7059–7071.

Rahmati, O., H. R. Pourghasemi, and A. M. Melesse. 2016. "Application of GIS-based Data Driven Random Forest and Maximum Entropy Models for Groundwater Potential Mapping: A Case Study at Mehran Region, Iran." *Catena*. 137:360–372.

Rajaveni, S. P., K. Brindha, and L. Elango. 2017. "Geological and Geomorphological Controls on Groundwater Occurrence in a Hard Rock Region." *Applied Water Science*. 7(3):1377–1389.

Recknagel, F. 2001. "Applications of Machine Learning to Ecological Modelling." *Ecological Modelling*. 146(1–3):303–310.

Regmi, N. R., J. R. Giardino, E. V. McDonald, and J. D. Vitek. 2015. "A Review of Mass Movement Processes and Risk in the Critical Zone of Earth." *Developments in Earth Surface Processes*. 19:319–362.

Rodriguez-Galiano, V. F., and M. Chica-Rivas. 2014. "Evaluation of Different Machine Learning Methods for Land Cover Mapping of a Mediterranean Area Using Multi-seasonal Landsat Images and Digital Terrain Models." *International Journal of Digital Earth*. 7(6):492–509.

Samui, P. 2008. "Slope Stability Analysis: A Support Vector Machine Approach." *Environmental Geology*. 56(2):255.

Schapire, R. E. 2003. "The Boosting Approach to Machine Learning: An Overview." In *Nonlinear Estimation and Classification*. Denison, D.D., Hansen, M.H., Holmes, C.C., Mallick, B., Yu.B. (eds.) New York: Springer; p. 149–171. doi:10.1007/978-0-387-21579-2_9.

Shafizadeh-Moghadam, H., R. Valavi, H. Shahabi, K. Chapi, and A. Shirzadi. 2018. "Novel Forecasting Approaches Using Combination of Machine Learning and Statistical Models for Flood Susceptibility Mapping." *Journal of Environmental Management*. 217:1–11. doi:10.1016/j.jenvman.2018.03.089.

Shih, S. F., J. A. Doolittle, D. L. Myhre, and G. W. Schellentrager. 1986. "Using Radar for Groundwater Investigation." *Journal of Irrigation and Drainage Engineering*. 112(2):110–118.

Tahmassebipoor, N., O. Rahmati, F. Noormohamadi, and S. Lee. 2016. "Spatial Analysis of Groundwater Potential Using Weights-of-evidence and Evidential Belief Function Models and Remote Sensing." *Arabian Journal of Geosciences*. 9:79. doi:10.1007/s12517-015-2166-z.

Tehrany, M. S., B. Pradhan, and M. N. Jebur. 2014. "Flood Susceptibility Mapping Using a Novel Ensemble Weights-of-evidence and Support Vector Machine Models in GIS." *Journal of Hydrology*. 512:332–343. doi:10.1016/j.jhydrol.2014.03.008.

Tehrany, M. S., B. Pradhan, S. Mansor, and N. Ahmad. 2015. "Flood Susceptibility Assessment Using GIS-based Support Vector Machine Model with Different Kernel Types". *Catena*. 125:91–101. doi:10.1016/j.catena.2014.10.017.

Yao, X., L. G. Tham, and F. C. Dai. 2008. "Landslide Susceptibility Mapping Based on Support Vector Machine: A Case Study on Natural Slopes of Hong Kong, China." *Geomorphology*. 101(4):572–582.

Youssef, A. M., H. R. Pourghasemi, Z. S. Pourtaghi, and M. M. Al-Katheeri. 2016. "Landslide Susceptibility Mapping Using Random Forest, Boosted Regression Tree, Classification and Regression Tree, and General Linear Models and Comparison of Their Performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia." *Landslides*. 13(5):839–856.

Endorsement Certificate from the Mentor & Host Institute

This is to certify that:

I. The applicant, **Dr. PANKAJ PRASAD**, will assume full responsibility for implementing the project.

II. The fellowship will start from the date on which the fellow joins University/Institute where he/she implements the fellowship. The mentor will send the joining report to the SERB. SERB will release the funds on receipt of the joining report.

III. The applicant, if selected as SERB-N PDF, will be governed by the rules and regulations of the University/ Institute and will be under administrative control of the University/ Institute for the duration of the Fellowship.

The grant-in-aid by the Science & Engineering Research Board (SERB) will be used to meet the expenditure on the project and for the period for which the project has been sanctioned as indicated in the sanction letter/ order.

No administrative or other liability will be attached to the Science & Engineering Research Board (SERB) at the end of the Fellowship.

VI. The University/ Institute will provide basic infrastructure and other required facilities to the fellow for undertaking the research objectives.

VII. The University/ Institute will take into its books all assets received under this sanction and its disposal would be at the discretion of Science & Engineering Research Board (SERB).

VIII. University/ Institute assume to undertake the financial and other management responsibilities of the project.

IX. The University/ Institute shall settle the financial accounts to the SERB as per the prescribed guidelines within three months from the date of termination of the Fellowship.

Dated: 7th August 2023

Signature of the Mentor:

Name & Designation: *Dr. C. Jeganathan, Professor, Department of Remote Sensing, Birla Institute of Technology (BIT), Mesra, Ranchi – 835215, Jharkhand.*

Dated:

Signature of the Registrar of University/Head of Institute

Seal of the Institution:

Registrar
Birla Institute of Technology
Mesra, Ranchi

GOA UNIVERSITY

This is to certify that

Pankaj Prasad

son of

Shri Pradip Prasad

and

Smt. Sabita Prasad

from School of Earth, Ocean & Atmospheric Sciences

has been conferred the degree of

Doctor of Philosophy

in Earth Science

having passed the qualifying examination held in July, 2022

Given under the seal of the University.

Declaration of Result : 18th July, 2022
Taleigao Plateau - Goa

/201711822

2020 - 12995

Vice - Chancellor

Date of Issue 20/10/2022

G4-016

G07481-0270

West Bengal Board Of Secondary Education

MADHYAMIK PARIKSHA (SECONDARY EXAMINATION), 2008

Certified that **PANKAJ PRASAD**

Son / ~~Daughter~~ / Ward of **PRADIP PRASAD**

appearing from / **XX CHAMPDANI NIBARAN MUKHOPADHYAY VIDYAMANDIR**

Whose date of birth is **TWENTY SIXTH** day of **FEBRUARY** One Thousand Nine Hundred and **NINETY THREE** duly passed the Madhyamik Pariksha (Secondary Examination) held in the month of February, 2008 and was placed in the **FIRST** division.

Kolkata-700 016,

Dated, **the 28th May, 2008.**

Swapan Kumar Sarcar
Secretary

Mamata Roy
President

PP0482

018217

Undertaking by the Principal Investigator

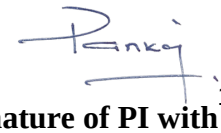
To

The Secretary
SERB, New Delhi

Sir

I Dr. Pankaj Prasad

herby certify that the research proposal titled Land subsidence assessment using geospatial (UAV, DInSAR), artificial intelligence and GPR techniques in Coal mining regions of East India submitted for possible funding by SERB, New Delhi is my original idea and has not been copied/taken verbatim from anyone or from any other sources. I further certify that this proposal has been checked for plagiarism through a plagiarism detection tool i.e. iThenticate approved by the Institute and the contents are original and not copied/taken from any one or many other sources. I am aware of the UGCs Regulations on prevention of Plagiarism i.e. University Grant Commission (Promotion of Academic Integrity and Prevention of Plagiarism in Higher Educational Institutions) Regulation, 2018. I also declare that there are no plagiarism charges established or pending against me in the last five years. If the funding agency notices any plagiarism or any other discrepancies in the above proposal of mine, I would abide by whatsoever action taken against me by SERB, as deemed necessary.



10.08.2023

Signature of PI with date

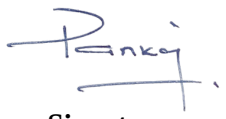
Name / designation

Dr. Pankaj Prasad
Project Fellow
Birla Institute of Technology, Mesra

Undertaking by the Fellow

I, ---Dr. Pankaj Prasad---------, Son of Shri. -Pradip Prasad, resident of ~~7~~88c,11g, Ram Mohan Sarani, Baidyabati, Hooghly-712222, West Bengal agree to undertake the following, If I am offered the SERB N-PDF
`

1. **I shall abide by the rules and regulations of SERB during the entire tenure of the fellowship.**
2. **I shall also abide by the rules, discipline of the institution where I will be implementing my fellowship**
3. **I shall devote full time to research work during the tenure of the fellowship**
4. **I shall prepare the progress report at the end of each year and communicate the same to SERB through the mentor**
5. **I shall send two copies of the consolidated progress report at the end of the fellowship period.**
6. **I further state that I shall have no claim whatsoever for regular/permanent absorption on expiry of the fellowship.**



Date: 10.08.2023 **Signature**

Ecological Informatics 68 (2022) 101522

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Evaluation and comparison of the earth observing sensors in land cover/land use studies using machine learning algorithms

Pankaj Prasad [a,c,*], Victor Joseph Loveson [a], Priyankar Chandra [b], Mahender Kotha [c]

[a] Geological Oceanography Division, CSIR- National Institute of Oceanography, Dona Paula 403004, Goa, India
[b] Department of Geography, Institute of Science, Banaras Hindu University, Varanasi 221005, Uttar Pradesh, India
[c] School of Earth, Ocean and Atmospheric Sciences, Goa University, Taleigao 403001, Goa, India

ARTICLE INFO

Keywords:
Land cover/land use
Mapping
Wetland
Remote sensing
Machine learning

ABSTRACT

The rapid transformation of land cover/land use (LCLU) is a strong indication of global environmental change. In order to monitor LCLU through maps, a significant dataset and robust technique are necessary. Thus, the primary objective of the current research is to evaluate and compare the efficiency of several notable satellite sensors including Landsat-8 (L-8), Sentinel-2 (S-2), Sentinel-1 (S-1), combined Sentinel-1 and Sentinel-2 (S-1-2), LISS III (L-3), and LISS IV (L-4) for LCLU mapping applying random forest (RF), logit boost (LB), stochastic gradient boosting (SGB), artificial neural network (ANN), and K-nearest neighbor (KNN) models. For this purpose, 300 samples for each of the six LCLU classes have been selected based on field survey and high resolution Cartosat-3 images. The classification accuracy namely producer accuracy (PA), user accuracy (UA), overall accuracy (OA) and kappa coefficient have been calculated from the confusion matrix of the applied models. This results show the highest accuracy has been derived from the integration of S-1-2 datasets followed by S-2, L-8, L-3, L-4, and S-1. On the other hand, LB model is the most consistent and efficient in comparison with other models for all the datasets. Regarding importance of variable, SWIR band is repeatedly the most crucial factor while blue band is the least significant variable. From this comparative assessment of sensors, it has been found that high spatial and spectral resolutions along with combination of satellite datasets are required to get better accuracy rather than only high spatial resolution in regional scale mapping. The present study strongly advocates the use of combined S-1-2 data together with the application of LB model for LCLU classification.

1. Introduction

At present, LCLU maps are essential documents for sustainable use and management of natural resources at global, regional, landscape, and local scales (Clerici et al., 2017; Letourneau et al., 2012; Lillesand et al., 2008; Pandey et al., 2021; Tavares et al., 2019). The terms land-cover and land-use concerned with ground surface, have become buzzwords. Specifically, land-use is the inevitable outcome of man-nature interaction whereas land-cover represents the physical condition of land surface. The information of LCLU is necessary for various studies namely urban planning, wetland, biogeochemical cycles, measuring land and shoreline erosion, desertification, forestry, agricultural monitoring and policy making, hazard monitoring, and also for climate change studies (Anderson, 1976; Clerici et al., 2017; Diengdoh et al., 2020; Gemitzi, 2021; Nguyen and Henebry, 2019; Noi and Kappas, 2018; Schirpke et al., 2012; Talukdar et al., 2020). Thus, a suitable dataset along with robust method is required for the preparation of high precision LCLU map.

Over the pre-remote sensing era, LCLU classification was mainly based on field surveys (terrestrial mapping). After that, the photogrammetric method (aerial photo mapping) came into practice. These conventional methods were laborious, time consuming, costly, and inconvenient for remote and dynamic landscape regions (Adam et al., 2014). The launch of Landsat mission in 1970s overcame many disadvantages of traditional methods and broke new ground in remote sensing based classification. The development of remote sensing techniques offers a vast amount of satellite datasets of the entire earth's surface at both large and small scales. Remote sensing has emerged as a time-saving, cost-effective and more accurate technique which has been an important source of LCLU information. The remote sensing data along with the field survey has increased the confidence of producer in interpreting maps (Anderson, 1976). By integrating remote sensing with

* Corresponding author at: Geological Oceanography Division, CSIR- National Institute of Oceanography, Dona Paula 403004, Goa, India.
E-mail address: ppankaj.earthscience@gmail.com (P. Prasad).

https://doi.org/10.1016/j.ecoinf.2021.101522
Received 23 August 2021; Received in revised form 5 December 2021; Accepted 6 December 2021
Available online 11 December 2021

geographic information system, LCLU classification and evaluation are being done efficiently and widely (Lu and Weng, 2007).

Earlier, a multitude of studies have been carried out on LCLU using different satellite datasets, viz. ASTER (Xu et al., 2004; Yüksel et al., 2008), MODIS (Thenkabail et al., 2005; Zhan et al., 2002), SPOT (Deng et al., 2009; Willhauck et al., 2000), AWIFS (Kandrika and Roy, 2008; Panigrahy et al., 2009), Landsat (Güler et al., 2007; Ololade et al., 2008). In recent times, the precision of LCLU classification has been enhanced spectacularly with the advancement of spatial, spectral, and temporal resolutions of satellite image (Adam et al., 2014; Nguyen and Henebry, 2019; Pandey et al., 2021). The open data policy of the United States Geological Survey (USGS) and European Space Agency (ESA) has unlocked the floodgate of spatial data to the research community. There has been a marked increase in the use of Landsat-8 data for LCLU classification, natural resource and hazard mapping since 2013 due to the availability of Landsat-8 OLI (L-8) with better spectral characteristics than the earlier Landsat (1–7) sensors (Forkuor et al., 2018). ESA has launched Sentinel missions for different fields of research. Recently, Sentinel-2 (S-2) data has become very popular in land use mapping because of its high precision results (Tavares et al., 2019). On the other side, Sentinel-1 (S-1) has been launched for the purpose of monitoring regions under high cloud coverage using radar images (Tavares et al., 2019). The use of LISS III (L-3) and LISS IV (L-4) sensors in LCLU study is not as widespread as the other sensors. However, these datasets have very high spatial resolution. Many researchers have attempted to compare different optical sensors for land use study (Chander et al., 2007; Chauhan and Dwivedi, 2008; Deng et al., 2008; Forkuor et al., 2018; Ghayour et al., 2021). Few studies have been done integrating the optical and radar datasets (Clerici et al., 2017; Denize et al., 2019; Deus, 2016; Mishra et al., 2019; Steinhausen et al., 2018).

There are several common methods for image classification of remotely sensed data namely parallelepiped, K-means, ISODATA, minimum distance, and maximum likelihood. These methods often fail to identify the class in complex physical and biological landscapes (Adam et al., 2014; Gong et al., 2011; Lu and Weng, 2007). At present, machine learning algorithms have improved the classification outputs of LCLU mapping manyfold compared to conventional methods (Ghayour et al., 2021). The main advantages of the machine learning algorithms are: the ability to handle large datasets, no need of any assumptions for data distribution, ease to interpret the decision rules, and high learning capability from the input datasets (Gong et al., 2011). The machine learning methods including random forest (RF), k-nearest neighbor (KNN), naïve bayesian, boosted regression tree, support vector machine, extreme gradient boosting, mahalanobis distance, stochastic gradient boosting (SGB), logitboost (LB), and artificial neural network (ANN) have been in use for LCLU classification (Adam et al., 2014; Clerici et al., 2017; Noi and Kappas, 2018; Steinhausen et al., 2018; Denize et al., 2019; Abdi, 2020; Talukdar et al., 2020; Ghayour et al., 2021). These algorithms have been enormously applied in natural resource and hazard mapping such as groundwater potentiality (Chen et al., 2019; Naghibi et al., 2019), wetland (Hird et al., 2017; Maxwell et al., 2016), flood susceptibility (Chapi et al., 2017; Khosravi et al., 2018; Prasad et al., 2021a), soil erosion (Arabameri et al., 2018; Pourghasemi et al., 2020), landslide susceptibility (Bui et al., 2019; Chen et al., 2020). However, the application of machine learning approaches in land-use mapping is still limited. Among the aforesaid models, RF, LB, SGB, ANN, and KNN, have been chosen for the present study. These classification methods were selected on the basis of their high performance and precision compared to other machine learning algorithms (Abdi, 2020; Adam et al., 2014; Breiman, 2001; Oh et al., 2019).

Consulting literatures and considering the advantages of using different sensors, the present study has been undertaken to compare and evaluate the L-8, S-2, S-1, combined S-1-2, L-3, and L-4 sensors for LCLU mapping in a single platform. As far as the authors are concerned, no research comparing the efficiency of five sensors (high to medium-high resolution) with the help of five machine learning models for LCLU study.

In view of above identified research gap, three research objectives have been formulated in this study: i) to examine the comparative precision of applied satellite sensors datasets, ii) to select the significant models for LCLU mapping, and iii) to evaluate the bands and indices of multi-sensor data. The outcomes of the study will help decision makers to implement sustainable land use planning and land resources management in the study region. Even the conceptual framework of this study can be useful in different parts of the world for LCLU mapping.

2. Study area

The area of investigation is located in the heterogamous landscape of the central west coast of India (Fig. 1). The altitudinal variation of the area is between 0 m to 450 m. The climate of the study region is subtropical type with average annual rainfall of 3000 mm and mean annual temperature variation of 25 °C (Prasad et al., 2020). Wetlands are one of the main attractions of this coastal region owing to the rich diversity of flora and fauna (Ministry of Environment and Forests, 2010). Water bodies of the study area include rivers (Vaghotan, Deogad, Achara, Gad, Karli, Mochemad, Terekhol), estuaries, and lakes. The major parts of the region are covered with thick evergreen forest. The uneven distribution of the forest is manifested by its high concentration along the boundary of the Western Ghats and its scattered distribution in coastal belt. Most of the forest mainly found in the poor shallow soils on steep hill slopes is generally unfit for cultivation. There are few cities namely Vaibhavwadi, Malvan, Kankavli, Vengurla, Kudal, Oros, and Devgad and numbers of small towns are located in the study area. The area is well-connected with railway, national highway, state highway, and ferry transport. Most of the Deccan basalts made portion of the study area are barren land which is not suitable for the agriculture. The

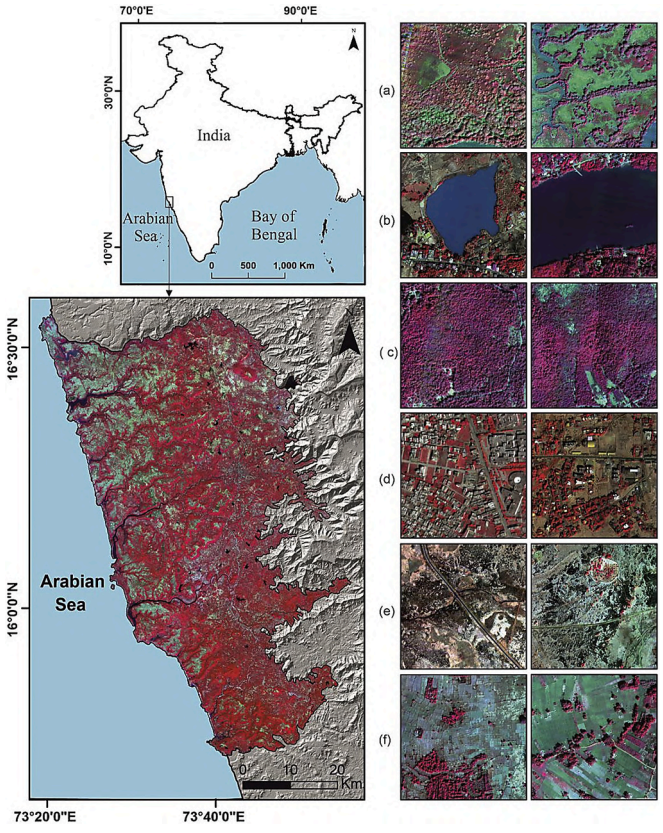


Fig. 1. Location map and different types of LCLU classes (a) wetland, (b) water body, (c) forest, (d) built-up, (e) barren land, (f) agricultural land extracted from high resolution Cartosat-3 image of the study area used for inventory.

inhabitants are mainly engaged in agriculture, fisheries, and tourism (District Mining Officer, 2017). However, major parts of the agricultural area are devoid of irrigation facilities.

3. Materials and methods

The methodology of the current work is presented in Fig. 2.

3.1. Image acquisition and preprocessing

Selection of satellite images is primary and vital step of LCLU mapping. Seasonal variation, cloud, and haze increase the classification error and lead to misinterpretation. To avoid such difficulty, cloud and haze-free satellite imageries of same month were taken in this study. Both optical (L-8, L-3, L-4, S-2, C-3) and radar (S-1) images were used for this study. The characteristics of the bands of the sensors are described in Table 1.

3.1.1. Landsat-8 operational land imager (L-8)

The L-8 data was downloaded from the USGS earth explorer (https://earthexplorer.usgs.gov/) on 17 December 2020. A single tile of the image covering the area of interest was pre-processed using semi-automatic classification plugin (Dark Object Subtraction method for surface reflectance values) in QGIS platform.

3.1.2. LISS III and LISS IV (L-3 and L-4)

The L-3 and L-4 datasets were purchased from the National Remote Sensing Centre of Indian space research organization (NRSC-ISRO) (https://www.nrsc.gov.in/) on 07 December 2020. The study area was entirely covered with the three scenes of each dataset. The goal of these missions is to produce satellite data for integrated land and water resource management. The L-3 and L-4 sensors have 70 km orbital swath in all the bands with 24 days temporal resolution. Both L-3 and L-4 data were ortho-rectified including internal and terrain relief errors, sensor altitude, and scale variation. ATCOR tool (compatible with ERDAS IMAGINE software) was used for pre-processing of L-3 and L-4 datasets.

3.1.3. Sentinel-2 (S-2)

The S-2 data was acquired on 15 December 2020 from the European space agency. For level 1c data no geometric and radiometric corrections is needed but it requires atmospheric correction (Abdi, 2020) for which Sen2Cor plugin (compatible with SNAP software) was used. The bands having 20 m spatial resolution were transformed into 10 m applying the nearest neighbor algorithm.

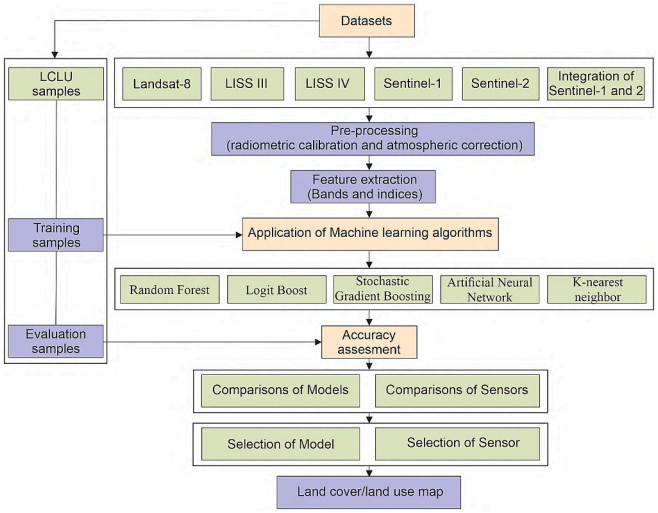


Fig. 2. Flowchart of the methodology.

3.1.4. Sentinel-1 (S-1)

The S-1 images were obtained on 18 December 2020 in ground range detected (GRD) interferometric wide swath mode (IW). The available polarizations of the area of interest were VH (vertically transmitted horizontally received signal) and VV (vertically transmitted vertically received signal). The pre-processing of SAR images is pre-requisite for LCLU mapping. The pre-processing involves several steps viz. subset, apply orbit file, thermal noise removal, border noise removal, calibration, speckle filtering, and range-doppler terrain correction. In the first step, two subsets from each SAR image covering the whole research area were carved out to minimize the processing time. The apply orbit file method helps to get the correct details of the imagery. Thermal noise removal was applied to improve the backscattering signal of the images. The low intensity noise and false values on image edges were erased using border noise removal. Calibration transformed the DN values into radiometrically calibrated sigma nought (σ^0) backscatter values based on the following equation (Filipponi, 2019):

$$\sigma^0 = \frac{|DN|}{A_i^2} \tag{1}$$

here DN represents the digital number of backscattering intensity and A_i^2 indicates the absolute calibration constant, available in metadata. Speckle filtering is applied to enhance the quality of data by minimizing the speckle noise. For this, refined Lee filter (window of 7×7 pixels) was employed due to its superiority in comparison with other speckle filter (Denize et al., 2019; Filipponi, 2019). The topographic distortion of the image was corrected using the SRTM digital elevation model (with Universal Transverse Mercator zone 43 N) and bilinear interpolation algorithm. At the end of image pre-processing, sigma nought values of each pixel were transformed into decibel (dB) using a logarithmic transformation. All the procedures of pre-processing of SAR images were done with the help of Sentinel-1 toolbox in SNAP software.

3.1.5. Cartosat-3 (C-3)

The C-3 images were purchased from the NRSC-ISRO. These images were used for the inventory of LCLU classes on account of high spatial resolution (1.1 m). To avoid bias, month of the images (05.12.2020) were maintained same with the other sensors. The collected C-3 data was orthorectified.

3.2. Input variables of the optical and SAR images

The input variables of the satellite images vary from sensor to sensor due to different spectral bands of the imageries in LCLU mapping. In this study, these factors were selected based on literature survey. For L-8, six spectral bands and three indices were considered at 30 m spatial resolution. In the case of S-2, a total of thirteen variables were selected for the study. Regarding L-3, the available four bands along with three indices were chosen. The L-4 sensor has only three bands and hence only one index is possible to prepare. The study did not take into account the atmospheric (coastal aerosol, water vapour, cirrus) and thermal factors owing to their sensitivity to aerosol and clouds which cause noise in classification (Clerici et al., 2017). Several indices namely GNDVI, MNDWI, NDBI were prepared for classification. The applied indices were calculated as

$$GNDVI = \frac{NIR - Green}{NIR + Green} \tag{2}$$

$$MNDWI = \frac{Green - SWIR}{Green + SWIR} \tag{3}$$

$$NDBI = \frac{SWIR - NIR}{SWIR + NIR} \tag{4}$$

From S-1 (SAR image), a total of eleven parameters were extracted for classification task. The backscattering of the SAR image is affected by

Table 1
Characteristics of the satellite sensors.

Sentinel-2				Sentinel-1		
Bands	Central Wavelength (nm)	Spatial Resolution (m)	Date	Polarizations and Indices	Spatial Resolution (m)	Date
Blue (2)	0.4900	10		VV	10	
Green (3)	0.5600	10		VH	10	
Red (4)	0.6650	10		VV Contrast	10	
Red edge (5)	0.7050	20		VV Mean	10	
Red edge (6)	0.7400	20		VV Variance	10	
Red edge (7)	0.7830	20		VV Correlation	10	
NIR (8)	0.8420	10	15.12.2020 (2 Scenes)	VH Contrast	10	18.12.2020 (2 Scenes)
Red edge (8A)	0.8650	20		VH Mean	10	
SWIR (11)	1.6100	20		VH Variance	10	
SWIR (12)	2.1900	20		VH Correlation	10	
GNDVI	–	10		VV/VH	10	
MNDWI	–	10				
NDBI	–	10				

Landsat-8				LISS IV			
Bands & Indices	Central Wavelength (nm)	Spatial Resolution (m)	Date	Bands & Indices	Central Wavelength (nm)	Spatial Resolution (m)	Date
Blue (2)	0.4826	30		Green (2)	0.560	24	
Green (3)	0.5613	30		Red (3)	0.655	24	
Red (4)	0.6546	30		NIR (4)	0.820	24	
NIR (5)	0.8646	30	17.12.2020 (1 Scenes)	SWIR (5)	1.625	24	07.12.2020 (3 Scenes)
SWIR (6)	1.6090	30		GNDVI	–	24	
SWIR (7)	2.2010	30		MNDWI	–	24	
GNDVI	–	30		NDBI	–	24	
MNDWI	–	30					
NDBI	–	30					

LISS III				Cartosat-3			
Bands & Indices	Central Wavelength (nm)	Spatial Resolution (m)	Date	Bands & Indices	Central Wavelength (nm)	Spatial Resolution (m)	Date
Green (2)	0.560	24		Blue (1)	0.48	1.1	
Red (3)	0.655	24		Green (2)	0.56	1.1	
NIR (4)	0.820	24		Red (3)	0.65	1.1	
SWIR (5)	1.625	24	07.12.2020 (3 Scenes)	NIR (4)	0.82	1.1	05.12.2020 (3 Scenes)
GNDVI	–	24					
MNDWI	–	24					
NDBI	–	24					

the texture of the earth surface which stores useful information to improve the LCLU classification (Deus, 2016; Pavanelli et al., 2018). For this, grey level co-occurring matrix (GLCM) was employed to obtain the texture variance, contrast, mean, and correlation. GLCM statistics with a 7 × 7 window size was applied to both polarizations (VV and VH) generating eight indices for model classification. These indices have been used in many studies as input variable for land cover mapping (Clerici et al., 2017; Deus, 2016; Pavanelli et al., 2018; Tavares et al., 2019). The equations for GLCM textural measures are as follows (Deus, 2016).

$$Mean = \sum_{i,j=0}^{N-1} iP_{i,j} \tag{5}$$

$$Variance = \sum_{i,j=0}^{N-1} iP_{i,j}(i-u)^2 \tag{6}$$

$$Contrast = \sum_{i,j=0}^{N-1} iP_{i,j}(i-j)^2 \tag{7}$$

$$Correlation = \frac{\sum_{i,j=0}^{N-1} iP_{i,j} - u_x u_y}{\sigma_x \sigma_y} \tag{8}$$

where p(i,j) represents normalized grey tone spatial dependence matrix such as SUM (i,j = 0, N − 1) (P(i,j)) = 1; i and j indicate rows and columns respectively, μ symbolizes mean, N denotes the number of distinct grey levels in the quantized image, u_x, u_y, σ_x and σ_y are the means and standard deviations of p_x and p_y, respectively.

3.3. Reference data

It is important to collect the training and testing samples for image classification. In this study, LCLU classes were recognized from the high resolution orthorectified C-3 images, google earth image, and field visit. A total of 1800 points were identified and each class was formed with 300 samples in order to avoid biases of the classes. The datasets were divided into training and testing classes at the ratio of 70:30, respectively. In the present study, six LCLU classes such as wetlands, water bodies, forest, built-up, barren land, and agricultural land were considered based on literature review and extensive field survey.

3.4. Image classification methods

Random forest (RF) is an ensemble based supervised classification approach that combines bagging and random subspace method (Adam et al., 2014; Breiman, 2001; Prasad et al., 2020). The method constructs a large number of decision trees to get output of the class. Here no assumption is needed for the input and output variables. The main merit of this method is that it can handle different types of data such as satellite data (Breiman, 2001; Talukdar et al., 2020). RF algorithm requires

two tuning parameters namely number of trees (ntree), and number of variables in each split (mtry).

Logit boost (LB) is a minor modification of popular AdaBoost model to improve the statistical results (Friedman and Tibshirani, 2000; Oh et al., 2019; Tehrany et al., 2019). The model uses the additive logistic regression function by minimizing the logistic loss for classification of the datasets. LB model is designed to solve the overfitting problem and to reduce bias and variance using logitboost algorithm. The main advantage of the LB method is that it especially specialized in handling noisy data. The LB model has been extensively used in computer and medical sciences with high precision results.

Stochastic gradient boosting (SGB) is a hybrid machine learning method employed in different classification and regression tasks. The model merges the bagging and boosting approaches to optimize the better predictive performance (Friedman, 2002; Zhou et al., 2016). In this context, the model uses a set of data at each step of boosting process rather than whole dataset. The SGB model includes interaction depth (number of trees), bagging fraction (fraction of training data), shrinkage rate (learning speed), and training fraction (loss function). As, boosting procedure uses small number of trees at each stage of classification so that it can avoid the overfitting issue (Lawrence et al., 2004; Moisen et al., 2006).

Artificial neural network (ANN) is a popular non-linear machine learning model to solve complex classification problems without any assumption. ANN method is highly self-learning and adaptive method with the ability to identify and generalize complex datasets (Ghayour et al., 2021; Gong et al., 2011; Yu and Chen, 2020). For the prediction, ANN creates a method from the input layers (Kalantar et al., 2018; Tan et al., 2021). In general, ANN consists of three layers including input layers (bands and indices of satellite data), hidden layers (calculations of neurons), and output layers (LCLU classes). In this study, number of neurons and hidden layers were obtained from the trial and error process.

K-nearest neighbor (KNN) is a nonparametric supervised algorithm that ensures no need of class density function prediction (Naghibi and Dashtpagerdi, 2017; Shahabi et al., 2020). KNN method is very easy to understand and apply. The algorithm assigns class to samples on the basis of the most similar samples in the training data (Motevalli et al., 2019). For the classification, KNN method measures the distance of the unknown cell point to the closest neighboring points in high-

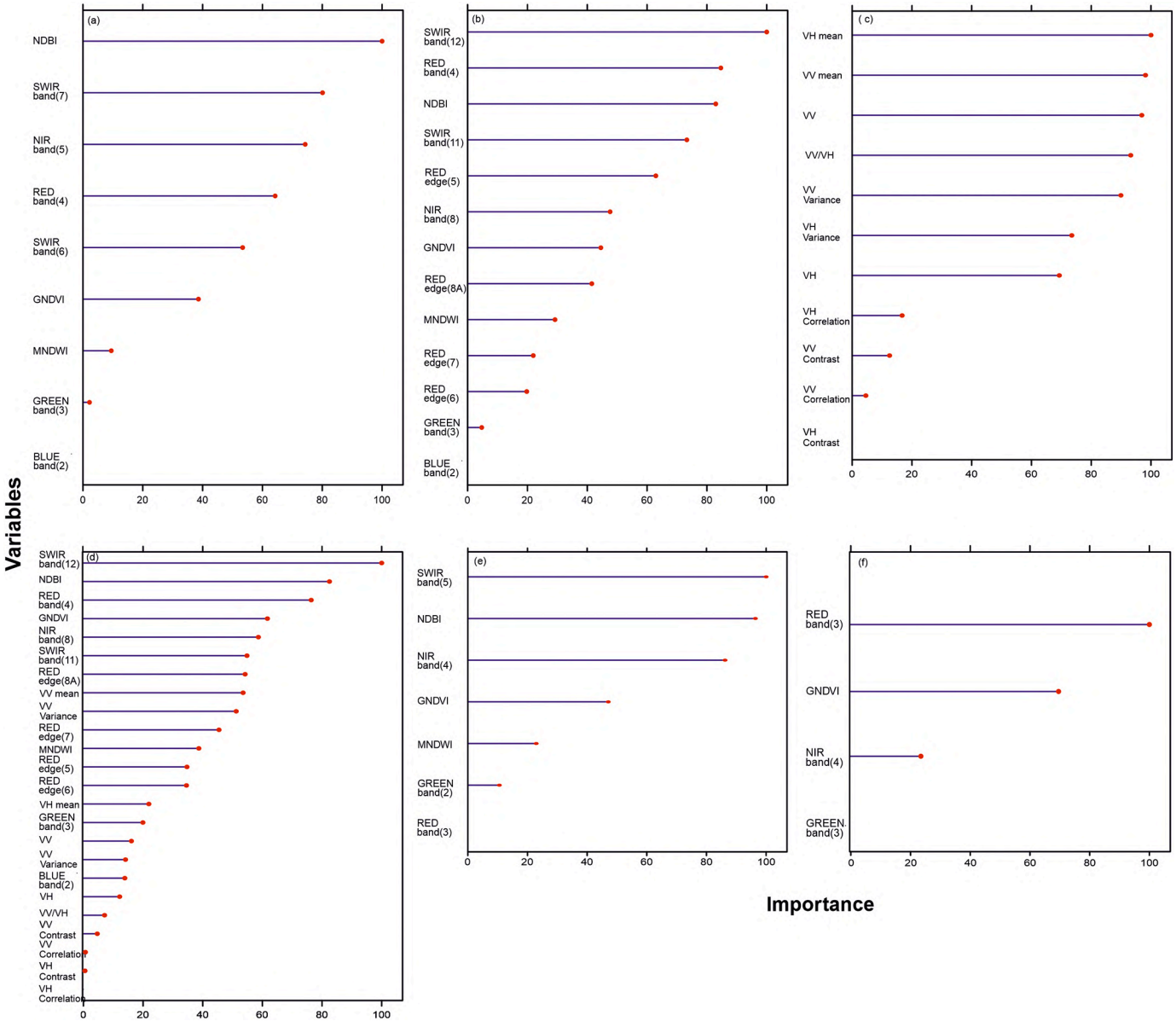


Fig. 3. Variables importance of each satellite sensor (a) Landsat-8, (b) Sentinel-2, (c) Sentinel-1, (d) combined Sentine-l-2, (e) LISS III, (f) LISS IV.

dimensional feature space based on the given K value (Avand et al., 2019). Euclidean distance method was used to calculate the distance between the points. In this study, tuning parameters were fixed in accordance with the highest precision of the model. The applied models were operated in R environment using several packages.

3.5. Accuracy assessment

The accuracy assessment is the vital step in LCLU classification to determine the appropriate method for sustainable land management (Deng et al., 2008). Among the accuracy assessment metrics, user's accuracy, producer's accuracy, overall accuracy, and kappa coefficient are widely used methods in LCLU mapping (Deus, 2016; Steinhausen et al., 2018). The information of these measurements was available in Noi and Kappas (2018); and Whyte et al. (2018). The accuracy for each class was calculated from the tabulated square confusion matrix of the LCLU classes.

4. Results

4.1. Variable importance metrics

In LCLU classification, bands, polarization, and indices play a key role to distinguish different classes. In Fig. 3, the contributing factors of the multi-satellite sensors have been shown using the variable important function of RF model. For L-8, the influencing variables in ascending order were NDBI, SWIR-2, NIR, red, SWIR-1, GNDVI, MNDWI, green, and blue (Fig. 3a). Among the thirteen variables of S-2 imagery, SWIR-2, Red, NDBI, were exceptionally important for classification followed by SWIR-1, Red edge (5), NIR, GNDVI, vegetation red, MNDWI, Red edge (7), Red edge (6), Green, and blue (Fig. 3b). Regarding S-1 data, VH-mean was of the highest significance succeeded by VV-mean, VV, VV/VH, VV variance, VH variance, VH, VH correlation, VV, contrast, VV correlation, VH contrast (Fig. 3c). In the case of combined of S-1-2 factors, the most affecting variables were SWIR-2, NDBI, Red, GNDVI, NIR, SWIR-1, vegetation red, and VV mean whereas the least important variables were VH correlation, VH, contrast, VV correlation, VV contrast, VV/VH, VH, blue, and VH variance (Fig. 3d). The SWIR, NDBI, and NIR factors of the L-3 image were more effective to segregate the LCLU classes as compared to GNDVI, MNDWI, green, and RED (Fig. 3e). Out of four factors of L-4 data, red was the most crucial in terms of importance followed by GNDVI, NIR, and green (Fig. 3f). From the comprehensive results considering all the satellite sensors, it was seen that the factors SWIR, NDBI, red edge, NIR, red, VV and VH mean, VV, were more reliable than the blue, green, VH and VV contrast, VV and VH correlation variables were less significant for class recognization. Noticeably, SWIR was the only band that consistently appeared as the most influential factor (L-8, S-2, Combined S-1-2, L-3). In most of the LCLU studies (Abdi, 2020; Sinha and Tiwari, 2018; Tavares et al., 2019), SWIR band has been recognized as the principal determinant for improving the classification because its higher electromagnetic wavelength helped to discriminate the LCLU types. Here, it is worthy to note that the blue band failed to separate the classes because of its lower wavelengths.

4.2. Comparisons of LCLU classes

In the current study, six LCLU classes including forest, agricultural land, barren land, built-up, and wetlands were selected for mapping. The PA and UA of each LCLU types influenced the OA and kappa coefficient of the respective classes (Table 2 and Fig. 4). From this table, it can be observed that the LB model held the highest PA and UA with respect to almost all the satellite sensors. For this very reason, only the results of PA and UA from LB model have been considered to be discussed. The PA of the wetland class were 7.14% (S-1), 12.8% (L-3), 56.7% (S-1-2), 60% (S-2), 64.3 (L-8) and UA were 28.6% (L-3), 50% (S-1), 61.8% (S-1-2), 67.2% (L-8), 76.4 (S-2). In the case of water bodies, the highest PA value was same as UA value (84.5%) resulted from the integration of S-1-2 sensors which indicated the reference and classified data correctly identified the class. Regarding forest class, PA value crossed 85% mark (except S-1) and UA value was more than 70% (except L-3 and S-1). The built-up recorded its maximum PA and UA as much as 90.5% and 86.4% respectively, as a result of the combination of S-1-2 sensors whereas the lowest PA and UA were 21% and 27.6% respectively, from the L-4 sensor. For the barren land, PA values varied from 55.2% (L-3) to 82.8% (S-1-2) while UA values ranged from 57.5% (S-1-2) to 90.2% (L-3). In the agricultural class, PA was more than 70% for all the sensors where L-4 attained the maximum of 89.2% but UA was less accurate (62.3%) compared to the PA. All the classes except water body and wetland were well classified. Spectral reflectance values were incorporated to eliminate confusion between these LCLU classes. The similar result was found in the research of Shimabukuro et al. (2020) where forest land, agricultural land and water body were concordant with their spectral reflectance. On the other side, water body and wetland classes were incongruent with their spectral characteristics.

4.3. Comparison of the satellite sensors and classifiers

The selection of a significant satellite dataset is a fundamental task for effective image classification. In the present research, a comprehensive framework for LCLU classification using multi-sensor satellite images was developed. The derived results (UA, PA, OA, kappa coefficient) of the satellite sensors from the confusion matrix of the validation dataset using several machine learning models were summarized in Table 2 and Fig. 5. For the L-8 sensor, all the models (except KNN) have more or less similar OA (74.5–75.7%) and kappa coefficient (0.69–0.71). The S-2 sensor achieved the maximum OA (78.3) and kappa value (0.739) from the LB model while KNN obtained the minimum OA (68.9) and kappa value (0.627). Regarding SAR data of S-1 sensor, it was observed that the highest OA was 51.4% in LB model and lowest OA was 39.8% in ANN model. The results of the integration of S-1-2 sensors showed 81.7% OA (LB) followed by 80.6% (RF), 78.9% (SGB), 45.2% (KNN), and 39.4% (ANN) whereas interestingly, Kappa values of the models were in the same sequence as of OA i.e. 0.779 (LB), 0.767 (RF), 0.746 (SGB), 0.344 (KNN), and 0.282 (ANN). In the case of L-3 sensor, OA ranged from 54 to 60.8% while kappa coefficient varied from 0.451–0.527 across the employed models. In regards to L-4 sensor, OA values were confined within a range of 42.6% to 52% and kappa values lay between 0.31 and 0.420. From the results, it was found that the integration of S-1 and-2 resulted in the achievement of the highest OA and Kappa values succeeded by individual S-2, and L-8. On the other side, S-1, L-4 and L-3 were unfavorable based on the model accuracy. Among the applied models, the LB model had highest OA and kappa coefficient followed by SGB, RF, ANN, and KNN.

4.4. Land cover/land use (LCLU) map

LCLU map is required for various applications and preparing this type of map in heterogeneous landscape especially in coastal environment is a challenging task. In recent years, the remarkable advancement of data science and remote sensing technology has been instrumental in solving many environmental and socio-economic problems. Although, the geo-environmental conditions, characteristics of data, and purpose of application affect the suitability of any mapping method. For this, the current research applied diverse input data and different methods to prepare appropriate land use map. Keeping in view the results of the satellite datasets and methods, integrated S-1-2 as input data and LB, SGB, RF models were selected for LCLU mapping because of high performance (Fig. 6). The area under each class has been shown in Table 3. The minimal variation in area of the respective classes signifies the reliable performance of the models (Prasad et al., 2020). With respect to the three models, the intra-class areal differences lay between 3%–6%

Table 2
Different accuracies of the sensors in LCLU classification (6 classes).

Landsat-8

	RF		LB		SGB		ANN		KNN	
Classes	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA
Wetland	67.4	63.3	64.3	67.2	68.5	69.2	66.3	64.9	46.7	51.2
Water body	67.0	84.2	54.3	82.6	62.1	80.0	67.0	84.2	58.3	69.8
Forest	91.3	76.0	90.5	72.0	90.0	75.0	93.8	75.8	86.3	67.0
Built-up	74.2	73.4	78.9	81.2	80.7	79.8	81.7	86.4	64.5	63.2
Barren land	73.4	87.3	73.5	83.6	74.5	85.4	72.3	82.9	72.3	85.0
Agriculture	79.5	68.1	85.1	65.5	83.3	67.0	73.1	60.0	78.2	66.3
OA	74.8		74.5		75.7		75.2		66.8	
K	0.70		0.69		0.71		0.70		0.60	

Sentinel-2

	RF		LB		SGB		ANN		KNN	
Classes	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA
Wetland	64.1	66.3	60.0	76.4	63.0	64.4	57.6	69.7	47.8	59.5
Water body	85.4	83.2	84.4	80.9	84.5	78.4	86.4	84.0	78.6	73.6
Forest	88.8	83.5	86.8	81.5	85.0	82.9	87.5	74.5	83.8	72.0
Built-up	76.3	81.6	80.0	81.1	79.6	87.1	87.1	88.0	59.1	75.3
Barren land	70.2	84.6	75.6	75.6	73.4	85.2	75.5	84.5	67.0	74.1
Agriculture	79.5	65.2	80.6	73.0	80.8	69.2	74.4	65.9	79.5	59.0
OA	77.2		78.3		77.6		78.1		68.9	
K	0.727		0.739		0.731		0.738		0.627	

Sentinel-1

	RF		LB		SGB		ANN		KNN	
Classes	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA
Wetland	35.9	33.7	7.14	50.0	32.6	32.3	10.8	23.8	26.1	30.0
Water body	34.0	58.3	40.0	80.0	35.9	64.9	21.3	64.7	35.9	48.0
Forest	50.0	38.0	63.4	38.8	57.5	43.8	76.3	26.2	50.0	34.8
Built-up	61.3	50.9	63.0	44.7	59.1	49.1	32.3	58.8	45.2	43.8
Barren land	59.8	70.9	56.3	69.0	64.9	68.5	42.5	70.2	60.6	74.0
Agriculture	52.6	47.7	73.7	44.7	55.1	51.1	66.7	42.3	61.5	50.5
OA	48.5		51.4		50.3		39.8		45.9	
K	0.383		0.409		0.406		0.286		0.353	

Sentinel-1 and-2

	RF		LB		SGB		ANN		KNN	
Classes	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA
Wetland	66.3	68.5	56.7	61.8	63.0	62.4	0	0	25.0	29.5
Water body	82.5	86.7	84.5	84.5	82.5	80.2	26.2	60.0	36.9	50.0
Forest	90.0	80.0	90.4	80.5	85.0	84.0	93.8	23.0	48.8	33.0
Built-up	88.2	83.7	90.5	86.4	85.0	89.8	0	0	41.9	43.8
Barren land	74.5	88.6	82.8	85.7	80.9	83.5	71.3	65.0	61.7	69.9
Agriculture	83.3	75.6	78.1	84.8	76.9	74.1	56.4	67.7	60.2	49.0
OA	80.6		81.7		78.9		39.4		45.2	
K	0.767		0.779		0.746		0.282		0.344	

LISS III

	RF		LB		SGB		ANN		KNN	
Classes	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA
Wetland	32.6	38.4	12.8	28.6	36.9	40.9	23.9	30.9	36.9	43.6
Water body	37.8	63.9	57.9	70.2	39.8	70.6	25.2	66.6	42.7	55.7
Forest	82.5	55.6	88.7	58.2	82.5	55.9	83.7	51.1	76.2	58.7
Built-up	58.0	51.9	58.3	50.9	57.0	48.6	61.2	45.6	50.5	45.2
Barren land	64.8	78.2	55.2	90.2	59.6	80.0	61.7	82.9	61.7	78.4
Agriculture	75.6	59.0	86.3	58.1	80.7	61.7	61.7	59.6	74.4	57.4
OA	57.2		60.8		58.0		54.0		55.9	
K	0.484		0.527		0.497		0.451		0.472	

LISS IV

	RF		LB		SGB		ANN		KNN	
Classes	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA

(*continued on next page*)

Table 2 (*continued*)

LISS IV										
	RF		LB		SGB		ANN		KNN	
Classes	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA
Wetland	22.8	22.8	9.80	45.5	22.8	27.6	27.2	26.0	15.2	20.9
Water body	37.8	51.3	49.1	51.9	38.8	54.8	34.0	70.0	39.8	56.2
Forest	80.0	54.7	83.9	52.2	80.0	52.9	90.0	50.4	78.8	50.0
Built-up	17.2	23.2	21.0	27.6	20.4	36.5	10.7	34.4	19.4	34.0
Barren land	50.0	50.5	60.5	57.5	59.6	51.9	63.8	58.8	56.4	50.5
Agriculture	55.1	46.2	89.2	62.3	73.1	51.8	70.5	45.1	74.4	50.0
OA	42.6		52.0		47.6		47.6		45.7	
K	0.313		0.420		0.373		0.375		0.351	

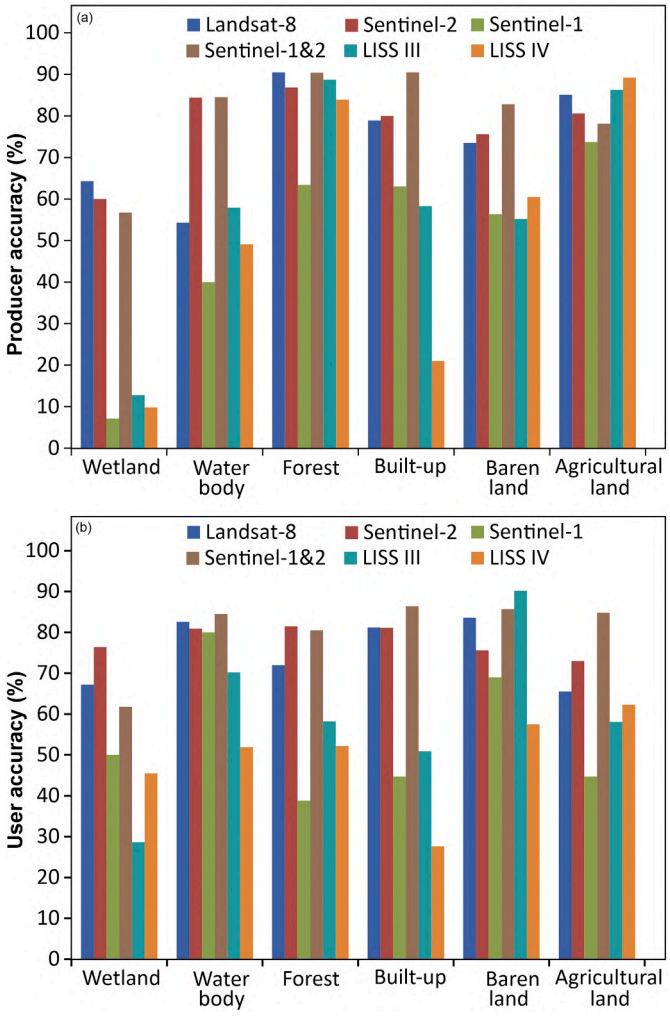


Fig. 4. (a) Producer and (b) User accuracy for each LCLU class of different satellite sensors from LB model.

which justified the selection of these models for LCLU mapping.

5. Discussion

In the previous works, researchers have attempted to compare different satellite sensors for LCLU classification (Chander et al., 2007; Deng et al., 2008; Denize et al., 2019; Whyte et al., 2018; Mansaray et al., 2020; Ghayour et al., 2021). These studies are confined to only two or three sensors. In the present research, five satellite sensors including optical and microwave sensors with low to high spatial and spectral resolutions were taken for comparative assessment. Each sensor has unique properties which influence the classification accuracy. In LCLU mapping, spatial and spectral resolutions are the main controlling factors that enhance the overall accuracy whereas temporal resolution is important for change detection mapping. Spatial resolution gives information about the extent of features while spectral resolution provides information of the condition of features (Lu and Weng, 2007; Pandey et al., 2021). The selection and accuracy of the sensors primarily depend on the spatial complexity of the area of investigation, image resolution (spatial and spectral), training size, and degree of autocorrelation between land use classes (Chen et al., 2004; Chen and Stow, 2002). Spatial complexity is mainly scale dependent as it changes with generalization of spatial and spectral information of the earth surface (Cola, 1994; Papadimitriou, 2020). Mishra et al. (2019) used five different sensors for evaluation of textural features in improving LCLU classification accuracy of heterogeneous landscape applying supervised SVM classifier. For spatially heterogeneous classes, a comparatively large number of training pixels are needed for the purpose of deriving representative training statistics. On the other hand, a small number of pixels may be adequate for spectrally homogeneous classes. In general, spatial resolution of the image affects the spectral heterogeneity since it generates mixed pixels. The number of these mixed pixels tends to increase with coarser spatial resolution of the image and vice versa (Rocchini, 2007). The comparative studies of L-8 and S-2 showed that the S-2 was more advantageous in contrast with L-8 sensor (Forkuor et al., 2018; Ghayour et al., 2021; Sekertekin et al., 2017; Topaloğlu et al., 2016). It matched with the present results where S-2 dataset had slightly higher accuracy (2 to 4%) than L-8 based on applied machine learning models. Forkuor et al. (2018) compared the spectral bands of L-8 and S-2, and proposed that the narrower bands of S-2 facilitated the separation of LCLU classes. Besides, presence of three red edge bands and medium-high spatial resolution could enhance the ability of detection and discrimination of specific land surface properties. The results demonstrated that the higher accuracy of S-2 was due to higher spectral (12 bands) and spatial (10 m) resolutions. On the other hand, despite having higher spatial resolution, L-4 (5.8 m) and L-3 (24 m) datasets failed to increase accuracy because of low spectral resolution (3 and 4 bands, respectively). On the contrary, L-8 (30 m) produced higher accuracy (15–22% than L-3 and L-4) owing to availability of more spectral bands (6). The comparative results of L-4 and L-3 highlighted that the L-3 had better precision which attributed to the presence of SWIR band (the most influential factor). In addition, the availability of SWIR band helped built two indices (MNDWI and NDBI). But L-4 does not have this feature. From the research of Aplin (2003), it was evident that the very high resolution data may not be able to differentiate the classes in complex landscapes. In the case of independent SAR data, accuracy was very low by reason of lowest sensitivity in identifying the class types. The heterogeneity and complexity of the earth surface render microwave data less desirable in comparison with optical data for LCLU study. However, microwave data is sensitive to structural features of the terrain which can be helpful to interpret the LCLU types (Henderson and Xia, 1997; Whyte et al., 2018). Since, the application of the optical or radar dataset leads to misclassification, they may not improve the precision of model alone (Deus,

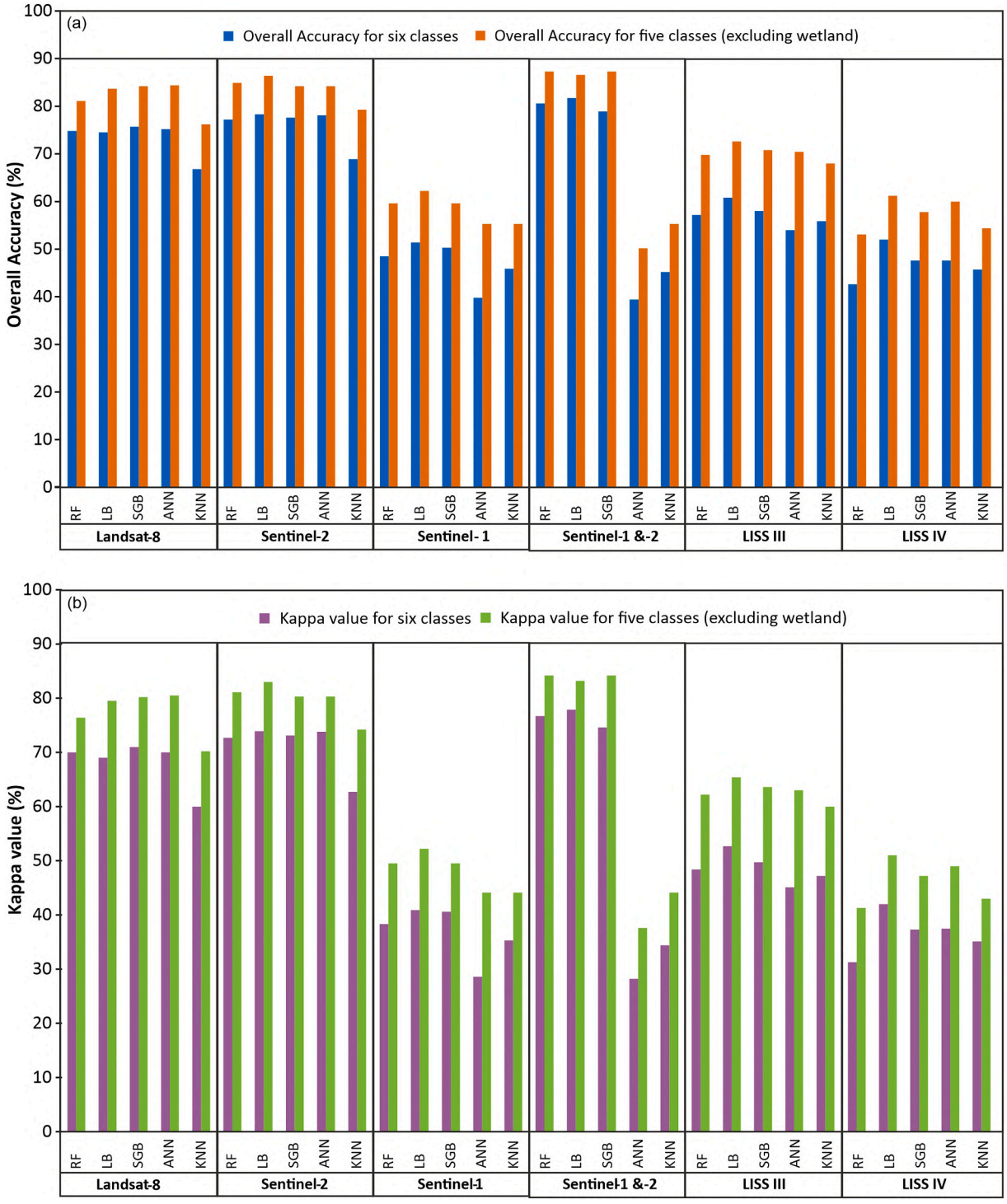


Fig. 5. (a) Overall accuracy and (b) Kappa coefficient of each satellite sensor using various machine learning models.

2016; Lu et al., 2011). Besides, integration of multi-sensor data effectively discriminate different classes, produce detailed feature information, and adds missing information which enhances the model accuracy (Clerici et al., 2017; Deus, 2016; Mansaray et al., 2020; Pandey et al., 2021; Tavares et al., 2019). In the study of Clerici et al. (2017), overall accuracy was 30% from S-1 data, 72.5% from S-2 data and 88.75% when it combined. In the same vein of the previous research, the present study also achieved maximum accuracy combining S-1-2.

In spite of wide accessibility of the satellite images from various earth observing sensors, there is a strong need to examine the robustness and advantages of the machine learning classifiers in LCLU mapping. Each Machine learning model applies different statistical, optimisation and probabilistic methods to make the effective pattern for predicting from a large, complex and unstructured datasets (Uddin et al., 2019). The precision of the models to some extent depends on the quantity and quality of the input datasets (Prasad et al., 2021b). In the current research five machine learning models were applied to select the appropriate model for each satellite sensor. All the models used the 10 cross validation method with five repetitions to avoid bias and error. In the comparative studies of machine learning algorithms, a model is said

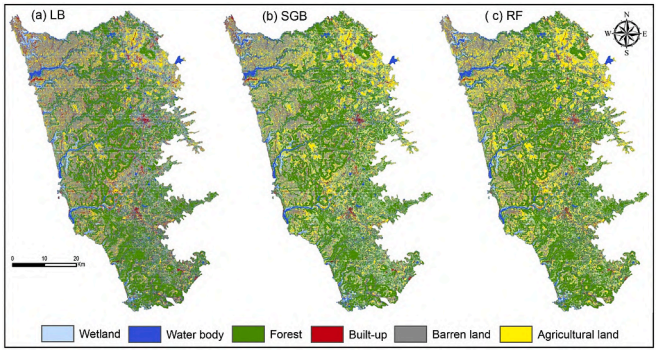


Fig. 6. LCLU maps derived from the combined Sentinel-1-2 dataset employing the (a) LB, (b) SGB, and (c) RF models.

Table 3
Percentage of area under different LCLU classes.

Area/Classes	Area in % (SGB)	Area in % (RF)	Area in % (LB)
Wetland	24.51	22.64	20.82
Water body	8.73	8.31	11.66
Forest	35.60	37.07	32.25
Built-up	5.23	6.82	8.40
Barren land	12.23	10.37	17.37
Agricultural land	13.70	14.80	9.49

to be robust if it performs in a consistent manner (Abdi, 2020). The outputs of LB, SGB, and RF models were consistent with least variation in accuracy while ANN and KNN perform inconsistently with each satellite sensor. The better performance of the LB model may be because of optimizing the multinomial likelihood which facilitates to figure out the multiclass problems (Cai et al., 2006; Pham et al., 2016; Sun et al., 2014). Perhaps, LB model has not been used before in LCLU classification. However, LB model was employed in natural hazard mapping with high precision (Al-Najjar et al., 2019; Fadhillah et al., 2020; Kadavi et al., 2018; Oh et al., 2019). In contrast, the ANN model poorly performed with S-1 and combined of S-1-2 datasets. In the other studies also (Navale and Haldar, 2019; Zhang and Xu, 2018), ANN had failed to gain better accuracy from the SAR data. In these studies, water body and built-up classes were totally misclassified by the integrated of S-1-2 datasets which made ANN model reduce the accuracy. The underperformance of the KNN model for each of the dataset is attributed to inability of the algorithm to properly train the training data. The performance of the models may vary with the study region because of diverse spatial complexity and also with the variety of input data. In the study of Prasad et al. (2020) the efficiency of the applied models in primary research area remained almost unchanged in other alike geo-environmental regions too.

Among the LCLU classes of the study area, wetland has been highly misclassified. According to Lyon (2001) three indicators should be considered for the delineation of wetland. These criteria are: (a) the soils must be hydric or waterlogged, (b) the soils must exhibit indicators of wetland hydrologic conditions resulting from flooding or ponding of water and/or saturation of soils, and (c) the typical wetland plants must comprise 50% or more of the plants found on the site. Junk et al. (2014) defined the boundary of wetland by the margin of waterlogged or fluctuation of water levels or permanently flooded region with wetland plants. In general, among the LCLU classes, water body has more or less similar spectral information due to homogenous characteristics which produced very high accuracy (Ghayour et al., 2021; Shen et al., 2020). The discrimination of water-related classes such as water body, wetland are difficult because of similar kinds of spectral responses that led to less accuracy (Whyte et al., 2018). These classes are very sensitive particularly in the coastal area for tidal and wave influences. The reference data of each class and satellite datasets were acquired in different times that caused ambiguous classification of water body and wetland. In spite of this, speckle is a common issue in monotypic cover types and mixed areas of wetland class because of intrinsic spectral variation, diversity of plant species, and fine scale patchiness (Dronova, 2015). There remains a certain amount of uncertainty in LCLU classification but identifying and understanding the problem is necessary (Canters, 1997; Lu and Weng, 2007). That's why, further investigation of LCLU classes excluding wetland class was conducted to reexamine the accuracy of the classes. Table 3S depicts that the PA, UA, OA and kappa coefficient increased to a great extent. These results confirmed that the misclassification of the wetland reduced the overall accuracy of the classification. Such shortcoming urges the need for a separate wetland map using high spatial and spectral resolution and large number of input features (plant species of wetland and water samples) for better training of the models. Rocchini (2007) has investigated the impacts of various spectral and spatial resolutions (Quickbird, Aster, and Landsat ETM+) on plant species richness assessment in wetland area of Italy. He concluded that detailed scale (spatial and spectral) research is appropriate for higher spatial complexity region such as wetland area.

In the present study, the validity of the sensors has not been tested in other regions. Thus, more researches are needed to appraise the efficiency of integrated of S-1 and S-2 for LCLU classification. The current research only deals with post-monsoon satellite images. Therefore, it is recommended that the further research can be carried out with multitemporal images of the study region to verify the model and sensor precision in LCLU mapping.

6. Conclusion

This study is the first of its kind to assess and compare a total of five satellite sensors including optical and SAR data using five machine learning algorithms for LCLU mapping. Besides, to the best of the authors' knowledge, LB model has not been applied before in land use study. It is clear from the sensor's suitability assessment that the integration of S-1-2 datasets produced the best classification accuracy while unsatisfactory result came from the S-1, L-4, and L-3 datasets. This comparative result demonstrates that the multiple satellite data influenced the processing unevenly because of having different spatial and spectral resolutions, eventually resulting in the change of classification accuracy of the models. It is evident from the findings of the selected models performance, LB model consistently (Prasad et al., 2021b) outperformed the other models in case of all the datasets for multi-class LCLU mapping. The result also points out the limitations of the applied models to identify the wetland class distinguishing it from water bodies in coastal region. In terms of variable importance, SWIR band was the most influential factor for LCLU classification while blue band was the least significant variable.

In future, further research can be continued in this direction including hyperspectral satellite data together with topographical data for the improvement in LCLU class discrimination and mapping for regional environmental monitoring.

The LCLU map of this study can be helpful as input data for various purposes for example coastal zone management, land suitability and capability mapping, and also for hydrological modeling to support flood management. Furthermore, the methodological framework of this research may be useful in different parts of the world for LCLU mapping.

Data availability statement

The additional dataset is included in the supplementary section.

Declaration of Competing Interest

All authors are declared no actual or potential conflict of interest including any financial, personal or other relationships with other

people or organizations.

Acknowledgements

We acknowledge the financial support from the Mecon Limited (Project no. SSP-3249) and University Grant Commission (3160 NET-June 2015). We also acknowledge the NASA, ESA, and ISRO for providing the satellite data. The authors are thankful to the Director CSIR-NIO for time to support. Field assistance from the survey team members is thankfully acknowledged. The NIO contribution number is 6844.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2021.101522.

References

Abdi, A.M., 2020. Land cover and land use classification performance of machine learning algorithms in a boreal landscape using Sentinel-2 data. GISci. Remote Sens. 57 (1), 1–20. https://doi.org/10.1080/15481603.2019.1650447.

Adam, E., Mutanga, O., Odindi, J., Abdel-Rahman, E.M., 2014. Land-use/cover classification in a heterogeneous coastal landscape using RapidEye imagery: evaluating the performance of random forest and support vector machines classifiers. Int. J. Remote Sens. 35 (10), 3440–3458. https://doi.org/10.1080/01431161.2014.903435.

Al-Najjar, H.A., Kalantar, B., Pradhan, B., Saeidi, V., 2019, October. Conditioning factor determination for mapping and prediction of landslide susceptibility using machine learning algorithms. In: Earth Resources and Environmental Remote Sensing/GIS Applications X, vol. 11156. International Society for Optics and Photonics, p. 111560K.

Anderson, J.R., 1976. A Land Use and Land Cover Classification System for Use with Remote Sensor Data, vol. 964. US Government Printing Office.

Aplin, P., 2003. Comparison of simulated IKONOS and SPOT HRV imagery for classifying urban areas. Remot. Sens. Cities 23–45.

Arabameri, A., Rezaei, K., Pourghasemi, H.R., Lee, S., Yamani, M., 2018. GIS-based gully erosion susceptibility mapping: a comparison among three data-driven models and AHP knowledge-based technique. Environ. Earth Sci. 77 (17), 628. https://doi.org/10.1007/s12665-018-7808-5.

Avand, M., Janizadeh, S., Naghibi, S.A., Pourghasemi, H.R., Khosrobeigi Bozchaloei, S., Blaschke, T., 2019. A comparative assessment of random forest and k-nearest neighbor classifiers for gully erosion susceptibility mapping. Water 11 (10), 2076. https://doi.org/10.3390/w11102076.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Bui, D.T., Shirzadi, A., Shahabi, H., Geertsema, M., Omidvar, E., Clague, J.J., Thai Pham, B., Dou, J., Talebpour Asl, D., Bin Ahmad, B., Lee, S., 2019. New ensemble models for shallow landslide susceptibility modeling in a semi-arid watershed. Forests 10 (9), 743. https://doi.org/10.3390/f10090743.

Cai, Y.D., Feng, K.Y., Lu, W.C., Chou, K.C., 2006. Using LogitBoost classifier to predict protein structural classes. J. Theor. Biol. 238 (1), 172–176. https://doi.org/10.1016/j.jtbi.2005.05.03.

Canters, F., 1997. Evaluating the uncertainty of area estimates derived from fuuy land-cover classification. Photogramm. Eng. Remote. Sens. 63 (4), 403–414.

Chander, G., Coan, M.J., Scaramuzza, P.L., 2007. Evaluation and comparison of the IRS-P6 and the Landsat sensors. IEEE Trans. Geosci. Remote Sens. 46 (1), 209–221. https://doi.org/10.1109/TGRS.2007.907426.

Chapi, K., Singh, V.P., Shirzadi, A., Shahabi, H., Bui, D.T., Pham, B.T., Khosravi, K., 2017. A novel hybrid artificial intelligence approach for flood susceptibility assessment. Environ. Model. Softw. 95, 229–245. https://doi.org/10.1016/j.envsoft.2017.06.012.

Chauhan, H.B., Dwivedi, R.M., 2008. Inter sensor comparison between RESOURCESAT LISS III, LISS IV and AWiFS with reference to coastal landuse/landcover studies. Int. J. Appl. Earth Obs. Geoinf. 10 (2), 181–185. https://doi.org/10.1016/j.jag.2007.10.007.

Chen, D., Stow, D., 2002. The effect of training strategies on supervised classification at different spatial resolutions. Photogramm. Eng. Remote. Sens. 68 (11), 1155–1162.

Chen, D., Stow, D.A., Gong, P., 2004. Examining the effect of spatial resolution and texture window size on classification accuracy: an urban environment case. Int. J. Remote Sens. 25 (11), 2177–2192. https://doi.org/10.1080/01431160310001618464.

Chen, W., Panahi, M., Khosravi, K., Pourghasemi, H.R., Rezaie, F., Parvinnezhad, D., 2019. Spatial prediction of groundwater potentiality using ANFIS ensembled with teaching-learning-based and biogeography-based optimization. J. Hydrol. 572, 435–448. https://doi.org/10.1016/j.jhydrol.2019.03.013.

Chen, W., Chen, Y., Tsangaratos, P., Ilia, I., Wang, X., 2020. Combining evolutionary algorithms and machine learning models in landslide susceptibility assessments. Remote Sens. 12 (23), 3854. https://doi.org/10.3390/rs12233854.

Clerici, N., Valbuena Calderón, C.A., Posada, J.M., 2017. Fusion of sentinel-1A and sentinel-2A data for land cover mapping: a case study in the lower Magdalena region, Colombia. J. Maps 13 (2), 718–726. https://doi.org/10.1080/17445647.2017.1372316.

Cola, L.D., 1994. Simulating and mapping spatial complexity using multi-scale techniques. Int. J. Geogr. Inf. Syst. 8 (5), 411–427. https://doi.org/10.1080/02693799408902011.

Deng, J.S., Wang, K., Deng, Y.H., Qi, G.J., 2008. PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data. Int. J. Remote Sens. 29 (16), 4823–4838. https://doi.org/10.1080/01431160801950162.

Deng, J.S., Wang, K., Hong, Y., Qi, J.G., 2009. Spatio-temporal dynamics and evolution of land use change and landscape pattern in response to rapid urbanization. Landsc. Urban Plan. 92 (3–4), 187–198. https://doi.org/10.1016/j.landurbplan.2009.05.001.

Denize, J., Hubert-Moy, L., Betbeder, J., Corgne, S., Baudry, J., Pottier, E., 2019. Evaluation of using sentinel-1 and-2 time-series to identify winter land use in agricultural landscapes. Remote Sens. 11 (1), 37. https://doi.org/10.3390/rs11010037.

Deus, D., 2016. Integration of ALOS PALSAR and landsat data for land cover and forest mapping in northern tanzania. Land 5 (4), 43. https://doi.org/10.3390/land5040043.

Diengdoh, V.L., Ondei, S., Hunt, M., Brook, B.W., 2020. A validated ensemble method for multinomial land-cover classification. Ecol. Inform. 56, 101065.

District Mining Officer, 2017. District Survey Report, Maharashtra.

Dronova, I., 2015. Object-based image analysis in wetland research: a review. Remote Sens. 7 (5), 6380–6413. https://doi.org/10.3390/rs70506380.

Fadhillah, M.F., Achmad, A.R., Lee, C.W., 2020. Integration of InSAR time-series data and GIS to assess land subsidence along Subway lines in the Seoul metropolitan area, South Korea. Remote Sens. 12 (21), 3505. https://doi.org/10.3390/rs12213505.

Filipponi, F., 2019. Sentinel-1 GRD preprocessing workflow. Multidiscip. Dig. Publish. Inst. Proceed. 18 (1), 11.

Forkuor, G., Dimobe, K., Serme, I., Tondoh, J.E., 2018. Landsat-8 vs. Sentinel-2: examining the added value of sentinel-2's red-edge bands to land-use and land-cover mapping in Burkina Faso. GISci. Remote Sens. 55 (3), 331–354. https://doi.org/10.1080/15481603.2017.1370169.

Friedman, J.H., 2002. Stochastic gradient boosting. Comput. Stat. Data Anal. 38 (4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2.

Friedman, J.H., Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). Ann. Stat. 28 (2), 337–407.

Gemitzi, A., 2021. Predicting land cover changes using a CA Markov model under different shared socioeconomic pathways in Greece. GISci. Remote Sens. 1-17 https://doi.org/10.1080/15481603.2021.1885235.

Ghayour, L., Neshat, A., Paryani, S., Shahabi, H., Shirzadi, A., Chen, W., Al-Ansari, N., Geertsema, M., Pourmehdi Amiri, M., Gholamnia, M., Dou, J., 2021. Performance evaluation of Sentinel-2 and Landsat 8 OLI data for land cover/use classification using a comparison between machine learning algorithms. Remote Sens. 13 (7), 1349. https://doi.org/10.3390/rs13071349.

Gong, B., Im, J., Mountrakis, G., 2011. An artificial immune network approach to multi-sensor land use/land cover classification. Remote Sens. Environ. 115 (2), 600–614. https://doi.org/10.1016/j.rse.2010.10.005.

Güler, M., Yomralıoğlu, T., Reis, S., 2007. Using landsat data to determine land use/land cover changes in Samsun, Turkey. Environ. Monit. Assess. 127 (1), 155–167. https://doi.org/10.1007/s10661-006-9270-1.

Henderson, F.M., Xia, Z.G., 1997. SAR applications in human settlement detection, population estimation and urban land use pattern analysis: a status report. IEEE Trans. Geosci. Remote Sens. 35 (1), 79–85.

Hird, J.N., DeLancey, E.R., McDermid, G.J., Kariyeva, J., 2017. Google earth engine, open-access satellite data, and machine learning in support of large-area probabilistic wetland mapping. Remote Sens. 9 (12), 1315. https://doi.org/10.3390/rs9121315.

Junk, W.J., Piedade, M.T.F., Lourival, R., Wittmann, F., Kandus, P., Lacerda, L.D., Bozelli, R.L., Esteves, F.D.A., Nunes da Cunha, C., Maltchik, L., Schöngart, J., 2014. Brazilian wetlands: their definition, delineation, and classification for research, sustainable management, and protection. Aquat. Conserv. Mar. Freshwat. Ecosyst. 24 (1), 5–22. https://doi.org/10.1002/aqc.2386.

Kadavi, P.R., Lee, C.W., Lee, S., 2018. Application of ensemble-based machine learning models to landslide susceptibility mapping. Remote Sens. 10 (8), 1252. https://doi.org/10.3390/rs10081252.

Kalantar, B., Pradhan, B., Naghibi, S.A., Motevalli, A., Mansor, S., 2018. Assessment of the effects of training data selection on the landslide susceptibility mapping: a comparison between support vector machine (SVM), logistic regression (LR) and artificial neural networks (ANN). Geom. Nat. Haz. Risk 9 (1), 49–69. https://doi.org/10.1080/19475705.2017.1407368.

Kandrika, S., Roy, P.S., 2008. Land use land cover classification of Orissa using multi-temporal IRS-P6 awifs data: a decision tree approach. Int. J. Appl. Earth Obs. Geoinf. 10 (2), 186–193. https://doi.org/10.1016/j.jag.2007.10.003.

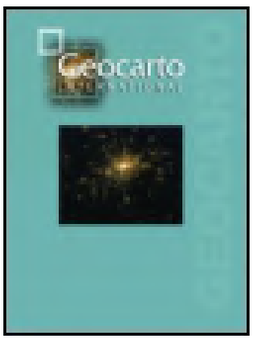
Khosravi, K., Pham, B.T., Chapi, K., Shirzadi, A., Shahabi, H., Revhaug, I., Prakash, I., Bui, D.T., 2018. A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at Haraz watershed, northern Iran. Sci. Total Environ. 627, 744–755. https://doi.org/10.1016/j.scitotenv.2018.01.266.

Lawrence, R., Bunn, A., Powell, S., Zambon, M., 2004. Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. Remote Sens. Environ. 90 (3), 331–336. https://doi.org/10.1016/j.rse.2004.01.007.

Letourneau, A., Verburg, P.H., Stehfest, E., 2012. A land-use systems approach to represent land-use dynamics at continental and global scales. Environ. Model. Softw. 33, 61–79. https://doi.org/10.1016/j.envsoft.2012.01.007.

Lillesand, T., Kiefer, R.W., Chipman, J., 2008. Remote Sensing and Image Interpretation. John Wiley & Sons.
Lu, D., Weng, Q., 2007. A survey of image classification methods and techniques for improving classification performance. Int. J. Remote Sens. 28 (5), 823–870. https://doi.org/10.1080/01431160600746456.
Lu, D., Li, G., Moran, E., Dutra, L., Batistella, M., 2011. A comparison of multisensor integration methods for land cover classification in the Brazilian Amazon. GISci. Remote Sens. 48 (3), 345–370. https://doi.org/10.2747/1548-1603.48.3.345.
Lyon, J.G., 2001. Wetland Landscape Characterization: GIS, Remote Sensing and Image Analysis. An Arbor Press.
Mansaray, L.R., Wang, F., Huang, J., Yang, L., Kanu, A.S., 2020. Accuracies of support vector machine and random forest in rice mapping with sentinel-1A, Landsat-8 and sentinel-2A datasets. Geocarto Int. 35 (10), 1088–1108.
Maxwell, A.E., Warner, T.A., Strager, M.P., 2016. Predicting palustrine wetland probability using random forest machine learning and digital elevation data-derived terrain variables. Photogramm. Eng. Remote. Sens. 82 (6), 437–447. https://doi.org/10.14358/PERS.82.6.437.
Ministry of Environment and Forests, 2010. Wetlands Conservation and Management. https://sindhudurg.nic.in/en/document-category/wetland/.
Mishra, V.N., Prasad, R., Rai, P.K., Vishwakarma, A.K., Arora, A., 2019. Performance evaluation of textural features in improving land use/land cover classification accuracy of heterogeneous landscape using multi-sensor remote sensing data. Earth Sci. Inf. 12 (1), 71–86. https://doi.org/10.1007/s12145-018-0369-z.
Moisen, G.G., Freeman, E.A., Blackard, J.A., Frescino, T.S., Zimmermann, N.E., Edwards Jr., T.C., 2006. Predicting tree species presence and basal area in Utah: a comparison of stochastic gradient boosting, generalized additive models, and tree-based methods. Ecol. Model. 199 (2), 176–187. https://doi.org/10.1016/j.ecolmodel.2006.05.021.
Motevalli, A., Naghibi, S.A., Hashemi, H., Berndtsson, R., Pradhan, B., Gholami, V., 2019. Inverse method using boosted regression tree and k-nearest neighbor to quantify effects of point and non-point source nitrate pollution in groundwater. J. Clean. Prod. 228, 1248–1263. https://doi.org/10.1016/j.jclepro.2019.04.293.
Naghibi, S.A., Dashtpagerdi, M.M., 2017. Evaluation of four supervised learning methods for groundwater spring potential mapping in Khalkhal region (Iran) using GIS-based features. Hydrogeol. J. 25 (1), 169–189. https://doi.org/10.1007/s10040-016-1466-z.
Naghibi, S.A., Dolatkordestani, M., Rezaei, A., Amouzegari, P., Heravi, M.T., Kalantar, B., Pradhan, B., 2019. Application of rotation forest with decision trees as base classifier and a novel ensemble model in spatial modeling of groundwater potential. Environ. Monit. Assess. 191 (4), 248. https://doi.org/10.1007/s10661-019-7362-y.
Navale, A., Haldar, D., 2019. Evaluation of machine learning algorithms to sentinel SAR data. Spat. Inf. Res. 1-11 https://doi.org/10.1007/s41324-019-00296-8.
Nguyen, L.H., Henebry, G.M., 2019. Characterizing land use/land cover using multi-sensor time series from the perspective of land surface phenology. Remote Sens. 11 (14), 1677. https://doi.org/10.3390/rs11141677.
Noi, T.P., Kappas, M., 2018. Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery. Sensors 18 (1), 18. https://doi.org/10.3390/s18010018.
Oh, H.J., Syifa, M., Lee, C.W., Lee, S., 2019. Land subsidence susceptibility mapping using bayesian, functional, and meta-ensemble machine learning models. Appl. Sci. 9 (6), 1248. https://doi.org/10.3390/app9061248.
Ololade, O., Annegarn, H.J., Limpitlaw, D., Kneen, M.A., 2008, July. Land-use/cover mapping and change detection in the Rustenburg mining region using Landsat images. In: IGARSS 2008-2008 IEEE International Geoscience and Remote Sensing Symposium (Vol. 4, pp. IV-818). IEEE. https://doi.org/10.1109/IGARSS.2008.4779848.
Pandey, P.C., Koutsias, N., Petropoulos, G.P., Srivastava, P.K., Ben Dor, E., 2021. Land use/land cover in view of earth observation: data sources, input dimensions, and classifiers—a review of the state of the art. Geocarto Int. 1-32 https://doi.org/10.1080/10106049.2019.1629647.
Panigrahy, R.K., Ray, S.S., Panigrahy, S., 2009. Study on the utility of IRS-P6 AWIFS SWIR band for crop discrimination and classification. J. Ind. Soc. Remote Sens. 37 (2), 325–333. https://doi.org/10.1007/s12524-009-0026-6.
Papadimitriou, F., 2020. Spatial Complexity: Theory Mathematical Methods and Applications. Springer.
Pavanelli, J.A.P., Santos, J.R.D., Galvão, L.S., Xaud, M., Xaud, H.A.M., 2018. PALSAR-2/ALOS-2 and OLI/LANDSAT-8 data integration for land use and land cover mapping in northern Brazilian Amazon. Bol. Ciênc. Geodésicas 24 (2), 250–269. https://doi.org/10.1590/s1982-21702018000200017.
Pham, B.T., Bui, D.T., Dholakia, M.B., Prakash, I., Pham, H.V., 2016. 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. Geotech. Geol. Eng. 34 (6), 1807–1824. https://doi.org/10.1007/s10706-016-9990-0.
Pourghasemi, H.R., Sadhasivam, N., Kariminejad, N., Collins, A.L., 2020. Gully erosion spatial modelling: role of machine learning algorithms in selection of the best controlling factors and modelling process. Geosci. Front. https://doi.org/10.1016/j.gsf.2020.03.005.
Prasad, Pankaj, Loveson Joseph, Victor, Chandra, Priyankar, Kotha, Mahender, 2021b. Artificial intelligence approaches for spatial prediction of landslides in mountainous regions of western India. Environmental Earth Sciences 80 (21), 1–20. https://doi.org/10.1007/s12665-021-10033-w.
Prasad, P., Loveson, V.J., Das, B., Kotha, M., 2021a. Novel ensemble machine learning models in flood susceptibility mapping. Geocarto Int. 1-23 https://doi.org/10.1080/10106049.2021.1892209.
Prasad, P., Loveson, V.J., Kotha, M., Yadav, R., 2020. Application of machine learning techniques in groundwater potential mapping along the west coast of India. GISci. Remote Sens. 57 (6), 735–752. https://doi.org/10.1080/15481603.2020.1794104.
Rocchini, D., 2007. Effects of spatial and spectral resolution in estimating ecosystem α-diversity by satellite imagery. Remote Sens. Environ. 111 (4), 423–434. https://doi.org/10.1016/j.rse.2007.03.018.
Schirpke, U., Leitinger, G., Tappeiner, U., Tasser, E., 2012. SPA-LUCC: developing land-use/cover scenarios in mountain landscapes. Ecol. Inform. 12, 68–76.
Sekertekin, A., Marangoz, A.M., Akcin, H., 2017. Pixel-based classification analysis of land use land cover using Sentinel-2 and Landsat-8 data. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci 42, 91–93.
Shahabi, H., Shirzadi, A., Ghaderi, K., Omidvar, E., Al-Ansari, N., Clague, J.J., Geertsema, M., Khosravi, K., Amini, A., Bahrami, S., Rahmati, O., 2020. Flood detection and susceptibility mapping using Sentinel-1 remote sensing data and a machine learning approach: hybrid intelligence of bagging ensemble based on K-nearest neighbor classifier. Remote Sens. 12 (2), 266. https://doi.org/10.3390/rs12020266.
Shen, X., Liu, B., Jiang, M., Lu, X., 2020. Marshland loss warms local land surface temperature in China. Geophys. Res. Lett. 47 (6) https://doi.org/10.1029/2020GL087648 e2020GL087648.
Shimabukuro, Y.E., Arai, E., Duarte, V., Dutra, A.C., Cassol, H.L.G., Sano, E.E., Hoffmann, T.B., 2020. Discriminating land use and land cover classes in Brazil based on the annual PROBA-V 100 m time series. IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens. 13, 3409–3420. https://doi.org/10.1109/JSTARS.2020.2994893.
Sinha, S.K., Tiwari, L.K., 2018. Enhancement of image classification for forest encroachment mapping with destriped SWIR band in the wavelet domain. IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens. 11 (7), 2276–2281. https://doi.org/10.1109/JSTARS.2018.2814838.
Steinhausen, M.J., Wagner, P.D., Narasimhan, B., Waske, B., 2018. Combining Sentinel-1 and Sentinel-2 data for improved land use and land cover mapping of monsoon regions. Int. J. Appl. Earth Obs. Geoinf. 73, 595–604. https://doi.org/10.1016/j.jag.2018.08.011.
Sun, P., Reid, M.D., Zhou, J., 2014. An improved multiclass LogitBoost using adaptive-one-vs-one. Mach. Learn. 97 (3), 295–326. https://doi.org/10.1007/s10994-014-5434-3.
Talukdar, S., Singha, P., Mahato, S., Pal, S., Liou, Y.A., Rahman, A., 2020. Land-use land-cover classification by machine learning classifiers for satellite observations—A review. Remote Sens. 12 (7), 1135. https://doi.org/10.3390/rs12071135.
Tan, J., Zuo, J., Xie, X., Ding, M., Xu, Z., Zhou, F., 2021. MLAs land cover mapping performance across varying geomorphology with Landsat OLI-8 and minimum human intervention. Ecol. Inform. 61, 101227.
Tavares, P.A., Beltrão, N.E.S., Guimarães, U.S., Teodoro, A.C., 2019. Integration of sentinel-1 and sentinel-2 for classification and LULC mapping in the urban area of Belém, eastern Brazilian Amazon. Sensors 19 (5), 1140. https://doi.org/10.3390/s19051140.
Tehrany, M.S., Jones, S., Shabani, F., Martínez-Álvarez, F., Bui, D.T., 2019. A novel ensemble modeling approach for the spatial prediction of tropical forest fire susceptibility using logitboost machine learning classifier and multi-source geospatial data. Theor. Appl. Climatol. 137, 637–653. https://doi.org/10.1007/s10661-019-7362-y.
Thenkabail, P.S., Schull, M., Turral, H., 2005. Ganges and Indus river basin land use/land cover (LULC) and irrigated area mapping using continuous streams of MODIS data. Remote Sens. Environ. 95 (3), 317–341. https://doi.org/10.1016/j.rse.2004.12.018.
Topaloğlu, R.H., Sertel, E., Musaoğlu, N., 2016. Assessment of classification accuracies of sentinel-2 and landsat-8 data for land cover/use mapping. In: International Archives of the Photogrammetry, Remote Sensing & Spatial INFORMATION Sciences, p. 41.
Uddin, S., Khan, A., Hossain, M.E., Moni, M.A., 2019. Comparing different supervised machine learning algorithms for disease prediction. BMC Med. Inform. Decision Mak. 19 (1), 1–16. https://doi.org/10.1186/s12911-019-1004-8.
Whyte, A., Ferentinos, K.P., Petropoulos, G.P., 2018. A new synergistic approach for monitoring wetlands using Sentinels-1 and 2 data with object-based machine learning algorithms. Environ. Model. Softw. 104, 40–54. https://doi.org/10.1016/j.envsoft.2018.01.023.
Willhauck, G., Schneider, T., De Kok, R., Ammer, U., 2000, July. Comparison of object oriented classification techniques and standard image analysis for the use of change detection between SPOT multispectral satellite images and aerial photos. In: Proceedings of XIX ISPRS Congress, vol. 33. IAPRS, Amsterdam, pp. 35–42.
Xu, B., Gong, P., Biging, G., Liang, S., Seto, E., Spear, R., 2004. Snail density prediction for schistosomiasis control using IKONOS and ASTER images. Photogramm. Eng. Remote. Sens. 70 (11), 1285–1294. https://doi.org/10.1080/01431160412331291233.
Yu, C., Chen, J., 2020. Landslide susceptibility mapping using the slope unit for southeastern Helong City, Jilin Province, China: a comparison of ANN and SVM. Symmetry 12 (6), 1047. https://doi.org/10.3390/sym12061047.
Yüksel, A., Akay, A.E., Gundogan, R., 2008. Using ASTER imagery in land use/cover classification of eastern Mediterranean landscapes according to CORINE land cover project. Sensors 8 (2), 1237–1251. https://doi.org/10.3390/s8021287.
Zhan, X., Sohlberg, R.A., Townshend, J.R.G., DiMiceli, C., Carroll, M.L., Eastman, J.C., Hansen, M.C., DeFries, R.S., 2002. Detection of land cover changes using MODIS 250

m data. Remote Sens. Environ. 83 (1–2), 336–350. https://doi.org/10.1016/S0034-4257(02)00081-0.
Zhang, H., Xu, R., 2018. Exploring the optimal integration levels between SAR and optical data for better urban land cover mapping in the Pearl River Delta. Int. J. Appl. Earth Obs. Geoinf. 64, 87–95. https://doi.org/10.1016/j.jag.2017.08.013.
Zhou, J., Shi, X.Z., Huang, R.D., Qiu, X.Y., Chen, C., 2016. Feasibility of stochastic gradient boosting approach for predicting rockburst damage in burst-prone mines. Trans. Nonferrous Metals Soc. China 26 (7), 1938–1.



Geocarto International

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tgei20

Novel ensemble machine learning models in flood susceptibility mapping

Pankaj Prasad, Victor Joseph Loveson, Bappa Das & Mahender Kotha

To cite this article: Pankaj Prasad, Victor Joseph Loveson, Bappa Das & Mahender Kotha (2021): Novel ensemble machine learning models in flood susceptibility mapping, Geocarto International, DOI: 10.1080/10106049.2021.1892209

To link to this article: https://doi.org/10.1080/10106049.2021.1892209

GEOCARTO INTERNATIONAL
https://doi.org/10.1080/10106049.2021.1892209

Taylor & Francis
Taylor & Francis Group

Novel ensemble machine learning models in flood susceptibility mapping

Pankaj Prasad[a,b], Victor Joseph Loveson[a], Bappa Das[c] and Mahender Kotha[b]

[a]Geological Oceanography Division, CSIR - National Institute of Oceanography, Dona Paula, Goa, India; [b]School of Earth, Ocean and Atmospheric Sciences, Goa University, Taleigao, Goa, India; [c]Natural Resource Management Section, ICAR-Central Coastal Agricultural Research Institute, Old Goa, Goa, India

ABSTRACT

The research aims to propose the new ensemble models by combining the machine learning techniques, such as rotation forest (RF), nearest shrunken centroids (NSC), k-nearest neighbour (KNN), boosted regression tree (BRT), and logitboost (LB) with the base classifier adabag (AB) for flood susceptibility mapping (FSM). The proposed models were implemented in the central west coast of India, which is vulnerable to flood events. For flood inventory mapping, a total of 210 flood localities were identified. Twelve effective factors were selected using the boruta algorithm for FSM. The area under the receiver operating characteristics (AUROC) curve and other statistical measures (sensitivity, specificity, accuracy, kappa, root mean square error (RMSE), and mean absolute error (MAE)) were employed to estimate and compare the success rate of the approaches. The validation results of the individual models in terms of AUC value were AB (92.74%) >RF (91.50%) >BRT (90.75%) >LB (89.07%) >NSC (88.97%) >KNN (83.88%), whereas the ensemble models showed that the AB-RF (94%) was of the highest prediction efficiency followed by, AB-KNN (93.33%), AB-NSC (93.02%), AB-LB (92.83%), and AB-BRT (92.64%). The outcomes of the ensemble models established that the AB is more appropriate to increase the accuracy of different single models. Therefore, this study can be useful for proper planning and management of the study area and flood hazard mapping in alike geographic environment.

ARTICLE HISTORY
Received 4 October 2020
Accepted 1 February 2021

KEYWORDS
Flood hazard; geographical information system; boruta approach; adabag; central west coast of India

1. Introduction

Among the several types of natural hazards, flood is the most destructive phenomenon worldwide which has continuously increased since the last three decades (Youssef et al. 2011; Khosravi et al. 2016b; Termeh et al. 2018). The main reason for flood is heavy rainfall in short time duration, which may also occur due to the rock slide, debris flow, and levee breaks (Khosravi et al. 2018). The floods are broadly categorized into five types, namely coastal, flash, fluvial, urban, and pluvial floods. The occurrence and types of floods

CONTACT Pankaj Prasad ppankaj.earthscience@gmail.com; ppankaj@nio.org

are related to the differential interplay of various topographical, hydrological, and atmospheric factors. The central west coast of India experiences many devastating flood events frequently during the monsoon seasons. These flood hazards inflict human health and fatalities, economic losses, and severe destruction of the environment. Therefore, generating flood susceptibility map following appropriate methods is the critical objective of researchers and planners to reduce future flood damages.

There have been ample studies made by the researchers to prepare the FSM using different methods. Most of the traditional one-dimensional hydrological methods are no longer capable of flood modelling because of the earth's dynamic and complex characteristics (Sahoo et al. 2009; Khosravi et al. 2018). Such methods are mostly employed in a smaller area with a large amount of high precision data. In recent times, the development of geospatial techniques and a better understanding of the effective factors for flood play an essential role in flood modelling and prediction in a large area with time and cost-effectiveness (Haq et al. 2012). The combined methods of remote sensing, geographical information system, and statistical techniques were well-documented for the demarcation of flood vulnerable zones (Tehrany et al. 2013). The previous studies used various statistical methods including frequency ratio (Cao et al. 2016; Arabameri et al. 2020a), logistic regression (Tehrany et al. 2014b; Nandi et al. 2016), certainty factor (Arabameri et al. 2019), and weight of evidence (Rahmati et al. 2016) for flood prediction modelling. However, the strict assumptions in statistical modelling such as homoscedasticity (errors is not associated with predictor values), normally distributed predictand, and linear relationship reduced the model accuracy, whereas machine learning models can overcome the statistical drawbacks and can predict complex non-linear structures (Benediktsson et al. 1990; Mason and Baddour 2008; Tehrany et al. 2013; Shafizadeh-Moghadam et al. 2017; Costache and Bui 2019). The predictive accuracy of machine learning approaches has upgraded significantly compared to statistical methods in FSM (Chen et al. 2018; Costache and Bui 2019). Examples of the well-known machine learning models in flood studies are adaptive network-based fuzzy inference system (Bui et al. 2018a; Ahmadlou et al. 2019), Adaboost (Al-Abadi 2018; Bui et al. 2019b), random forest (Zhao et al. 2018; Hong et al. 2018c), boosted regression tree (Rahmati and Pourghasemi 2017; Shafizadeh-Moghadam et al. 2018), classification and regression tree (Shafizadeh-Moghadam et al. 2018; Choubin et al. 2019), support vector machine (Tehrany et al. 2014a; Hong et al. 2018c), fuzzy logic (Pulvirenti et al. 2011; Sahana and Patel 2019), artificial neural network (Elsafi 2014; Kourgialas and Karatzas 2017; Ngo et al. 2018; Zhao et al. 2018), rotation forest (Al-Abadi 2018; Costache and Bui 2019), and k-nearest neighbour (Liu et al. 2016; Shahabi et al. 2020).

Over the last few years, the ensemble model is getting popular in natural hazards modelling, such as flood (Mahmoud and Gan 2018; Costache and Bui 2019), landslide (Hong et al. 2018a; Bui et al. 2019a), drought (Roodposhti et al. 2017; Zhang et al. 2019), land subsidence (Oh et al. 2019; Arabameri et al. 2020b), and gully erosion (Bui et al. 2019c; Arabameri et al. 2020a). The ensemble model combines the different base classifiers into one predictive model (Hong et al. 2018a). The advantage of the ensemble model compared to single model is that it obtains higher goodness-of-fit and prediction accuracy by eliminating the weaknesses of the single model (Alotaibi and Sasi 2016; Bui et al. 2019c). However, the use of ensemble models in flood susceptibility mapping is still limited. Besides, no research work has been done on FSM using ensemble model on the west coast of India.

Due to the increase of flood events, there is a need to implement of new models to enhance the prediction capability in flood hazard (Hong et al. 2018b). For that, the primary objective of the current study is to produce and compare the novel ensemble models such as RF, NSC, KNN, LB, and BRT, with adabag as the base classifier for flood

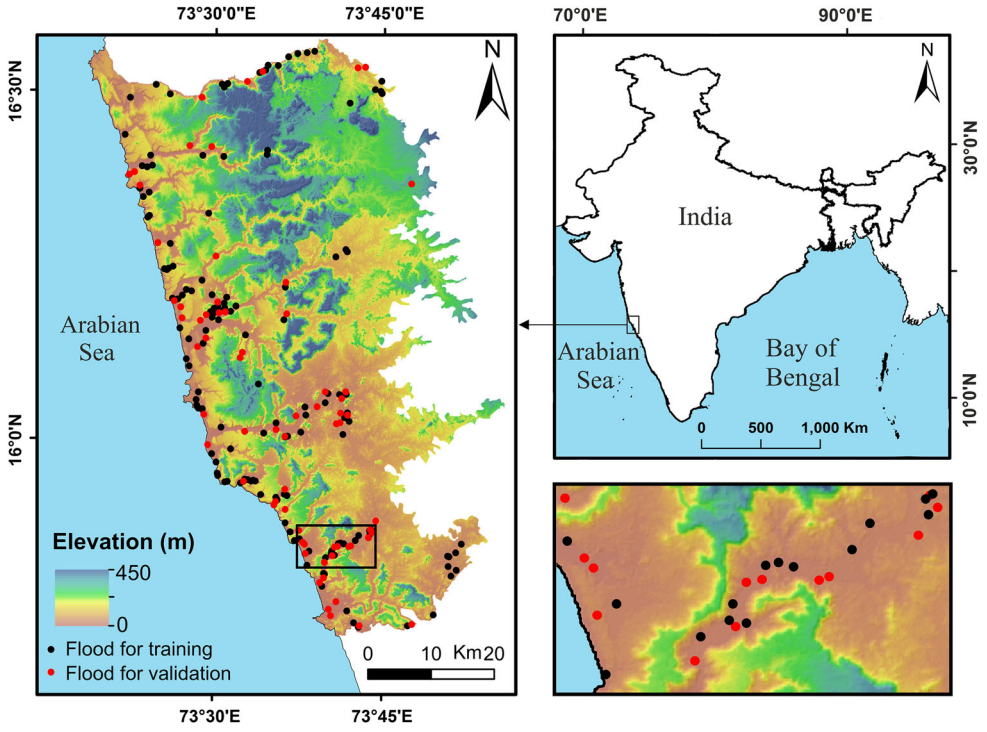


Figure 1. Location map of the study area.

susceptibility mapping along the central west coast of India. The foremost advantage of the adabag model is that the ensemble keeps growing until it gets properly developed. The other classification models of this research were used for their high precision in different studies as an individual model. In the present research, ensemble methods have been used for their novelty and higher prediction accuracy. To the best of our knowledge, the proposed models have not been used before in any natural hazard mapping. The results of these models can be useful for sustainable management of flood prone regions.

2. Study area and GIS database

2.1. Study area

The study area of about 3177 km^2 lying between the latitude 15°43′11.43″N to 16°33′45.63″N and longitude 73°18′36.53″E to 73°55′50.07″E is located along the central west coast of India. The area under study is geographically bounded by rivers, Western Ghats, and the shoreline of the west coast of India (Figure 1). From the above mean sea level, the altitude ranges from 0 m at the seashore to 450 m at landward hill ranges. The slope of the area is flat to moderate.

Geologically, the study region is covered by Archean, Dharwar, Kallaadgi, Sahyadri, and Quaternary group of formation. Concerning the geomorphological settings, the area is composed of pediment-pediplain complex, coastal plain, and dissected plateau. The principal rivers of the research area are Waghotan, Sukhandi, Tillari, Karli, and Gad, which witness flood at least once a year. Besides, the creeks are also playing a role in the seashore area. The climatic condition of the study place is sub-tropical. The maximum temperature of the area reaches 40 °C in the summer season (March-May), and the

minimum temperature remains around 15 °C in the winter season (December-January). The mean annual precipitation is around 3000 mm. The rainfall mainly occurs during June to September months due to the arrival of southwest monsoon. The flood occurrences in the area of interest are frequent due to heavy rainfall in a short period, cyclone, and high waves. In 2005, most of the area was devastated by floods and the subsequent loss as estimated was more than 90 million rupees. Keeping in view the severity of flood events and the concern of people life as well as resources, the study region has been selected as the site for FSM.

2.2. GIS database

2.2.1. Flood inventory mapping

To prepare the flood susceptibility map of an area, identifying and analysing the past flood locations are mandatory (Merz et al. 2007; Khosravi et al. 2016b). Therefore, an inventory map is regarded as a fundamental aspect to define the relationships between flood locations and effective factors (Tehrany et al. 2014a; Rahmati et al. 2016). In this research, the inventory map contained 210 flood locations which were recognized from the previous records, satellite images, and field surveys (Figure 1). The previous years (2005, 2009, and 2019) high-magnitude flood locations were collected from the department of disaster management, Maharashtra state and examined from the Landsat TM and OLI images before and during flood events. From these images, open water regions were delineated using the modified normalised difference water index (MNDWI) method (Xu 2006; Sahana and Patel 2019). It is formulated as

$$MNDWI = \frac{Green - MIR}{Green + MIR} \tag{1}$$

here the green and MIR (middle infrared) represent the bands of the images.

In the MNDWI method, all types of water body were extracted. The threshold value for the classification of water and non-water body is 0 for the images. To demarcate the only flooded area, permanent water bodies were taken out from the vector images. After that, these locations were matched with the past flood records. Additionally, an equal numbers of non-flood affected points were identified from previous records, topographical maps and Google earth images with reference to literature knowledge (Khosravi et al. 2016a; Mojaddadi et al. 2017). The presence and absence of flood occurrences indicated by the values 1 and 0, respectively. The flood inventory map was classified into two groups viz. training and testing in the proportion of 70% (147 flood locations) and 30% (63 flood locations), respectively (Figure 1).

2.2.2. Flood effective parameters

The flood effective variables are required for preparing the FSM (Rahmati et al. 2016). Concerning heterogeneity of the earth environment, the conditioning factors of the flood may vary from place to place (Tehrany et al. 2013). For present work, a total number of 12 variables, including elevation, aspect, slope, topographical roughness index (TRI), topographical wetness index (TWI), stream power index (SPI), distance from the rivers, rainfall, normalized difference vegetation index (NDVI), soil, geomorphology, and lithology were considered to create the thematic layers (Figure 2a–l). The SRTM DEM data of 30 m resolution was utilized for the preparation of topographical thematic layers. Landsat 8 OLI image was applied to generate the NDVI map. The categorical factors (rainfall, soil, geomorphology, geology) were produced from several data, source of which has been

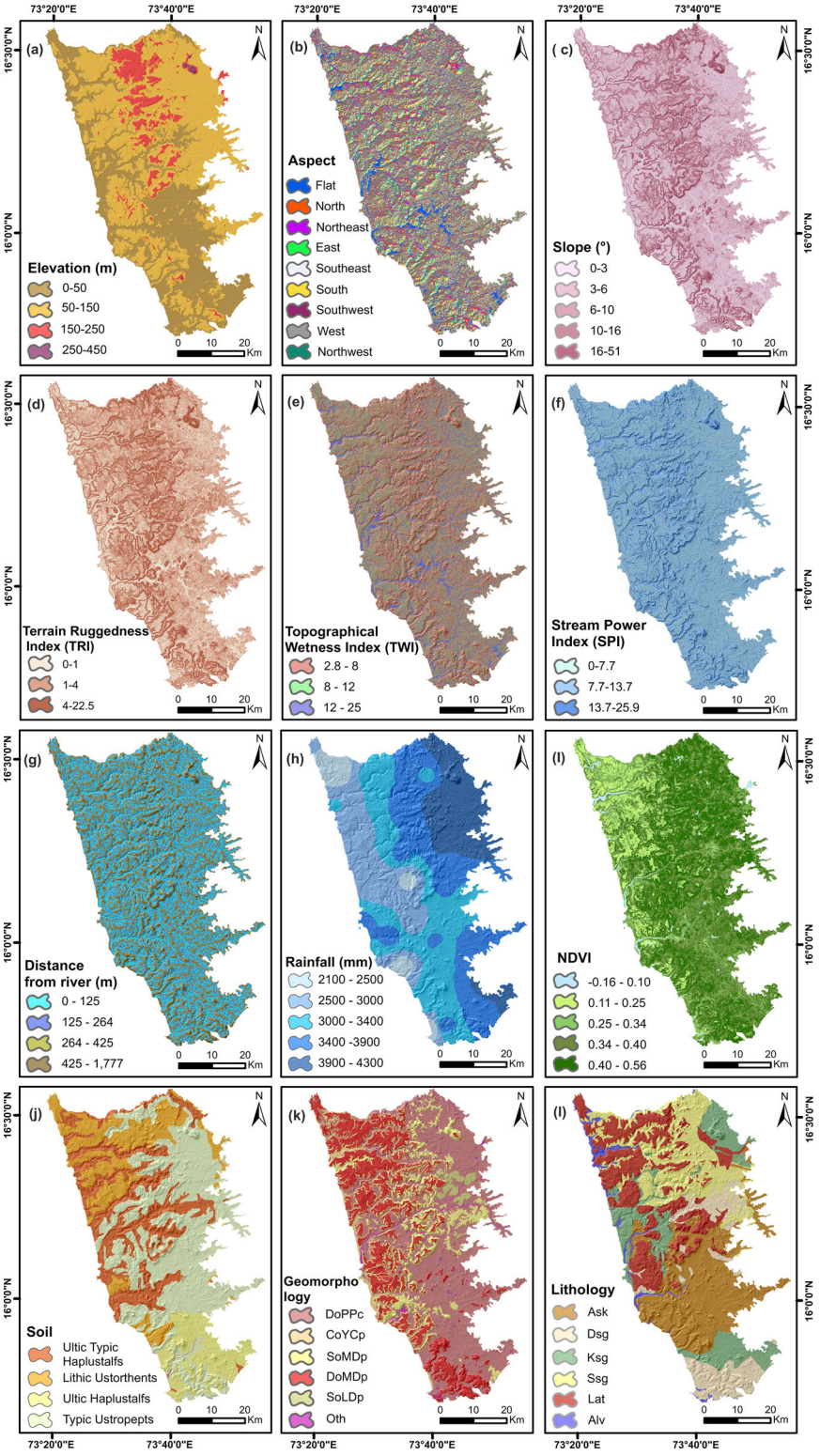


Figure 2. Thematic layers of the groundwater affecting factors (a) elevation, (b) aspect, (c) slope, (d) TRI, (e) TWI, (f) SPI, (g) distance from rivers, (h) rainfall, (i) NDVI, (j) soil, (k) geomorphology, (l) lithology.

mentioned later. All thematic maps were prepared and classified in the Arc GIS environment. The input and output maps of the study area were converted into 30 × 30 m cell size with 2201 columns and 3159 rows.

The high accuracy topographic data directly influences the results of modelling (Tehrany et al. 2013). Many researchers inferred that altitude is a vital factor for flood modelling (Khosravi et al. 2018). It is known fact that the low elevation and flat areas are prone to flood occurrence (Tehrany et al. 2013). Another affecting factor aspect was extracted from DEM. Aspect layer was divided into nine classes. TRI is a critical morphological factor in flood modelling which was produced from the DEM applying Equation (2) (Riley et al. 1999; Mojaddadi et al. 2017).

$$TRI = Abs(maxm^2 - minm^2) \tag{2}$$

here maxm and minm represent the maximum and minimum elevations.

TWI and SPI factors, which are related to soil moisture, affect hydrological conditions (Tehrany et al. 2015; Hong et al. 2018c). Slope map was prepared from the DEM and classified into five categories. Generally, due to earth's gravitational force, flow of water get accelerated in a higher slope area leading to flood in gentle slope area (Hong et al. 2018c). TWI defines the spatial distribution of wetness status of an area (Moore et al. 1991). SPI represents the index of erosive power of overland flow at a particular point on the surface (Moore and Grayson 1991; Khosravi et al. 2016b). TWI and SPI were derived from the following equations (Moore et al. 1991).

$$TWI = In(A/\tan\beta) \tag{3}$$

$$SPI = A \times \tan\beta \tag{4}$$

where A and β are the flow accumulation and slope gradient at each pixel, respectively.

The distance from the stream has a consequential role in flood occurrence (Khosravi et al. 2018). Four buffer zones (0–125, 125–264, 264–425, 425–1777 m) were delineated from the streams with the help of Arc GIS software. Generally, an inverse relationship is exist between flood location and vegetation density, which regulates the water flow (Tehrany et al. 2015). NDVI layer was produced from the Landsat 8 OLI image of 19 October 2016. The following formula was used to extract the surface vegetation density and coverage in the area of interest.

$$NDVI = (NIR - R)/(NIR + R) \tag{5}$$

where NIR and R are the surface reflectance of near infrared and red bands of the satellite image respectively. The range of the extracted value lay between −0.16 − 0.56.

The rainfall amount, intensity, duration, and distribution are critical to cause the flood in any area (Bracken et al. 2008). Rainfall data for the years of 2013–2018 were collected from the rain gauze stations of Maharashtra state (http://maharain.gov.in/). The inverse distance weighted method was applied to prepare the rainfall distribution map.

Four soil groups were demarcated, such as 1) Uth 2) Lu 3) Uh, and 4) Tu, using the base layer of National Bureau of Soil Survey and Land use planning (NBSS&LUP). The thematic layer of geomorphology at 1:50000 scale was prepared on the basis of satellite image, NRSC base map, and extensive field survey. The map was reclassified into six categories: 1) DoPPc 2) CoYCp 3) SoMDp 4) DoMDp 5) SoLDp, 6) Oth. The lithology is an important parameter for hydrological variation and sediment production (Miller et al. 1990; Khosravi et al. 2016a). The lithology map was grouped into six major classes, including Ask, Dsg, Ksg, Ssg, Lat, and Alv, using the lithology map prepared by the

Table 1. Details of soil, geomorphology, and lithology factors.

Parameters and code	Name	Characteristics
Soil		
Uth	Ultic typic haplustalfs	Moderately deep, well drained, loamy soils on gently sloping valley lands with moderate erosion
Lu	Lithic ustorthents	Very shallow, well drained, loamy soils on moderately sloping undulating lands with mesas and narrow valleys with severe erosion and moderate stoniness
Uh	Ultic haplustalfs	Moderately deep, well drained, loamy soils on moderately sloping undulating lands with mesas and narrow valleys with moderate erosion and moderate stoniness
Tu	Typic ustropepts	Slightly deep, well drained, loamy soils on moderately sloping undulating lands with mesas and narrow valley with severe erosion and moderate stoniness
Geomorphology		
DoPPc	Denudational origin-pediment pediplain complex	Pediment, pediplain, residual hill, residual mound, gullied land, rolling plain, wash plain
CoYCp	Coastal origin-younger coastal plain	Beach ridge, tidal flat, dune ridge, tidal creek, beach, sea cliff, sand dunes, spit, estuarine mudflat
SoMDp	Structural origin moderately dissected plateau	Dome, basin, scarp, ridge, hogback, cuesta, mesa, butte
DoMDp	Denudational origin moderately dissected plateau	Plateau top, valley, scarp, mesa, butte
SoLDp	Structural origin-low dissected plateau	Ridge, valley, basin, mesa, butte, scarp
Oth		
Lithology		
Ask	Archean schist and gneisses	Granite gneiss, quartzite, meta-gabbro, amphibole schist
Dsg	Dharwarsupergroup	Meta greywacke, metabasalt, granite
Ksg	Kalladgisupergroup	Sedimentary quartzite, shale
Ssg	Sahyadrisupergroup	Unclassified flows, Aa flow, Megacryst flow
Lat	Laterite	
Alv	Alluvium	

Geological Survey of India (GSI). The description of the different classes of soil, geomorphology, and lithology was mentioned in Table 1.

3. Methodology

The methodology of the current work to achieve the objective is displayed in Figure 3. The illustration of the methods has been described in the succeeding sections:

3.1. Boruta feature selection method

In the machine learning models, feature selection is a prerequisite step to determine the key variables that reduce the high computation times and risk of overfitting. This study used the boruta algorithm for feature selection. The algorithm is based on the wrapper method, which uses the random forest classifier (Kursa and Rudnicki 2010; Amiri et al. 2019). These steps were followed for the algorithm: i) At first, it duplicates all independent variables and shuffles the added attributes to remove their correlations with the response. ii) Then, the random forest classifier calculates the Z score values and find out the maximum Z score among shadow attributes (MZSA). iii) At last, the variables having value more than MZSA were selected as casual factors in flood modelling. Since, the boruta approach is an upgraded form of the variable important function of random forest, it

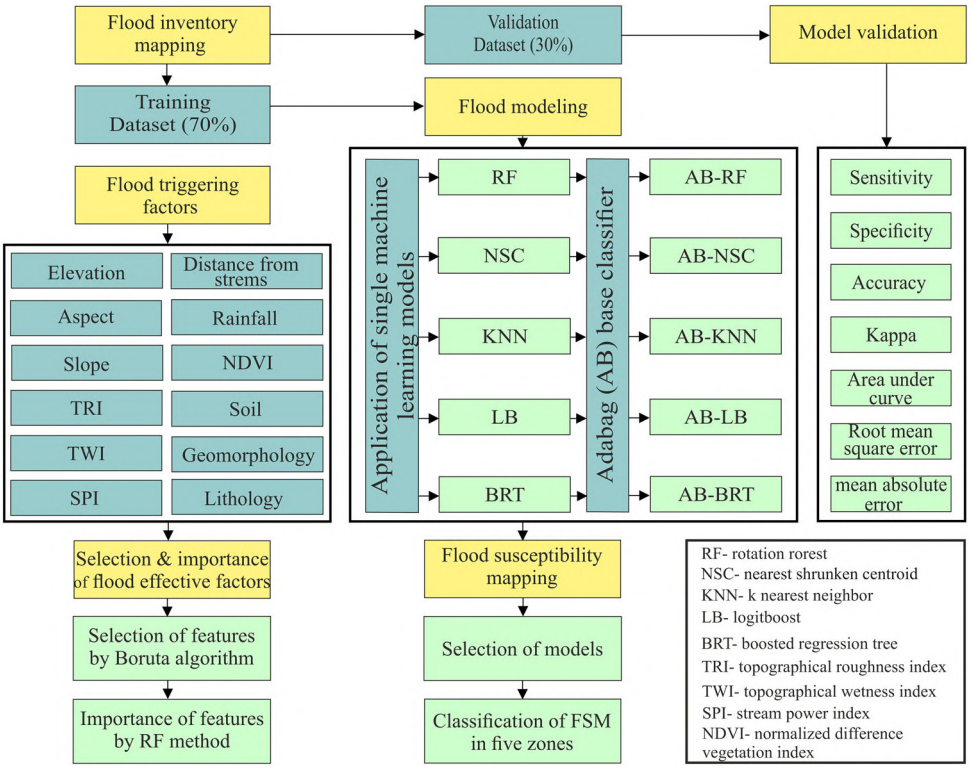


Figure 3. Flowchart of the methodology.

has the advantage in eliminating all the irrelevant variables (Kursa and Rudnicki 2010; Amiri et al. 2019).

3.2. Weight of Evidence (WoE) floods and feature inter-relationship method

The WoE approach is a quantitative 'data-driven' method for assuming the flood events (Armaş 2012). Many research works have utilized WoE model for spatial mapping using such as mineral, groundwater, flood, soil, and landslide mapping (Khosravi et al. 2016a; Chen et al. 2018). In the current research, the model was applied to define the relationship between conditioning factors and flood events. The WoE method measures the positive (W_i^+) and negative (W_i^-) weights for each flood influencing variable (A) as follows.

$$W_i^+ = In\left[\frac{p(B|A)}{P(B|\overline{A}}\right] \tag{6}$$

$$W_i^- = In\left[\frac{P(\overline{B}|A)}{P(\overline{B}|\overline{A}}\right] \tag{7}$$

$$\mathrm{W_c} = \mathrm{W_i^+} + \mathrm{W_i^-} \tag{8}$$

In this calculation, B and $\overline{B}$ denote the presence and absence of a specific class of flood affecting factors, respectively, whereas A and $\bar{A}$ are the presence and absence of flood events, respectively. The weight contrast (W_c) between W_i^+ and W_i^- indicates the spatial

correlation between the decisive factor and flood locations. The W is computed for the standard deviation as stated below:

$$S(C) = \sqrt{S^2\left(W_i^+\right) + S^2\left(W_i^-\right)} \tag{9}$$

$$S^2 W_i^+ = \frac{1}{P(B|A)} + \frac{1}{P(B|\overline{A})} \tag{10}$$

$$S^2 W_i^- = \frac{1}{P(\overline{B}|A)} + \frac{1}{P(\overline{B}|\overline{A})} \tag{11}$$

In the Equations (10) and (11), $S^2 W_i^+$ and $S^2 W_i^-$ are the variance of W_i^+ and W_i^-, respectively. The final weight (W_{final}) is measured as follows

$$W_{final} = \frac{W_c}{S(C)} \tag{12}$$

3.3. Application of machine learning models

3.3.1. Adabag base classifier

Adabag model is invented with the properties of bagging and boosting methods using classification trees as individual classifiers (Alfaro et al. 2013; Alotaibi and Sasi 2016). The algorithm consists of a total of eight functions. Each method (boosting and bagging) has three functions i) to make the boosting (or bagging) classifier and classify the training dataset; ii) to predict the new dataset on the basis of training dataset; iii) to estimate the accuracy of dataset by cross-validation. At last, the margin and error evolution are calculated. The main advantage of this model is to detect the overfitting and to handle multiclass tasks. Previously, the model was used in different scientific fields, namely automated content analysis, quality controls of signals, economics, finance and advances (Alfaro et al. 2013).

3.3.2. RF, NSC, KNN, LB, and BRT ensembles

Rotation forest is an ensemble learning method, introduced by Rodriguez et al. (2006). RF is the combination of random subspace and bagging method with principal component analysis (PCA) (Nguyen et al. 2017). In this technique, the training data are randomly segregated into K subsets through PCA. Each tree of RF model is trained by entire dataset with a rotated feature space, but the decision trees are individually created for the base classifiers (Rodriguez et al. 2006; Naghibi et al. 2019).

Nearest shrunken centroids method has been used in computer sciences and medical sciences with very high-precision (Tibshirani et al. 2003; Wang and Zhu 2007; Pardo and Sberveglieri 2008). However, it has not been used in FSM. The advantage of NSC method is to execute and evaluate the model performance. The classification of the model is based on the nearest centroid which computes the standardized centroid for each of the class (Wang and Zhu 2007; Pardo and Sberveglieri 2008).

K-nearest neighbour is a widely accepted classification method used in data mining application (Liu et al. 2016; Shahabi et al. 2020). It is a nonparametric and history matching algorithm that resample the present data from historical data (Lu et al. 2016). This is vital for hydrological modelling such as floods, stream flow where no prior knowledge of data is required. For the present work, euclidean distance technique was used to assess the closeness between the present feature vector and historical samples.

Logit boost method is a boosting technique proposed by Friedman et al. (2000). It is a modification of adaboost model to reduce the bias and variance (Oh et al. 2019). LB approach applies the additive logistic regression function for the classification. LB model can handle the multiclass problems. In LB approach, the vector values represent the number of input factors, and two output classes of flood and non-flood occurrences.

Boosted regression tree approach comprises both decision trees and boosting algorithms and is applied for the purpose classification and regression tasks (Elith et al. 2008; Kordestani et al. 2019). BRT uses boosting technique to merge different trees to improve the prediction (Elith et al. 2008). The algorithm can replace missing data of any type of predictive variables (numeric, binary, categorical). BRT model has three parameters namely shrinkage, interaction depth, and the number of trees for tuning the model.

Ensemble method is an advanced technique in predictive modelling. It integrates outputs of several models as input for enhancing precision of the final model (Naghibi et al. 2017). There are three most preferred ensemble methods, including boosting, bagging and stacking. Boosting method assembles homogeneous type of multiple models which learn to fix their earlier prediction error. Bagging generates a similar kind of multiple models from several subsamples of the training dataset (Brownlee 2014). In the present study, stacking method was used to create different types of multiple models with the combination of the best prediction of the prior models. All the models were implemented in the R environment using different packages.

3.4. Validation of flood susceptibility mapping

The applied individual and ensemble models were evaluated using statistical measures including sensitivity, specificity, accuracy, kappa, RMSE, MAE, and AUROC. All these quantitative metrics were constituted for training (70%) and testing (30%) datasets based on four types of possible consequences (PC) including, true positive (TP), false positive (FP), true negative (TN), and false negative (FN). TP and TN correctly classified the pixels of flood and non-flood whereas FP and FN erroneously separated the pixels of the classes. Based on four PC, the above statistical criteria were formulated as:

$$Sensitivity = \frac{TP}{TP + FN} \tag{13}$$

$$Specificity = \frac{TN}{TN + FP} \tag{14}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

$$Kappa\ \ index(K) = \frac{P_{obs} - P_{\exp}}{1 - P_{\exp}} \tag{16}$$

where, $P_{obs} = TP + TN$ and $P_{exp} = ((TP + FN)(TP + FP) + (FP + TN)(FN + TN))$

The present study also employed RMSE and MAE for cost functions to examine the error of the models. The smaller value of the RMSE and MAE indicate better accuracy of the model.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_p - X_a)^2} \tag{17}$$

$$MAE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|X_p - X_a\right|} \tag{18}$$

Where n is the total number of flood events, X_p and X_a are the predicted value (from training or testing datasets) and the actual value (probability of flood models), respectively.

The ROC curve is a popular and reliable method to evaluate the quality of probabilistic models (Chapi et al. 2017). ROC is a two-dimensional plot where the x-axis and y-axis represent false positive rate (1- specificity) and true positive rate (sensitivity), respectively. AUROC curve is a numerical index of model efficiency and their capability of assuming flood event and non-event. It value ranges between 0.5 (inaccurate) and 1 (accurate) (Chapi et al. 2017). AUC can be formulated as follows

$$AUC = \sum TP + \sum TN/(K+S) \tag{19}$$

where K and S are the total numbers of flood events and non-flood events, respectively.

4. Results and discussion

4.1. Feature selection by boruta algorithm

The first step of the flood hazard mapping is to select the significant geo-environmental factors. The strategy of suitable feature selection in machine learning techniques is to improve the efficiency of models by removing the irrelevant variables (Khosravi et al. 2018; Bui et al. 2019b). In this research, boruta approach was employed for the feature extraction. The outcomes from the boruta showed that the twelve variables were significant for the present study where the elevation was the most significant variable followed by NDVI, geomorphology, geology, soil, TRI, slope, TWI, aspect, rainfall, distance from streams, and SPI (Figure 4a). In contrast, the land-use and curvature had no significant impact on flood occurrence. Therefore, these factors were discarded to improve the performance of the models (Table 2). It was seen that the curvature and landuse had minimal importance for flood mapping in the research of Tehrany et al. (2015), Chapi et al.

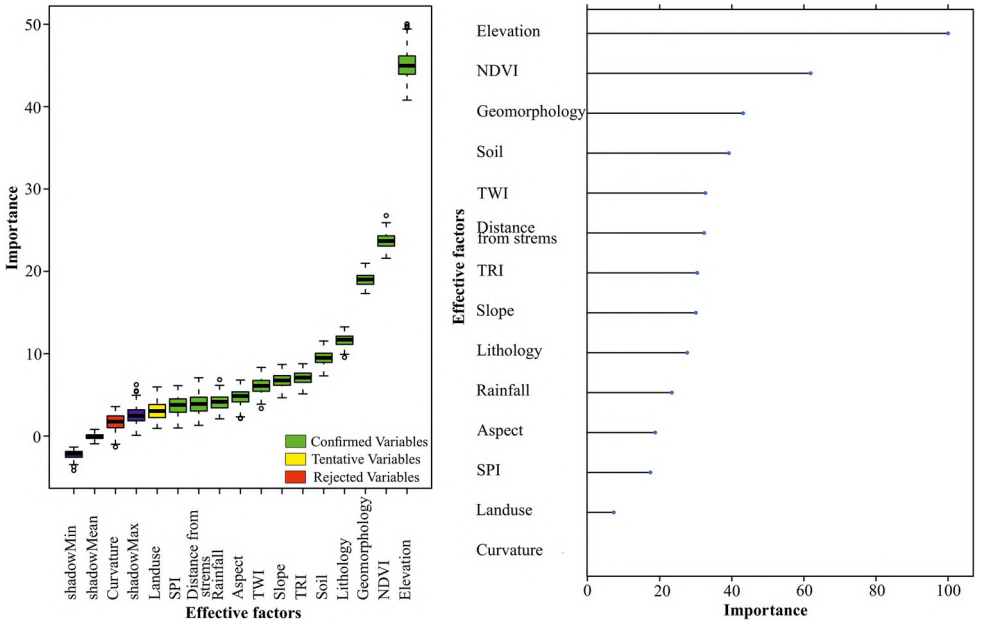


Figure 4. (a) Feature selection by boruta algorithm method, and (b) Variables importance in flood susceptibility mapping.

Table 2. Feature selection by Boruta algorithm.

Conditioning factors/ importance	Mean importance	Median importance	min Importance	Max importance	Norm hits	Decision
Elevation	45.09	44.96	40.79	50.04	1	Confirmed
NDVI	23.7	23.68	21.58	26.76	1	Confirmed
Geomorphology	18.96	19	17.3	20.97	1	Confirmed
Lithology	11.64	11.7	9.55	13.25	1	Confirmed
Soil	9.54	9.48	7.3	11.53	1	Confirmed
TRI	7.09	7.09	5.11	8.77	1	Confirmed
Slope	6.73	6.75	4.64	8.69	1	Confirmed
TWI	6.01	6.09	3.34	8.33	0.98	Confirmed
Aspect	4.75	4.85	2.12	6.79	0.91	Confirmed
Rainfall	4.18	4.15	2.08	6.84	0.88	Confirmed
Distance from rivers	3.86	3.89	1.3	7.07	0.8	Confirmed
SPI	3.71	3.77	0.97	6.11	0.76	Confirmed
Landuse	3.15	3.03	0.93	5.96	0.63	Tentative
Curvature	1.65	1.75	−1.33	3.57	0.26	Rejected

(2017), Khosravi et al. (2018), Costache et al. (2020a). The variable important function of RF method was used for measuring the effectiveness of each feature. In this context, elevation held the maximum importance followed by NDVI, and geomorphology, while curvature was of lowest importance followed by landuse, and SPI (Figure 4b). Similarly, in the study of Das (2019) and, it was found that the elevation, geomorphology had greater importance and curvature was least responsible for the flood occurrences in Ulhas basin, northern west coast of India. The elevation and geomorphology of the study area had greatly influenced the flood events because a huge amount of water during the flood events had come from the upslope area of the Western Ghats region, leading to the inundation of the lower elevation of coastal and pediment-pediplain. Kale (2003), Kumar et al. (2018), Arora et al. (2019, 2021), and Ahmadlou et al. (2021) described geomorphology as the major factor for flooding. The previous studies revealed that the importance of geo-environmental factors do not remain consistent everywhere due to the heterogeneity of the earth, applied methods, and different source of the datasets (Chapi et al. 2017; Khosravi et al. 2018; Shafizadeh-Moghadam et al. 2018; Costache and Bui 2020; Costache et al. 2020a). Even, the relative importance of flood conditioning factors may vary with spatial scale. It is evident from the study of Costache (2019b) and Costache et al. (2020a) that the influence of those factors in a larger area deviates from its subset area. On the other side, it was also found from the works of Khosravi et al. (2016b); Shafizadeh-Moghadam et al. (2018), Arora et al. (2019, 2021), and Costache (2019a); Costache et al. (2020b) that the order of flood influential variables in terms of importance was almost unaltered in the same respective study regions owing to constant spatial dimension and invariable geo-environmental conditions. In the current research, the result of RF was almost matched with the result of the boruta algorithm (Figure 4a,b). By implication, the robustness of boruta approach was verified for FSM in the present study.

4.2. Spatial interaction of floods and conditioning factors by WoE model

The spatial interaction between flood hazard occurrences and conditioning factors is important for FSM (Khosravi et al. 2018). This result was assessed using the WoE model and was summarized in Table 3. As illustrated in the previous section, the final weight (C/SC) was considered for showing the above relationship. The positive value of the C/SC indicates the high correlation and vice versa. In the current research, the range of C/SC value was −9.53 to 12.78. Based on the value, it was noticed that the elevation class from

Table 3. Spatial relationship between each conditioning factor and flood locations using WoE model.

Parameters	Classes	% of area	% of inventory	W+	W−	C	SW+	SW−	SC	C/SC
Elevation (m)	0–50	34.29	96.06	1.01	−2.81	3.82	0.01	0.13	0.36	10.59
	50–150	56.36	3.94	−2.66	0.78	−3.44	0.13	0.01	0.36	−9.53
	150–250	8.87	0.00	0	0.09	0	0	0.00	0	0
	250–450	0.48	0.00	0	0.00	0	0	0.00	0	0
Aspect	Flat	4.82	23.65	1.59	−0.22	1.82	0.02	0.01	0.17	10.99
	North	11.04	9.85	−0.11	0.01	−0.13	0.05	0.01	0.24	−0.54
	Northeast	10.71	8.37	−0.25	0.03	−0.27	0.06	0.01	0.25	−1.07
	East	10.70	6.90	−0.44	0.04	−0.48	0.07	0.01	0.28	−1.74
	Southeast	11.59	14.29	0.21	−0.03	0.24	0.03	0.01	0.20	1.20
	South	12.37	7.88	−0.45	0.05	−0.50	0.06	0.01	0.26	−1.92
	Southwest	13.22	8.87	−0.40	0.05	−0.45	0.06	0.01	0.25	−1.82
	West	13.13	9.85	−0.29	0.04	−0.32	0.05	0.01	0.24	−1.38
	Northwest	12.42	10.34	−0.18	0.02	−0.21	0.05	0.01	0.23	−0.90
Slope (degree)	0–3	31.78	50.74	0.47	−0.33	0.79	0.01	0.01	0.14	5.65
	6–10	30.86	28.08	−0.09	0.04	−0.13	0.02	0.01	0.16	−0.86
	10–16	18.01	12.81	−0.34	0.06	−0.40	0.04	0.01	0.21	−1.92
	16–25	10.96	6.90	−0.46	0.04	−0.51	0.07	0.01	0.28	−1.83
	16–51	8.39	1.48	−1.74	0.07	−1.81	0.33	0.01	0.58	−3.11
TRI	0–1	27.98	51.72	0.61	−0.40	1.01	0.01	0.01	0.14	7.22
	1–4	55.16	42.86	−0.25	0.24	−0.50	0.01	0.01	0.14	−3.49
	4–22.5	16.86	5.42	−1.14	0.13	−1.26	0.09	0.01	0.31	−4.08
TWI	2.8–8	63.26	41.38	−0.42	0.47	−0.89	0.01	0.01	0.14	−6.25
	8–12	22.69	23.15	0.02	−0.01	0.03	0.02	0.01	0.17	0.15
	12–25	14.05	35.47	0.93	−0.29	1.21	0.01	0.01	0.15	8.26
SPI	0–7.7	35.60	24.63	−0.37	0.16	−0.53	0.02	0.01	0.16	−3.23
	7.7–13.7	40.37	53.20	0.28	−0.24	0.52	0.01	0.01	0.14	3.68
	13.7–25.9	24.03	22.17	−0.08	0.02	−0.11	0.02	0.01	0.17	−0.62
Distance from river (m)	0–125	38.13	56.16	0.39	−0.34	0.73	0.01	0.01	0.14	5.17
	125–264	30.30	32.02	0.05	−0.02	0.08	0.02	0.01	0.15	0.53
	264–425	22.80	7.88	−1.08	0.18	−1.26	0.06	0.01	0.26	−4.84
	425–1777	8.77	3.94	−0.75	0.05	−0.80	0.13	0.01	0.36	−2.21
Rainfall (mm)	2100–2500	4.94	9.36	0.64	−0.05	0.69	0.05	0.01	0.24	2.87
	2500–3000	19.98	33.00	0.50	−0.18	0.68	0.01	0.01	0.15	4.55
	3000–3400	29.14	34.48	0.17	−0.08	0.25	0.01	0.01	0.15	1.67
	3400–3900	28.53	14.29	−0.69	0.18	−0.87	0.03	0.01	0.20	−4.36
	3900–4300	17.41	8.87	−0.67	0.10	−0.77	0.06	0.01	0.25	−3.13
NDVI	(−0.16 − 0.10)	1.84	0.00	0	0.02	0	0	0.00	0	0
	(0.11–0.25)	9.06	10.84	0.18	−0.02	0.20	0.05	0.01	0.23	0.88
	(0.25–0.34)	16.10	53.69	1.20	−0.59	1.80	0.01	0.01	0.14	12.78
	(0.34–0.40)	34.23	31.03	−0.10	0.05	−0.15	0.02	0.01	0.15	−0.96
	(0.40–0.56)	38.77	4.43	−2.17	0.45	−2.61	0.11	0.01	0.34	−7.67
Soil	UlticTypicHaplustalfs	25.27	55.67	0.79	−0.52	1.31	0.01	0.01	0.14	9.28
	Lithic Ustorthents	18.15	6.90	−0.97	0.13	−1.10	0.07	0.01	0.28	−3.96
	UlticHaplustalfs	10.92	7.39	−0.39	0.04	−0.43	0.07	0.01	0.27	−1.60
	TypicUstropepts	45.64	30.05	−0.42	0.25	−0.67	0.02	0.01	0.15	−4.38
Geomorphology	DoPPc	49.76	81.77	0.50	−1.01	1.51	0.01	0.03	0.18	8.31
	CoYCp	2.25	10.34	1.53	−0.09	1.61	0.05	0.01	0.23	6.99
	SoMDp	18.65	5.42	−1.24	0.15	−1.39	0.09	0.01	0.31	−4.47
	DoMDp	25.96	0.00	0	0.30	0	0	0.00	0	0
	Oth	1.35	2.46	0.59	−0.01	0.61	0.20	0.01	0.45	1.34
	SoLDp	2.03	0.00	0	0.02	0	0	0.00	0	0
Lithology	Alv	2.28	6.90	1.10	−0.05	1.15	0.07	0.01	0.28	4.14
	Ksg	15.32	31.53	0.71	−0.21	0.92	0.02	0.01	0.15	6.11
	Lat	23.18	4.93	−1.56	0.22	−1.77	0.10	0.01	0.32	−5.47
	Dsg	11.21	12.32	0.09	−0.01	0.10	0.04	0.01	0.21	0.45
	Ask	27.14	33.00	0.19	−0.08	0.27	0.01	0.01	0.15	1.79
	Ssg	19.99	11.33	−0.58	0.10	−0.68	0.04	0.01	0.22	−3.08

0 to 50 m had the highest weight (10.59), while the class 50–100 m acquired the lowest weight of −9.53. For aspect class, the flat area had the maximum weight of 10.99, whereas other directions held negative weights except for southeast (1.20). In case of slope factor, the flood events mostly coincided with the lower slope areas (0–3°) having weight 5.65. The first class (0–1) of TRI had a positive weight (7.22) while other classes had negative weights. The C/SC values of TWI class increased with increasing wetness. With regards to the result of SPI, the second class (7.7–13.7) had the highest weight (3.68), where more than 50% of flood incidents happened in this class. The distance from streams showed that the 0–125 m experienced more than 50% flood occurrence with 5.17 C/SC value. Analysis of flood events and rainfall distribution revealed that the range of 2500–3000 mm had the greatest weight (4.55) among the other classes. The NDVI outputs displayed that the 0.25–0.34 class gained maximum weight (12.78) compared to all other subclasses of the factor. Among the different soil types, the Ultic Typic Haplustalfs soil was the only class that was positively correlated (9.28) with the flood events. Regarding geomorphology, the highest C/SC value was noted in the Denudational origin-pediment pediplain complex (8.31) followed by Coastal origin-younger coastal plain (6.99). For lithology, Kalladgi group of rocks and Alluvial achieved 6.11 and 4.14 weights, respectively, demonstrating high flood probability in these rock types.

4.3. Comparison and validation of flood susceptibility models

The development of model has still a very challenging task for the research community (Hong et al. 2018c). Numerous statistical and machine learning models have been implemented, but there is no universally accepted standard method for FSM. It is always necessary to apply the new methods and techniques in natural hazard mapping for better planning and management. The main objective of the present research was to establish the robustness of the novel adabag approach with the different machine learning model ensembles in FSM. Before that, the AB model has not been applied in any natural hazard mapping.

In the present research, all the individual models, namely RF, NSC, KNN, LB, and BRT were applied, and then ensembled with the adabag base classifier. All the models were trained through the 10-fold cross-validation method, whereby the model was trained ten times on ten distinct subsets of the training dataset to determine the best fitting model. Visually, the appearance of all the ensemble-based probability maps (except BRT) was more or less identical. However, visual assessment is not an appropriate method for validating the models (Tehrany et al. 2014a). For validation, the researchers use the AUROC curve and different statistical methods such as accuracy, kappa coefficient, sensitivity, specificity, RMSE, and MAE (Chapi et al. 2017; Bui et al. 2018b). AUROC method is a well-documented and accurate method to assess the success and prediction rates of the models (Pradhan and Lee 2010; Tehrany et al. 2014a, 2015). In this work, both AUROC and statistical metrics were applied for the model validation. It is necessary to examine and compare the efficiency of novel ensemble approach for both training and testing datasets. The training dataset indicates the robustness of the models, where the testing data represents the predictive power of the models (Hong et al. 2015; Khosravi et al. 2018).

In training dataset, performance of AB-RF model obtained the best goodness-of-fit, as quantified by AUC (0.976), accuracy (0.912), sensitivity (0.908), kappa (0.823). In addition, AB-RF model achieved the lowest values of RMSE (0.254) and MAE (0.064), which implied the lowest error in the model. On the other side, the BRT model exhibited the lowest performance followed by AB-NSC, AB-LB, AB-KNN. The highest advancement of the model fitness from individual to ensemble was recorded by the NSC model, in terms

Table 4. Performance of the models for training dataset.

Factors	RF	AB-RF	BRT	AB-BRT	LB	AB- LB	KNN	AB-KNN	NSC	AB-NSC
TP	135	138	127	133	130	135	119	134	130	132
TN	132	130	129	129	134	133	122	134	105	134
FP	10	12	13	13	8	9	20	8	37	8
FN	17	14	25	19	22	17	33	18	22	20
Overall	0.9082	0.9116	0.8707	0.8912	0.8980	0.9116	0.8197	0.9116	0.7993	0.9048
Kappa	0.8164	0.8230	0.7419	0.7824	0.7963	0.8233	0.6401	0.8233	0.5968	0.8098
Sensitivity	0.8882	0.9079	0.8355	0.8750	0.8553	0.8862	0.7829	0.8816	0.8553	0.8684
Specificity	0.9296	0.9155	0.9085	0.9085	0.9437	0.9366	0.8592	0.9437	0.7394	0.9437
AUC	0.9661	0.9757	0.9399	0.9639	0.9568	0.9723	0.9071	0.9668	0.8915	0.9633
RMSE	0.2657	0.2544	0.3174	0.2766	0.2841	0.2575	0.3492	0.2618	0.4654	0.2736
MAE	0.0706	0.0647	0.1007	0.0765	0.0807	0.0663	0.1219	0.0685	0.2166	0.0748

Table 5. Performance of the models for validation dataset.

Factors	RF	AB-RF	BRT	AB-BRT	LB	AB- LB	KNN	AB-KNN	NSC	AB-NSC
TP	48	49	48	49	49	51	42	49	49	50
TN	58	60	62	60	59	58	56	60	45	62
FP	10	13	06	8	09	10	12	8	23	6
FN	10	19	10	60	09	07	16	9	9	8
Overall	0.8413	0.8492	0.8730	0.8651	0.8571	0.8651	0.7778	0.8651	0.7460	0.8889
Kappa	0.6805	0.6969	0.7431	0.7281	0.7125	0.7295	0.5050	0.7281	0.4978	0.7758
Sensitivity	0.8276	0.8448	0.8276	0.8448	0.8448	0.8793	0.7241	0.8448	0.8448	0.8621
Specificity	0.8529	0.8529	0.9118	0.8823	0.8676	0.8529	0.8235	0.8824	0.6618	0.9118
AUC	0.9150	0.940	0.9075	0.9264	0.8907	0.9283	0.8388	0.9333	0.8897	0.9302
RMSE	0.3354	0.3213	0.3305	0.3220	0.3462	0.3157	0.3998	0.3087	0.4692	0.3107
MAE	0.1125	0.1032	0.1092	0.1036	0.1198	0.0999	0.1599	0.0953	0.2202	0.0965

of accuracy (+0.143), kappa (+0.278), sensitivity (+0.017), specificity (+0.25), AUC (+0.041), RMSE (–0.159) and MAE (–0.124) (Table 4; Figure 5).

In testing dataset, the AB-RF model had the highest AUC (0.940) indicating the probability of flood events and non-events which were correctly classified. On the other side, AB-NSC approach manifested the best efficiency, in respect of accuracy (0.889), kappa (0.776), and specificity (0.912). With regards to RMSE and MAE, the lowest value was achieved from the AB-KNN and AB-NSC models compared to other models. The biggest improvement from the individual to ensemble model was observed in AB-NSC model followed by AB-KNN, AB-BRT, AB-LB, and AB-RF models based on all evaluation criteria except AUC (Table 5; Figure 5). The AUC of KNN, NSC, LB, RF, and BRT models were increased by 10.04%, 4.05%, 3.76%, 2.50%, and 1.89%, respectively with the use of machine learning ensembles.

The overall findings from the training and testing datasets showed that the performance of NSC model enhanced significantly in comparison to other models. The NSC model has been effectively employed for highly accurate classification of DNA-microarray in many studies (Tibshirani et al. 2003; Dabney 2005; Wang and Zhu 2007). On the other side, the RF method had the highest AUC value for both training and testing datasets. Also, the RF method was found to have performed with the highest AUC value in other spatial mapping (Hong et al. 2018a; Bui et al. 2019a; Mosavi et al. 2020). Added to in other works on FSM using single model in the parts of west coast of India, the flood prediction rate was relatively low as compared to the present study (Das 2018, 2019). From the present research, it was confirmed that the efficiency of the single model upgraded with the ensemble of adabag base classifier for the prediction of flood susceptible areas. The ensemble classifier has the capability to reduce the bias and escape the overfitting issues against base classifier to enhance model performance (Kuncheva 2014; Hong et al. 2018a). The better performance of the AB model may be attributed

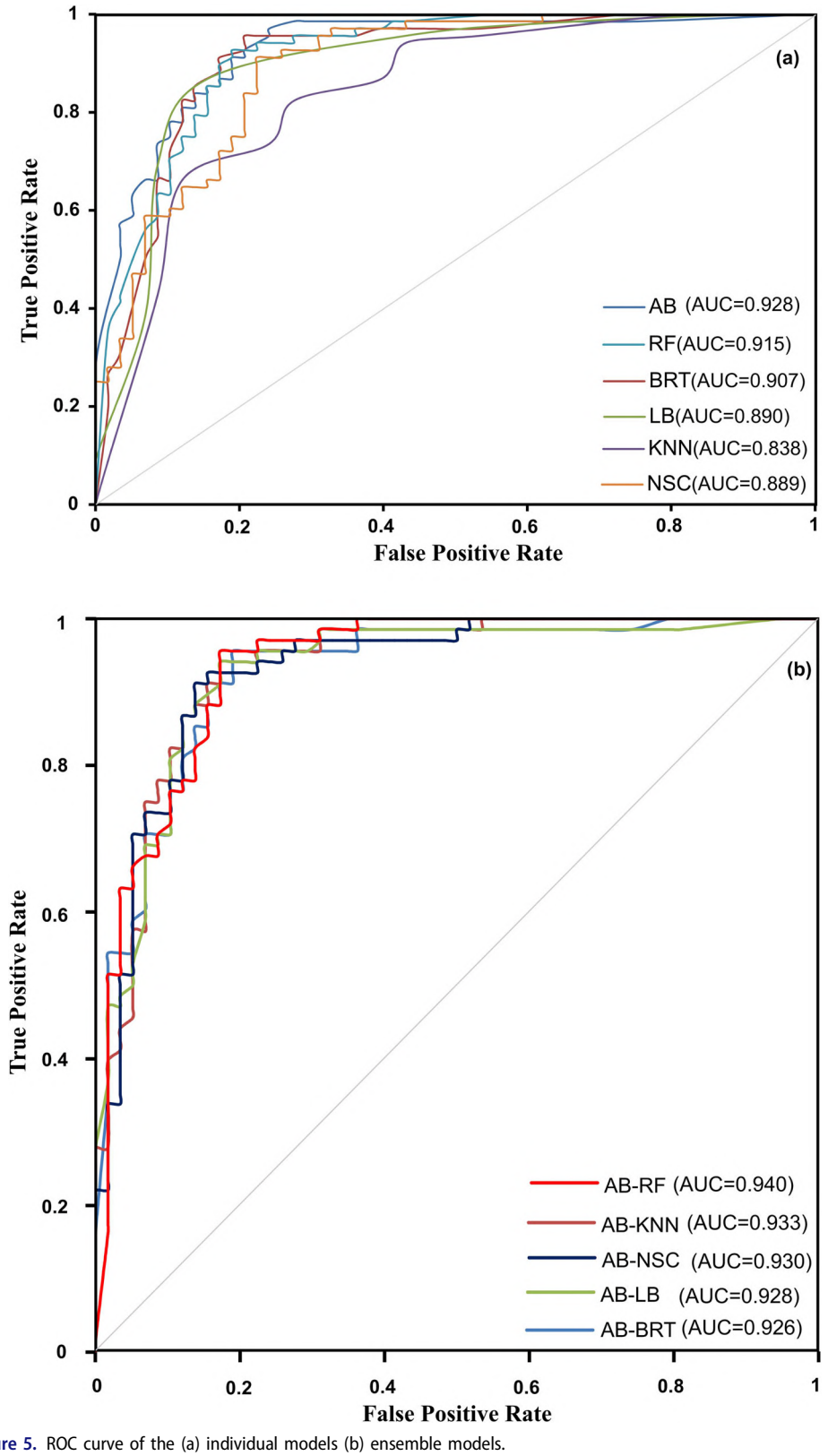


Figure 5. ROC curve of the (a) individual models (b) ensemble models.

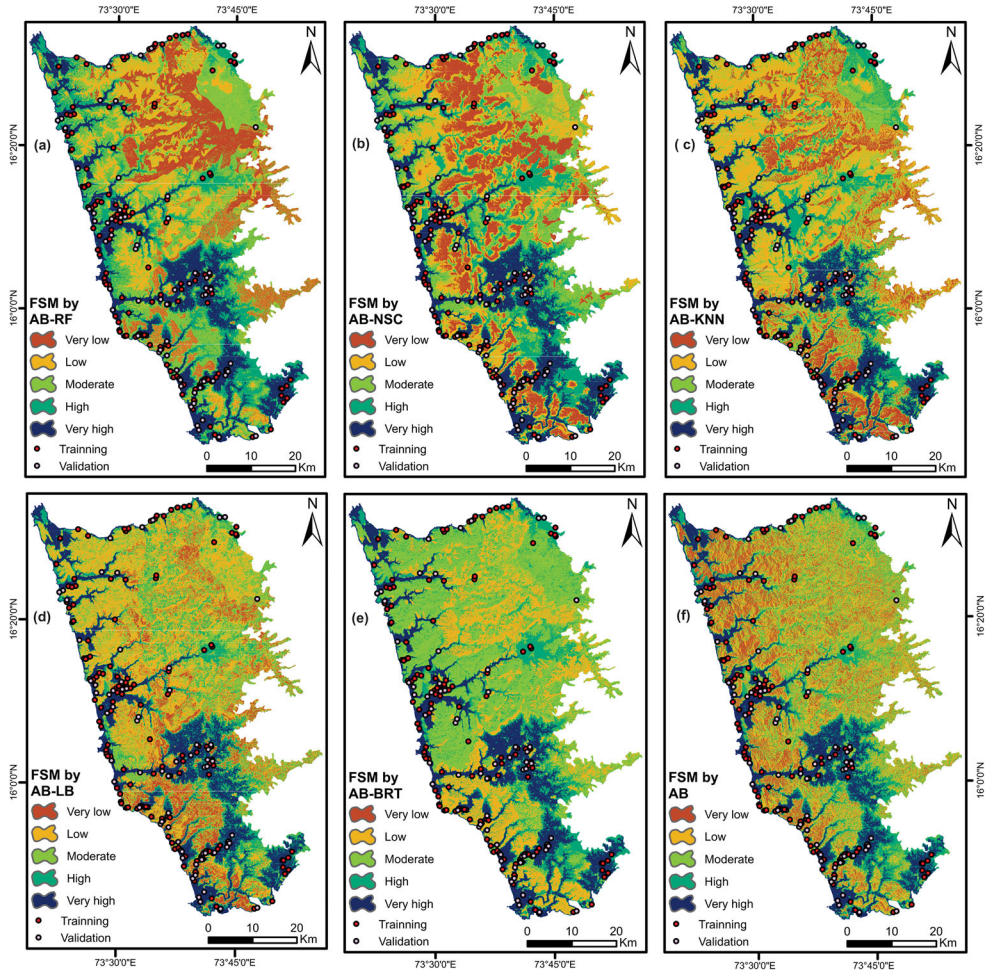


Figure 6. Flood susceptibility maps derived from (a) AB-RF, (b) AB-NSC, (c) AB-KNN, (d) AB-LB, (e) AB-BRT, and (f) AB models.

to its capability to analysis the evolution of error with the growth of ensemble classifier which helps to detect overfitting. It states that the ensemble continued to grow until the model developed enough (Alfaro et al. 2013). The ensemble of AB with individual model can be applied in other study areas as well as spatial prediction of gully erosion, groundwater potentiality, landslide susceptibility at a regional scale.

4.4. Flood susceptibility maps

It is essential to know the intensity of the flood susceptibility in an area, which could be more helpful in flood risk mitigation. For that, the probability maps of the models were classified in various zones. There are different methods for the classification viz. equal interval, min-max normalization, standard deviation, natural break, geometrical and quantile. The selection of the method is based on the nature of data and the aim of the research (Tehrany et al. 2015). For instance, equal interval and min-max normalization are applicable if the dataset has normal distribution. In case of standard deviation, it organizes the dataset into a specified number of groups. The natural break method is

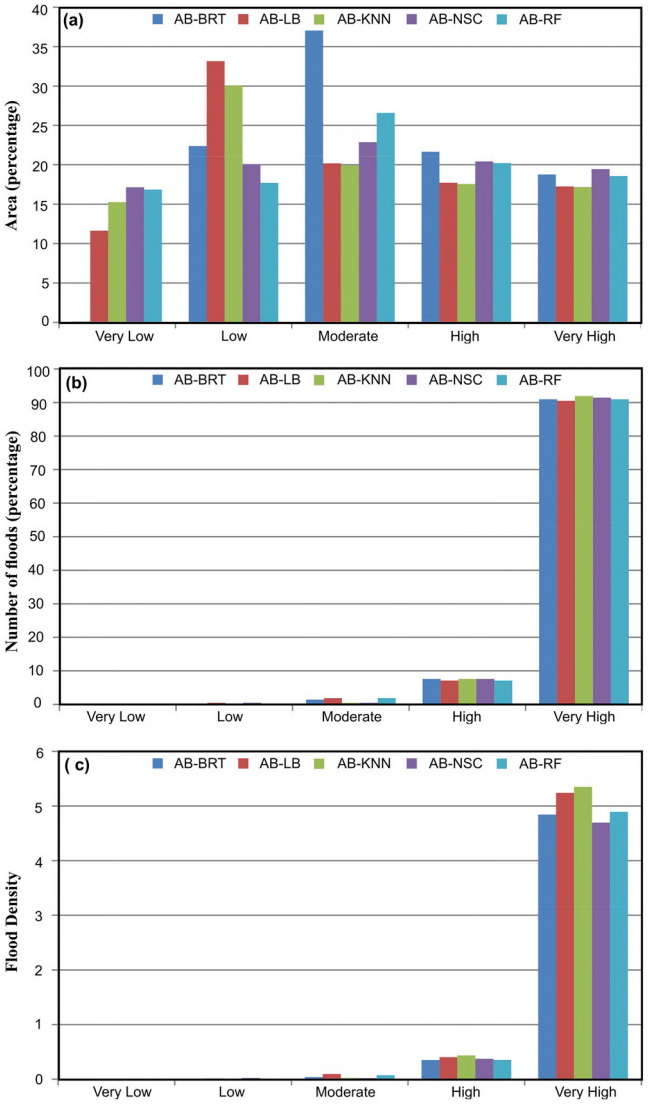


Figure 7. (a) Area, (b) flood events, and (c) flood density in each flood susceptibility classes of the models.

suitable when there is a sudden and abrupt change in the dataset. The geometric method is appropriate for continuous data. In the literature, for generating the susceptibility maps, has extensively been used as a standard classification method (Tehrany et al. 2015; Chapi et al. 2017). Hence, in this work, the quantile method was opted for generating flood susceptibility classes. Flood susceptible maps of the study area were classified into five zones, namely very low, low, moderate, high, and very high (Figure 6). It was observed that most of the south and south-eastern zones came under the high and very high vulnerable class. The region near river, confluence zone of the rivers or river and sea are also vulnerable to flood (Khosravi et al. 2016b), which is reflected in the present work. Also, it was found that the very low, low, and moderate susceptible zones mostly occupied the north, northeast, and eastern parts of the study area (Figure 6). In these zones, high elevation with dense vegetation is the main reason for lower probability of flood occurrences.

It is assumed that a better model always has the lowest variation in terms of area of high and very high classes (Naghibi et al. 2017; Prasad et al. 2020). The lowest variation of high and very high flood susceptibility classes were less than 4% and 5%, respectively, for all the models (Figure 7). The model fitness was also illustrated through the flood frequency in different susceptibility classes. Ideally, the flood events and flood density (FD) value should decrease from the very high susceptible areas to very low susceptible areas (Pradhan and Lee 2010; Chapi et al. 2017). In Figure 7b, the flood events were mostly associated with the very high (>90%) and high (>7%) susceptible classes for all the models. From this fact, it can be inferred that each model have had the capability to accurately classify the different flood susceptibility zones. FD is another criterion to verify the susceptibility classes of the various models. FD may be defined as the ratio of the total flood events to the total area in each susceptibility class. From the FD results, it was seen that all the models had high FD (> 4.5) in high susceptible zones whereas it became 0 in very low susceptible zones.

5. Conclusion

The present study introduced adabag model as a base classifier with five other ensemble classifiers including, RF, LB, NSC, KNN, and BRT. Furthermore, the prediction rates of individual models were compared with new ensemble approaches for FSM in the central west coast of India. The most vital features of the flood modelling were elevation, NDVI, geomorphology, soil, TWI, distance from streams, TRI, slope, lithology, rainfall, aspect, and SPI. The boruta approach was recognized as a suitable method for feature selection. The AB-RF model had the highest AUC, whereas the AB-NSC model recorded the highest improvement from single to ensemble model for both training and validation datasets. However, the outputs of the present study demonstrated that each novel ensemble model gave better accuracy compared to individual models.

In recent times, globally, the flood management program has been quite good, but there are still several areas that have not been appropriately planned. Therefore, the proposed models of the present work can be useful for FSM in similar geo-environmental settings across the world for future planning and development.

Acknowledgements

We acknowledge the support from the University Grant Commission (3160 NET-June 2015) to the first author. The authors are thankful to the Director CSIR-NIO for encouragement from time to time. The authors are also grateful to the anonymous reviewers for their critical comments and constructive suggestions to improve the manuscript. Field support from the survey team members is thankfully acknowledged. The NIO contribution number is 6674.

Disclosure statement

All authors are declared no actual or potential conflict of interest including any financial, personal or other relationships with other people or organizations.

ORCID

Pankaj Prasad http://orcid.org/0000-0002-3118-2201
Bappa Das http://orcid.org/0000-0003-1286-1492

References

Ahmadlou M, Al-Fugara AK, Al-Shabeeb AR, Arora A, Al-Adamat R, Pham QB, Al-Ansari N, Linh NTT, Sajedi H. 2021. Flood susceptibility mapping and assessment using a novel deep learning model combining multilayer perceptron and autoencoder neural networks. J Flood Risk Manage. 14(1):12683.

Ahmadlou M, Karimi M, Alizadeh S, Shirzadi A, Parvinnejhad D, Shahabi H, Panahi M. 2019. Flood susceptibility assessment using integration of adaptive network-based fuzzy inference system (ANFIS) and biogeography-based optimization (BBO) and BAT algorithms (BA). Geocarto Int. 34(11):1252–1272. https://doi.org//10.1080/10106049.2018.1474276.

Al-Abadi AM. 2018. Mapping flood susceptibility in an arid region of southern Iraq using ensemble machine learning classifiers: a comparative study. Arab J Geosci. 11(9):218.

Alfaro E, Gámez M, Garcia N. 2013. adabag: an R package for classification with boosting and bagging. J Stat Soft. 54(2):1–35.

Alotaibi NN, Sasi S. 2016. Tree-based ensemble models for predicting the ICU transfer of stroke inpatients. In 2016 International Conference on Data Science and Engineering (ICDSE); Aug 23–25; Cochin, India: IEEE United States. p. 1–6.

Amiri M, Pourghasemi HR, Ghanbarian GA, Afzali SF. 2019. Assessment of the importance of gully erosion effective factors using Boruta algorithm and its spatial modeling and mapping using three machine learning algorithms. Geoderma. 340:55–69.

Arabameri A, Asadi Nalivan O, Saha S, Roy J, Pradhan B, Tiefenbacher JP, Thi Ngo PT. 2020a. Novel ensemble approaches of machine learning techniques in modeling the gully erosion susceptibility. Remote Sensing. 12(11):1890. https://doi.org/10.3390/rs12111890.

Arabameri A, Pradhan B, Rezaei K. 2019. Gully erosion zonation mapping using integrated geographically weighted regression with certainty factor and random forest models in GIS. J Environ Manage. 232: 928–942. https://doi.org/10.1016/j.jenvman.2018.11.110.

Arabameri A, Saha S, Roy J, Tiefenbacher JP, Cerda A, Biggs T, Pradhan B, Ngo PTT, Collins AL. 2020b. A novel ensemble computational intelligence approach for the spatial prediction of land subsidence susceptibility. Sci Total Environ. 726:138595.

Arora A, Arabameri A, Pandey M, Siddiqui MA, Shukla UK, Bui DT, Mishra VN, Bhardwaj A. 2021. Optimization of state-of-the-art fuzzy-metaheuristic ANFIS-based machine learning models for flood susceptibility prediction mapping in the Middle Ganga Plain, India. Sci Total Environ. 750:141565.

Arora A, Pandey M, Siddiqui MA, Hong H, Mishra VN. 2019. Spatial flood susceptibility prediction in Middle Ganga Plain: comparison of frequency ratio and Shannon's entropy models. Geocarto Int. 1–32. https://doi.org/10.1080/10106049.2019.1687594.

Armaş I. 2012. Weights of evidence method for landslide susceptibility mapping Prahova subcarpathians. Nat Hazards. 60(3):937–950.

Benediktsson JA, Swain PH, Ersoy OK. 1990. Neural network approaches versus statistical methods in classification of multisource remote sensing data. IEEE Trans Geosci Remote Sensing. 28(4):540–552.

Bracken LJ, Cox NJ, Shannon J. 2008. The relationship between rainfall inputs and flood generation in south–east Spain. Hydrol Process. 22(5):683–696.

Brownlee J. 2014. Machine learning mastery. [Last accessed 2020 Aug 10]. http://machinelearningmastery.com/discover-feature-engineering-howtoengineer-features-and-how-to-getgood-at-it.

Bui DT, Khosravi K, Li S, Shahabi H, Panahi M, Singh VP, Chapi K, Shirzadi A, Panahi S, Chen W, et al. 2018b. New hybrids of anfis with several optimization algorithms for flood susceptibility modeling. Water. 10(9):1210.

Bui DT, Panahi M, Shahabi H, Singh VP, Shirzadi A, Chapi K, Khosravi K, Chen W, Panahi S, Li S, et al. 2018a. Novel hybrid evolutionary algorithms for spatial prediction of floods. Sci Rep. 8(1):1–14.

Bui DT, Shirzadi A, Shahabi H, Chapi K, Omidavr E, Pham BT, Talebpour Asl D, Khaledian H, Pradhan B, Panahi M, et al. 2019c. A novel ensemble artificial intelligence approach for gully erosion mapping in a semi-arid watershed (Iran). Sensors. 19(11):2444.

Bui DT, Shirzadi A, Shahabi H, Geertsema M, Omidvar E, Clague J, Thai Pham B, Dou J, Talebpour Asl D, Bin Ahmad B, et al. 2019a. New ensemble models for shallow landslide susceptibility modeling in a semi-arid watershed. Forests. 10(9):743.

Bui DT, Tsangaratos P, Ngo PTT, Pham TD, Pham BT. 2019b. Flash flood susceptibility modeling using an optimized fuzzy rule based feature selection technique and tree based ensemble methods. Sci Total Environ. 668:1038–1054.

Cao C, Xu P, Wang Y, Chen J, Zheng L, Niu C. 2016. Flash flood hazard susceptibility mapping using frequency ratio and statistical index methods in coalmine subsidence areas. Sustainability. 8(9):948.

Chapi K, Singh VP, Shirzadi A, Shahabi H, Bui DT, Pham BT, Khosravi K. 2017. A novel hybrid artificial intelligence approach for flood susceptibility assessment. Environ Model Software. 95:229–245.
Chen W, Li H, Hou E, Wang S, Wang G, Panahi M, Li T, Peng T, Guo C, Niu C, et al. 2018. GIS-based groundwater potential analysis using novel ensemble weights-of-evidence with logistic regression and functional tree models. Sci Total Environ. 634:853–867.
Choubin B, Moradi E, Golshan M, Adamowski J, Sajedi-Hosseini F, Mosavi A. 2019. An ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees, and support vector machines. Sci Total Environ. 651(Pt 2):2087–2096.
Costache R. 2019a. Flash-flood Potential Index mapping using weights of evidence, decision Trees models and their novel hybrid integration. Stoch Environ Res Risk Assess. 33(7):1375–1402.
Costache R. 2019b. Flash-flood potential assessment in the upper and middle sector of Prahova river catchment (Romania). A comparative approach between four hybrid models. Sci Total Environ. 659: 1115–1134.
Costache R, Bui DT. 2019. Spatial prediction of flood potential using new ensembles of bivariate statistics and artificial intelligence: a case study at the Putna river catchment of Romania. Sci Total Environ. 691:1098–1118.
Costache R, Bui DT. 2020. Identification of areas prone to flash-flood phenomena using multiple-criteria decision-making, bivariate statistics, machine learning and their ensembles. Sci Total Environ. 712: 136492.
Costache R, Hong H, Pham QB. 2020a. Comparative assessment of the flash-flood potential within small mountain catchments using bivariate statistics and their novel hybrid integration with machine learning models. Sci Total Environ. 711:134514.
Costache R, Ngo PTT, Bui DT. 2020b. Novel ensembles of deep learning neural network and statistical learning for flash-flood susceptibility mapping. Water. 12(6):1549.
Dabney AR. 2005. Classification of microarrays to nearest centroids. Bioinformatics. 21(22):4148–4154.
Das S. 2018. Geographic information system and AHP-based flood hazard zonation of Vaitarna basin, Maharashtra, India. Arab J Geosci. 11(19):576.
Das S. 2019. Geospatial mapping of flood susceptibility and hydro-geomorphic response to the floods in Ulhas basin, India. Remote Sens Appl: Soc Environ. 14:60–74.
Elith J, Leathwick JR, Hastie T. 2008. A working guide to boosted regression trees. J Anim Ecol. 77(4): 802–813.
Elsafi SH. 2014. Artificial neural networks (ANNs) for flood forecasting at Dongola Station in the River Nile, Sudan. Alexandria Eng J. 53(3):655–662.
Friedman J, Hastie T, Tibshirani R. 2000. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). Ann Statist. 28(2):337–407.
Haq M, Akhtar M, Muhammad S, Paras S, Rahmatullah J. 2012. Techniques of remote sensing and GIS for flood monitoring and damage assessment: a case study of Sindh province, Pakistan. Egyptian J Remote Sens Space Sci. 15(2):135–141.
Hong H, Liu J, Bui DT, Pradhan B, Acharya TD, Pham BT, Zhu AX, Chen W, Ahmad BB. 2018a. Landslide susceptibility mapping using J48 Decision Tree with AdaBoost, bagging and rotation forest ensembles in the Guangchang area (China). Catena. 163:399–413.
Hong H, Panahi M, Shirzadi A, Ma T, Liu J, Zhu AX, Chen W, Kougias I, Kazakis N. 2018b. Flood susceptibility assessment in Hengfeng area coupling adaptive neuro-fuzzy inference system with genetic algorithm and differential evolution. Sci Total Environ. 621:1124–1141.
Hong H, Pradhan B, Xu C, Bui DT. 2015. Spatial prediction of landslide hazard at the Yihuang area (China) using two-class kernel logistic regression, alternating decision tree and support vector machines. Catena. 133:266–281.
Hong H, Tsangaratos P, Ilia I, Liu J, Zhu AX, Chen W. 2018c. Application of fuzzy weight of evidence and data mining techniques in construction of flood susceptibility map of Poyang County, China. Sci Total Environ. 625:575–588.
Kale VS. 2003. Geomorphic effects of monsoon floods on Indian rivers. In: Mirza MMQ, Dixit A, Nishat A, editors. Flood problem and management in South Asia. Dordrecht: Springer; p. 65–84.
Khosravi K, Nohani E, Maroufinia E, Pourghasemi HR. 2016a. A GIS-based flood susceptibility assessment and its mapping in Iran: a comparison between frequency ratio and weights-of-evidence bivariate statistical models with multi-criteria decision-making technique. Nat Hazards. 83(2):947–987.
Khosravi K, Pham BT, Chapi K, Shirzadi A, Shahabi H, Revhaug I, Prakash I, Bui DT. 2018. A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at Haraz watershed, northern Iran. Sci Total Environ. 627:744–755.

Khosravi K, Pourghasemi HR, Chapi K, Bahri M. 2016b. Flash flood susceptibility analysis and its mapping using different bivariate models in Iran: a comparison between Shannon's entropy, statistical index, and weighting factor models. Environ Monit Assess. 188(12):656.
Kordestani MD, Naghibi SA, Hashemi H, Ahmadi K, Kalantar B, Pradhan B. 2019. Groundwater potential mapping using a novel data-mining ensemble model. Hydrogeol J. 27(1):211–224.
Kourgialas NN, Karatzas GP. 2017. A national scale flood hazard mapping methodology: the case of Greece–protection and adaptation policy approaches. Sci Total Environ. 601–602:441–452.
Kumar R, Singh R, Gautam H, Pandey MK. 2018. Flood hazard assessment of August 20, 2016 floods in Satna district, Madhya Pradesh, India. Remote Sens Appl: Soc Environ. 11:104–118.
Kuncheva LI. 2014. Combining pattern classifiers: methods and algorithms. New Jersey: John Wiley and Sons.
Kursa MB, Rudnicki WR. 2010. Feature selection with the Boruta package. J Stat Softw. 36(11):1–13.
Liu K, Li Z, Yao C, Chen J, Zhang K, Saifullah M. 2016. Coupling the k-nearest neighbor procedure with the Kalman filter for real-time updating of the hydraulic model in flood forecasting. Int J Sediment Res. 31(2):149–158.
Lu Y, Qin XS, Xie YJ. 2016. An integrated statistical and data-driven framework for supporting flood risk analysis under climate change. J Hydrol. 533:28–39.
Mahmoud SH, Gan TY. 2018. Multi-criteria approach to develop flood susceptibility maps in arid regions of Middle East. J Cleaner Prod. 196:216–229.
Mason SJ, Baddour O. 2008. Statistical modelling. In: Troccoli A, Harrison M, Anderson DLT, Mason SJ, editors. Seasonal climate: forecasting and managing risk. Dordrecht: Springer; p. 163–201.
Merz B, Thieken AH, Gocht M. 2007. Flood risk mapping at the local scale: concepts and challenges. In: Begum S, Stive MJF, Hall JW, editors. Flood risk management in Europe. Dordrecht: Springer; p. 231–251.
Miller JR, Ritter DF, Kochel CR. 1990. Morphometric assessment of lithologic controls on drainage basin evolution in the crawford upland, south-central Indiana. Am J Sci. 290(5):569–599.
Mojaddadi H, Pradhan B, Nampak H, Ahmad N, Ghazali AHB. 2017. Ensemble machine-learning-based geospatial approach for flood risk assessment using multi-sensor remote-sensing data and GIS. Geomat Nat Hazards Risk. 8(2):1080–1102.
Moore ID, Grayson RB. 1991. Terrain-based catchment partitioning and runoff prediction using vector elevation data. Water Resour Res. 27(6):1177–1191.
Moore ID, Grayson RB, Ladson AR. 1991. Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. Hydrol Process. 5(1):3–30.
Mosavi A, Hosseini FS, Choubin B, Goodarzi M, Dineva AA. 2020. Groundwater salinity susceptibility mapping using classifier ensemble and Bayesian machine learning models. IEEE Access. 8: 145564–145576.
Nandi A, Mandal A, Wilson M, Smith D. 2016. Flood hazard mapping in Jamaica using principal component analysis and logistic regression. Environ Earth Sci. 75(6):465.
Naghibi SA, Dolatkordestani M, Rezaei A, Amouzegari P, Heravi MT, Kalantar B, Pradhan B. 2019. Application of rotation forest with decision trees as base classifier and a novel ensemble model in spatial modeling of groundwater potential. Environ Monit Assess. 191(4):248.
Naghibi SA, Ahmadi K, Daneshi A. 2017. Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping. Water Resour Manage. 31(9):2761–2775.
Ngo PTT, Hoang ND, Pradhan B, Nguyen QK, Tran XT, Nguyen QM, Nguyen VN, Samui P, Tien Bui D. 2018. A novel hybrid swarm optimized multilayer neural network for spatial prediction of flash floods in tropical areas using sentinel-1 SAR imagery and geospatial data. Sensors. 18(11):3704.
Nguyen QK, Tien Bui D, Hoang ND, Trinh PT, Nguyen VH, Yilmaz I. 2017. A novel hybrid approach based on instance based learning classifier and rotation forest ensemble for spatial prediction of rainfall-induced shallow landslides using GIS. Sustainability. 9(5):813.
Oh HJ, Syifa M, Lee CW, Lee S. 2019. Land subsidence susceptibility mapping using bayesian, functional, and meta-ensemble machine learning models. Appl Sci. 9(6):1248.
Pardo M, Sberveglieri G. 2008. Random forests and nearest shrunken centroids for the classification of sensor array data. Sens Actuators B. 131(1):93–99.
Pradhan B, Lee S. 2010. Landslide susceptibility assessment and factor effect analysis: backpropagation artificial neural networks and their comparison with frequency ratio and bivariate logistic regression modelling. Environ Model Softw. 25(6):747–759.
Prasad P, Loveson VJ, Kotha M, Yadav R. 2020. Application of machine learning techniques in groundwater potential mapping along the west coast of India. GIScience Remote Sensing. 57(6):735–752.

Pulvirenti L, Pierdicca N, Chini M, Guerriero L. 2011. An algorithm for operational flood mapping from synthetic aperture radar (SAR) data based on the fuzzy logic. Nat Hazards Earth Syst Sci. 11(2): 529–540.
Rahmati O, Pourghasemi HR. 2017. Identification of critical flood prone areas in data-scarce and ungauged regions: a comparison of three data mining models. Water Resour Manage. 31(5):1473–1487.
Rahmati O, Pourghasemi HR, Zeinivand H. 2016. Flood susceptibility mapping using frequency ratio and weights-of-evidence models in the Golastan Province, Iran. Geocarto Int. 31(1):42–70.
Riley SJ, DeGloria SD, Elliot R. 1999. Index that quantifies topographic heterogeneity. Intermountain J Sci. 5(1–4):23–27.
Rodriguez JJ, Kuncheva LI, Alonso CJ. 2006. Rotation forest: a new classifier ensemble method. IEEE Trans Pattern Anal Mach Intell. 28(10):1619–1630.
Roodposhti MS, Safarrad T, Shahabi H. 2017. Drought sensitivity mapping using two one-class support vector machine algorithms. Atmos Res. 193:73–82.
Sahana M, Patel PP. 2019. A comparison of frequency ratio and fuzzy logic models for flood susceptibility assessment of the lower Kosi River Basin in India. Environ Earth Sci. 78(10):289.
Sahoo GB, Schladow SG, Reuter JE. 2009. Forecasting stream water temperature using regression analysis, artificial neural network, and chaotic non-linear dynamic models. J Hydrol. 378(3–4):325–342.
Shafizadeh-Moghadam H, Asghari A, Tayyebi A, Taleai M. 2017. Coupling machine learning, tree-based and statistical models with cellular automata to simulate urban growth. Comput Environ Urban Syst. 64:297–308.
Shafizadeh-Moghadam H, Valavi R, Shahabi H, Chapi K, Shirzadi A. 2018. Novel forecasting approaches using combination of machine learning and statistical models for flood susceptibility mapping. J Environ Manage. 217:1–11.
Shahabi H, Shirzadi A, Ghaderi K, Omidvar E, Al-Ansari N, Clague JJ, Geertsema M, Khosravi K, Amini A, Bahrami S, et al. 2020. Flood detection and susceptibility mapping using Sentinel-1 remote sensing data and a machine learning approach: hybrid intelligence of bagging ensemble based on K-nearest neighbor classifier. Remote Sensing. 12(2):266.
Tehrany MS, Lee MJ, Pradhan B, Jebur MN, Lee S. 2014b. Flood susceptibility mapping using integrated bivariate and multivariate statistical models. Environ Earth Sci. 72(10):4001–4015.
Tehrany MS, Pradhan B, Jebur MN. 2013. Spatial prediction of flood susceptible areas using rule based decision tree (DT) and a novel ensemble bivariate and multivariate statistical models in GIS. J Hydrol. 504:69–79.
Tehrany MS, Pradhan B, Jebur MN. 2014a. Flood susceptibility mapping using a novel ensemble weights-of-evidence and support vector machine models in GIS. J Hydrol. 512:332–343.
Tehrany MS, Pradhan B, Jebur MN. 2015. Flood susceptibility analysis and its verification using a novel ensemble support vector machine and frequency ratio method. Stoch Environ Res Risk Assess. 29(4): 1149–1165.
Termeh SVR, Kornejady A, Pourghasemi HR, Keesstra S. 2018. Flood susceptibility mapping using novel ensembles of adaptive neuro fuzzy inference system and metaheuristic algorithms. Sci Total Environ. 615:438–451.
Tibshirani R, Hastie T, Narasimhan B, Chu G. 2003. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. Statist Sci. 18(1):104–117.
Wang S, Zhu J. 2007. Improved centroids estimation for the nearest shrunken centroid classifier. Bioinformatics. 23(8):972–979.
Xu H. 2006. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. Int J Remote Sens. 27(14):3025–3033.
Youssef AM, Pradhan B, Hassan AM. 2011. Flash flood risk estimation along the St. Katherine road, southern Sinai, Egypt using GIS based morphometry and satellite imagery. Environ Earth Sci. 62(3): 611–623.
Zhang X, Dong Q, Chen J. 2019. Comparison of ensemble models for drought prediction based on climate indexes. Stoch Environ Res Risk Assess. 33(2):593–606.
Zhao G, Pang B, Xu Z, Yue J, Tu T. 2018. Mapping flood susceptibility in mountainous areas on a national scale in China. Sci Total Environ. 615:1133–1142.