Environmental Sciences Europe

RESEARCH

Open Access

Prediction of deforestation risk in north-east India: evaluating forest canopy density dynamics and spatial drivers through machine learning models

Sandipan Das[1*], Debarshi Ghosh[2*], Apurba Sarkar[3*], Uday Chatterjee[4], Sanjoy Mandal[5], Sujoy Kumar Malo[5], Saidur Rahaman[5], Mantu Das[5], Snehasish Saha[5] and Pradip Chouhan[6]

Abstract

Rapid population growth in North-East India has led to urban sprawl, agricultural expansion, and industrialization, resulting in significant deforestation. Numerous studies have utilized Forest Canopy Density (FCD) coupled with remote sensing and GIS to assess forest cover changes. This study evaluates how FCD changed from 2000 to 2020, highlighting deforestation patterns in North-East India. Machine learning models, including Binary Logistic Regression (BLR), Random Forest (RF), REP-Tree, and XGBoost Regression (XGBR), were used to identify areas at risk of deforestation. Influencing factors considered were forest density, barren land, agricultural land, urban areas, distance to roads, and topographical characteristics. Results indicated that proximity to roads notably increases deforestation risk. Higher densities of agricultural and urban land also contribute to greater deforestation rates, whereas increased forest and barren land densities reduce this risk. Among the tested models, Random Forest showed superior performance with a high True Positive Rate (TPR) and low False Positive Rate (FPR), effectively identifying high-risk deforestation zones. Analysis also revealed a concerning shift from high-density to low-density forests, signifying substantial forest cover loss and potential threats to biodiversity and ecosystem services. The findings emphasize the need for integrated land-use planning and targeted conservation, specifically addressing road proximity, to effectively combat deforestation in North-East India and similar regions.

Keywords Forest canopy density, Deforestation, Machine learning algorithms, Vegetation density, Ensemble model, Scaled shadow index

*Correspondence:
Sandipan Das
sandipan@sig.ac.in; sandipanraj2002@gmail.com
Debarshi Ghosh
wetlanddeb@gmail.com
Apurba Sarkar
apurbasarkarugb@gmail.com
[1] Symbiosis Institute of Geoinformatics, Symbiosis International (Deemed University), Pune, Maharashtra 411016, India
[2] Department of Geography, Dhupguri Girls' College, Dhupguri, West Bengal, India
[3] Department of Geography, Gour Banga University, Malda, West Bengal, India
[4] Department of Geography, Bhatter College (Autonomous), Chaulia, West Bengal, India
[5] Department of Geography and Applied Geography, University of North Bengal, Rajarammohunpur, Siliguri, West Bengal, India
[6] Department of Geography, University of Gour Banga, Malda, India

Springer Open

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Introduction

Forests cover approximately 31% of Earth's terrestrial surface and provide critical ecosystem services such as biodiversity conservation, carbon sequestration, climate moderation, and hydrological regulation [49, 52]. Tropical forests, in particular, have faced alarming declines due to intensifying anthropogenic pressures, including agriculture, grazing, timber extraction, and infrastructure development [51, 94]. These pressures lead to a loss of approximately 13.5 million hectares of tropical forest annually. Urbanization and infrastructure development alone account for approximately 80% and 15% of global deforestation, respectively, with agriculture contributing the remaining 5% [43, 84]. Recent national-scale analyses show fractional vegetation cover (FVC) rises with rainfall but declines with higher land-surface temperature and urbanization (CNLI), underscoring a vegetation–urbanization trade-off in South Asia [2]. Deforestation is driven by both proximate causes, such as logging, mining, grazing, and fires, and underlying factors like socioeconomic policies, market dynamics, and land-use changes [84]. Consequences include biodiversity loss [87], soil erosion, hydrological disturbances, microclimatic changes, and increased atmospheric CO_2 concentrations contributing to global warming [50, 92]. While many studies emphasize statistical associations between deforestation and its drivers, fewer explore the causal ecological processes underpinning forest canopy density (FCD) changes, such as climate patterns, soil moisture, and disturbance regimes [6, 45]. Mechanistic evidence from Khyber Pakhtunkhwa links interannual NDVI variability directly to temperature, precipitation, and surface latent heat flux, highlighting strong climatic control on canopy greenness [3, 42].

India, a mega-diverse nation with approximately 29% forest cover, exemplifies these deforestation challenges. North-East India, a critical part of the Indo-Myanmar global biodiversity hotspot, is characterized by diverse forest types, steep mountainous terrain, and highly variable rainfall patterns. Unique regional pressures, including traditional shifting cultivation (Jhum), forest fires, and expanding infrastructure, exacerbate deforestation and forest degradation. In analogousto mountain conifer systems, burn severity mapped via NBR/dNBR (Normalized Burn Ratio and Differenced Normalized Burn Ratio) and machine-learning (RF, XGBoost, logistic regression) accurately delineated high-risk zones, illustrating a transferable Earth Observation (EO)-driven template for susceptibility mapping [60, 63].

These activities impact the forest canopy through cyclical clearing and regrowth, soil nutrient depletion, and heightened erosion susceptibility. Despite the ecological significance of North-East India's forests, this region is underrepresented in global deforestation research. Most existing analyses are primarily descriptive or correlational, emphasizing forest cover change rates or simple demographic associations rather than mechanistic drivers. This research gap highlights the need for integrated, mechanistic approaches capable of not only mapping deforestation but also revealing ecological processes driving canopy loss [7, 12, 35].

This study employs advanced remote sensing technologies combined with machine learning (ML) models [82, 85, 94] to address these gaps. Comparable MODIS-based frameworks have mapped Pakistan-wide FVC trajectories (2003–2020) and ranked drivers, with Random Forest explaining ~ 89% of variance using rainfall,land surface temperature (LST), and convolutional neural network (CNLI) [2]. We use high-resolution FCD mapping spanning two decades (2000–2020) and apply multiple ML models—Binary Logistic Regression (BLR), Random Forest (RF), REP-Tree, and Gradient Boosted Regression (XGBR) models—to predict deforestation risk zones [77]. RF and REP-Tree are included for their ability to handle complex, high-dimensional ecological datasets efficiently and robustly [65]. RF, in particular, is justified by its strong predictive accuracy, capacity to manage large datasets, and inherent ability to quantify predictor importance [56, 86]. REP-Tree complements RF by offering transparency in model construction and ease of interpretation [26, 93]. Our models integrate multispectral vegetation indices such as the Advanced Vegetation Index (AVI), Bare Soil Index (BSI), Shadow Index (SI), and Vegetation Density (VD), alongside topographical elevation, slope and anthropogenic factors (proximity to roads). This comprehensive predictor set helps identify causal linkages, for instance, how topographic conditions might mitigate fire risks or how infrastructure proximity may facilitate deforestation. The specific objectives of this study are to: (a) quantify spatiotemporal changes in Forest Canopy Density (FCD) across North-East India from 2000 to 2020, (b) predict Deforestation Potential Zones(DPZ) using ML models based on biophysical and anthropogenic variables and evaluate their performance; and (c) bridge empirical findings and actionable conservation insights. Our findings aim to inform regional land-use planning, climate adaptation strategies, and policy interventions crucial for forest ecosystem resilience in North-East India.

Study area

The study area encompasses the north-eastern states of India-Assam, Meghalaya, Arunachal Pradesh, Nagaland, Tripura, Manipur, and Mizoram collectively known as the "Seven Sisters" [64] (Fig. 1). This geographically distinct region, located between latitudes 22°00′N and

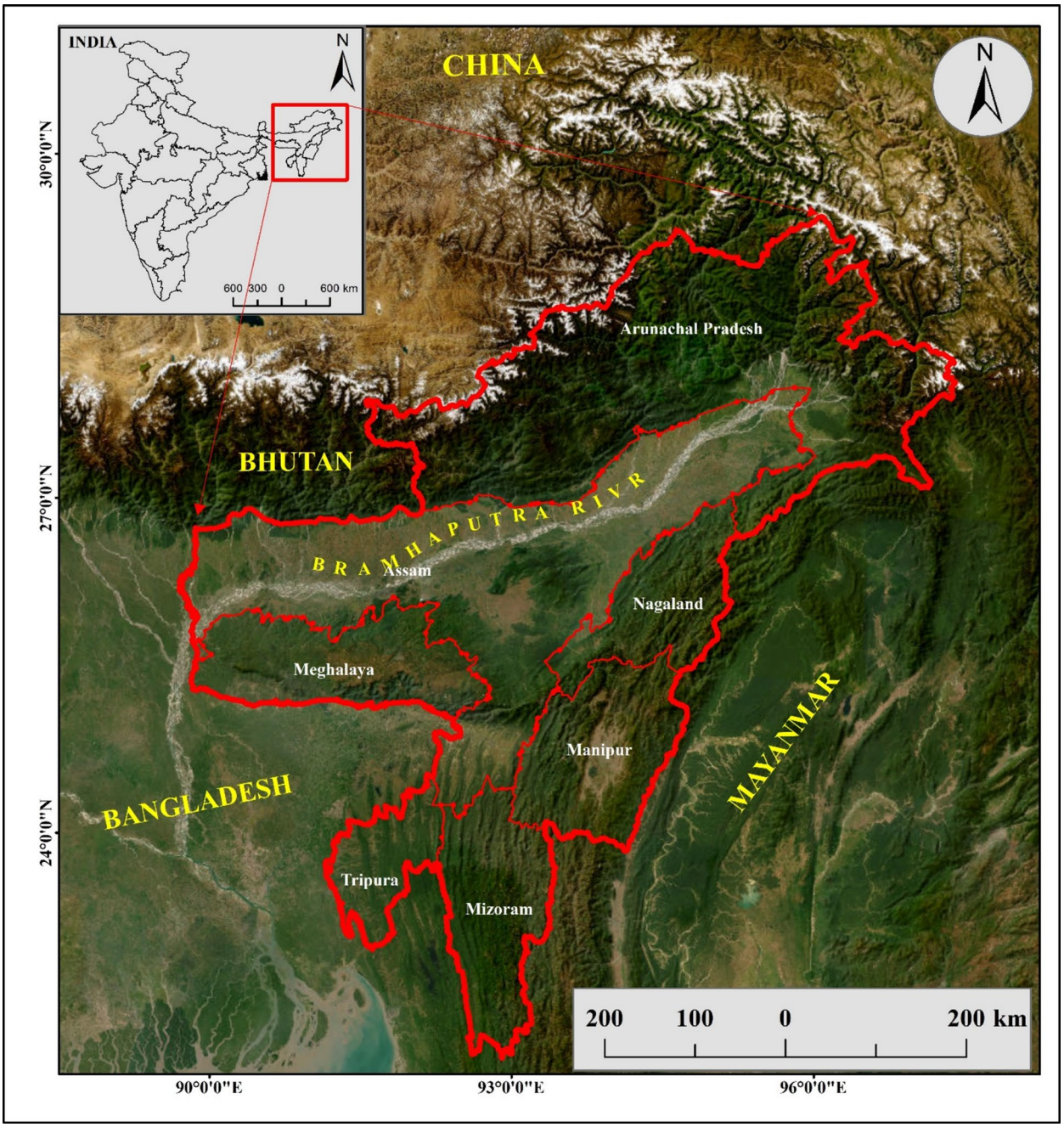

Fig. 1 Location of the study area

29°05′N and longitudes 88°00′E and 97°30′E, shares international borders with Bangladesh, Bhutan, China, and Myanmar. Renowned for its extraordinary ecological diversity, North-East India is a vital part of the Indo-Myanmar biodiversity hotspot, one of the world's 25 globally recognized biodiversity hotspots. Characterized by complex mountainous terrains and deep valleys, the region exhibits remarkable elevation gradients, rising sharply from low-lying Brahmaputra plains (~20 m) to rugged peaks exceeding 6000 m in Arunachal Pradesh [13, 19, 36]. These elevation variations create diverse climatic zones ranging from tropical and subtropical in the plains and foothills to temperate and alpine conditions in higher elevations. The region experiences a predominantly humid subtropical climate, characterized by abundant rainfall during the monsoon season, typically

between June and September, with average annual precipitation ranging from ≈2000 mm in plains to over 4000 mm in some mountainous locations such as Cherrapunji in Meghalaya, one of the wettest places on Earth. Temperature patterns also vary widely, with annual averages ranging from 10 °C to 15 °C in mountainous regions to 25 °C to 30 °C in the plains.

According to Koppen's climatic classification, North-East India predominantly falls under the humid subtropical climate zone (Cwa), with regions of high elevation classified as subtropical highland (Cwb) and Alpine climates (ET). This intricate climatic and topographic mosaic supports an exceptionally high level of biodiversity, with the region harbouring approximately half of India's floral diversity. Over 7500 flowering plant species, including 700 orchids, 63 bamboo species, 64 citrus varieties, and around 28 conifers, thrive within these ecosystems, complemented by diverse assemblages of mosses, ferns, and lichens [33]. Remarkably, about one-third of the region's vegetation is endemic, meaning it occurs nowhere else on the planet. Notable endemic and rare plant species include the fast-growing Pinus kesiya, which thrives at elevations between 900 and 1800 m; Cycas pectinata, reported from the Kamrup district of Assam; and Gnetum gnemon, found in the Khasi Hills and Barak Valley [33]. The region is also home to globally significant fauna, serving as a crucial habitat for many endangered and charismatic wildlife species such as the one-horned Indian rhinoceros, clouded leopard, and numerous endemic bird species [58]. However, this ecological treasure is under constant threat due to anthropogenic pressures, primarily driven by shifting cultivation practices locally known as 'Jhum', extensive deforestation for agricultural expansion, mining, infrastructure development, and urbanization.

Sources and database

To create the Forest Canopy Density (FCD) and Deforestation Probability Maps (DPM) of the study area, this research utilized a diverse array of data sources. These included Landsat TM & OLI imagery, COP Digital Elevation Model (DEM), Google Earth Pro, and Earth Explorer data, which are freely accessible. Additionally, field data were selected for validation purposes. Table 1 lists the data sources, and Table 2 provides a thorough description of the path/row, bands, spatial resolution, brightness, and reflectance of satellite pictures, as well as the SOI topography sheets. The SPSS 22 software was used to run the BLR model. The Gradient Boosted Regression Model (XGBR), Random Forest, and Rep-Tree models were run in R Studio using CRAN packages. All statistical analyses and machine learning models were implemented using the R programming language (version 4.3.1). The Random Forest (RF) model was executed with the randomForest package, REP-Tree was implemented using the RWeka package, and XGBoost regression was performed using the xgboost package. Binary Logistic Regression (BLR) analyses were conducted using SPSS version 22.0.

Table 1 Data sources

Data layers	Data format	Data Source
Advanced Vegetation Index (AVI), Bare Soil Index (BSI), Shadow Index (SI)	Raster grid	LANDSAT/LC08/C02/T1_L2 LANDSAT/LT05/C02/T1_TOA https://earthexplorer.usgs.gov
Slope, Elevation	Raster grid	COP DEM 30 m (Digital Elevation Model) 10.5270/ESA-c5d3d65
Distance to Settlement (DFST)	Point	Land Cover Type (MCD12Q1) Version 6.1 10.5067/MODIS/MCD12Q1.061
Distance To Road (DTR)	Polyline	https://www.openstreetmap.org/export
Distance To Railway (DTRL)	Polyline	Google Earth Pro
Agricultural Land Density (AD)	Raster grid	Land Cover Type (MCD12Q1) Version 6.1 10.5067/MODIS/MCD12Q1.061
Population Density (PD)	Raster grid	https://www.worldpop.org/ 10.1038/sdata.2015.45
Distance from Burn Patches (DBP)	Raster grid	The Terra and Aqua combined MCD64A1 Version 6.1 https://lpdaac.usgs.gov
Forest Density (FD)	Raster grid	Land Cover Type (MCD12Q1) Version 6.1 10.5067/MODIS/MCD12Q1.061
Distance to Dam (DDam)	Raster gird	Google Earth Pro
Edge Density of Forest covers (ED)	Raster gird	Land Cover Type (MCD12Q1) Version 6.1 10.5067/MODIS/MCD12Q1.061
Deforestation Validation	Pixel data	Analysis of satellite imageries and field validation

Table 2 Details of data used in this study

Data	Path/row	Band Details	Spatial resolution(m)	Radiance(max)	Reflectance
LANDSAT 8 OLI TIRS (Source: https://earthexplorer.usgs.gov)	139/43	Band 1 = coastal aerosol (0.43–0.45um)	30	766.38531	1.210700 (max) –0.099980 (min)
	139/43	Band 2 = Blue (0.45–0.51um)	30	784.78790	0.311577 (max)-0.002423 (min)
		Band 3 = Green (0.53–0.59um)	30	723.17535	0.651225 (max)-0.005003 (min)
		Band 4 = Red (0.64–0.67um)	30	609.82220	0.556060 (max)-0.002464 (min)
		Band 5 = NIR (085–0.88um)	30	373.18076	0.671423 (max)-0.004588 (min)
		Band 6 = SWIR1 (1.57–1.65um)	30	92.80666	0.452189 (max)-0.005540 (min)
		Band 7 = SWRI2 (2.11–2.29um)	15	31.28081	–
		Band 8 = Panchromatic (0.50–0.68um)	30	690.15088	0.629659 (max)-0.005724 (min)
		Band 9 = Cirrus (1.36–1.38um)	100	145.84750	
		Band 10 = Thermal infrared1 (10.6–11.19um)	100	22.00180	
		Band 11 = Thermal infrared2 (11.5–12.51um)	30	22.00180	
			30	193.0	
			30	365.0	
			30	264.0	
LANDSAT 5 TM (source: https://earthexplorer.usgs.gov)		Band 1 = Blue (0.45–0.52um)	30	221.0	
		Band 2 = Green (0.52–0.60um)	30	30.2	
		Band 3 = Red (0.63–0.69um)	120	15.303	
		Band 4 = NIR (0.76–0.90um)	30	16.5	
		Band 5 = SWIR1 (1.55–1.75um)			
		Band 6 = Thermal (10.40–12.50um)			
		Band 7 = SWIR2 (2.08–2.35um) No. (72 O/12, 72 P/5, 72 P/6, 72 P/13, 72 P/14)			

Data processing and visualization tasks were performed in R, with additional spatial analyses and image pre-processing carried out in ArcGIS 10.3 and ENVI 6.1.

Methods and materials

The process of mapping deforestation probability in the study area using Forest Canopy Density (FCD) and binary logistic regression comprised six stages (Fig. 2): (i) we assembled thematic data layers describing potential drivers of deforestation; the variables and units are: Elevation (m); Forest Density (FD, %); Agricultural Density (AD, %); Distance from Dam (DDam, in meter); Barren Land Density (BLD, %); Distance from Burn Patches (DBP, in meter); Population Density (PD, persons km^2); Slope (°); Distance to Road (DTR, in meter); Distance to Railway (DTRL, in meter); Distance from Settlement (DFST, in meter); and Edge Density of forest patches (ED, in meter ha^{-1}); (ii) we examined collinearity among predictors; (iii) we specified the predictive modelling framework; (iv) following preparation of the FCD maps, the deforestation area for the entire North-East was delineated and used as the dependent variable for probability modelling; (v) Binary Logistic Regression (BLR), Gradient-Boosted Regression (XGBR), Random Forest, and REP-Tree models were fitted; (vi) model performance was validated using the Receiver Operating Characteristic (ROC) curve, efficiency metrics, the True Skill Statistic (TSS), and the Kappa coefficient.

Drivers of deforestation

Deforestation probability models select variables like elevation, forest density, agricultural density, distance from dam, barren land density, distance from burn patches, population density, slope, distance to road, distance to railway, distance from settlement, and edge density of forest patches [76, 78]. Slope influences deforestation by affecting invasive species spread and vegetation cover. Elevation helps identify deforestation-prone areas, with lower altitudes affected by grazing, wood smuggling, agricultural expansion, and human occupation. Distance from forest edges impacts deforestation rates, with greater distances reducing human activity. Fragmented forest edges are more prone to encroachment. Forest density, indicating the number of trees in an area, assesses forest health and deforestation probabilities [11]. Barren land density and agricultural density provide insights into deforestation extent and human impact, guiding conservation efforts. Road and rail networks are significant contributors, with proximity increasing risk due to forest fragmentation [79]. Settlement expansion inversely relates to forest cover, as higher density increases land and resource demand, leading to deforestation. Remote sensing and GIS-based machine learning models, including binary logistic regression, gradient boosted regression, random forest, and Rep-Tree models, predict deforestation zones [21, 69]. This study validates deforestation factors using statistical methods, offering

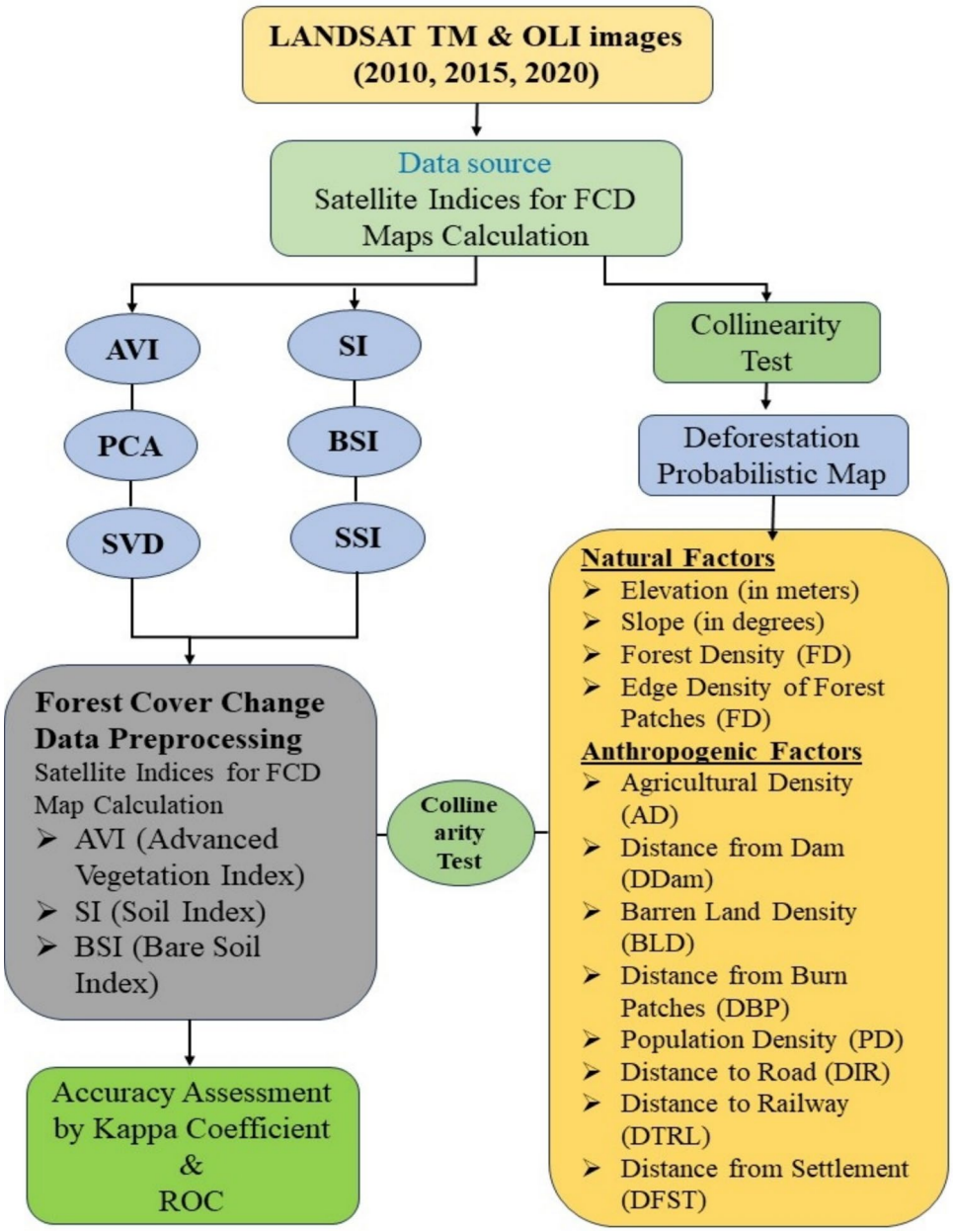

Fig. 2 Methodological flow diagram of the work [71, 76, 78]

valuable insights for future research and effective forest management and conservation [88].

Deforested area demarcation based on forest canopy density (FCD)

The Forest Canopy Density (FCD) index determines deforested area boundaries, assessing vegetation health and distribution via vegetation, bare soil, shadow indices, and temperature [38, 76, 78]. Creating FCD maps in a GIS environment involves a sequential process using four indices: Advanced Vegetation Index (AVI), Bare Soil Index (BSI), Shadow Index (SI), and Scaled Shadow Index (SSI). These indices differentiate dense forests from degraded areas. Satellite images undergo pre-processing in ENVI (6.1) and Arc GIS (10.3) for AVI, BSI, SI, and SSI operations, including radiometric and geometric corrections and atmospheric adjustments. Landsat image processing involves techniques like Dark Object Subtraction (DOS), Enhanced DOS, Solar Spectrum Vector (SSV), and the Empirical Line Method (ELM) to reduce

atmospheric effects [37]. For all spectral index calculations, band references correspond to the Landsat sensor in use (Table 2). These steps generate FCD and deforestation probability maps, essential for monitoring temporal and spatial vegetation quality.

Advanced vegetation index (AVI)

AVI is a useful tool for assessing vegetation health as it takes into account both the red and near-infrared bands. By calculating the AVI using the given equation, it is possible to determine the healthy vegetation area based on canopy density [73, 74]. This information can be valuable for various applications such as monitoring crop health or assessing ecosystem vitality. AVI has been calculated using the following equation:

$$\mathrm{AVI} = [(\mathrm{B4} + 1) \times (256 - \mathrm{B3}) \times (\mathrm{B4} - \mathrm{B3})]^{\frac{1}{3}} \quad (1)$$

In this context, B3 is indicative of the red band, while B4 corresponds to the near-infrared (NIR) band. A higher value of the Advanced Vegetation Index (AVI) signifies more extensive forest coverage.

Bare soil index (BSI)

The Bare Soil Index (BSI) is a tool used to identify healthy vegetation areas in exposed soil areas. It uses SWIR1, Red, NIR, and Blue bands to measure soil mineral composition and vegetation condition. A higher BSI value indicates more exposed soil, low vegetation density, and more deforested areas. The BSI ranges from 0 to 200, with higher values indicating more deforestation [79]. The index combines information from different bands to assess soil mineral composition and vegetation condition, providing a comprehensive measure of deforestation. BSI has been calculated using the following equation:

$$\mathrm{BSI} = \left[\frac{(\mathrm{SWIR1} + \mathrm{RED}) - (\mathrm{NIR} + \mathrm{BLUE})}{(\mathrm{SWIR1} + \mathrm{RED}) + (\mathrm{NIR} + \mathrm{BLUE})}\right] \quad (2)$$

Shadow index (SI)

The Shadow Index (SI) effectively identifies dense vegetation and canopy cover, indicating mature and healthy forests, and provides insights into biodiversity and ecological stability by evaluating the spectral and thermal properties of forest shadows [10, 11, 31]. The constants used in the spectral indices, such as the value 256 in the Shadow Index (SI), are related to the 8-bit radiometric resolution of Landsat satellite data. Since Landsat imagery typically encodes pixel values in an 8-bit format, the data range is from 0 to 255, resulting in 256 discrete levels of brightness [20, 28]. Therefore, the constant 256 serves as a baseline offset to invert reflectance values for the identification of shadow areas effectively. The Scaled Shadow Index (SSI), a linear transformation of SI, ranges from 0 to 100%, with higher values indicating greater canopy shadow, differentiating canopy cover from ground vegetation [71]. The SI is calculated using the following equation:

$$\mathrm{SI} = [(256 - \mathrm{B1}) \times (256 - \mathrm{B2}) \times (256 - \mathrm{B3})]^{\frac{1}{3}} \quad (3)$$

Scaled shadow index (SSI)

The Scaled Shadow Index (SSI), derived from a linear transformation of the Shadow Index (SI), is a standardized metric that ranges from 0 to 100%, indicating the extent of forest shadows [71]. A higher SSI value denotes greater canopy shadow, indicative of denser forest cover, whereas a lower SSI value signifies minimal shadow, characteristic of open or sparsely vegetated areas [47]. Normalization across indices such as AVI, BSI, SI, and SSI is applied to standardize the values within a consistent range (usually 0–100%) to facilitate direct comparisons across different years and sensor platforms, thus enabling accurate temporal analysis of forest canopy dynamics [66, 91, 95]. The SSI is calculated following the equation below:

$$\mathrm{SSI} = 100 \times \frac{(SI - SImin)}{(SImax - SImin)} \quad (4)$$

Vegetation density (VD)

We used Principal Component Analysis (PCA) to derive (i.e., integrate) a single Vegetation Density (VD) metric by fusing two spectral indices -Advanced Vegetation Index (AVI) and Bare Soil Index (BSI). Only these two variables were included for PCA in the ARC-GIS 10.8 environment. Before analysis, AVI and BSI were standardized to comparable scales, and a two-band stack AVI and BSI was decomposed into orthogonal components based on eigenvalues/eigenvectors [24]. We used the ArcGIS Spatial Analyst-Principal Components tool to derive an integrated Vegetation Density (VD) metric by fusing two indices: Advanced Vegetation Index (AVI) and Bare Soil Index (BSI). Only these two rasters were included in the PCA. Both inputs were co-registered, clipped to a common extent, and stacked as a two-band raster AVI, BSI. The tool was run with Number of principal components = 2, producing (i) a two-band multiband raster (PC1, PC2) and (ii) an ASCII statistics file with eigenvalues, eigenvectors (loadings), and percent variance explained. PC1, which captured the dominant variance and showed the expected sign pattern (positive loading on AVI; negative loading on BSI), was interpreted as the vegetation signal and used as the VD surface. Where

necessary, the PC1 sign was oriented so that larger values indicate denser vegetation [23, 41, 91]. VD was then min–max rescaled to 0–100% to enable temporal and spatial comparisons across years [24, 96].

Forest canopy density (FCD)

The Forest Canopy Density (FCD) [24] evaluates the quality and health of forested regions by combining Vegetation Density (VD) and Scaled Shadow Index (SSI) values, ranging from 0 to 100% [30, 71, 76, 78, 81]. A higher FCD value indicates a denser and healthier forest canopy, while a lower value suggests the opposite [71]. The FCD can be determined by utilizing the equation, which is as follows:

$$\mathrm{FCD} = \left(\frac{(\mathrm{VD} \times \mathrm{SSI}) + 1}{1/2} \right) - 1 \tag{5}$$

Identification of deforested regions using forest canopy density (FCD)

FCD maps for the years 2000, 2010, 2015, and 2021 were created using indices like AVI, SSI, and VD, with a 70% threshold to categorize areas as forest (1) or non-forest (0) [76, 78, 80]. The reclassified FCD maps were subtracted to produce a Forest Change Map (FCM) for 2000–2021, used as the dependent variable in the BLR model to investigate the relationship between forest changes and factors like population density, land use, and climate variables (Fig. 3). This model helps identify drivers of deforestation or afforestation, providing valuable insights for effective forest management and conservation.

Techniques used for mapping the probability of deforestation

Three unique modelling methodologies were selected to generate a deforestation probability map for the basin in which the research was conducted. These methods make use of a mixture of probabilistic and machine learning techniques, both of which contribute to an increase in the predictability and precision of the results. The three models that were chosen are as follows:

Binary logistic regression (BLR) This study uses the Binary Logistic Regression (BLR) model for predictive analysis, focusing on the relationship between a binary

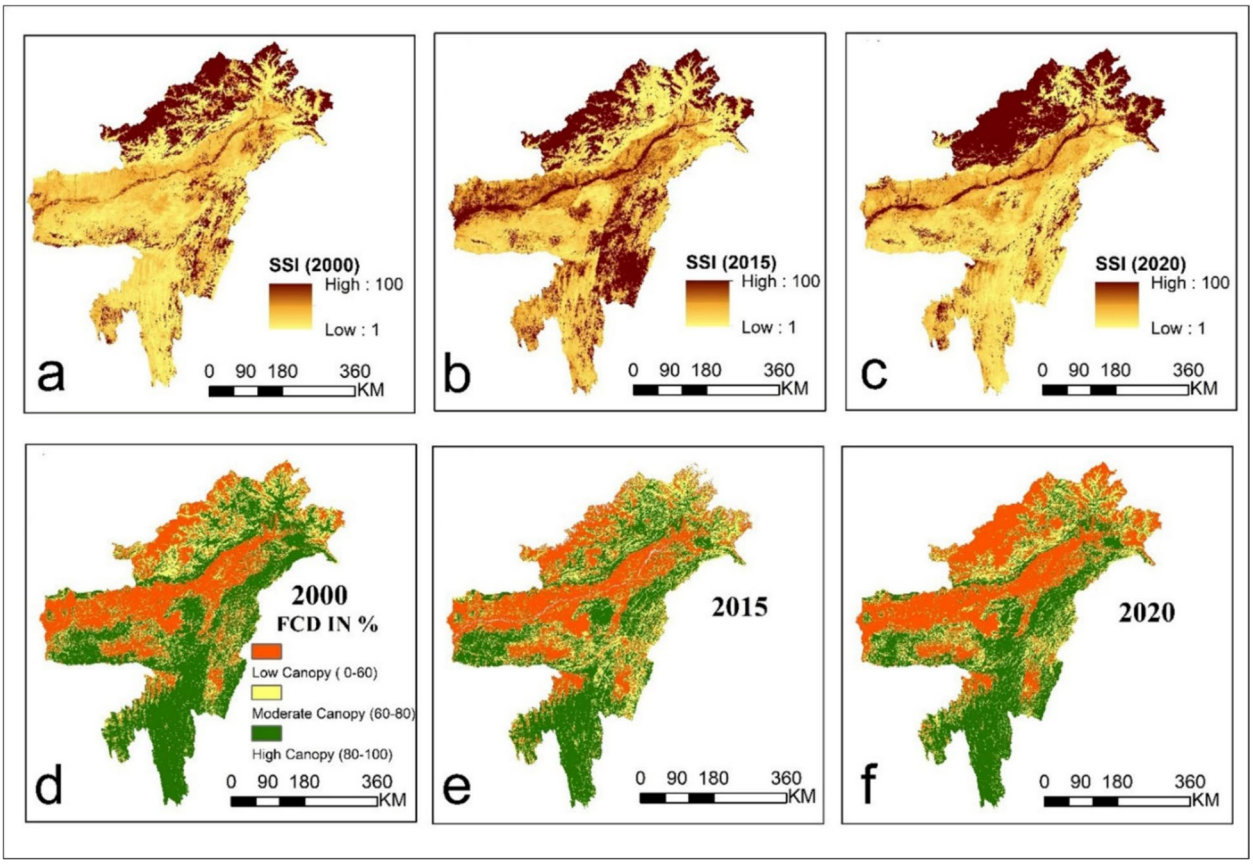

Fig. 3 Forest canopy density maps of **a** SSI of 2000, **b** SSI of 2015, **c** SSI of 2020, **d** FCD% of 2000, **e** FCD % of 2015, **f** FCD% of 2020

dependent variable (deforestation) and independent variables (deforestation determinants) [15, 39, 70]. The study collected 1000 data points from both deforestation and non-deforestation sites, with 70% allocated to the training dataset and 30% reserved for model validation. The BLR model transforms the probability of deforestation incidence into a logit value, ensuring a continuous dependent variable and an infinite new dependent variables. The model equation is:

$$Y = \text{Logit}(P) = \text{Ln}\left(\frac{P}{1-P}\right) = C0 + C1 \times X1 + C2 \times X2 + \ldots + Cn \times Xn \quad (6)$$

The model uses all selected spatial data layers to calculate coefficient values. The study aims to predict deforestation probabilities based on the chosen independent variables, using a dataset of 7000 points for both deforestation and non-deforestation sites.

Algorithms for machine learning Machine learning algorithms play an important part in calculating the possibilities of deforestation and offer superior forecasting capabilities. The modelling method has been enhanced with the application of the machine learning techniques listed below:

XGBoost (gradient boosted regression) XGBoost is an ensemble learning algorithm that makes use of decision trees and gradient-boosting techniques [16, 72]. It excels in handling intricate interactions among the data and was selected for its capacity to produce reliable and accurate predictions regarding the likelihood of deforestation. XGBoost Equation for binary classification-

$$F(x) = \frac{1}{(1 + \exp(-[\alpha + \Sigma T\_i(x)])} \quad (7)$$

where F(x) represents the predicted probability of the positive class, α is the initial model prediction, and T_i(x) are the individual decision trees. Variables. This study implemented the XGBoost model using the 'XGBoost' package in R 4.3.1 software.

Random forest (RF) Random Forest (RF) is a machine learning ensemble model that creates multiple decision trees for classification, enhancing precision by aggregating individual trees and introducing diversity through data replication and variable alteration. This study used the RF model with 600 trees and a split variable of 4 in R 4.3.1, evaluating performance with error metrics like mean decrease in accuracy and mean decrease in Gini impurity [89].

REP-tree REP-Tree, or Reduced Error Pruning Tree, simplifies decision tree models for better interpretability and computing efficiency, employing methods like information gain, entropy, and variance minimization [67, 68]. It addresses overfitting with post-pruning techniques and represents binary classification for deforestation probability as Y is 0 if P(deforestation) < 0.5; 1 if P(deforestation) > = 0.5. The 'RWeka' package in R allows us to work with Weka classifiers for REP-Tree. This provides flexibility for both scripting and interactive data analysis. In this study, a decision threshold of 0.5 was adopted to convert continuous deforestation probabilities from the REP-Tree model into binary deforestation (1) or non-deforestation (0) classifications. This threshold was selected as it represents a balanced default criterion widely accepted in ecological and spatial modelling, ensuring equal emphasis on both sensitivity (true positives) and specificity (true negatives) [55]. Using this midpoint (0.5) threshold minimizes bias toward either class in cases where there is no compelling ecological reason to favour sensitivity or specificity disproportionately, thus offering a fair and interpretable classification outcome [29, 62].

Validation

Comprehensive validation metrics such as the Receiver Operating Characteristic (ROC) curve, Area Under the Curve (AUC), True Skill Statistics (TSS), Efficiency (E), and Taylor Diagrams are employed collectively due to their distinct strengths and complementary perspectives in model performance evaluation. The ROC curve is a widely used graphical tool that illustrates the relationship between the True Positive Rate (TPR, sensitivity) and the False Positive Rate (FPR, 1-specificity) across various classification thresholds. The standard ROC threshold of 0.5 was adopted for converting predicted probabilities into binary outcomes (deforested vs. non-deforested). This threshold is conventionally used to balance sensitivity (TPR) and specificity (1-FPR), ensuring an equitable trade-off between false positives and false negatives in ecological modelling contexts [29]. It is commonly utilized in modelling natural phenomena such as landslides, soil erosion, and groundwater potential assessments [1, 27, 44]. Mathematically, these rates are calculated as:

$$TPR = \frac{TP}{(TP + FN)} \quad (8)$$

$$FPR = \frac{FP}{(FP + TN)} \tag{9}$$

where True Positives (TP), False Negatives (FN), False Positives (FP), and True Negatives (TN) are denoted by the letters TP, FN, FP, and TN, respectively. The ROC curve's effectiveness is quantified through the Area Under the Curve (AUC) metric. AUC provides a singular measure of overall model accuracy across all thresholds. The predictive accuracy based on AUC can be classified as follows: excellent (0.9–1.0), very good (0.8–0.9), good (0.7–0.8), moderate (0.6–0.7), and weak (0.5–0.6). Higher AUC values indicate superior model performance. However, reliance on AUC alone is insufficient since it does not fully account for the practical accuracy of classifications at specific thresholds. Therefore, threshold-dependent metrics like Efficiency (E) and True Skill Statistics (TSS) are incorporated. Efficiency (E), calculated as the ratio of correctly classified instances (TP and TN) to total cases, measures overall predictive accuracy:

$$E = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{10}$$

True Skill Statistics (TSS), defined as the difference between TPR and FPR, evaluates a model's discriminatory ability independently of prevalence and dataset size:

$$TSS = (TPR - FPR) \tag{11}$$

While E indicates the overall correctness of predictions, TSS provides a balanced assessment, particularly valuable when modeling rare or unevenly distributed phenomena. Additionally, Taylor Diagrams offer another complementary graphical method for evaluating predictive models. These diagrams succinctly display the relationship between observed and simulated data, emphasizing correlation, Root Mean Square Error (RMSE), and variance. A perfect simulation is indicated by a correlation of 1 and an RMSE of 0. Deviations from this ideal indicate discrepancies in predictive accuracy, correlation, and variability. The combination of Taylor Diagrams with metrics such as RMSE, correlation, and Mean Absolute Percentage Error (MAPE) provides a multi-dimensional and thorough analysis, highlighting accuracy, variability, and agreement between observed and simulated data sets (Table 8).

Investigation of collinearity study

The outcome of the study of collinearity reveals that there is no such multi-collinearity problem in the variables since all of the parameters have VIF values that are greater than 5 and tolerance values that are greater than 0.1 (Table 3). Because of this, each variable may be considered independent, and it is possible to utilize it in a predictive analysis using models such as BLR, XGBR, RF, and REP-Tree. Illustrates the values of tolerance and VIF for each independent variable individually. These numbers give proof that the variables that were chosen to be independent do not display high degrees of correlation with one another. This leads one to believe that they may be incorporated into a predictive study without causing any issues regarding the accuracy of the models being affected by multicollinearity.

Table 3 Collinearity statistics of selected variables

Variables	Collinearity statistics	
	Tolerance	VIF
Elevation (in meters)	0.370	2.702
Forest land Density (FD)	0.670	1.492
Agricultural Land Density (AD)	0.325	3.081
Distance To Dam (DDam)	0.710	1.409
Barren Land Density (BLD)	0.399	2.504
Distance to Burn Patches (DBP)	0.256	3.905
Population Density (PD)	0.818	1.222
Slope (in degrees)	0.773	1.293
Distance to road (DTR)	0.685	2.741
Distance to Railway lines (DTRL)	0.871	2.627
Distance to Settlement (DFST)	0.647	1.238
Forest Edge Density (ED)	0.387	1.892

Results

Forest cover changes over time and space using the FCD model

In 2000, a substantial portion of the region's forests maintained a high canopy density (80–100% FCD), whereas by 2020 this category had shrunk considerably. Specifically, the area under high-density forest dropped from roughly 46.7% in 2000 to 38.4% in 2020 (Table 4), indicating the loss or thinning of mature, closed-canopy forests. Correspondingly, low-density forest (FCD 0–60%) spread dramatically—from about 31.9% of the forest area in 2000 up to 39.6% by 2020. This rise in low-density cover signifies that many previously dense forest stands have transitioned into more open, fragmented states. The moderate-density forest class (FCD 60–80%) showed a fluctuating pattern: it initially increased slightly (peaking around 2015 at ~ 25% of forest area) but declined to ~ 22% by 2020, suggesting localized regrowth in the mid-2010s followed by renewed disturbances. The net loss of high-density forests (over 8% of forest area) and concomitant gain in low-density areas point to incipient forest degradation on a regional scale, with potentially irreversible consequences for biodiversity and ecosystem services if the trend continues.

Table 4 Spatial–temporal changes of forest cover from 2000 to 2020

	2000		2015		2020			
FCD Class	**Area**	**Area %**	**Area**	**Area %**	**Area**	**Area %**	**Change in Km**	**Change in %**
Low	81,306.25	31.87889695	87,635.75	35.0073202	100,984	39.59500828	−19,677.75	−7.71611133
Moderate	54,571.25	21.39652555	62,999.75	25.1661271	56,045	21.97479045	−1473.75	−0.57826491
High	119,169.75	46.7245775	99,700	39.8265528	98,013.25	38.43020127	21,156.5	8.294376234
Total	255,047.25	100	255,047.25	100	255,042.25	100		

Spatially, the geographic pattern of FCD change highlights the influence of elevation and associated climatic conditions. High-elevation mountainous zones (e.g., parts of Arunachal Pradesh) generally retained higher canopy densities, whereas many lower elevation areas—notably the Brahmaputra Valley and the foothills—showed marked declines in FCD. For instance, the southern and south-eastern states (e.g., Mizoram, Meghalaya, parts of Assam's hill districts) had extensive healthy vegetation in 2000 (indicated by high AVI values approaching 400), but by 2020, much of this vegetation showed reduced greenness (AVI upper range dropping to ~ 148) and canopy cover. In contrast, lowland agricultural zones along the Brahmaputra River and its floodplains already exhibited low canopy density early on and remained deforested (or further expanded in agriculture) over time. These observations suggest that lower elevations with more variable or seasonal precipitation regimes are especially vulnerable to canopy cover loss. The Bare Soil Index (BSI) ranged from 37 to 122, with lower values signifying denser forests, effectively highlighting areas of bare soil and deforestation (Fig. 4d, e, f). The Shadow Index (SI) ranged from 0 to 100, with higher values in the northern and north-western regions and lower values in the southern, central, and eastern areas (except for 2015) (Fig. 3a, b, c). Vegetation Density (VD) presented a similar pattern, with high VD in the eastern and southern parts and low VD in the north (Fig. 5a, b, c). These four indices, along with forest cover dynamics maps created by integrating relevant data, depicted forest concentration in the region. The Forest Canopy Density (FCD) maps from 2000 to 2020 differentiated low (0–60), moderate (60–80), and high (80–100) canopy densities (Fig. 3d, e, f), based on the coverage extent. The progression of FCD from 2000 to 2017 is detailed in Table 4.

Spatial analysis of Forest Canopy Density (FCD) from 2000 to 2020 highlights substantial inter-state variation across Northeast India. Mizoram, for instance, experienced significant canopy deterioration, shifting extensively from high-density forest (80–100% FCD) in 2000 to predominantly moderate (60–80%) and low-density (0–60%) stands by 2020. Similarly, Meghalaya exhibited pronounced fragmentation, with many previously dense forest areas transitioning to lower canopy classes, particularly evident by 2020. In contrast, Arunachal Pradesh largely preserved its high-density forest cover, particularly at higher elevations, suggesting topographic constraints reduced anthropogenic impacts. Assam presented distinct regional contrasts: while its hill districts mirrored the degradation patterns observed in Mizoram and Meghalaya, the lowland areas of the Brahmaputra Valley maintained consistently low canopy densities throughout the period, with further reductions visible by 2020. Transient increases in moderate-density cover around 2015 in parts of Meghalaya and Nagaland were short-lived, reverting to lower density levels by 2020.

Analysis of deforestation probability through binary logistic regression (BLR)

To determine deforestation probability, binary logistic regression was employed using twelve independent variables and previously deforested areas between 2000 and 2015 as a dependent variable. The deforestation map (Fig. 5), produced by the FCD study from 2000 to 2020, is the dependent variable. A score closer to 1 indicates a very high probability of deforestation, whereas a value closer to 0 indicates the reverse. These five categories: very low, low, moderate, high, and very high represent 20.07%, 35.01%, 29.78%, 12.52%, and 2.62% of the area respectively, based on the results of the BLR model for the risk of deforestation (Table 5 and Fig. 6). There is a high to very high risk of deforestation in the vicinity of 4.97% of the total forested land.

Analysis of deforestation probability through Random Forest (RF)

In this study, deforestation probability was assessed using a binary dependent variable: areas classified as either 'deforested' or 'non-deforested'. The analysis, conducted in R, considered a range of conditioning factors, both numerical and categorical, to determine their influence on deforestation. The Random Forest (RF) algorithm, an ensemble learning method known for its robustness in classification and regression tasks, was employed for this purpose. The RF model, which constructs multiple decision trees and outputs the mode or mean prediction, was

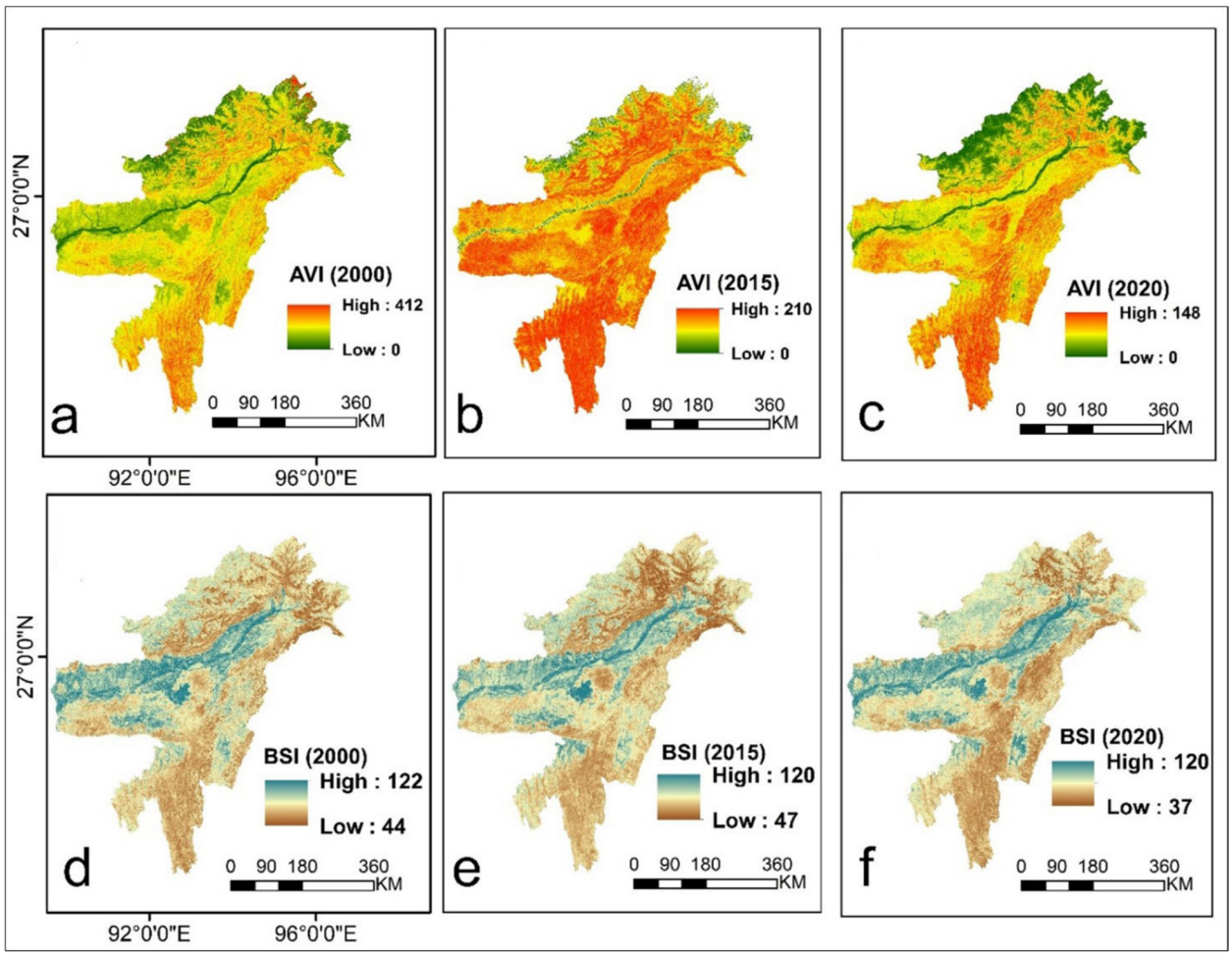

Fig. 4 Maps utilized for generating FCD Maps including **a** AVI from 2000, **b** AVI from 2015, **c** AVI from 2020; **d** BSI from 2000, **e** BSI from 2015, **f** BSI from 2020

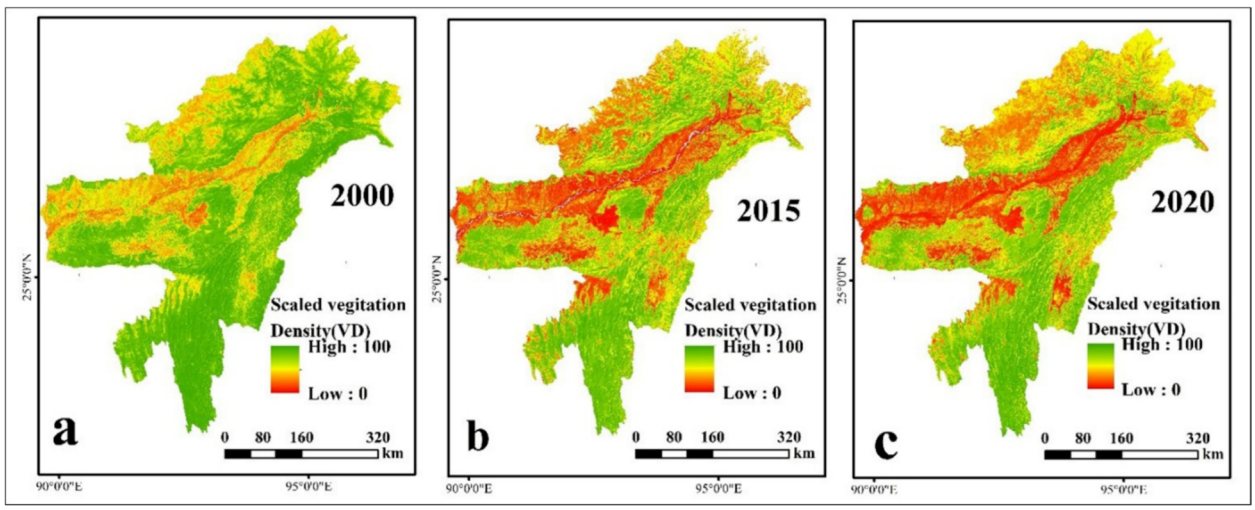

Fig. 5 Maps employed for generating FCD maps **a** VD from 2000, **b** VD from 2015, **c** VD from 2020 (inside the maps - the 'Scaled Vegitation' spelling is a problem, I am supplying new maps with correction)

Table 5 Probability of deforestation in forested regions

CLASS	Rep-Tree Model			Random Forest Model			XGBR Model			BLR Model			Total Forested Area
	Area	% of total forested area	% of total area	Area	% of total forested area	% of total area	Area	% of total forested area	% of total area	Area	% of total forested area	% of total area	
Very low	52,716.82	34.20	20.67	78,892.99	51.16	30.93	37,807.17	24.56	14.82	30,096.93	20.07	11.80	154,194.66
Low	41,192.84	26.72	16.15	37,110.35	24.07	14.55	50,019.46	32.50	19.61	52,001.38	35.01	20.39	
Moderate	24,837.39	16.11	9.74	18,376.50	11.92	7.20	34,845.67	22.55	13.61	44,678.23	29.78	17.52	
High	29,175.64	18.90	11.42	11,854.94	7.69	4.65	23,330.21	15.07	9.09	20,141.23	12.52	8.05	
Very high	6271.98	4.07	2.46	7959.87	5.16	3.12	8192.15	5.32	3.21	7276.89	2.62	3.01	

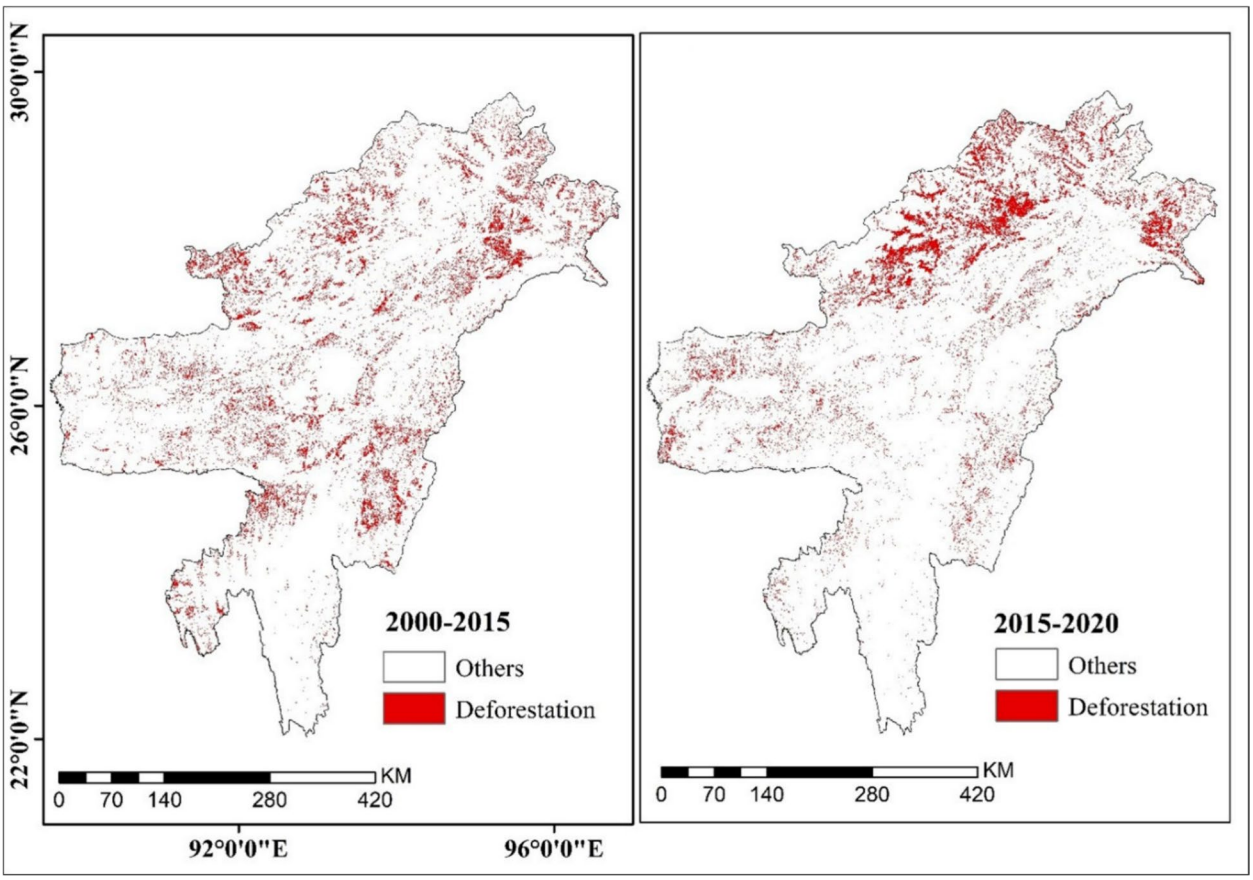

Fig. 6 Thematic layer illustrating changes in forest cover, created by comparing forest canopy density maps from 2000 and 2020

utilized to analyse the probability of deforestation. This analysis revealed that in the study area's total forested area, 51.16% fell into the very high deforestation probability category, followed by 24.07% in the high category, 11.92% in the moderate, 7.69% in the low, and 5.16% in the very low category. Notably, the model identified specific areas of concern: the northern zone exhibited a few patches with a very high probability of deforestation, and similar isolated areas were observed in the southern part of the region. The analysis determined that 3.12% of the entire region fell into the very high deforestation probability zone (DPZ), while 4.65% was classified as high DPZ.

Analysis of deforestation probability through REP-Tree model

The Deforestation Probability Zones (DPZ) have been identified by applying REP-Tree using the deforestation conditioning variables. This strategy has made it possible to delineate the DPZ by using decision tree pruning to identify the best possible trees. As in the past, the results have been divided into five groups: very high, high, moderate, low, and very low DPZs correspond to 34.20%, 26.72%, 16.11%, 18.90%, and 4.07% of the forest area (Table 5). To determine the precise positions of DPZs, the model's output was converted to a GIS raster format (Fig. 7). The very high probability patches are found in the upper Brahmaputra valley and in the centre of the state of Mizoram (Fig. 8). A total of 7959.87 km^2 is classified as having a very high likelihood.

Probability analysis of deforestation using XGBoost regression

Using a set of deforestation conditioning factors, the study used the Gradient Boosted Regression (XGBR) ensemble technique to analyse the probability of deforestation. The Deforestation Probability Zones (DPZ) could be identified with this method via R software (Fig. 9). The XGBR model's primary benefit is its capacity to use decision trees and optimize them through pruning to identify the most precise forecast. The results were categorized into five groups, which reflected the conclusions of the earlier binary logistic regression. Five categories—very high, high, moderate, low, and very low Deforestation Probability Zones were identified, and Table 5 shows the

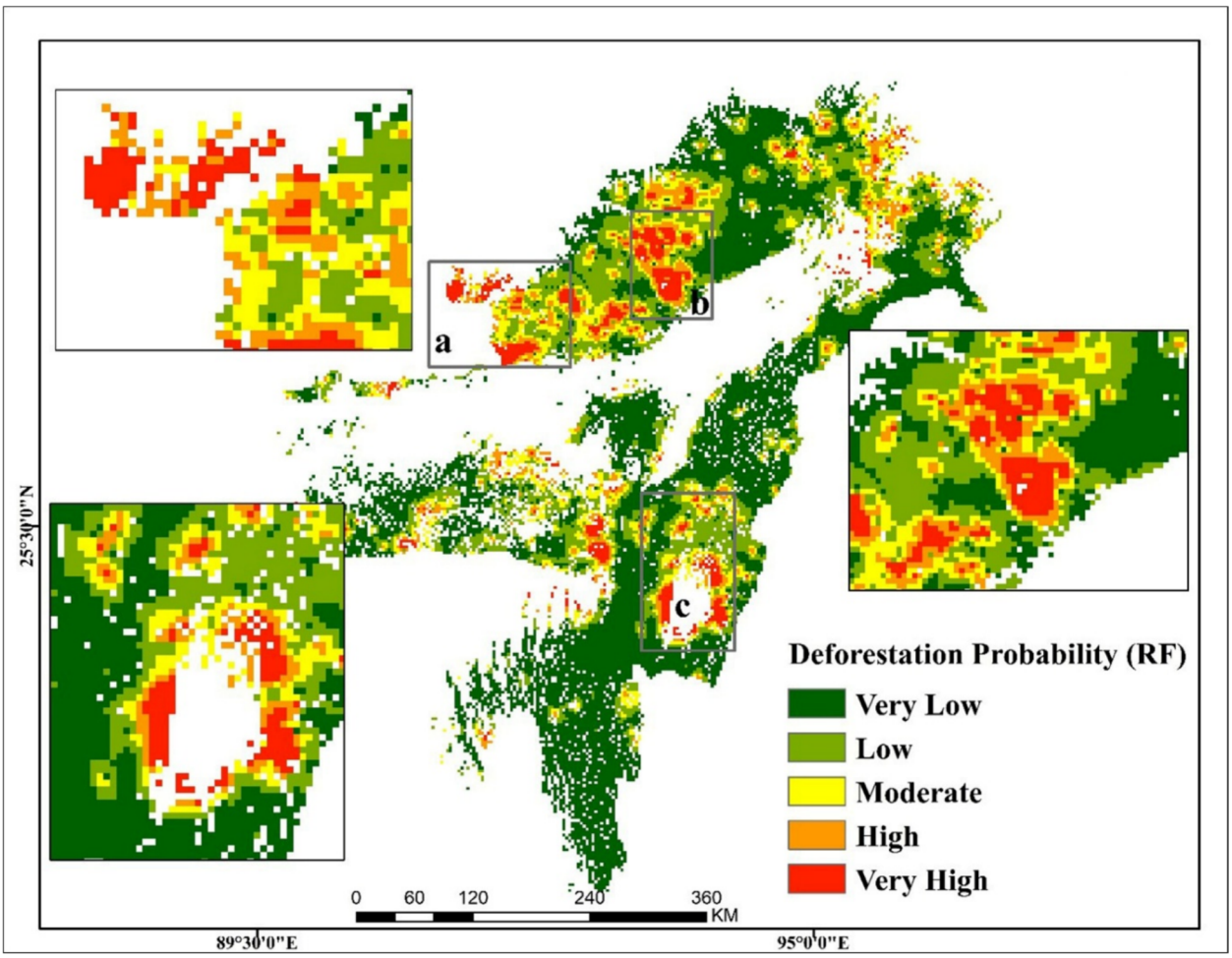

Fig. 7 Deforestation probability map based on RF

percentages of wooded regions assigned to each category. The analysis's most noteworthy result was the DPZs' geographic dispersion. The study found areas in Arunachal Pradesh with a very high probability of deforestation in the upper Brahmaputra valley, as shown graphically in (Fig. 9). The extent to which the very high probability category covered a huge area of roughly 7276.89 km^2 was particularly noteworthy. This indicates that a sizable chunk of the research region is highly vulnerable to deforestation.

Validation of FCD maps

Validation of FCD of 2000, 2015, and 2020 was performed using the kappa statistics using 1000 ground truth points. The FCD maps' Kappa coefficient values for 2000, 2015, and 2020 are 0.78, 0.83, and 0.85, respectively (Table 6). These values suggest that FCD maps are a good fit for modelling the probability of deforestation. Estimating the receiver operating characteristics (ROC) of FCD maps has also been used to assess their accuracy and dependability (Fig. 10). This method is frequently employed to assess a diagnostic test's accuracy. There are two different probability function types on the ROC curve: true positive, which represents an event response that is accurately anticipated, and false positive, which represents an event response that is incorrectly predicted. Here, with the aid of Google Earth imagery, the ROC curves (Fig. 11) were created using randomly selected validation datasets from deforestation and non-deforestation locations (Table 7). The findings indicate that the AUC of the FCD maps for the years 2000, 2015, and 2020 are, respectively, 72.60, 74.54, and 74.60 (Table 8 and Fig. 10). This implies the validity and acceptance of the FCD maps, which are further applicable to the evaluation of deforestation (Tables 9 and 10).

Validation of deforestation probability models

In the study, a number of metrics, such as the AUC of receiver operating characteristics (ROC), efficiency (E), true skill statistics (TSS), and the Kappa coefficient, were

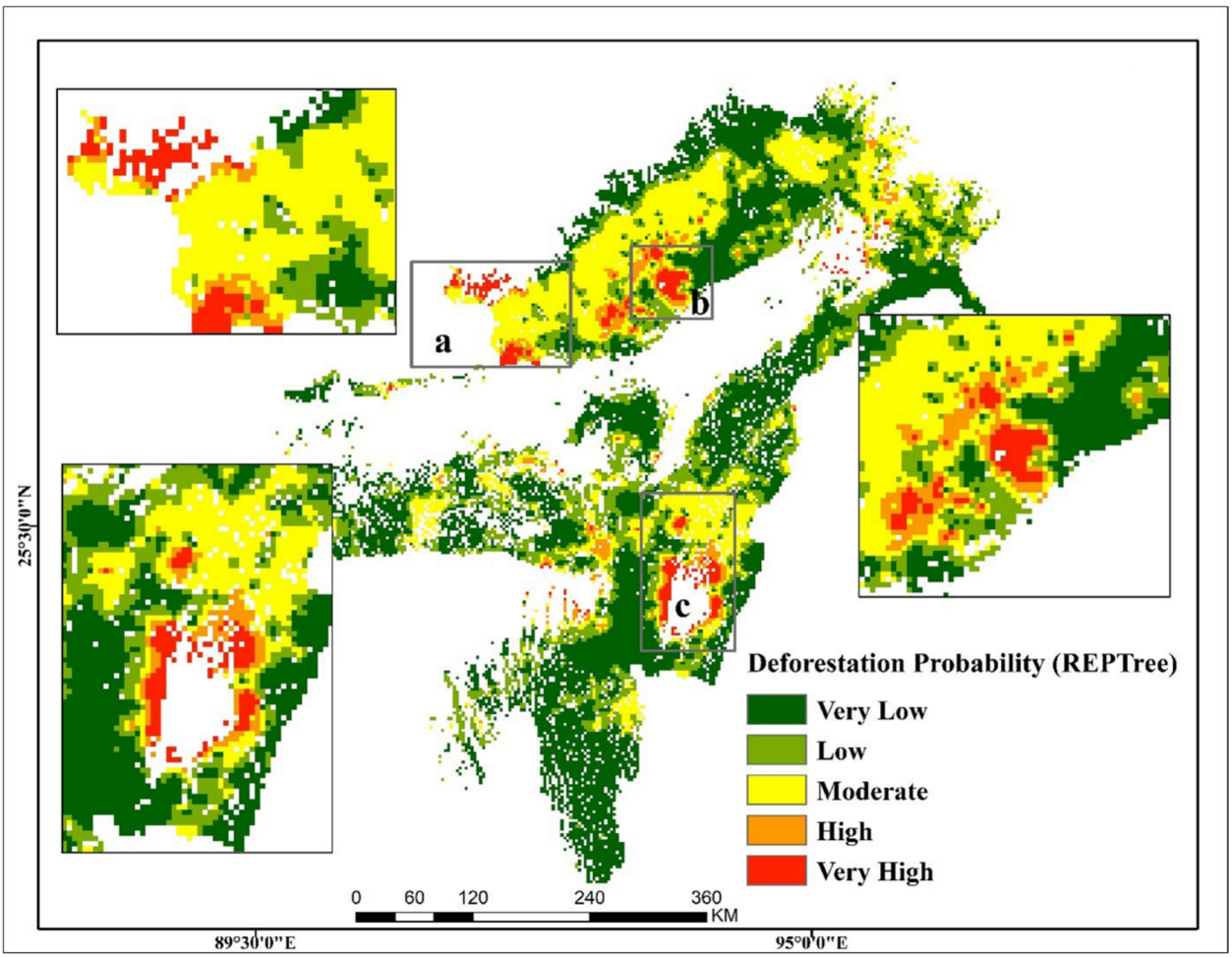

Fig. 8 Deforestation probability map based on REPTree

used to evaluate the accuracy and reliability of Binary Logistic Regression (BLR) (Table 11), Random Forest (RF), REPTree, and XGBoost. To ascertain how well these models estimate the risk of deforestation in the research area, evaluations were carried out using a dataset that was validated using 30% of the total. For the BLR, RF, and REPTree models, the sensitivity (also known as the True Positive Rate or TPR) values were given as 0.68, 0.988, and 0.86, respectively [75]. The models' accuracy in identifying areas as probable zones for deforestation is indicated by these figures, wherein REPTree exhibits the maximum sensitivity. Furthermore, BLR, RF, and REPTree's False Positive Rate (FPR) values of 0.32, 0.012, and 0.14, respectively, demonstrated the models' capacity to avoid incorrectly categorizing non-deforestation areas. Once more, REPTree's efficacy in reducing false positives was outstanding. Furthermore, these models' robustness in forecasting Deforestation Probability Zones (DPZs) was demonstrated by their efficiency and TSS values [4]. With the maximum efficiency and TSS in this situation, REPTree came out as the model that could accurately detect both DPZs and non-DPZs. The Kappa coefficients indicate the degree of agreement between the models' predictions and actual observations. With a Kappa coefficient of 0.824, REPTree's predictions showed a high degree of agreement. When all of these validation criteria are taken into account, the overall findings show that the machine learning models RF, XGBR, and REPTree, in particular, are more capable of accurately forecasting deforestation areas than the probabilistic BLR model. In summary, both the RF and REPTree models perform well in mapping the probability of deforestation, with REPTree demonstrating the most striking results across a range of evaluation parameters. Random Forest has the highest coefficient of determination (R^2) but a higher RMSE compared to BLR. Consider the trade-off between accuracy and simplicity. If interpretability is crucial, the BLR model may be preferred despite lacking an MAPE value (Table 11). The Random Forest (RF) model exhibits a Root Mean Square Error (RMSE) below

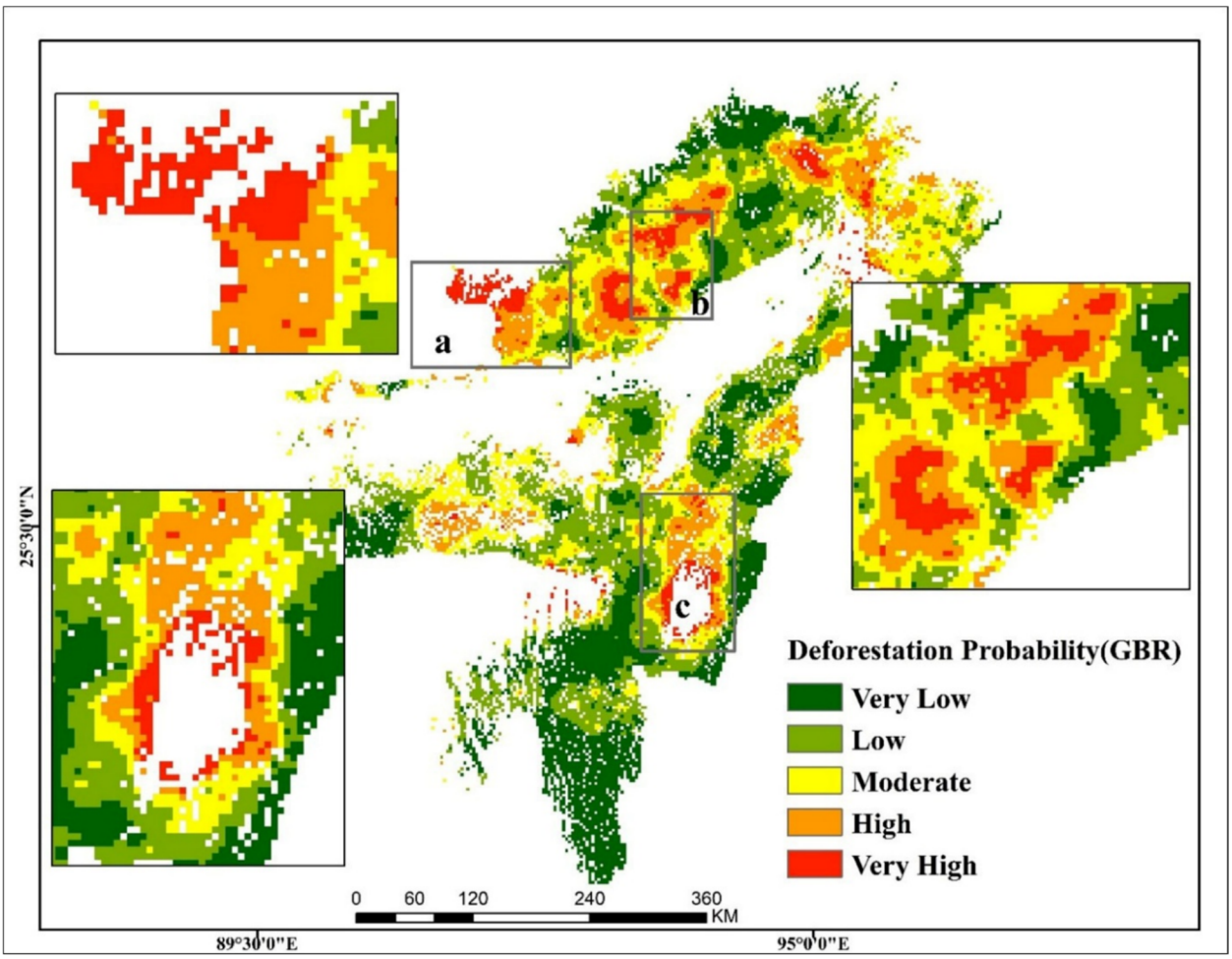

Fig. 9 Deforestation probability map based on XGBR

Table 6 Regional Distribution of Deforestation Risk Zones (DRZ), % High and Very High Probability ,and key drivers in North-East India (2000–2020)

Region/District	% High + Very High Risk	Trends
Karbi Anglong/Dima Hasao, Assam	> 15% of Assam total	Road-driven, peri-urban expansion
Guwahati (peri-urban), Assam	~ 5%	Urban sprawl
West Tripura/Agartala	> 30%	Urbanization, infrastructure
Mon/Tuensang, Nagaland	> 35%	Jhum-driven loss
Serchhip/Aizawl, Mizoram	> 30%	Shift from high to low density 2000–20
Ukhrul/Churachandpur, Manipur	> 28%	Shifting cultivation
Garo Hills, Meghalaya	> 35%	Village/foothill open-up
Brahmaputra Valley, Assam	> 40% (low FCD)	Floodplain/foothill degradation
Arunachal Pradesh (high hills)	< 15% (loss); > 85% high FCD retained	Terrain-driven protection

0.17, indicating good accuracy. The correlation is nearly above 95%, further emphasizing the model's agreement with the observed data (Fig. 12a).

Discussion

Deforestation, recognized globally as a primary contributor to environmental degradation, biodiversity loss, and climate change, is a pivotal concern within ecological and

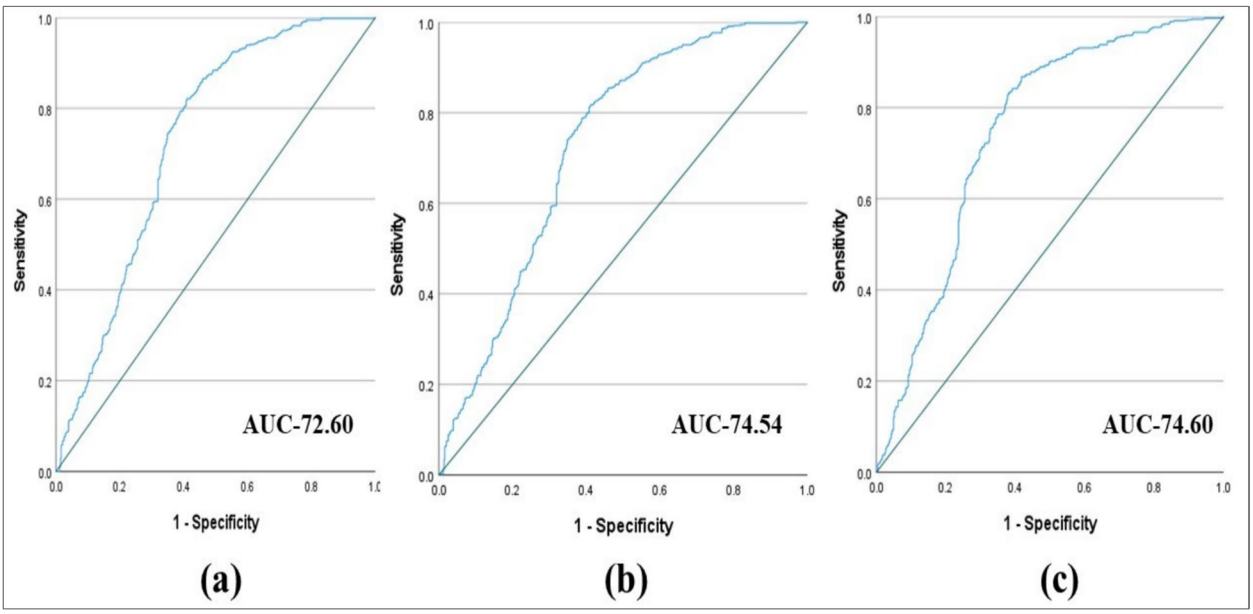

Fig. 10 ROC curves for validating Forest Canopy Density (FCD) maps **a** 2000, **b** 2015 and **c** 2020

environmental research [5, 25, 32, 47]. Effective deforestation probability mapping has emerged as a crucial tool for anticipating deforestation hotspots, allowing proactive management and conservation strategies [54, 73, 74, 80, 90]. Using the observed FCD changes as a baseline, we modelled deforestation probability across the region, producing deforestation risk maps for the present and near future. The models categorize the landscape into Deforestation Potential Zones (DPZ), ranging from very low to very high risk. Our predictive models clearly identify proximity to roads, expanding settlements, and shifting cultivation practices, particularly prevalent in Mizoram, Assam, and Manipur, as primary drivers of forest degradation [10, 46, 59]. Recent regional studies emphasize specific policy actions, such as establishing legally enforced buffer zones around high-density forest fragments to curb encroachment and maintain ecological corridors [9, 10, 17]. Restrictions or moratoriums on new road construction near sensitive forest areas are also critical, as such infrastructure significantly heightens deforestation risk by facilitating logging and human settlement expansion [10, 48].

The Binary Logistic Regression (BLR) model, for example, classified about 34.2% of the forest area as high to very high risk (combined) and roughly 22.97% as moderate risk, with the remainder in low-risk categories (Table 5). Notably, the BLR-based risk map indicated that approximately 5% of the total forest area falls in the high or very high risk zones, predominantly clustering around human-disturbed landscapes. Hotspots of deforestation risk were identified near expanding settlements and infrastructure corridors, for instance, along major highways and roads cutting through forested hills and in the vicinity of urbanizing centres in Assam and Tripura. In Assam, high-risk zones cluster around the outskirts of Guwahati and along road corridors through the Karbi Anglong hills, whereas in Tripura, deforestation hotspots appear near Agartala and across the state's eastern hill ranges. In addition, portions of the traditional shifting cultivation belts in states like Nagaland, Mizoram, and parts of Manipur showed elevated risk, reflecting ongoing cycles of clearing. For example, eastern Nagaland (notably Mon and Tuensang districts) and central Mizoram (around Serchhip district) correspond to these high-risk jhum areas (Table 6). Similarly, in Manipur, the hill districts such as Ukhrul and Tamenglong exhibit elevated deforestation probabilities linked to shifting cultivation. For example, in Assam, the hill districts of Karbi Anglong and Dima Hasao alone account for more than 15% of the state's total high and very high deforestation risk zones (Table 6 and Fig. 8), with peri-urban areas around Guwahati contributing an additional 5% due to rapid expansion. In Tripura, over 30% of forest patches near Agartala and the bordering West Tripura district are classified as moderate to high risk, reflecting urban encroachment and infrastructure growth [83].

These patterns affirm that anthropogenic forces are compounding the inherent vulnerability of certain forest areas. Our risk maps (Figs. 7, 8, 9 and 11) vividly illustrate this compounding effect—the most vulnerable zones are

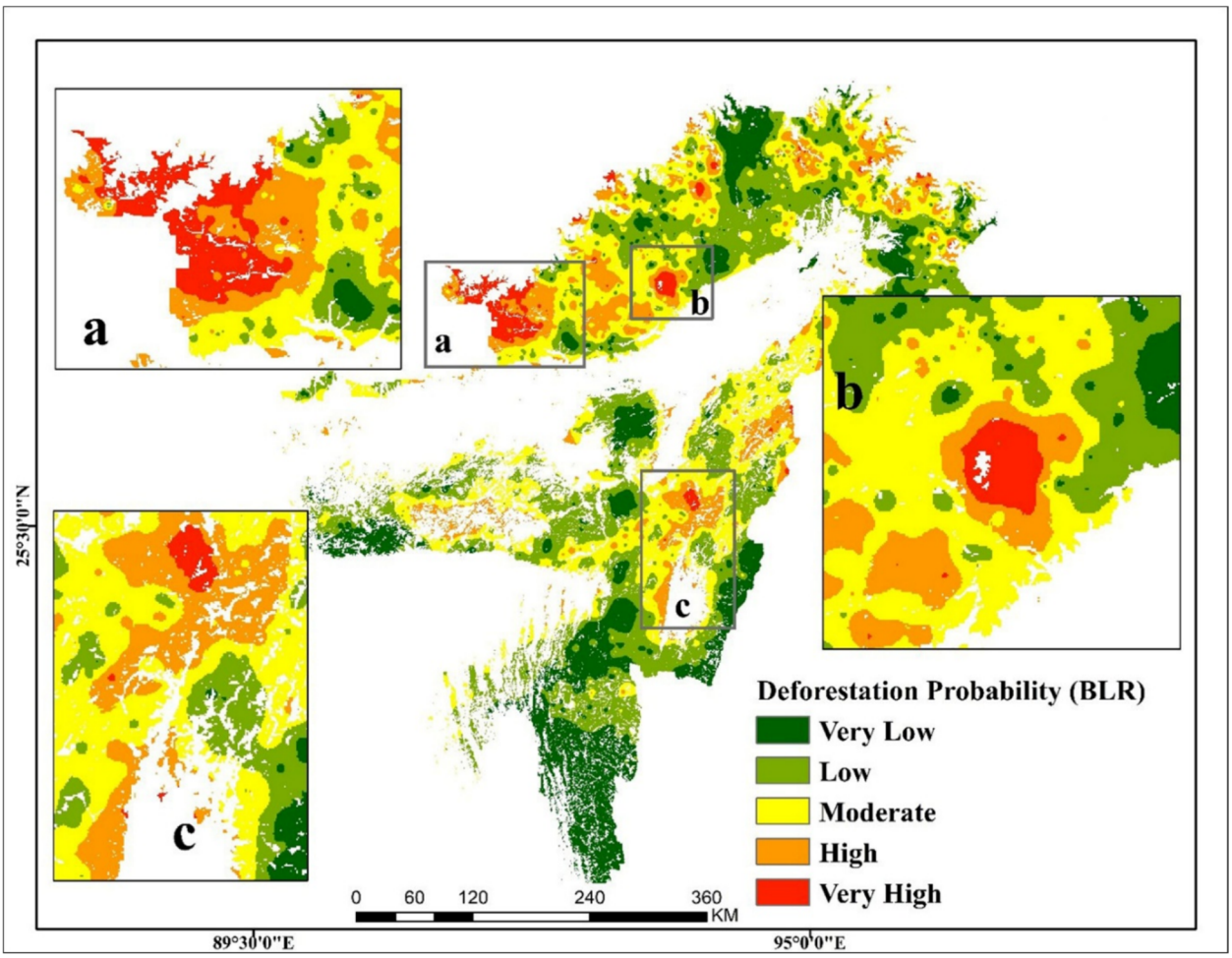

Fig. 11 Deforestation probability map based on BLR

Table 7 Outcome of the Receiver Operating Characteristic (ROC) analysis FCD and deforestation probability maps

Models	AUC	STD. Error	ASYMPTOTIC SIG	ASYMPTOTIC 95% Confidence interval	
				Lower bound	Upper bound
BLR model	0.68	0.008	0	0.665	0.696
XGBR model	0.818	0.007	0	0.804	0.832
Random forest	0.988	0.001	0	0.986	0.99
REP tree	0.86	0.006	0	0.849	0.872

Table 8 Values of true positive rate (TPR), false positive rate (FPR), efficiency (E), true skill statistic (TSS) and Kappa co-efficient

Metrics	TRP	FPR	TSS	E	K
REP Tree	0.86	0.14	0.72	0.897	0.79
Random	0.85	0.15	0.976	0.885	0.81
BLR	0.68	0.32	0.36	0.712	0.78
XGBR	0.818	0.182	0.636	0.864	0.83

Table 9 Model accuracy check

Metrics	R2	RMSE	MAPE
Rep_Tree	0.758	0.1384	4.2874
Random Forest	0.892	0.1710	7.0331
BLR	0.874	0.2868	NA
XGBR	0.7895	0.2916	6.9405

Table 10 Logistic regression coefficients for different Variables

Variables in the Equation									
SL No.	**Variables**	**B**	**S.E**	**Wald**	**df**	**Sig**	**Exp(B)**	**95% C.I.for EXP(B)**	
								Lower	**Upper**
1	Forest Density	0	0.002	0.018	1	0.894	1	0.997	1.004
2	Barren Land Density	−0.04	0.01	16.726	1	0	0.961	0.943	0.98
3	Agricultural Density	0.009	0.002	18.449	1	0	1.009	1.005	1.013
4	Proximity to Burn	0	0	0.506	1	0.477	1	1	1
5	Distance from DAM	−0.085	0.07	1.511	1	0.219	0.918	0.801	1.052
6	Elevation	0.001	0	228.138	1	0	1.001	1.001	1.001
7	Population Density	0	0	0.592	1	0.442	1	1	1
8	Distance From Railway	−0.68	0.161	17.804	1	0	0.507	0.369	0.695
9	Distance from Road	−3.964	0.967	16.806	1	0	0.019	0.003	0.126
10	Settlement density	−0.014	0.004	12.027	1	0.001	0.986	0.979	0.994
11	Distance From Urban Patches	0	0	10.468	1	0.001	1	1	1
12	Constant	−0.947	0.152	39.037	1	0	0.388		

Table 11 Relative importance of the factors calculated by RF model

Variables	**Mean decrease Gini**
Elevation (in meters)	25.8754
Forest land Density (FD)	12.5478
Agricultural Land Density (AD)	87.5412
Distance To Dam (DDam)	19.2671
Barren Land Density (BLD)	48.4691
Distance to Burn Patches (DBP)	69.4872
Population Density (PD)	74.2891
Slope (in degrees)	65.8643
Distance to road (DTR)	55.8391
Distance to Railway lines (DTRL)	62.1284
Distance to Settlement (DFST)	66.4136
Forest Edge Density (ED)	14.2803

those where environmental susceptibility and human pressure converge, for example, gentle foothill slopes that are both farmable and close to villages or roads [22, 61]. For clarity and completeness, we report all area-related percentages both as a share of total forest area and as a share of the overall study region. This approach helps readers understand deforestation risk not just within the forests themselves, but also in relation to the entire landscape. The rise in low-density forests often corresponds with significant ecosystem degradation, potentially exacerbating carbon emissions, biodiversity loss, and hydrological disruption. Moderate-density forests exhibited minor fluctuations, initially increasing to 25.17% in 2015 and subsequently decreasing to 21.97% in 2020, reflecting complex, localized land-use transitions involving regrowth and deforestation. Most concerning was the sharp decline in high-density forests, from 46.72% in 2000 to 38.43% by 2020, underscoring a significant loss of mature, ecologically vital forest ecosystems [8].

Over 2000–2020, we observed a clear trajectory of forests transitioning from high canopy cover to more open, low-density states. The sharp increase in low-density forest cover (nearly 8% increase in area) and concomitant fragmentation of high-density patches could signify the onset of irreversible degradation processes in some areas. This is consistent with broader tropical forest research, noting that once deforestation exceeds ~ 20–30% of a landscape, regional rainfall can decline and dry seasons lengthen, further inhibiting forest recovery [57, 97]. While our study area still retains significant forest cover, the quality and continuity of that forest are diminishing, raising alarms about ecosystem resilience going forward. Crucially, the interaction of climatic variability and human land-use emerges as a driving force behind these canopy changes. North-East India's monsoon climate, with its seasonal concentration of rainfall, means that forests here must endure a dry period each year. In a pristine system, dense forests can buffer climatic swings, retain moisture, and regulate local temperature. However, human disturbances exacerbate climate stress through multiple feedback loops. One feedback loop identified is the impact of shifting cultivation (Jhum) practices (Fig. 13f) under changing climatic conditions. Another critical feedback involves infrastructure-driven deforestation and its environmental repercussions. The expansion of roads, as identified in our models, greatly

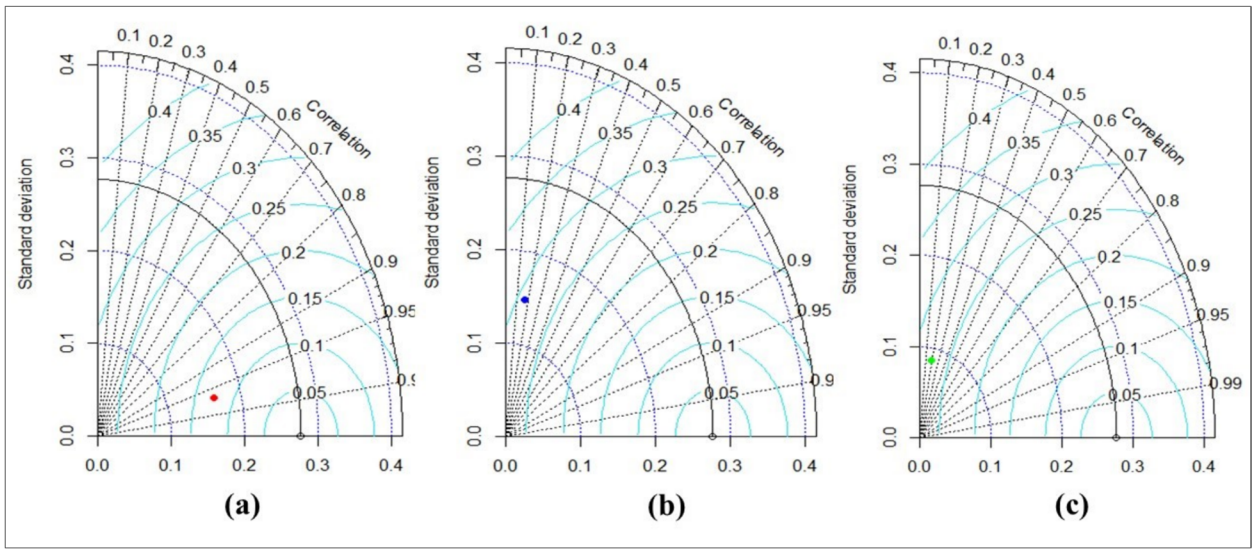

Fig. 12 Taylor's diagram (only based on machine learning models) showing reliability of the models **a**. Random Forest (RF), **b**. Rep Tree and c. Gradient Boosted regression (XGBR)

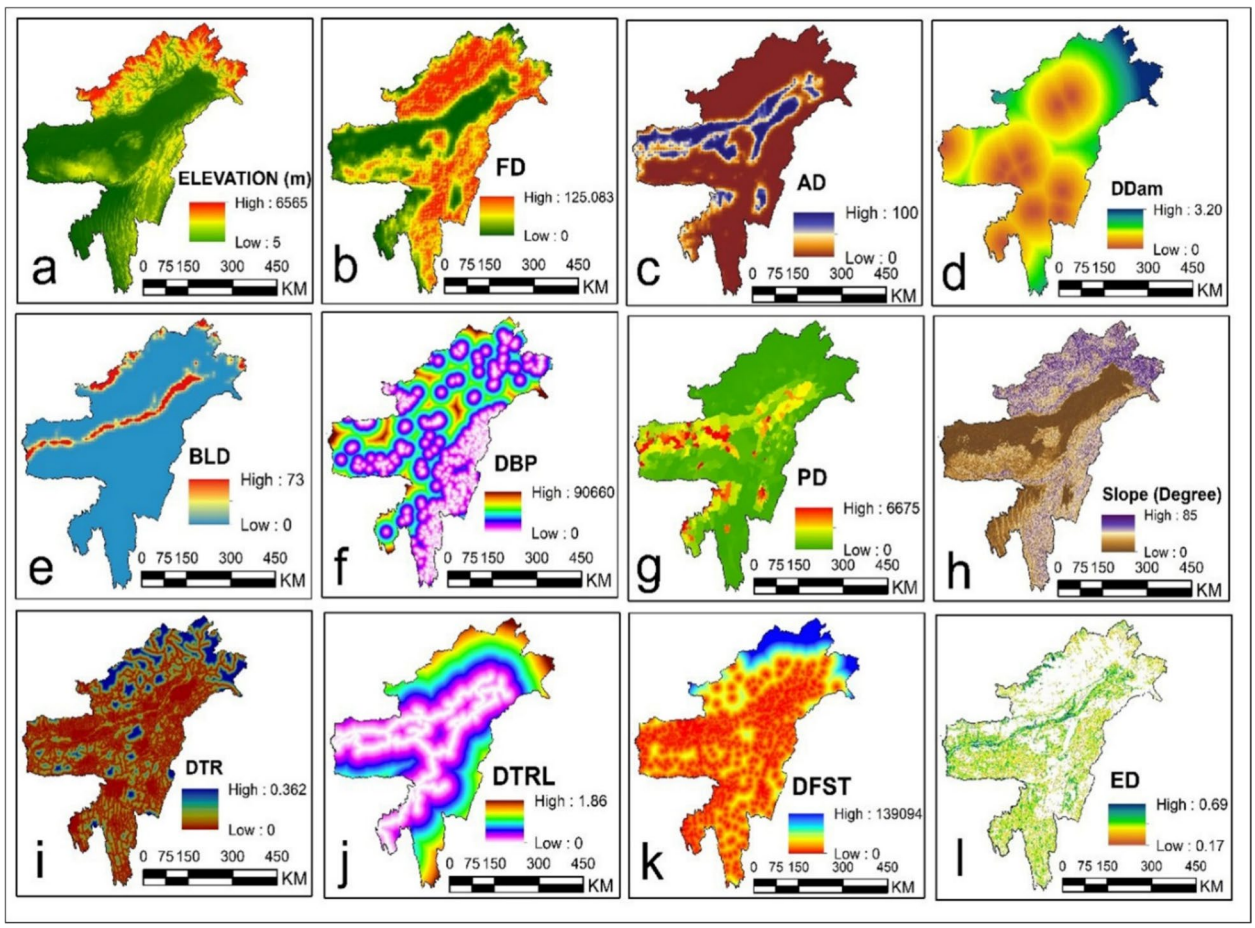

Fig. 13 Regression explanatory variables: **a** Elevation(m); **b** Forest Density (LCC); **c** Agricultural Density; **d** Distance from DAM; **e** Baren land density; **f** DBP; **g** Population Density; **h** Slope (Degree); **i** Distance from Road; **j** Distance from railway; **k** Distance from Settlement; **l** ED

elevates deforestation risk by facilitating logging and settlement in previously remote forests. Tropical forest ecosystems often require a minimum canopy cover to maintain internal humidity and facilitate seed dispersal [48]. When canopy cover falls below that minimum, we can see cascading effects—seedling mortality rises, key animal dispersers vanish, and the forest can tip into an alternative state (e.g., fern thickets or *Imperata* sp.). In North-East India, such transitions are of grave concern; anecdotal evidence from field observations and local forestry reports already points to areas where repeated jhum and logging have given way to alang-alang (*Imperata* sp.) grasslands that show no signs of reverting to forest [32] (Fig. 14).

A comprehensive conservation approach thus should encompass—(1) legally mandated buffer zones, (2) road construction restrictions, (3) sustainable shifting cultivation incentives, and (4) targeted ecological restoration efforts. Such integrated strategies, informed by high-resolution spatial analyses and grounded in recent research, will be crucial for maintaining the ecological integrity and resilience of forests in North-East India in the face of escalating anthropogenic and climatic pressures [9, 10, 46, 48]. Incentivizing sustainable shifting cultivation by integrating traditional ecological knowledge into formal management plans can significantly improve forest regeneration cycles, reducing long-term ecological damage [8–10, 46, 48]. Effective forest conservation also requires engaging local communities through joint forest management, participatory monitoring, and equitable benefit-sharing mechanisms, enhancing compliance and long-term conservation success [9, 10, 46, 48]. Within this integrated framework, the REPTree model demonstrated the highest predictive accuracy (True Positive Rate of 0.86), allowing for precise identification of vulnerable deforestation areas, while the Random Forest (RF) model offered similarly robust and interpretable predictions (TPR of 0.855) with a low incidence of false positives.

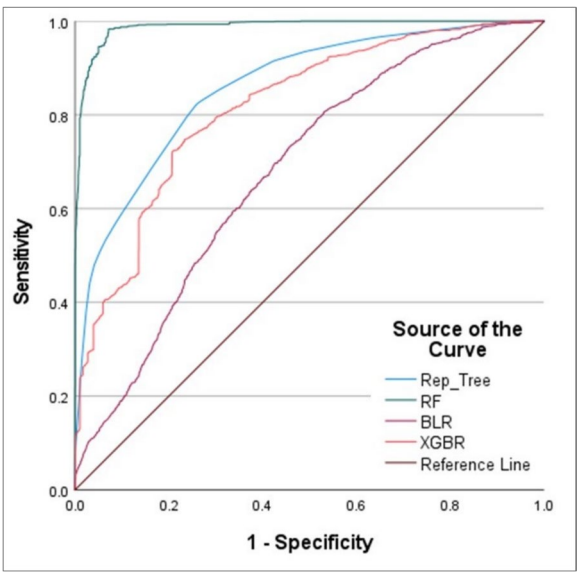

Fig. 14 ROC curves for validating Deforestation Probability Maps

These findings are consistent with the broader body of ecological modelling research, which underscores the reliability and practical utility of machine learning approaches, particularly REPTree and Random Forest, for large-scale deforestation prediction and spatial prioritization [18, 46]. The insights of this study underline the importance of model selection tailored to specific research goals and policy contexts [72]. Temporal trends in deforestation probabilities revealed mixed outcomes: the substantial reduction in the 'High' risk class suggests successful localized conservation or policy initiatives. However, the slight increase in the 'Very High' risk category indicates persistent pressures on forests, necessitating ongoing policy refinements and intensified conservation measures. By overlaying deforestation risk maps with protected area boundaries (e.g., Dampa Tiger Reserve, Karbi Anglong Reserve Forest), managers can pinpoint new encroachment fronts. Our analysis shows that over 28% of the forest in Karbi Anglong's buffer zone and 32% around Dampa are at high or very high risk (Table 6, [10, 14, 46]). NGOs can use these outputs to focus restoration in districts like Serchhip and Aizawl, where more than 30% of the forest has shifted from high to low density. This data-driven approach enables targeted protection and efficient resource allocation in vulnerable areas.

Conclusion

This study analysed Forest Canopy Density (FCD) dynamics and quantified deforestation probability across North-East India (2000–2020). High-density forest declined from 46.72% (2000) to 38.43% (2020), while low-density forest increased from 31.9% to 39.6%, indicating ecological degradation and loss of ecosystem services. Predictive modelling identified ~ 4.97% of forest area as high to very-high risk, driven mainly by proximity to roads, agricultural expansion, and urban development. Among the evaluated models, REP-Tree and Random Forest achieved the best predictive performance (highest accuracy, lowest false positives). The resulting risk maps and thresholds are operationally actionable: districts where > 30% of forest falls into high or very-high risk (e.g., Karbi Anglong, Mon, Serchhip) should be prioritised for enforcement, community outreach, restoration financing (e.g., Green India Mission, Compensatory Afforestation Fund Management & Planning Authority, India or CAMPA), and infrastructure siting

that avoids further fragmentation. Adopting a multi-model framework that can be updated over time represents a safe no-regrets strategy for conservation under uncertainty [34, 40, 53]. Such a framework emphasizes strengthening buffers around key habitat fragments, preventing new road encroachment in hotspot cells, promoting community forestry and agroforestry in surrounding lands, and applying targeted fire management. Together, these actions generate ecological and social benefits across a wide range of possible futures, even if risk estimates shift as new data become available. In adaptation and conservation planning, no-regrets actions are those that yield benefits under many future scenarios; ensemble modelling and periodic retraining with new satellite and socio-economic data support this robustness. Overall, our results demonstrate that integrating remote sensing, Forest Canopy Density modelling, and machine learning provides both diagnostic insights into predictive capacity for anticipating future deforestation risks. This approach linking empirical data with forward-looking risk mapping offers a valuable framework for policymakers and practitioners seeking to conserve North-East India's ecologically vital forests amid intensifying anthropogenic and climatic pressures.

Author contributions

Debarshi Ghosh, Apurba Sarkar: Conceptualization, Data curation, Methodology, Visualization, Supervision, Writing—original draft. Sanjoy Mandal: Visualization, Investigation, Writing—review and editing. Uday Chatterjee, Sujoy Kumar Malo: Data curation, Formal analysis. Saidur Rahaman, Sandipan Das: Methodology, Visualization, Validation, Project administration. Mantu Das: Writing—review and editing. Snehasish Saha, Pradip Chouhan: Methodology, Investigation, Validation, Writing–original draft.

Funding

Open access funding provided by Symbiosis International (Deemed University). No specific grant from funding agencies like public, commercial, or not-for-profit sectors are provided for this research.

Data availability

No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

On behalf of all the authors, I, as the corresponding author, give you consent to the publication.

Competing interests

The authors declare no competing interests.

Received: 4 June 2025 Accepted: 24 September 2025
Published online: 07 November 2025

References

1. Abu El-Magd SA, Ismael IS, El-Sabri MAS, Abdo MS, Farhat HI (2023) Integrated machine learning–based model and WQI for groundwater quality assessment: ML, geospatial, and hydro-index approaches. Environ Sci Pollut Res Int 30(18):53862–53875. https://doi.org/10.1007/s11356-023-25938-1
2. Ahmad S, Mehmood K, Rehman A, Ur N, Muhammad S, Shahzad F, Hussain K, Luo M, Alarfaj AA, Ali S, Razzaq W (2024) Environmental and Sustainability Indicators Unveiling fractional vegetation cover dynamics: a spatiotemporal analysis using MODIS NDVI and machine learning. Environ Sustain Ind. https://doi.org/10.1016/j.indic.2024.100485
3. Ahmad S, Mehmood K, Imran S, Raza H, Pfautsch S, Shah M, Jamjareegulgarn P, Shahzad F, Alarfaj AA, Ali S, Razzaq W, Dube T (2025) Ecological Informatics Spatiotemporal analysis of surface Urban Heat Island intensity and the role of vegetation in six major Pakistani cities. Ecol Inform 85(December 2024):102986. https://doi.org/10.1016/j.ecoinf.2024.102986
4. Allouche O, Tsoar A, Kadmon R (2006) Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). J Appl Ecol 43(6):1223–1232. https://doi.org/10.1111/j.1365-2664.2006.01214.x
5. Aquilas NA, Mukong AK, Kimengsi JN, Ngangnchi FH (2022) Economic activities and deforestation in the Congo basin: an environmental kuznets curve framework analysis. Environ Chall 8:100553. https://doi.org/10.1016/j.envc.2022.100553
6. Barber CP, Cochrane MA, Souza CM, Laurance WF (2014) Roads, deforestation, and the mitigating effect of protected areas in the Amazon. Biol Conserv 177:203–209. https://doi.org/10.1016/j.biocon.2014.07.004
7. Barona E, Ramankutty N, Hyman G, Coomes OT (2010) The role of pasture and soybean in deforestation of the Brazilian Amazon. Environ Res Lett 5(2):24002. https://doi.org/10.1088/1748-9326/5/2/024002
8. Bauters M, Vercleyen O, Vanlauwe B, Six J, Bonyoma B, Badjoko H, Hubau W, Hoyt A, Boudin M, Verbeeck H, Boeckx P (2019) Long-term recovery of the functional community assembly and carbon pools in an African tropical forest succession. Biotropica 51(3):319–329. https://doi.org/10.1111/btp.12647
9. Bengal W, Mondal I, Thakur S, Juliev M, De TK (2021) Comparative analysis of forest canopy mapping methods. Environ Dev Sustain 23(10):15157–15182. https://doi.org/10.1007/s10668-021-01291-6
10. Bera B, Bhattacharjee S, Sengupta N, Saha S (2021) Dynamics of deforestation and forest degradation hotspots applying geo-spatial techniques, apalchand forest in terai belt of Himalayan foothills: conservation priorities of forest ecosystem. Remote Sens Appl Soc Environ 22(April):100510. https://doi.org/10.1016/j.rsase.2021.100510
11. Bera B, Saha S, Bhattacharjee S (2020) Forest cover dynamics (1998 to 2019) and prediction of deforestation probability using binary logistic regression (BLR) model of Silabati watershed, India. Trees Forests People 2(September):100034. https://doi.org/10.1016/j.tfp.2020.100034
12. Bhattacharya RK, Das Chatterjee N, Das K (2021) Land use and land cover change and its resultant erosion susceptible level: an appraisal using RUSLE and logistic regression in a tropical plateau basin of West Bengal, India. Environ Dev Sustain. https://doi.org/10.1007/s10668-020-00628-x
13. Borgohain S, Das J, Saraf AK, Singh G, Baral SS (2017) Structural controls on topography and river morphodynamics in Upper Assam Valley, India. Geodinam Acta 29(1):62–69. https://doi.org/10.1080/09853111.2017.1313090
14. Brun C, Cook AR, Lee JSH, Wich SA, Koh LP, Carrasco LR (2015) Analysis of deforestation and protected area effectiveness in Indonesia: a comparison of Bayesian spatial models. Glob Environ Change 31:285–295. https://doi.org/10.1016/j.gloenvcha.2015.02.004
15. Buya S, Tongkumchum P, Owusu BE (2020) Modelling of land-use change in Thailand using binary logistic regression and multinomial logistic regression. Arab J Geosci. https://doi.org/10.1007/s12517-020-05451-2
16. Cai J, Xu K, Zhu Y, Hu F, Li L (2020) Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest. Appl Energy 262(November 2019):114566. https://doi.org/10.1016/j.apenergy.2020.114566
17. Curtis PG, Slay CM, Harris NL, Tyukavina A, Hansen MC (2018) Supplementary Materials: classifying drivers of global forest loss. Science 361(6407):1108–1111

18. Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random forests for classification in ecology. Ecology 88(11):2783–2792. https://doi.org/10.1890/07-0539.1
19. Das G, Patra JK, Singdevsachan SK, Gouda S, Shin HS (2016) Diversity of traditional and fermented foods of the Seven Sister states of India and their nutritional and nutraceutical potential: a review. Front Life Sci 9(4):292–312. https://doi.org/10.1080/21553769.2016.1249032
20. DeVries B, Pratihast AK, Verbesselt J, Kooistra L, Herold M (2016) Characterizing forest change using community-based monitoring data and Landsat time series. PLoS ONE 11(3):e0147121. https://doi.org/10.1371/journal.pone.0147121
21. Dominguez D, del Villar LD, Pantoja O, González-Rodríguez M (2022) Forecasting amazon rain-forest deforestation using a hybrid machine learning model. Sustainability (Switzerland) 14(2):1–18. https://doi.org/10.3390/su14020691
22. Du Y, Ding Y, Li Z, Cao G (2015) The role of hazard vulnerability assessments in disaster preparedness and prevention in China. Mil Med Res 2(1):27. https://doi.org/10.1186/s40779-015-0059-9
23. Danoedoro P, Gupita DD (2022) Combining pan-sharpening and forest cover density transformation methods for vegetation mapping using Landsat-8 satellite imagery. Int J Adv Sci Eng Inf Technol. https://doi.org/10.18517/ijaseit.12.3.12514
24. Deka J, Tripathi OP, Khan ML (2013) Implementation of forest canopy density model to monitor tropical deforestation. J Indian Soc Remote Sens 41(2):469–475. https://doi.org/10.1007/s12524-012-0224-5
25. Ellwanger J, Kulmann-Leal B, Kaminski V, Valverde J, Gorini da Veiga A, Spilki F, Fearnside P, Caesar L, Giatti L, Wallau G, Almeida S, BORBA M, Pousada da Hora V, Chies J (2020) Beyond diversity loss and climate change: Impacts of Amazon deforestation on infectious diseases and public health. Anais Da Academia Brasileira de Ciências 92:20191375. https://doi.org/10.1590/0001-3765202020191375
26. Erb K-H, Lauk C, Kastner T, Mayer A, Theurl MC, Haberl H (2016) Exploring the biophysical option space for feeding the world without deforestation. Nat Commun 7(1):11382. https://doi.org/10.1038/ncomms11382
27. Fang K, Tang H, Li C, Su X, An P, Sun S (2023) Centrifuge modelling of landslides and landslide hazard mitigation: a review. Geosci Front 14(1):101493. https://doi.org/10.1016/j.gsf.2022.101493
28. Fauzi AI, Harto AB, Hakim DM, Perdana RS (2019) Analisis degradasi penutup hutan di perkotaan menggunakan model forest canopy density studi kasus: kota bandar lampung. J Mineral, Energi, Dan Lingkungan. https://doi.org/10.31315/jmel.v3i2.3057
29. Freeman EA, Moisen GG (2008) A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. Ecol Modell 217(1–2):48–58. https://doi.org/10.1016/j.ecolmodel.2008.05.015
30. Gao Q, Yu M (2021) Canopy density and roughness differentiate resistance of a tropical dry forest to major hurricane damage. Remote Sens. https://doi.org/10.3390/rs13122262
31. Gayen A, Saha S (2017) Deforestation probable area predicted by logistic regression in Pathro River basin : a tributary of Ajay River. Spat Inf Res. https://doi.org/10.1007/s41324-017-0151-1
32. Geist HJ, Lambin EF (2002) Proximate causes and underlying driving forces of tropical deforestation. Bioscience 52(2):143–150. https://doi.org/10.1641/0006-3568(2002)052[0143:PCAUDF]2.0.CO;2
33. Global Forest Watch (2021) India Deforestation Rates and Statistics. Global Forest Watch
34. Hallegatte S (2009) Strategies to adapt to an uncertain climate change. Glob Environ Change 19(2):240–247. https://doi.org/10.1016/j.gloenvcha.2008.12.003
35. Hansen MC, Potapov PV, Moore R, Hancher M, Turubanova SA, Tyukavina A, Thau D, Stehman SV, Goetz SJ, Loveland TR, Kommareddy A, Egorov A, Chini L, Justice CO, Townshend JRG (2013) High-resolution global maps of 21st-century forest cover change. Science 342(6160):850–853. https://doi.org/10.1126/science.1244693
36. Hazarika N, Das AK, Borah SB (2015) Assessing land-use changes driven by river dynamics in chronically flood affected Upper Brahmaputra plains, India, using RS-GIS techniques. Egypt J Remote Sensing Space Sci 18(1):107–118. https://doi.org/10.1016/j.ejrs.2015.02.001
37. He F, Li S, Zhang X (2015) A spatially explicit reconstruction of forest cover in China over 1700–2000. Glob Planet Change 131:73–81. https://doi.org/10.1016/j.gloplacha.2015.05.008
38. Himayah S, Hartono, Danoedoro P (2016) The Utilization of Landsat 8 Multitemporal Imagery and Forest Canopy Density (FCD) Model for Forest Reclamation Priority of Natural Disaster Areas at Kelud Mountain, East Java. In: IOP Conference Series: Earth and Environmental Science. https://doi.org/10.1088/1755-1315/47/1/012043
39. Hu X, Wu C, Hong W, Qiu R, Li J, Hong T (2014) Forest cover change and its drivers in the upstream area of the Minjiang River, China. Ecol Indic 46:121–128. https://doi.org/10.1016/j.ecolind.2014.06.015
40. IPCC (2014) Part A: Global and Sectoral Aspects. (Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change). Climate Change 2014: Impacts, Adaptation, and Vulnerability 1132
41. Jensen JR, Clarke KC (2016) Digital image processing a remote sensing perspective * pearson series in geographic information science
42. Jiang Z, Huete AR, Chen J, Chen Y, Li J, Yan G, Zhang X (2006) Analysis of NDVI and scaled difference vegetation index retrievals of vegetation fraction. Remote Sens Environ 101(3):366–378. https://doi.org/10.1016/j.rse.2006.01.003
43. Kahn JR, McDonald JA (1995) Third-world debt and tropical deforestation. Ecol Econ. https://doi.org/10.1016/0921-8009(94)00024-P
44. Kalantar B, Al-Najjar HAH, Pradhan B, Saeidi V, Halin AA, Ueda N, Naghibi SA (2019) Optimized conditioning factors using machine learning techniques for groundwater potential mapping. Water. https://doi.org/10.3390/w11091909
45. Kay H, Santoro M, Cartus O, Bunting P, Lucas R (2021) Exploring the relationship between forest canopy height and canopy density from spaceborne lidar observations. Remote Sens 13(24):1–15. https://doi.org/10.3390/rs13244961
46. Kayet N, Pathak K, Kumar S, Singh CP, Chowdary VM, Chakrabarty A, Sinha N, Shaik I, Ghosh A (2021) Deforestation susceptibility assessment and prediction in hilltop mining-affected forest region. J Environ Manag 289(November 2020):112504. https://doi.org/10.1016/j.jenvman.2021.112504
47. Kumar M, Rawat SPS, Singh H, Ravindranath NH, Kalra N (2018) Dynamic forest vegetation models for predicting impacts of climate change on forests: an Indian perspective. Indian J Forestry 41(1):1–12. https://doi.org/10.54207/bsmps1000-2018-f7l9y5
48. Laurance WF, Goosem M, Laurance SGW (2009) Impacts of roads and linear clearings on tropical forests. Trends Ecol Evol 24(12):659–669. https://doi.org/10.1016/j.tree.2009.06.009
49. Lawrence D, Coe M, Walker W, Verchot L, Vandecar K (2022) The unseen effects of deforestation: biophysical effects on climate. Front Forests Glob Change. https://doi.org/10.3389/ffgc.2022.756115
50. Lee DS, Fahey DW, Skowron A, Allen MR, Burkhardt U, Chen Q, Doherty SJ, Freeman S, Forster PM, Fuglestvedt J, Gettelman A, De León RR, Lim LL, Lund MT, Millar RJ, Owen B, Penner JE, Pitari G, Prather MJ, Wilcox LJ (2021) The contribution of global aviation to anthropogenic climate forcing for 2000 to 2018. Atmos Environ. https://doi.org/10.1016/j.atmosenv.2020.117834
51. Lenatti M, Moreno-Sánchez PA, Polo EM, Mollura M, Barbieri R, Paglialonga A (2022) Evaluation of machine learning algorithms and explainability techniques to detect hearing loss from a speech-in-noise screening test. Am J Audiol. https://doi.org/10.1044/2022_AJA-21-00194
52. Levy SA, Cammelli F, Munger J, Gibbs HK, Garrett RD (2023) Deforestation in the Brazilian Amazon could be halved by scaling up the implementation of zero-deforestation cattle commitments. Glob Environ Change. https://doi.org/10.1016/j.gloenvcha.2023.102671
53. Lempert RJ, Popper SW, Bankes SC (2003) Shaping the next one hundred years
54. Luu C, Bui QD, Costache R, Nguyen LT, Nguyen TT, Van Phong T, Van Le H, Pham BT (2021) Flood-prone area mapping using machine learning techniques: a case study of Quang Binh province, Vietnam. Nat Hazards 108(3):3229–3251. https://doi.org/10.1007/s11069-021-04821-7
55. Liu C, Berry PM, Dawson TP, Pearson RG (2005) Selecting thresholds of occurrence in the prediction of species distributions. Ecography 28(3):385–393. https://doi.org/10.1111/j.0906-7590.2005.03957.x
56. Lohani S, Dilts TE, Weisberg PJ, Null SE, Hogan ZS (2020) Rapidly accelerating deforestation in Cambodia's Mekong River basin: a comparative analysis of spatial patterns and drivers. Water. https://doi.org/10.3390/W12082191

57. Malhi Y, Román-Cuesta RM (2008) Analysis of lacunarity and scales of spatial homogeneity in IKONOS images of Amazonian tropical forest canopies. Remote Sens Environ 112(5):2074–2087. https://doi.org/10.1016/j.rse.2008.01.009
58. Mathipi V, de Mandal S, Chawngthu Z, Lalfelpuii R, Kumar NS, Lalthanzara H (2020) Diversity and metabolic potential of earthworm gut microbiota in Indo-Myanmar biodiversity hotspot. J Pure Appl Microbiol. https://doi.org/10.22207/JPAM.14.2.48
59. Mageswary G, Karthikeyan D (2019) Statistical based feature selection and ensemble model for network intrusion detection using data mining technique. Int J Recent Technol Eng (IJRTE) 8:858–864. https://doi.org/10.35940/ijrte.C4049.098319
60. Mayfield HJ, Smith C, Gallagher M, Hockings M (2020) Considerations for selecting a machine learning technique for predicting deforestation. Environ Model Softw 131:104741. https://doi.org/10.1016/j.envsoft.2020.104741
61. Merghadi A, Yunus AP, Dou J, Whiteley J, ThaiPham B, Bui DT, Avtar R, Abderrahmane B (2020) Machine learning methods for landslide susceptibility studies: a comparative overview of algorithm performance. Earth Sci Rev 207:103225. https://doi.org/10.1016/j.earscirev.2020.103225
62. Merufinia E, Sharafati A, Abghari H, Hassanzadeh Y·(2023) On the simulation of streamflow using hybrid tree-based machine learning models: a case study of Kurkursar basin, Iran. Arab J Geosci 16:28. https://doi.org/10.1007/s12517-022-11045-x
63. Mehmood K, Anees SA, Rehman A, Tariq A, Liu Q, Muhammad S, Rabbi F, Pan S, Hatamleh WA (2024) Assessing forest cover changes and fragmentation in the Himalayan temperate region: implications for forest conservation and management. J For Res 35(1):1–14. https://doi.org/10.1007/s11676-024-01734-6
64. Mitra A, Khan A (2017) Green tourism management in India-'A 3D study of the seven sisters states of North-East with special reference to eco-tourism'. Int J Innov Res Sci Eng Technol 6(4):6923–6932
65. Naseem R, Khan B, Ahmad A, Almogren A, Jabeen S, Hayat B, Shah MA (2020) Investigating tree family machine learning techniques for a predictive system to unveil software defects. Complexity. https://doi.org/10.1155/2020/6688075
66. Nazir N, Ahmad S (2018) Forest land conversion dynamics: a case of Pakistan. Environ Dev Sustain 20(1):389–405. https://doi.org/10.1007/s10668-016-9887-3
67. Nivesh S, Negi D, Kashyap PS, Aggarwal S, Singh B, Saran B, Sawant PN, Sihag P (2022) Prediction of river discharge of Kesinga sub-catchment of Mahanadi basin using machine learning approaches. Arab J Geosci. https://doi.org/10.1007/s12517-022-10555-y
68. Pal S, Paul S (2020) Assessing wetland habitat vulnerability in moribund Ganges delta using bivariate models and machine learning algorithms. Ecol Indic 119(August):106866. https://doi.org/10.1016/j.ecolind.2020.106866
69. Pasha SV, Dadhwal VK (2024) National analysis on variations in estimates of forest cover dynamics over India (2001–2020) using multiple techniques and data sources. Spat Inf Res 32(4):451–461. https://doi.org/10.1007/s41324-024-00570-4
70. Pir Bavaghar M (2015) Deforestation modelling using logistic regression and GIS. J Forest Sci 61(5):193–199
71. Rikimaru A (2002) Tropical forest cover density mapping. Trop Ecol 43(1):39–47
72. Rodríguez-Pérez R, Bajorath J (2020) Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. J Comput Aided Mol Des 34(10):1013–1026. https://doi.org/10.1007/s10822-020-00314-0
73. Roy A, Das SK, Tripathi AK, Singh NU, Barman HK (2015) Biodiversity in North East India and their conservation. Prog Agric 15(2):182. https://doi.org/10.5958/0976-4615.2015.00005.8
74. Roy PS, Behera MD, Murthy MSR, Roy A, Singh S, Kushwaha SPS, Jha CS, Sudhakar S, Joshi PK, Reddy CS, Gupta S, Pujar G, Dutt CBS, Srivastava VK, Porwal MC, Tripathi P, Singh JS, Chitale V, Skidmore AK, Ramachandran RM (2015) New vegetation type map of India prepared using satellite remote sensing: comparison with global vegetation maps and utilities. Int J Appl Earth Obs Geoinf 39:142–159. https://doi.org/10.1016/j.jag.2015.03.003
75. Roy PS, Saran S (2004) Biodiversity information system for North East India. Geocarto Int 19(3):73–80. https://doi.org/10.1080/10106040408542320
76. Saha A, Ghosh M, Pal SC, Chowdhuri I, Chakrabortty R, Roy P, Das B, Malik S (2021) Assessment of forest cover dynamics using forest canopy density model in sali river basin: a spill channel of damodar river. Springer International Publishing, In Environmental Science and Engineering. https://doi.org/10.1007/978-3-030-56542-8_15
77. Saha S, Bhattacharjee S, Shit PK, Sengupta N, Bera B (2022) Deforestation probability assessment using integrated machine learning algorithms of Eastern Himalayan foothills (India). Resour Conserv Recycl Adv. https://doi.org/10.1016/j.rcradv.2022.200077
78. Saha S, Paul GC, Pradhan B, Abdul Maulud KN, Alamri AM (2021) Integrating multilayer perceptron neural nets with hybrid ensemble classifiers for deforestation probability assessment in Eastern India. Geomat Nat Hazards Risk 12(1):29–62. https://doi.org/10.1080/19475705.2020.1860139
79. Saha S, Saha M, Mukherjee K, Arabameri A, Ngo PTT, Paul GC (2020) Predicting the deforestation probability using the binary logistic regression, random forest, ensemble rotational forest, REPTree: a case study at the Gumani River Basin, India. Sci Total Environ 730:139197. https://doi.org/10.1016/j.scitotenv.2020.139197
80. Sahana M, Hong H, Sajjad H, Liu J, Zhu AX (2018) Assessing deforestation susceptibility to forest ecosystem in Rudraprayag district, India using fragmentation approach and frequency ratio model. Sci Total Environ 627:1264–1275. https://doi.org/10.1016/j.scitotenv.2018.01.290
81. Sahana M, Sajjad H, Ahmed R (2015) Assessing spatio-temporal health of forest cover using forest canopy density model and forest fragmentation approach in Sundarban reserve forest, India. Mod Earth Syst Environ 1(4):1–10. https://doi.org/10.1007/s40808-015-0043-0
82. Sboui T, Saidi S, Lakti A (2023) A machine-learning-based approach to predict deforestation related to oil palm: conceptual framework and experimental evaluation. Appl Sci (Switzerland). https://doi.org/10.3390/app13031772
83. Seto KC, Reenberg A, Boone CG, Fragkias M, Haase D, Langanke T, Marcotullio P, Munroe DK, Olah B, Simon D (2012) Urban land teleconnections and sustainability. Proc Natl Acad Sci USA 109(20):7687–7692. https://doi.org/10.1073/pnas.1117622109
84. Silva AMda, Rodgers J (2018) Deforestation across the world: causes and alternatives for mitigating. Int J Environ Sci Dev. https://doi.org/10.18178/ijesd.2018.9.3.1075
85. Studer S, Bui TB, Drescher C, Hanuschkin A, Winkler L, Peters S, Müller KR (2021) Towards CRISP-ML(Q): a machine learning process model with quality assurance methodology. Mach Learn Knowl Extr. https://doi.org/10.3390/make3020020
86. Taubert F, Fischer R, Groeneveld J, Lehmann S, Müller MS, Rödig E, Wiegand T, Huth A (2018) Global patterns of tropical forest fragmentation. Nature. https://doi.org/10.1038/nature25508
87. Trancoso R (2021) Changing Amazon deforestation patterns: urgent need to restore command and control policies and market interventions. Environ Res Lett. https://doi.org/10.1088/1748-9326/abee4c
88. Tarazona Y, Zabala A, Pons X, Broquetas A, Nowosad J, Zurqani HA (2021) Fusing Landsat and SAR data for mapping tropical deforestation through machine learning classification and the PVts-β non-seasonal detection approach. Can J Remote Sens 47(5):677–696. https://doi.org/10.1080/07038992.2021.1941823
89. Wang F, Sahana M, Pahlevanzadeh B, Chandra Pal S, Kumar Shit P, Piran MJ, Janizadeh S, Band SS, Mosavi A (2021) Applying different resampling strategies in machine learning models to predict head-cut gully erosion susceptibility. Alexandria Eng J 60(6):5813–5829. https://doi.org/10.1016/j.aej.2021.04.026
90. Werf GRVD, Morton DC, Defries RS, Olivier JGJ, Kasibhatla PS, Jackson RB, Collatz GJ (2009) CO2 emissions from forest loss. Nat Geosci. https://doi.org/10.1038/ngeo671
91. Xi Y, Tian Q, Zhang W, Zhang Z, Tong X, Brandt M, Fensholt R (2022) Quantifying understory vegetation density using multi-temporal Sentinel-2 and GEDI LiDAR data. GISci Remote Sens. https://doi.org/10.1080/15481603.2022.2148338
92. Yoro KO, Daramola MO (2020) Chapter 1-CO2 emission sources, greenhouse gases, and the global warming effect. In: Rahimpour MR, Farsi M, M. A. B. T.-A. in Makarem CC (eds.). Woodhead Publishing pp 3–28. https://doi.org/10.1016/B978-0-12-819657-1.00001-3
93. Yu R, Ren L, Luo Y (2021) Early detection of pine wilt disease in *Pinus tabuliformis* in North China using a field portable spectrometer and

UAV-based hyperspectral imagery. For Ecosyst. https://doi.org/10.1186/s40663-021-00328-6
94. Yuh YG, Tracz W, Matthews HD, Turner SE (2023) Application of machine learning approaches for land cover monitoring in northern Cameroon. Ecol Informatics. https://doi.org/10.1016/j.ecoinf.2022.101955
95. Zarco-Tejada PJ, Miller JR, Morales A, Berjón A, Agüera J (2004) Hyperspectral indices and model simulation for chlorophyll estimation in open-canopy tree crops. Remote Sens Environ. https://doi.org/10.1016/j.rse.2004.01.017
96. Zhang Y, Wang G, Zhuang H, Wang L, Innes JL, Ma K (2021) Integrating hotspots for endemic, threatened and rare species supports the identification of priority areas for vascular plants in SW China. For Ecol Manage. https://doi.org/10.1016/j.foreco.2021.118952
97. Zimmermann A, Church M (2001) Channel morphology, gradient profiles and bed stresses during flood in a step–pool channel. Geomorphology 40(3):311–327. https://doi.org/10.1016/S0169-555X(01)00057-5

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.