Engineering Geology 355 (2025) 108229

Contents lists available at ScienceDirect

Engineering Geology

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

Geology, geomorphology and Vs30 based site classification of the Himalayan region using a stacked model

Harish Thakur, P. Anbazhagan *

Department of Civil Engineering, Indian Institute of Science, Bengaluru 560012, India

A R T I C L E I N F O

Keywords:
Seismic site classification
Vs30
Geology aspects
Geomorphological predictors
Georeferencing
Stacked ensemble learning

A B S T R A C T

Seismic site classification maps are essential for earthquake hazard assessment and risk mitigation. This study presents a novel methodological framework for developing site classification maps for the Himalayan region by integrating geological information, geomorphological parameters, and shear wave velocity (Vs30) measurements within a machine-learning architecture. We compiled a comprehensive Vs30 database comprising 3077 data points from 39 studies across the Himalayan region, utilizing twelve predictor variables derived from digital elevation models and geological maps. Three individual machine learning models were developed and optimized, achieving Mean Absolute Errors (MAEs) between 48 and 53 m/s and Root Mean Squared Errors (RMSEs) between 70 and 77 m/s on their test set. A Stacked Ensemble Model (SEM) was developed using meta-learning to optimally combine the three base models, resulting in test set performance metrics of 53 m/s MAE, 0.38 R^2, and 78.2 m/s RMSE. The SEM delivers a 64.6 % overall improvement in prediction accuracy (MAE) compared to the existing slope-based method, with the most significant gains of 80.5 % error reduction occurring in the 180–259 m/s range representing typical urban soil conditions. Beyond average error reduction, the SEM shows improved error consistency through RMSE improvements (52.6 %) compared to the slope-based method. It confirms enhanced performance without additional prediction volatility, making it particularly valuable for engineering applications requiring accuracy and reliability. The model reveals significant spatial heterogeneity across the Himalayan region, with mountainous areas classified as site classes A and B (Vs30 > 760 m/s), foothills as classes C and D (360–760 m/s), and deltaic regions as classes D and E (Vs30 < 360 m/s). Urban-scale maps for Delhi, Dhaka, Kathmandu, and Guwahati demonstrate the approach's capability to capture local variations critical for seismic hazard assessment and building code implementation. This methodology overcomes the limitations of previous single-predictor approaches, providing a robust framework for regional-scale Vs30 prediction in geologically complex mountainous terrain.

1. Introduction

Site classification plays a critical role in earthquake hazard estimation and risk assessment. It provides essential information about site response and amplification characteristics, which inform preventive measures, urban planning decisions, and risk management strategies at both local and regional scales. Appropriate site classification contributes significantly to building stock safety and enables the development of more effective local building codes and design recommendations (BSSC, 2015). Furthermore, site classification serves as a foundation for estimating secondary earthquake effects, including site amplification and liquefaction potential at specific locations (Heath et al., 2020; Lin et al., 2021). This assessment capability is particularly valuable in seismically active regions such as the Himalayas, where several devastating earthquakes have occurred throughout history.

The Himalayan region has experienced numerous major seismic events, including the Kangra earthquake of 1905 (Mw 7.5), Bihar in 1934 (Mw 8.1), Assam in 1950 (Mw 8.6), Kashmir in 2005 (Mw 7.6), Sikkim in 2011 (Mw 6.9), and Nepal in 2015 (Mw 7.8). These events have resulted in substantial loss of human lives and extensive infrastructure damage (Kayal, 2008). The severity of earthquake-related losses can be significantly reduced through comprehensive preparedness and targeted mitigation measures, which fundamentally depend on understanding site conditions. Knowing site conditions is essential for earthquake preparedness, evaluating existing building inventory and planning future infrastructure development (Martínez-Segura et al.,

* Corresponding author.
E-mail addresses: harishthakur@iisc.ac.in (H. Thakur), anbazhagan@iisc.ac.in (P. Anbazhagan).

https://doi.org/10.1016/j.enggeo.2025.108229
Received 2 January 2025; Received in revised form 16 June 2025; Accepted 5 July 2025
Available online 7 July 2025

2024). Evidence from global earthquake events consistently demonstrates that site conditions substantially influence building response, resulting in damage patterns and associated economic losses (Trifunac, 2016; Panzera et al., 2018; Brando et al., 2020). Therefore, systematic assessment and classification of site conditions at local, regional, and national scales can enable the construction of more resilient structures and help prevent catastrophic failures. With this objective in mind, this study utilizes information from various sources to conduct site classification across the Himalayan region at a regional scale.

Past studies for seismic site classification based on parameters like shear wave velocity (Vs), geology, and natural frequency have been done worldwide (Walling et al., 2009; Yilar et al., 2017; Anbazhagan et al., 2019; Verdugo, 2019; Kim et al., 2021; Geyin and Maurer, 2023; Di Martino et al., 2024). These classification systems have been extensively applied in earthquake studies to assess seismic site effects, amplification potential and microzonation purposes (Mihalić et al., 2011; Molnar et al., 2020; Falcone et al., 2021).

Measuring shear wave velocity under field conditions typically employs various geotechnical and geophysical methods. Geotechnical approaches include cross-hole (CH) and down-hole (DH) testing (Kramer and Stewart, 2024), while geophysical methods encompass techniques such as Multichannel Analysis of Surface Waves (MASW) (Park et al., 1999) and Spectral Analysis of Surface Waves (SASW) (Nazarian and Stokoe, 1984). Additionally, shear wave velocity can be determined indirectly using correlations with Standard Penetration Test (SPT-N) values and Cone Penetration Test (CPT) measurements from existing or site-specific investigations (Akin et al., 2011; Ansary et al., 2023).

For site characterization based on shear wave velocity, the average velocity in the top 30 m of a site (Vs30) has emerged as a standard representative parameter derived from the site's Vs profile. While site-specific studies provide detailed assessments of local site effects, regional-level evaluations often necessitate reliance on simplified proxy-based methods (Vilanova et al., 2018). The following section presents a comprehensive overview of site classification studies conducted globally and specifically within the Himalayan region, utilizing Vs30 as the primary classification parameter.

1.1. Previous approaches to regional site classification

To overcome the limitations of direct measurement, researchers have developed proxy-based approaches that correlate Vs30 with readily available geological and topographical data. Matsuoka et al. (2005) developed site amplification maps for Japan using different geomorphologic features such as elevation, slope value, distance from the river, coastline, and mountain or hill. Wald and Allen (2007) pioneered a method correlating topographic slope with Vs30, which has been widely implemented due to the global availability of digital elevation models (DEMs). This approach is based on the observation that steeper slopes typically correspond to more competent materials with higher shear wave velocities. In comparison, gentler slopes generally indicate softer sediments with lower velocities. While efficient for global applications, this slope-based approach has demonstrated limitations in regions with complex geology (Thompson et al., 2014).

Subsequent studies have explored more sophisticated approaches combining multiple predictors. Thompson et al. (2014) developed a hybrid model incorporating geology and topography and measured Vs30 values for California. Vilanova et al. (2018) created a geologically based Vs30 model for Portugal, demonstrating improved performance over topography-based methods alone. Mori et al. (2020) utilized geomorphological parameters for developing a national Vs30 map for Italy. Crespo et al. (2022) gave a Vs30 estimation model in the Iberian Peninsula while using the site's topographic slope, geological age and lithology as proxies for estimation. For the Tehran region, Abbasnejadfard et al. (2023) utilized a Vs30 database (prepared using scientific reports, government agencies, and private engineering consultancy companies) to prepare a Vs30 map while utilizing different spatial interpolation methods.

More recently, machine learning techniques have shown considerable promise in integrating multiple predictors for Vs30 estimation, as Geyin and Maurer (2023) demonstrated for the United States and Puerto Rico. In China, Xie et al. (2016) developed a Vs30-based map for the Beijing plain area, utilizing proxies like slope and surficial geological features. They have highlighted that geology-based estimation of Vs30 is more reliable than slope-based estimation as the field values of Vs30 for their region were usually lower than those estimated by Wald and Allen's (2007) model. Liu et al. (2017) developed a multiscale random-field framework to develop a Vs30 map for the Suzhou city area. To account for spatial variability across different length scales correlations were used to develop the Vs30 maps. Li et al. (2019) used a slope-geological method to develop a site classification map for China. They have utilized geological features such as age, soil properties, and depositional area characteristics as predictors for Vs30. Zhang et al. (2023) have developed Vs30-based maps for different regions and provinces of China using parameters such as slope, surface texture, and landform features as input parameters.

In these different studies, maps have been produced at different resolutions (depending on the information available for Vs30 data and its predictors, such as geological and topographical data), which may not capture variation at the local scale for the sites where geological formations are diverse. In past studies, the areas covered for site classifications have soft soil sites with relatively flat terrains, and there is a general lack of Vs30 measurements for relatively steeper slopes, leading to underprediction of Vs30 corresponding to higher site classes (Geyin and Maurer, 2023).

Previous site classification efforts in the Himalayan region have primarily relied on interpolation techniques or single-parameter proxy methods. Sitharam et al. (2015) have prepared a site class map for the whole of India (where Vs30 is a site classification parameter based on the slope) using the Allen and Wald (2009) criterion, which is based on NEHRP site classes (Table 1). They used 10 km × 10 km resampled grids for the slope analysis and final Vs30 map preparation for their site class map. At the local scale, Rahman et al. (2016, 2018) have developed site characterization maps for Chittagong and Dhaka cities, respectively, using Holocene soil thickness as a predictor while employing an inverse distance weighing scheme for interpolation. Studies like Satyam and Rao (2008), Mahajan et al. (2012), and Anbazhagan et al. (2013) have utilized kriging for spatial interpolation of Vs30 values for Delhi, Jammu, and Lucknow cities, respectively. In a few studies for the Himalayan region, contour maps have been prepared for site classification (Kandpal et al., 2009; Naik et al., 2014; Bajaj and Anbazhagan, 2019; Riyaz and Singh, 2022).

These studies face three significant limitations. First, interpolation techniques are highly dependent on data density and may not accurately capture the complex spatial variations in geology and topography characteristic of the Himalayas. Second, single-parameter proxy methods fail to account for the combined effects of multiple factors influencing site response. Third, the resolution of existing maps is typically insufficient for detailed hazard assessment, particularly in areas with complex terrain where site conditions can vary significantly over short distances.

Table 1
Adopted Seismic Site Classification scheme in the present study based on Vs30 (shear wave velocity for top 30 m of the ground) and SPT-N as per National Earthquake Hazards Reduction Program or NEHRP (BSSC, 2004).

Site Class	Description	Vs30 (m/s)	SPT-N
A	Hard Rock	>1500	–
B	Rock	760–1500	>50
C	Very Dense Soil and Soft Rock	360–760	15–50
D	Stiff Soil	180–360	<15
E	Soft Clay Soil	<180	–

1.2. Research objectives

Despite the geological and seismic significance of the Himalayan region, comprehensive site classification maps that integrate multiple predictors and provide adequate spatial resolution remain unavailable. To address this gap, our study aims to:

1. Develop a comprehensive Vs30 database for the Himalayan region by compiling and standardizing data from diverse sources.
2. Identify and quantify the relative importance of geological and geomorphological predictors for Vs30 estimation.
3. Create and validate a robust machine-learning model that optimizes prediction accuracy across diverse terrain types.
4. Generate high-resolution site classification maps for the Himalayan region and key urban centres.

This study advances the field of seismic site classification by demonstrating an integrated approach that combines multiple data sources and machine learning techniques to overcome the limitations of previous methods. The resulting maps provide critical information for earthquake hazard assessment, infrastructure planning, and building code implementation in this seismically vulnerable region.

2. Geology and geomorphology of the himalayan region

The Himalayan region features a wide range of geological and geomorphological features that significantly influence seismic behaviour across different sites. The geomorphology of the Himalayan region is characterized by extreme variations in elevation, relief, and drainage patterns, directly influencing site conditions and seismic response. Mountains, valleys, canyons, gorges, rivers, floodplains and deltas are a few of the major surficial features of the region. At its base lies the Indo-Gangetic Plains, which encompass an area of 700,000 km^2 (Gangopadhyay, 2013). This plain lies at the foothills of the Himalayas and is bounded in the south by the Deccan Plateau. The Indo-Gangetic Plain is divided into two major river drainage systems - Indus in the West and Ganga-Brahmaputra towards the East; at the base of the Himalayas forms an alluvial floodplain with different soil types ranging from very fine clays to large boulders.

Sediments from the Pleistocene age and recent deposits of Indo-Gangetic alluvium cover the floodplains. Intermontane valleys, such as the Kathmandu Valley in Nepal and the Kashmir Valley in India, represent important population centres. These valleys are typically filled with thick sequences of lacustrine and fluvial sediments, which can significantly amplify seismic waves. The complex basin geometry and sediment heterogeneity create spatial variations in site response that require high-resolution characterization. Major rivers in this region are (in west to east direction) – Indus, Ganga, and Brahmaputra. These rivers have many tributaries, e.g., Sutlej, Beas, and Chenab for the Indus River, which in turn have sub-tributaries forming a larger river drainage network system of the region (Gangopadhyay, 2013). The major river basins for the region are – the Indus Basin (in the west), Ganga Basin (in Central Himalaya), Barak Basin (covering parts of India, Bangladesh and Myanmar), and Brahmaputra Basin (in the east). Quaternary deposits, in the form of alluvial fans and other fluvial depositional units, are common in the entire Himalayan foothills. These diverse geomorphological settings create a complex mosaic of site conditions across the Himalayan region that any single parameter cannot adequately characterize. Therefore, integrating geological and geomorphological information is essential for comprehensive site classification.

3. Dataset preparation

3.1. Past studies and reports

In past studies, methods of Vs30 estimation are cross-hole and downhole seismic, MASW, microtremor measurements, Multiple Simulations with One Receiver (MSOR), Horizontal to Vertical Spectral Ratio (HVSR), Ambient Noise Interferometry (ANI) and SPT-N-based correlations (Table 2). Data from 39 studies were collected with 3077 data points (Table 2). These Vs30 data points are shown in Fig. 1. Most of these studies are in India and Bangladesh, and only a few are in Nepal (De Risi et al., 2021) and Bhutan (Sarkar et al., 2021; Tempa et al., 2020). At the same time, this data covers different terrains and geographies and is useful to estimate Vs30 values at a macro & micro scale. Some past studies have developed Vs and SPT-N correlation using data collected from SPT and geophysical methods such as MASW, CH, DH, etc. These results have then been utilized to develop site characterization maps at a local level. In some cases, existing or newly developed SPT-N vs Vs relations were used to obtain the Vs range for the SPT-N value, which has been further used to get the Vs30 value for the other locations in the same study (Rahman et al., 2016; Rahman et al., 2018a).

3.2. Vs30 data types

Gilder et al. (2022) have classified Vs30 datasets into primary and secondary types for which rationale was adopted while considering features such as - the source of the data (direct, indirect and proxy-based), its reliability, depth of the profiles, data format type (such as point, grid or raster) and processing techniques used to generate such data. For the present study, shear wave velocity data collected from past studies can be divided based on their information relating to Vs30 calculation and the geolocation data format available. Studies were divided based on two geolocation information categories which are – G1: sites with reported/known lat and long values, and G2: sites with locations marked on a map (Table 2). In the case of G1 studies, information was usually provided in the main text or supplementary files of the studies. For Pandey et al. (2016) and Harinarayan and Kumar (2018) studies, site locations were obtained/cross-referenced by utilizing PESMOS (2024) seismic stations' location information.

3.3. Georeferencing for available maps

In a GIS environment such as QGIS (QGIS Development Team, 2023), georeferencing tools are available to assign spatial coordinates to the digital map image. In the present study, a portion of Vs30 data was extracted from the published maps in point format using georeferencing. To get the Vs30 data points and their spatial coordinates, we have used Metric Georeferencing, which utilizes geotagged locations as a reference/input and a mapped coordinated reference system (CRS) for georeferencing (Yao Xiaobai, 2020). This technique uses Ground Control Points (GCPs) to peg down the map onto the globe with a particular CRS. Usually, more than three points are given as input GCPs with respect to which position of all other points on the map is calculated.

Fig. 2 shows the schematics of the georeferencing procedure adopted in the present study. The following steps were taken for extracting Vs values and site coordinates for the test locations from available study maps with coordinates marked on their borders and with or without grids (such as Mahajan et al., 2007):

1. Grid drawing: Grids were drawn on the figure with reported Vs values (without grid).
2. Image overlay: This image was overlaid with the image showing test site locations on a map to get a final image showing site locations on a Vs map with a grid.
3. Georeferencing: The final image was then exported to the GIS interface, and site locations were geotagged using the Georeferencer toolbox available in the GIS interface using intersection points of grid lines as GCPs.
4. Coordinate extraction: Site location coordinates were finally exported from the georeferenced layer in the GIS interface.

Table 2
List of all studies considered for the dataset preparation with geolocation information categories (or GIC): G1 – sites with reported/known lat and long values, and G2 – Sites with locations marked on a map.

S/N	Study	Tests*	Site Characterization	GIC	Data	Region*
1	DST (2007)	SPT, HVSR	Vs30 (SPT-N based)	G1	200	Guwahati
2	Kuldeep Shekar et al. (2022)	HVSR, ANI	Vs30 (HVSR and ANI based)	G2	54	Assam
3	Shukla and Sil (2023)	HVSR	Vs30 (HVSR-based)	G1	7	Assam
4	Sarkar et al. (2021)	MASW	Vs30	G1	5	Phuentsholing
5	Tempa et al. (2020)	MASW	Vs30	G2	7	Bhutan
6	Kumar and Kumar (2023)	SPT	Vs30 (SPT-N based)	G1	48	BR
7	Rahman et al. (2016)	MASW, SSMM, DS, SPT	Vs30, Vs30 (SPT-N based)	G2	87	Chittagong
8	Rahman et al. (2018b)	MASW, SPT, DS	Vs30, Vs30 (SPT-N based)	G2	146	Dhaka
9	Ansary et al. (2023)	SPT, DS	Vs30 (SPT-N based)	G2	390	Dhaka
10	Satyam and Rao (2008)	MASW	Vs30	G1	117	DL
11	NCS (2016)	MASW, DS, CS, SPT	Vs30, Vs30 (SPT-N based)	G1	535	DL
12	Muthuganeisan and Raghukanth (2016)	MASW	Vs30	G1	73	HP
13	Mahajan and Kumar (2023)	MASW	Vs30	G1	67	Kangra
14	Sharma and Mahajan (2022)	HVSR, MSOR	Vs30	G2	82	Shimla
15	Puri and Jain (2021)	SPT-N	N30	G2	66	HR
16	Pandey et al. (2021)	MASW, HVSR	Vs30	G1	54	UK, UP
17	Bajaj and Anbazhagan (2019)	MASW	Vs30	G2	276	PB, HR, DL, UP, BR
18	Harinarayan and Kumar (2018)	HVSR	Vs30	G2	88	HP, HR, PB, DL, UK
19	Riyaz and Singh (2022)	MASW	Vs30	G1	21	Jammu
20	Zahoor et al. (2021)	MASW, SPT	Vs30, N30	G2	33	Srinagar
21	Mahajan et al. (2012)	MASW, HVSR	Vs30	G2	28	Jammu
22	Imam et al. (2023)	MASW	Vs30	G1	4	Jharkhand
23	Sil and Sitharam (2017)	MASW	Vs30	G2	25	Aizawl (Mizoram)
24	De Risi et al. (2021)	DS	Vs30	G1	15	Kathmandu
25	Bhutani and Naval (2020)	SPT	Vs30 (SPT-N based)	G1	45	Punjab
26	Kandpal et al. (2009)	SPT	Vs30 (SPT-N based)	G2	24	Chandigarh
27	Rahman et al. (2019)	SPT, DS, MASW	Vs30, Vs30 (SPT-N based)	G1	71	Sylhet
28	Rahman et al. (2018a)	SPT, DS, MASW	Vs30, Vs30 (SPT-N based)	G2	22	Sylhet
29	Sil and Sitharam (2014)	MASW, SPT	Vs30	G2	27	Agartala
30	Sharma et al. (2020)	1D-Active MASW, HVSR	Vs30	G1	6	Uttarakhand
31	Mahajan et al. (2007), Mahajan (2009)	2D- Active MASW	Vs30	G2	38	Dehradun
32	Pandey et al. (2016)	1D- Active MASW	Vs30	G1	12	UK
33	Naik et al. (2014)	SPT, DS	Vs30 (SPT-N based)	G2	12	Kanpur
34	Singh et al. (2020)	SPT	Vs30 (SPT-N based)	G1	22	Allahabad
35	Anbazhagan et al. (2013)	MASW, SPT	Vs30, Vs30 (SPT-N based)	G2	55	Lucknow (UP)
36	Chatterjee and Choudhury (2013)	SPT	Vs30 (SPT-N based)	G1	8	Kolkata
37	Bandyopadhyay et al. (2019)	SPT, DS	Vs30	G1	18	Kolkata
38	Nath et al. (2014)	SPT, DS, HVSR	Vs30	G1	218	Kolkata
39	Nath et al. (2022)	SPT, MASW, DS, HVSR	Vs30 (SPT-N based)	G2	71	West Bengal

* Abbreviations: SPT – Standard Penetration Test, HVSR – Horizontal to Vertical Spectral Ratio, MASW – Multichannel Analysis of Surface Waves, ANI – Ambient Noise Interferometry, DS – Downhole Seismic, CS – Crosshole Seismic, SSMM - Small Scale Microtremor Measurement, MSOR - Multiple Simulations with One Receiver. State name abbreviations: Bihar – BR, Delhi – DL, Himachal Pradesh – HP, Punjab – PB, Haryana – HR, Uttarakhand – UK, Uttar Pradesh – UP.

5. Value assignment: Vs values corresponding to each site location were read from the original Vs map reported in the study.

In some studies (DST, 2007; Mahajan et al., 2012; Bajaj and Anbazhagan, 2019; Riyaz and Singh, 2022), Vs30 values were given as contour maps. For such studies, the site location maps were overlaid over the contour map. Then, the Vs30 value of a particular site is calculated by taking an average of the contours within which it lies. Then, these images were georeferenced using the same procedure as described previously.

For DST (2007), extracted geolocations from contour maps matched well with the locations given for sites (but without Vs30 values) in the Annexure-VI of the study. For extraction of the Vs30 data points from Bajaj and Anbazhagan's (2019) study, after following the steps of georeferencing to get Vs30, the Vs30 ranges of different classes (as per NEHRP) were constrained according to the range given in Table 4 of the study. For Kuldeep Shekar et al. (2022), 64 locations were georeferenced, of which 10 locations had unreliable HVSR values and were not considered for Vs30 evaluation. In the study by Rahman et al. (2016) for Chittagong Vs30, values with site names are given in the manuscript, and location coordinates of these sites were obtained using the georeferencing procedure. Fig. 3 shows the spatial distribution of Vs30 data points compiled from these studies, which are predominantly concentrated in urban regions and large population centres.

3.4. Elevation and geological dataset

Predictors for Vs30 were determined using elevation (in the form of a digital elevation model or DEM) and geological dataset as input layer. Shuttle Radar Topographic Mission (SRTM) 30 m resolution DEM is utilized for elevation data and the calculation of other morphological predictors, as mentioned in Section 4.1. For the geological information of the region, we utilized geological maps by Wandrey (1998). Although finer resolution maps are available in the public domain, they were not used as the number of classes in such maps is high for the different sites, affecting model prediction capability as Vs30 information for finer distinct classes is usually unavailable, and it causes redundancy in the model due to the sparsity of predictors.

3.5. Dataset features

Fig. 4 shows a histogram of collected Vs30 datapoints with boundaries marked according to different NEHRP site classes (Table 1). We have drawn plots from the data collected from various studies to see how these sites' elevation and slope values relate to corresponding Vs30 values. A plot of Vs30 vs Elevation (Fig. 5a) shows that higher Vs30 values are observed over increasing elevation ranges, corresponding to higher slopes. A similar trend is observed in a Vs30 vs Slope plot (Fig. 5b), where over the range of slope values as elevation increases, there is a corresponding increase in Vs30 values. Matsuoka et al. (2006)

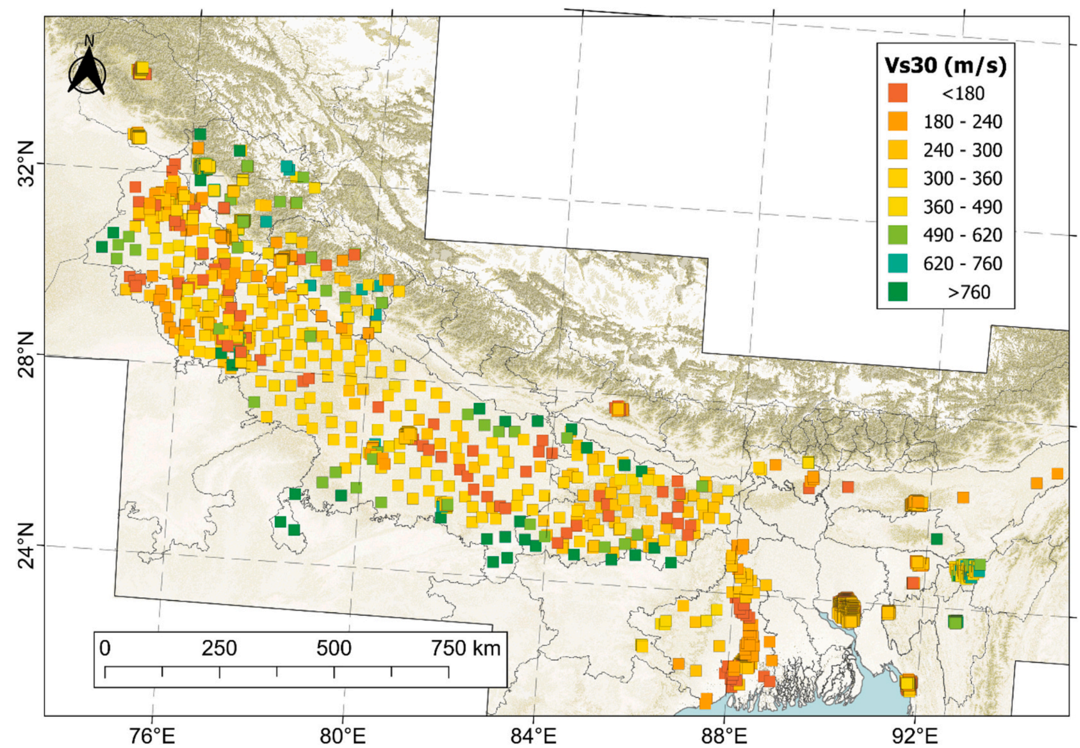


Fig. 1. A plot showing all the Vs30 data points for the region collected from past studies.

also made similar observations: "…the higher the elevation, the steeper the slope angle and the shorter the distance from the mountain or hill, Vs30 values become larger". Although such trends exist, no physics-based formulas relate these parameters with Vs30 (Geyin and Maurer, 2023). In past studies, empirical methods relying on regression-based approaches have been used to find predictors for Vs30. These rely on hypothesis testing to check the statistical importance of the estimated regression coefficients. Then, these coefficients were used to understand the effect of different predictors on Vs30. They suffer from the issue that as the number of predictors increases, it becomes difficult to explain the impact of different predictors in a physical sense. We circumvent this issue using ML approaches, which can provide meaningful insights and predictions without emphasizing the explanatory part.

4. Methodology

We developed a machine learning-based approach that integrates multiple geological and geomorphological predictors for Vs30 prediction across the Himalayan region to overcome the limitations of single-parameter proxy methods and interpolation techniques identified previously. To develop Vs30 site maps, the procedure adopted was as follows: (i) selection of model prediction parameters, (ii) extraction of prediction parameters using DEM at a certain resolution, (iii) prediction models' selection to create a stacked model, (iv) hyperparameter optimization of the stacked model, (v) modification of the stacked model to improve prediction for steep slopes, and (vi) Vs30 map preparation at different resolutions. Parameter extraction and map preparation were done using QGIS software (QGIS Development Team, 2023). Data preprocessing, model development, and analysis were done using MATLAB software (MATLAB, 2023) for the prediction model.

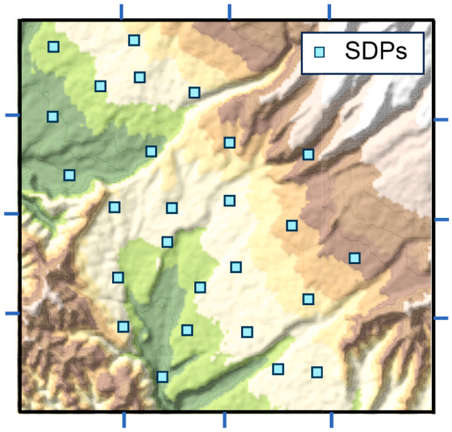
4.1. Model prediction parameters

For Vs30 prediction, we used geology, terrain heterogeneity, and geomorphology-based parameters to estimate Vs30 values using machine learning algorithms. A total of 12 parameters were used (Table 3), which are (1) elevation of the particular site, (2) slope values, (3) geological features, its (4) nearest distance to the river; geomorphological indices such as (5) Topographic Position Index or TPI, (6) Terrain Ruggedness Index or TRI; (7) Topographic Wetness Index or TWI; (8) roughness; (9) profile and (10) plan curvatures; (11) Vector Ruggedness Measure (VRM) and terrain classification (in terms of peaks, ridges, passes, channels pits and planes) using (12) morphometric features.

Next, we briefly describe a few of these features and the rationale behind adopting them. From the observation that the Vs30 value for a site relates to its elevation and slope, these two parameters were chosen as predictors for the modelling (Fig. 5a, b). Geological information of a site can play a crucial role in its shear wave velocity estimation (Rahman et al., 2019; Geyin and Maurer, 2023). Using the geological map (Wandrey, 1998), for Vs30 data points, a total of 13 geological classes were obtained. Fig. 6 shows the region's geological map, which is prepared by modifying the Wandrey (1998) geological map for South Asia. The legend of the figure geological classes obtained corresponding to the Vs30 datapoints locations have been mentioned. Fig. 7 shows the distribution of Vs30 data points across the four most common geological classes.

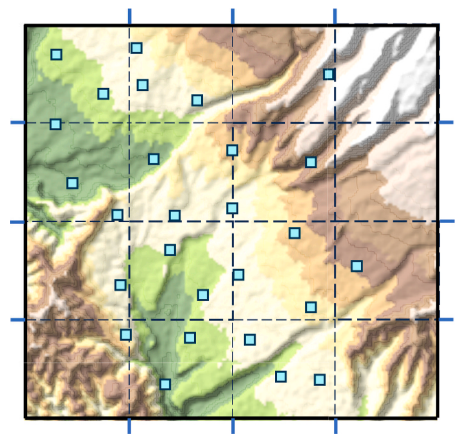
Most of the sites of Vs30 data points belong to Quaternary sediments sites, Precambrian rocks sites, Neogene sedimentary rocks sites and Quaternary sand and dunes sites. Another parameter indicating the nearest distance to the river has been used in past studies to correlate with the Vs30 (Matsuoka et al., 2006; Geyin and Maurer, 2023). It relies on the general observation that sites closer to the river network are likelier to belong to a lower site class than sites farther away. HydroSHEDS dataset (Lehner and Grill, 2013) has been utilized to calculate the closest distance to a river network.

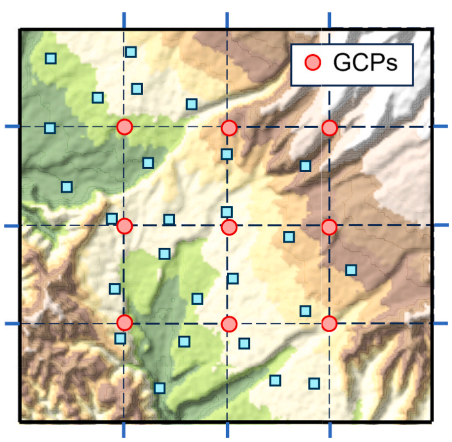
Similarly, the Terrain Wetness Index (TWI) is a parameter related to a landscape's water accumulation potential (Sörensen et al., 2006). Based on the assumption that the locations with more water accumulation potential relate to soil sites and their types compared to rocky sites, this parameter can contribute to Vs30 estimation. Topographic Position Index (TPI) is the difference between a cell elevation value and the average elevation of the neighbourhood around that cell (Jenness,



(a) Map with Vs30 Site data points (SDPs). Tick marks at the periphery indicate reported lat and long values.

(b) Gridded Map (lines were drawn on while connecting marking at the periphery of the image).





(c) Ground Control Points (GCPs) marked in the GIS interface for interpolation.

Fig. 2. Schematics of georeferencing procedure adopted in the present study for Vs30 datapoints extraction.

2006). The Terrain Ruggedness Index (TRI) parameter quantitatively measures topographic heterogeneity. It is calculated as the mean value of the absolute differences in elevation between a focal cell and its eight surrounding cells (Riley et al., 1999). TRI values indicate a mountainous region with high positive values, whereas values closer to zero indicate relatively flat grounds. The surface Roughness parameter captures surficial irregularities while utilizing a location's elevation data as input. It is measured as the largest difference in elevation for a central pixel while considering its surrounding cells. Profile and plan curvature are calculated parallel and perpendicular to the direction of the maximum slope to identify the convexity/concavity of the terrain. In the present case, the SRTM DEM model (with 30 m spatial resolution) and river network data from the HydroSHEDS database were utilized to calculate these parameters.

4.2. *Prediction parameter extraction*

A resampling scheme was used in the GIS environment to prepare maps at a certain scale, similar to Sitharam et al. (2015). Then, the following steps were followed to get all parameter values:

Step 1: In the GIS environment, all parameter values were calculated in the raster format utilizing the DEM base layer except the geological information layer, which was already available in the vector format. This step generated different raster layers for each parameter mentioned in Table 3.

Step 2: After resampling the base DEM to a particular resolution using a bilinear resampling technique, each pixel in the newly formed raster layer is converted to a vector datapoint, forming a final gridded vector DEM layer.

Step 3: All other layers resampled to the same resolution as the DEM layer were then sampled using the gridded vector DEM layer. This step generated a vector layer with an attribute table having all the

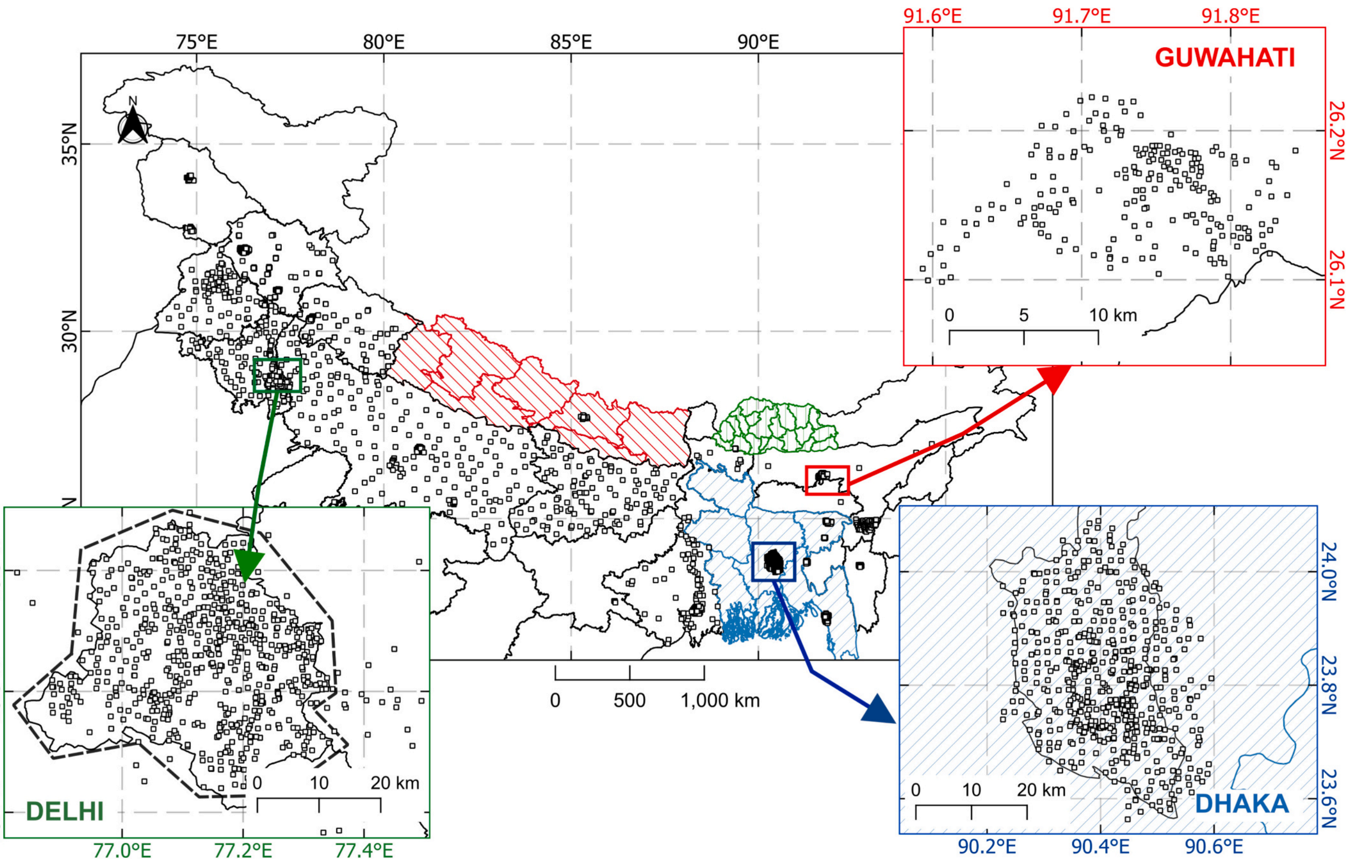


Fig. 3. Final Vs30 data points obtained for different locations in the Himalayan region using the geolocation data from reports, published articles and the metric georeferencing procedure. Inset maps show the spatial distribution of Vs30 data points in major cities of the region.

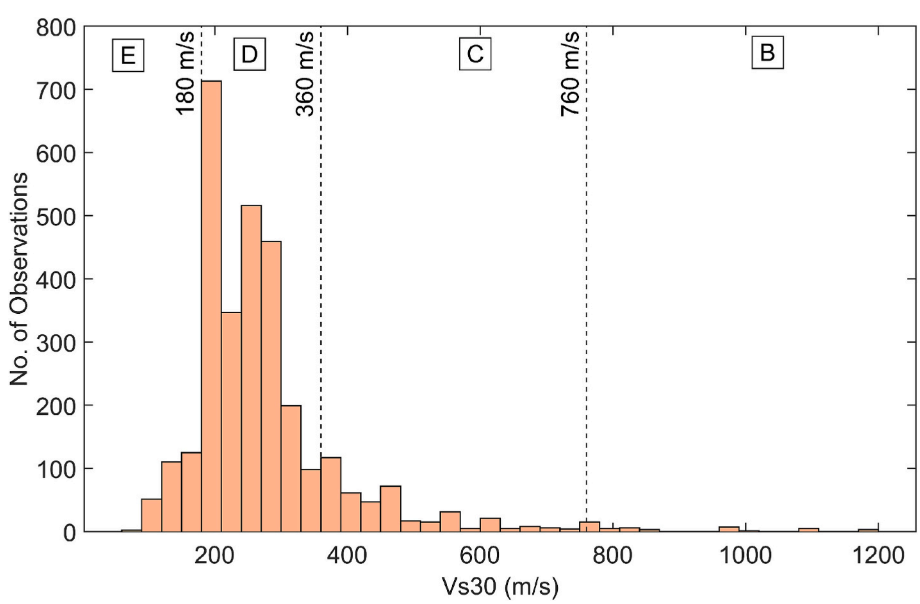


Fig. 4. Histogram of Vs30 datapoints with site class boundaries shown as vertical lines.

parameters (except geological information).

Step 4: Finally, geological information corresponding to each point in the vector layer (containing all other parameters information) is obtained by performing spatial joins in the GIS interface, which transfers the attribute of a base vector layer (here, geological data) to another layer.

This final extracted vector layer (with parameter information) with gridded points with known location coordinates was used to predict Vs30 values of their respective location.

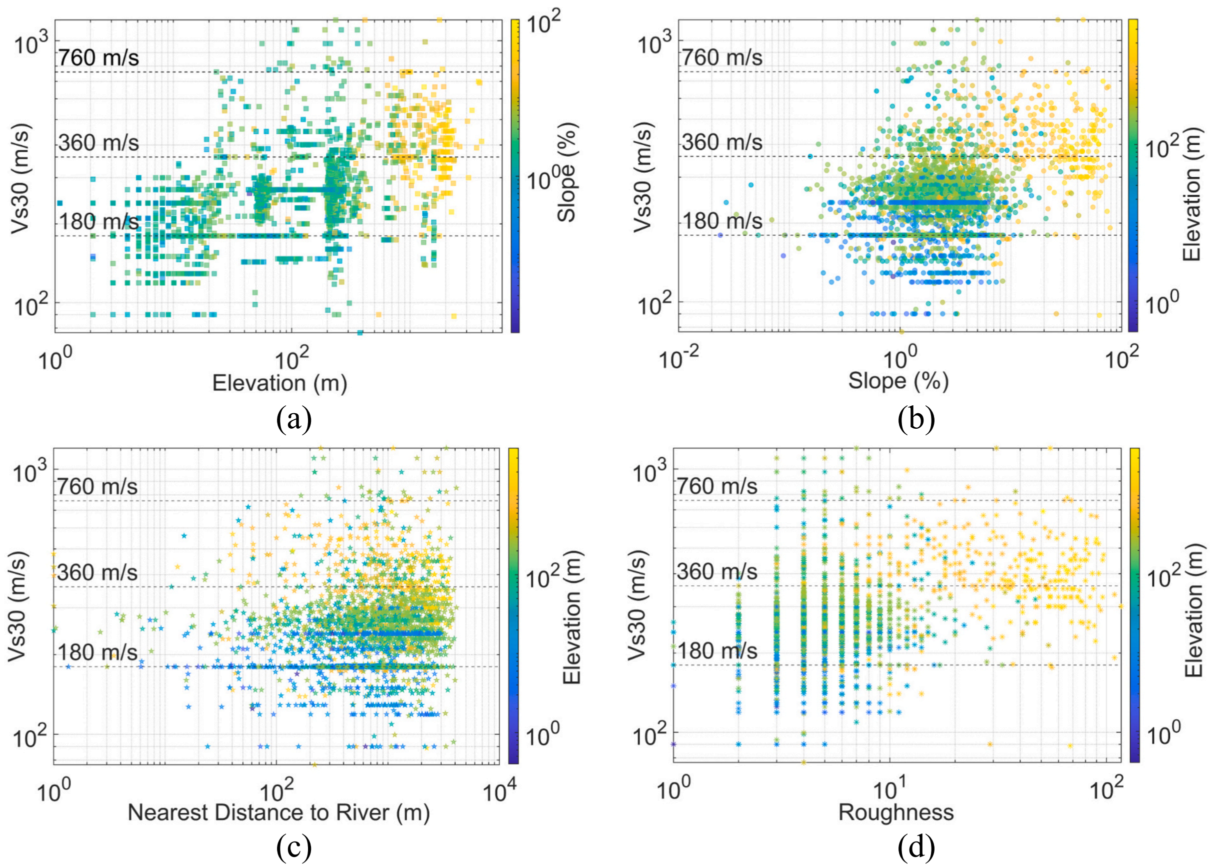


Fig. 5. Plots of Vs30 data points w.r.t. their (a) Elevation (in m), (b) Slope (in percentage), (c) Nearest Distance to River (in m), (d) Roughness values using SRTM-DEM. The colour ramp shows parameter values for the same locations.

Table 3
A list of all the parameters/predictors used for the Vs30 prediction.

Predictor (Units)	Abbrev.	Range
Elevation (m)	Elev	0–4014
Slope (%)	Slop	0–100.58
Geology	Geol	Categorical
Nearest distance to river (m)	Dist	0.0016–4106
Topographic Position Index	TPI	−18 to 15
Terrain Ruggedness Index	TRI	0–39
Topographic Wetness Index	TWI	8.54–17.51
Roughness	Rough	0–116
Profile Curvature	Prof_Curv	−0.0092 to 0.005683
Plan Curvature	Plan_Curv	−0.1288 to 0.19092
Vector Ruggedness Measure	VRM	0–0.0632
Morphometric Feature	MF	Categorical

4.3. Model selection, hyperparameter optimization and model stacking

As our dataset has only a few Vs30 data points corresponding to site classes B and A, we developed an ML model exclusively for predictions at sites with Vs30 values below 760 m/s (i.e., Site Classes below B). So, we removed the Vs30 data points corresponding to values greater than 760 m/s from our dataset, leading to 3022 data points from the initial value of 3077. For model training, Vs30 values and numerical regressors were transformed using BoxCox transformation in the collected dataset (Box and Cox, 1964). For geological information, which was in categorical form, the data was converted into dummy variables for the model development. Dummy variables were created while considering all the geological classes available for the Himalayan region from the map. The dataset was initially divided into training (Tr, 65 %), validation (Va, 20 %), and testing (Te, 15 %) sets. Then, different ML algorithms were trained together on the Tr dataset using 10-fold cross-validation, such as regression trees, support vector machines, Gaussian Process Regression (GPR) models (Rasmussen and Nickisch, 2010), ensemble techniques (boosting and bagging) (Breiman, 1996; Galar et al., 2011) and multi-layered Neural Networks. Va dataset is used to test the performance of these models.

In the next step, a Bayesian optimization algorithm is used for hyperparameter optimization for different models. Fifty trials were evaluated for each model during hyperparameter optimization while minimizing the initial model's root-mean-squared error (RMSE) values. Based on the prediction performance of the trained models on the test dataset (here Va), the top-performing three models with the best results based on accuracy measures such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and coefficient of determination (R^2) were opted for further hyperparameter optimization. Table 4 lists all the parameters used for a particular model's assessment. These top three performing models with optimized hyperparameters were used as base models for the stacked ensemble model. Fig. 8 shows the schematics of the first stage of the model development.

Model stacking is a common technique to improve prediction robustness and performance for unseen data. It works by combining the prediction results of different models. In this method, the best-performing models are usually stacked together (Vilalta and Drissi, 2002; Džeroski and Ženko, 2004). This technique relies on the fact that various ML techniques utilize different assumptions, number of model parameters, and tuning methods for model development. Their suitability for prediction may vary across different scenarios. The final model is an ensemble of all base models. The most common methods used for model stacking are bagging (Breiman, 1996), random forest (Ho, 1995), and boosting (Freund and Schapire, 1997). Fig. 9 shows the schematics of the procedure adopted for creating the Stacked Ensemble

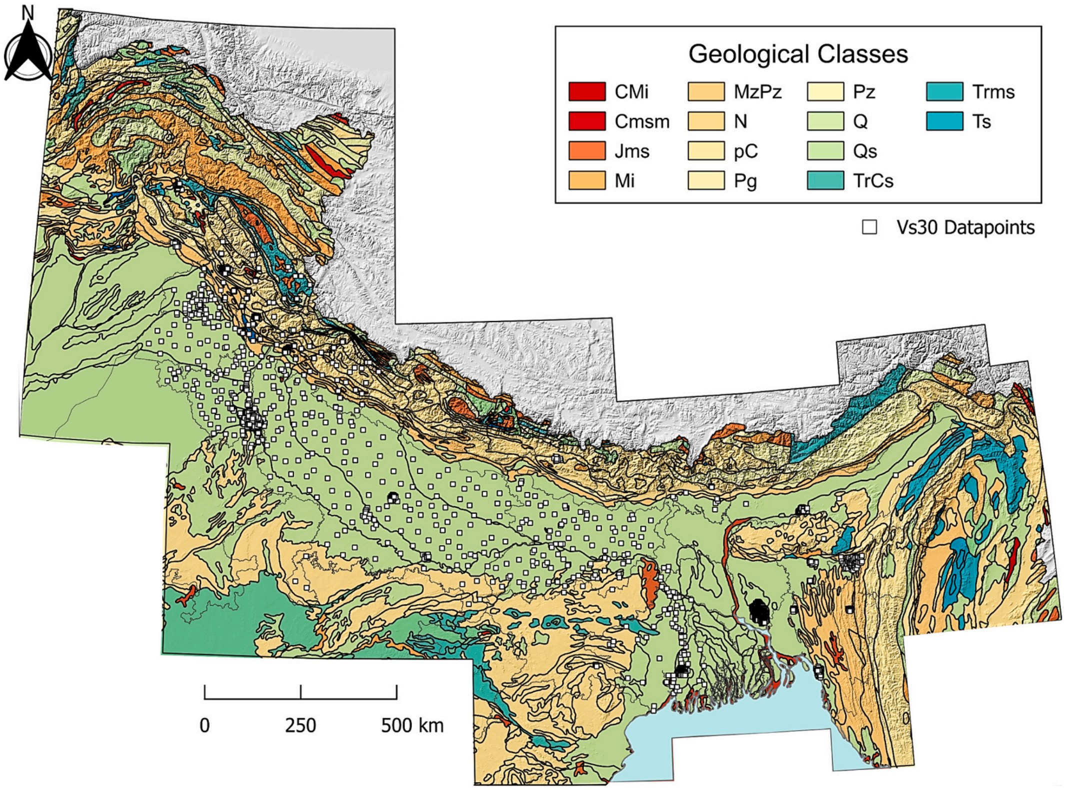


Fig. 6. Geological map considered for the present study (modified from Wandrey, 1998). The legend shows geological classes in which the Vs30 datapoints fall. Here CMi – Carboniferous Mesozoic rocks, Cmsm – Cambrian sedimentary and metamorphic rocks, Jms – Jurassic metamorphic and sedimentary rocks, Mi – Mesozoic igneous rocks, MzPz – Paleozoic and Mesozoic metamorphic rocks, N – Neogene sedimentary rocks, pC – undivided Precambrian rocks, Pg – Paleogene sedimentary rocks, Pz – Undifferentiated Paleozoic rocks, Q – Quaternary sediments, Qs – Quaternary sand and dunes, TrCs – Upper Carboniferious - Lower Triassic sedimentary rocks, Trms – Triassic metamorphic and sedimentary rocks, Ts – Tertiary sedimentary rocks.

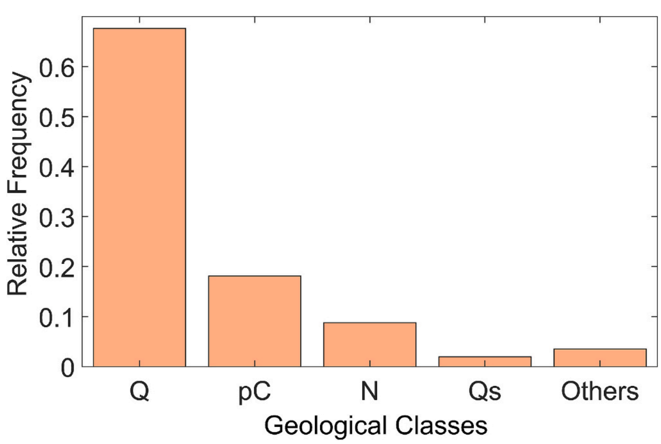


Fig. 7. Relative frequency of Vs30 data points with respect to geological classes. Here, Q – Quaternary sediments, pC – undivided Precambrian rocks, N – Neogene sedimentary rocks, Qs – Quaternary sand and dunes.

Model (SEM). In the present case, we have used an ensemble technique to stack three top-performing models. A 5-fold cross-validation strategy is used to train the stacked ensemble model. At this second stage of model development, predictions from base models on the Va dataset were utilized as training sets for the final stacked ensemble model. This final stacked model was tested using the Te dataset to evaluate its performance. A Bayesian optimization algorithm is used for hyper-parameter optimization of the ensemble.

The mean of absolute Shapley values (Lundberg and Lee, 2017) was used to understand the importance of different predictors on the model predictions. The Shapley value of an independent variable or feature for a data point explains the departure in corresponding predicted values from the average due to the variable, while its sign indicates the direction of the change. Shapley value of the i th feature for the query point x is calculated as (Lundberg and Lee, 2017):

Table 4
Accuracy measures are calculated and used to compare the different machine-learning models.

Model Performance Measure	Formula*
Root Mean Square Error (RMSE)	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}$
Mean Absolute Error (MAE)	$MAE = \sum_{i=1}^{n}\lvert y_i - \widehat{y}_i\rvert$
Coefficient of Determination (R^2)	$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$

* Notations: n – number of data points, y_i – observed value for i th datapoint, $\hat{y}$ – predicted value for i th datapoint, $\overline{y}$ - mean value of the observations.

$$\phi_i = \frac{1}{N}\sum_{S\subseteq F\setminus\{i\}}\left(\frac{(|S|!(N-|S|-1)!)}{(N-1)!}\right)\left[f_{S\cup\{i\}}\left(x_{S\cup\{i\}}\right) - f_S(x_S)\right] \quad (1)$$

here, N – number of all features or independent variables.

F – the set of all features or independent variables

S – a feature subset of F

$|S|$ - the cardinality (or number of elements) of the set S

$f_{S\cup\{i\}}$ – trained model with feature i present

f_S - trained model without feature i

x_S – Values of the input features in the set S

The mean value of the absolute Shapley values for a particular explanatory variable over a set of data points is calculated to give a

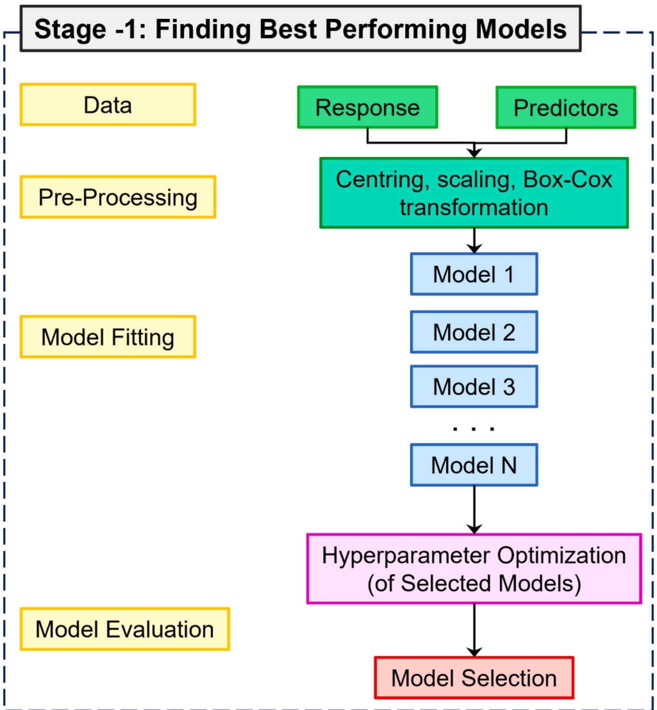


Fig. 8. A schematic for the first stage of the model building by selecting best-performing ML models.

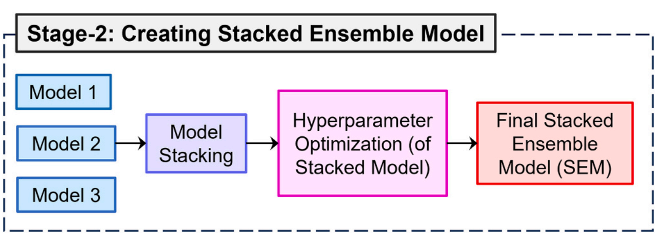


Fig. 9. Schematic diagram showing steps in creating the stacked ensemble model for Vs30 predictions.

magnitude of the variable's importance on the chosen model's predictions. Then predictor importance plots were prepared for all three models to understand the contribution of different predictors in the first stage and assess the impact of base models on the final SEM in the second stage.

4.4. Model modification and map preparation

Due to the lack of Vs30 datapoints corresponding to steeper slopes, prediction results from ML methods typically show better performance in site classes C, D, and E but underpredict the Vs30 values for site classes A and B (Geyin and Maurer, 2023). As discussed in earlier references (Matsuoka et al., 2005; Wald and Allen, 2007), steeper slope values are usually associated with site classes A and B (rocky terrains). Relying on observations from Allen and Wald (2009) (hereafter referred to as AW09) for higher slopes, we classify sites with slope values exceeding 14 degrees as site class B (Vs30 $>$ 760 m/s) in our map development.

For the Vs30 maps preparation at specific resolutions, values were resampled from base 30 m $\times$ 30 m SRTM using the nearest neighbour method for categorical data such as geological or morphometric features and bilinear interpolation for the continuous data values. In the nearest neighbour method, the value closest (in the input raster) to the cell on the output raster is assigned to it. In the bilinear interpolation technique, a weighted average of the four nearest input raster cells is assigned a new value to the output raster. In contrast, weights are assigned considering the distance between input cells and the centre of the cell in the output raster.

5. Results and discussion

Next, we present prediction results from best-performing models and the stacked ensemble model for validation and test sets. Then, we prepared Vs30 site-condition maps for the whole Himalayan region and a few sub-regions based on the Stacked Ensemble Model (SEM).

5.1. Stacked ensemble model's performance evaluation

Based on the prediction performance of the trained models (Table 5) on the test data, three models (two ensemble models - one boosted, another bagged, and a third GPR model with an exponential kernel function) with the best results were selected for further hyperparameter optimization. The Mean of Absolute Shapley values importance plot is shown in Fig. 10 for the three models, highlighting the importance ranking of different predictors on the model predictions. In Fig. 10, parameters other than those mentioned in Table 3 are geological categories. From these parameter importance plots, it can be observed that on average site Elevation, certain geological categories, terrain Slope value, Roughness, and Vector Ruggedness Measure (VRM) value of a site have the most influence on the prediction results.

The feature importance analysis in Fig. 10 reveals significant insights about the influence of geological data on Vs30 prediction across all three base models. Despite utilizing a relatively low-resolution geological map (Wandrey, 1998), geological classification emerges as one of the most influential predictor categories across all models, with several geological classes consistently ranking among the top predictors. Examining the mean absolute Shapley values more closely, we observe that certain geological classifications – particularly Quaternary sediments (Q), undivided Precambrian rocks (pC), Neogene sedimentary rocks (N), and Quaternary sand and dunes (Qs) – demonstrate consistently high importance rankings across all three models. This pattern correlates strongly with the distribution of our dataset, as shown in Fig. 7, where these same geological classes represent the most frequent classifications among our observation points. The high ranking of these geological classes relative to continuous variables like slope demonstrates geology's fundamental role in determining site response characteristics. While topographic parameters capture the surface expression of underlying geological conditions, the direct inclusion of geological classification provides critical information about site properties that cannot be inferred from topography alone. These findings underscore the value of integrating geological data into site classification models, even when available at relatively coarse resolution. The consistent importance of geological predictors across different modelling approaches suggests

Table 5
Results of different machine learning models' performance measures (Table 4) for validation (V) and test (T) sets. Here, GPR – Gaussian Process Regression, SVM – Support Vector Machines, and modifiers before these models are the type of their kernel function.

Model	RMSE (V)	R^2 (V)	MAE (V)	RMSE (T)	R^2 (T)	MAE (T)
Bagged Tree Ensemble	73.9	0.415	50.9	70.4	0.464	48.1
Boosted Tree Ensemble	74.5	0.406	50.4	70.2	0.466	46.6
Exponential GPR	78.7	0.337	55.8	76.5	0.367	52.6
Rational Quadratic GPR	79.0	0.333	55.9	76.9	0.360	53.0
Squared Exponential GPR	79.4	0.325	56.0	75.8	0.379	51.9
Matern 5/2 GPR	79.7	0.321	56.1	76.3	0.370	52.3
Medium Tree	80.5	0.306	55.6	76.3	0.370	53.4
Medium Gaussian SVM	80.6	0.304	54.8	77.3	0.353	51.3
Linear Regression	82.0	0.280	59.4	79.8	0.312	57.8
Fine Gaussian SVM	82.4	0.273	57.0	81.2	0.286	55.4

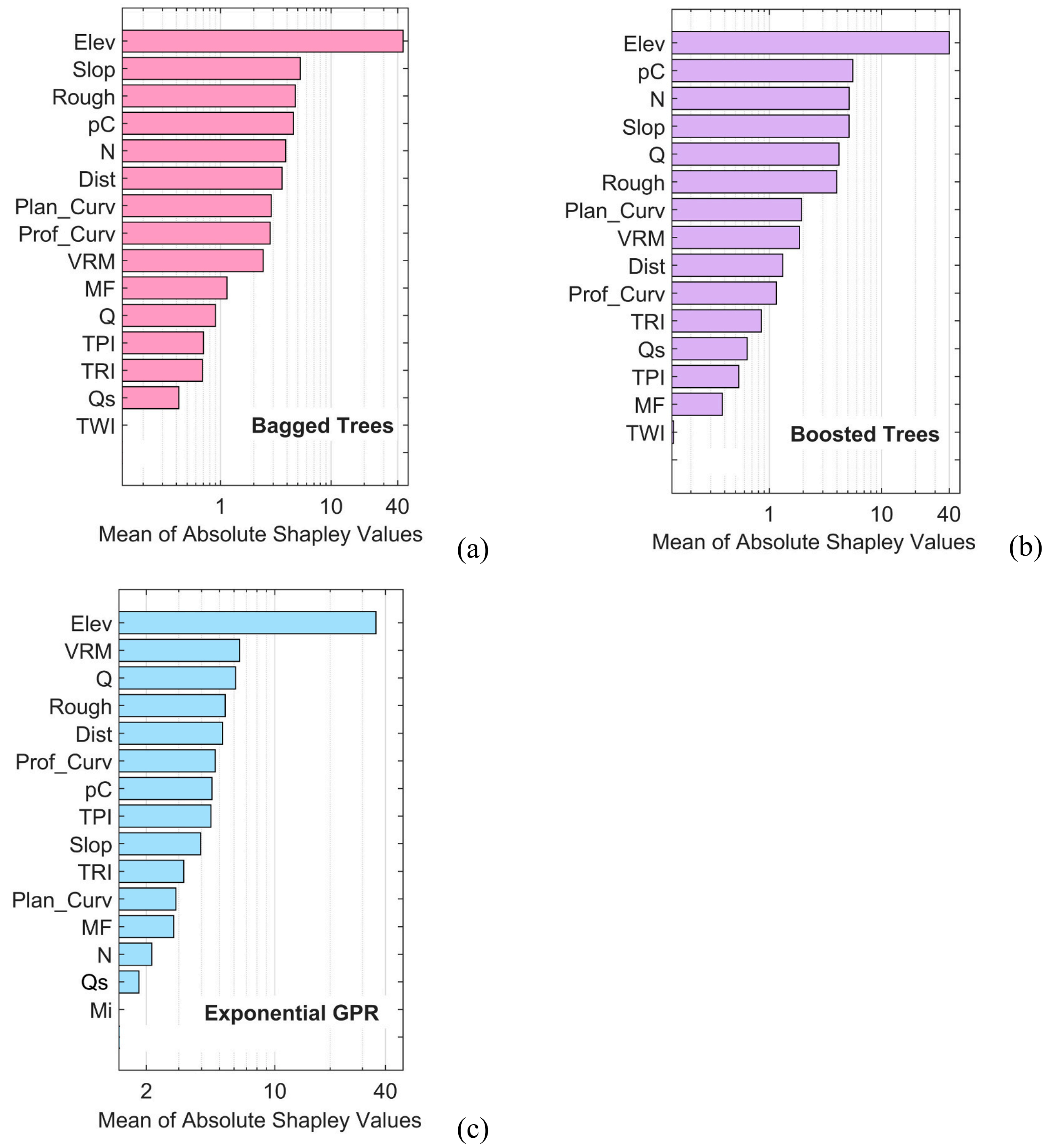


Fig. 10. Different Vs30 predictors' (Table 3) importance ranking based on the mean of absolute Shapley values (representing average impact on model output magnitude) for machine learning models - (a) Bagged trees, (b) Boosted trees, and (c) Exponential GPR, respectively. Parameters other than those mentioned in Table 3 are geological categories, as in Fig. 6.

that future site classification efforts could benefit from incorporating higher-resolution geological maps where available, potentially leading to further improvements in prediction accuracy.

Fig. 11 shows the prediction results for all three models. For all of them, the concentration of best prediction (less spread) is around the Vs30 values of 100 to 600 m/s. These models underpredict higher values of Vs30 corresponding to NEHRP site classes A and B. The Boosted Tree Ensemble model performs best on the test set with the lowest MAE (46.6) and highest R^2 (0.466), followed closely by the Bagged Tree Ensemble model. The Exponential GPR model shows the weakest performance of the three, with higher error metrics and lower R^2 values on both validation and test sets. All models perform better on the test set than the validation set, suggesting good generalization capability rather than overfitting. The relatively modest R^2 values (0.337–0.466) indicate the inherent complexity of predicting Vs30 values in the geologically diverse Himalayan region.

Fig. 12 shows the corresponding residual plots for the models. These residual plots show random scatter around the zero line. For higher values of predicted response, the scatter increases in the case of all three models. A comparison with the histogram plot in Fig. 4. shows that all models predict well around the Vs30 range, where most observation values lie. For the models to optimize their prediction, hyperparameters were tuned using a Bayesian optimization algorithm. In hyperparameter optimization, hyperparameters used for ensemble methods were the number of learning cycles, learn rate and minimum leaf size, and for the GPR model, these were sigma values of the model.

Fig. 13 presents the prediction results from the Stacked Ensemble Model (SEM) for both validation and test sets. The SEM demonstrates robust performance on the validation set, achieving an RMSE of 71.8 m/s and an R^2 of 0.445. When evaluated on the independent test dataset (Te), the final SEM, which was subsequently used for Vs30 prediction and map development, achieved an RMSE of 78.3 m/s and an R^2 of 0.38.

Fig. 14 illustrates the mean absolute Shapley values for the stacked model, revealing the relative influence of each base model on the final predictions. The boosted ensemble model exerts the strongest influence on the SEM predictions, followed by the bagged model and then the GPR

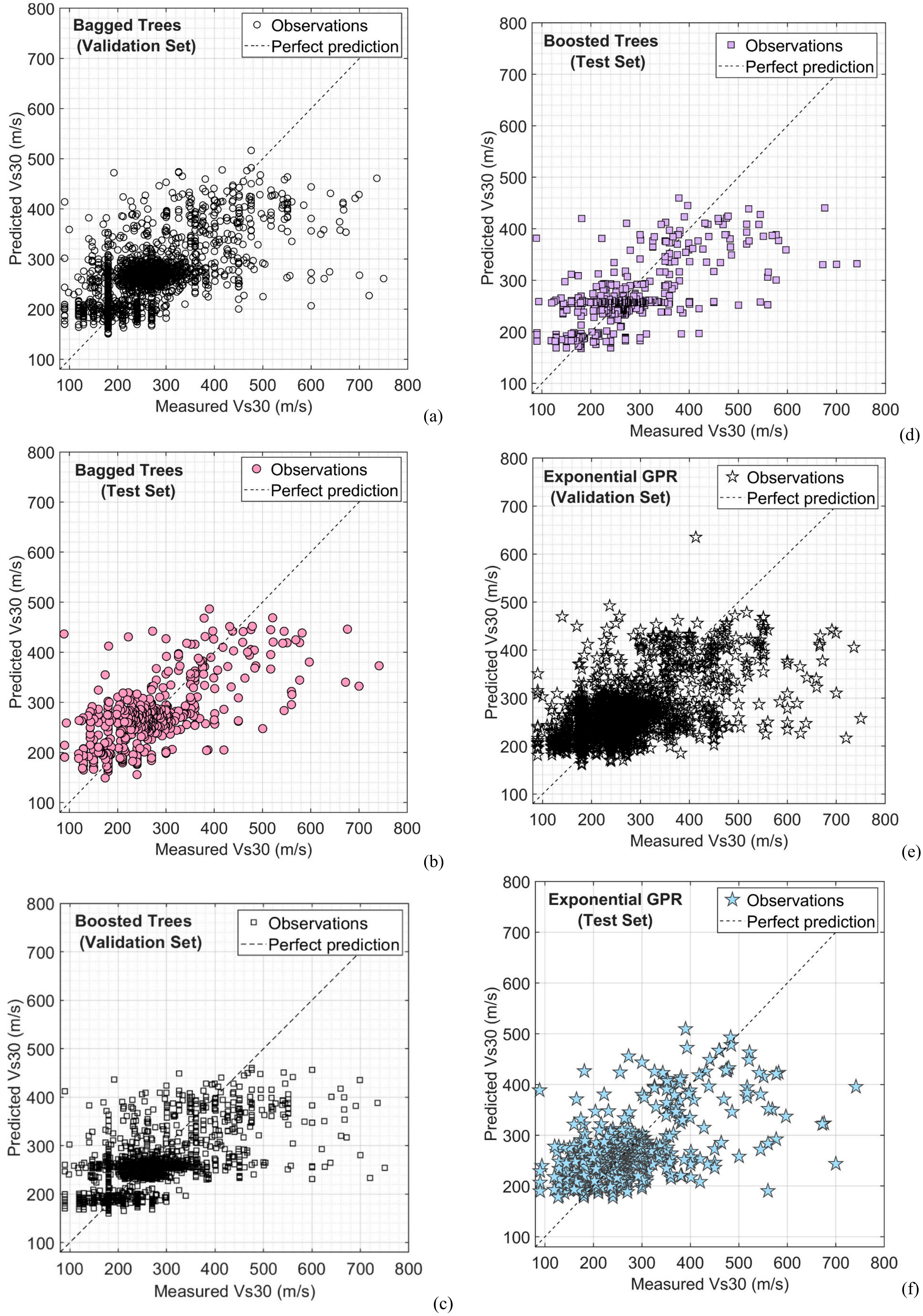


Fig. 11. Prediction results from best-performing ML models for validation and test sets. For validation set using (a) Bagged trees, (c) Boosted trees, and (e) Exponential GPR; and for test set using (b) Bagged trees, (d) Boosted trees, and (f) Exponential GPR, respectively.

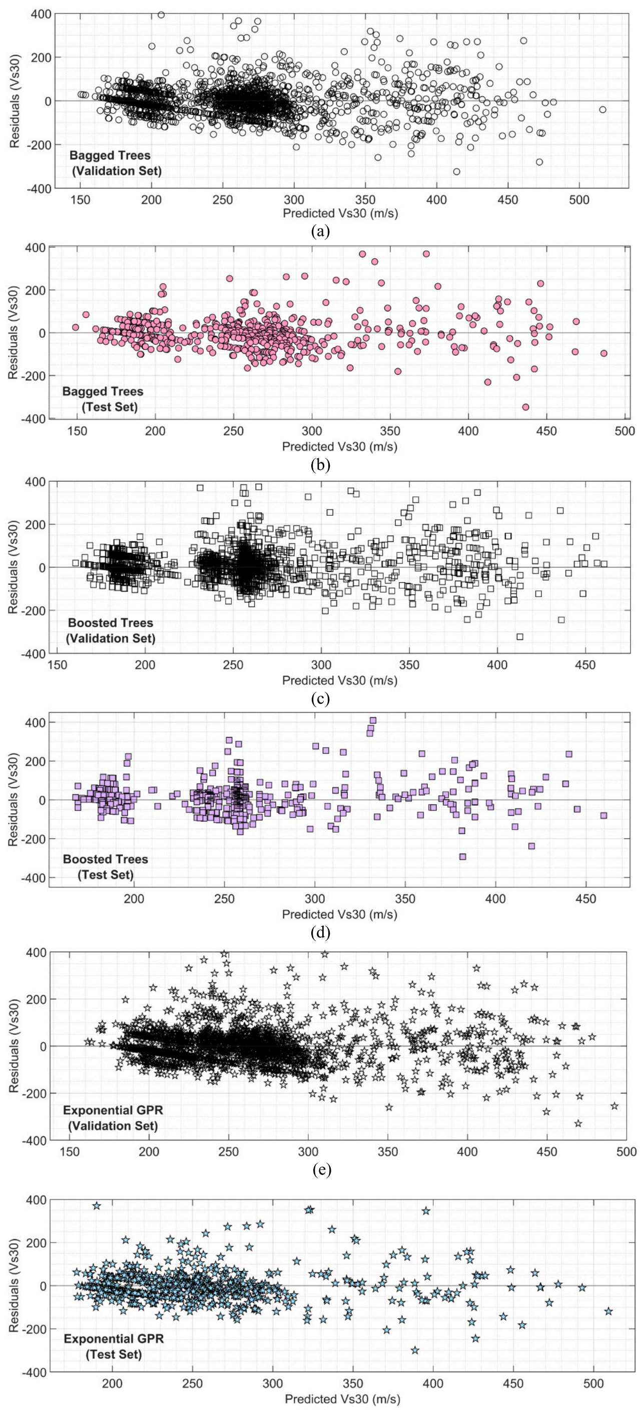


Fig. 12. Residuals from best-performing ML models for validation and test sets. For validation set using (a) Bagged trees, (c) Boosted trees, and (e) Exponential GPR; and for test set using (b) Bagged trees, (d) Boosted trees, and (f) Exponential GPR models, respectively.

model. This ranking pattern corresponds directly to the descending order of R^2 values observed for these individual models on the test set (Table 5).

The alignment between Shapley values and R^2 rankings can be attributed to two key factors. First, the meta-learner in the stacked ensemble model optimizes prediction accuracy by allocating greater weight to base models with superior predictive performance. Second, models with higher R^2 values typically demonstrate more consistent performance across different data subsets, making them more reliable

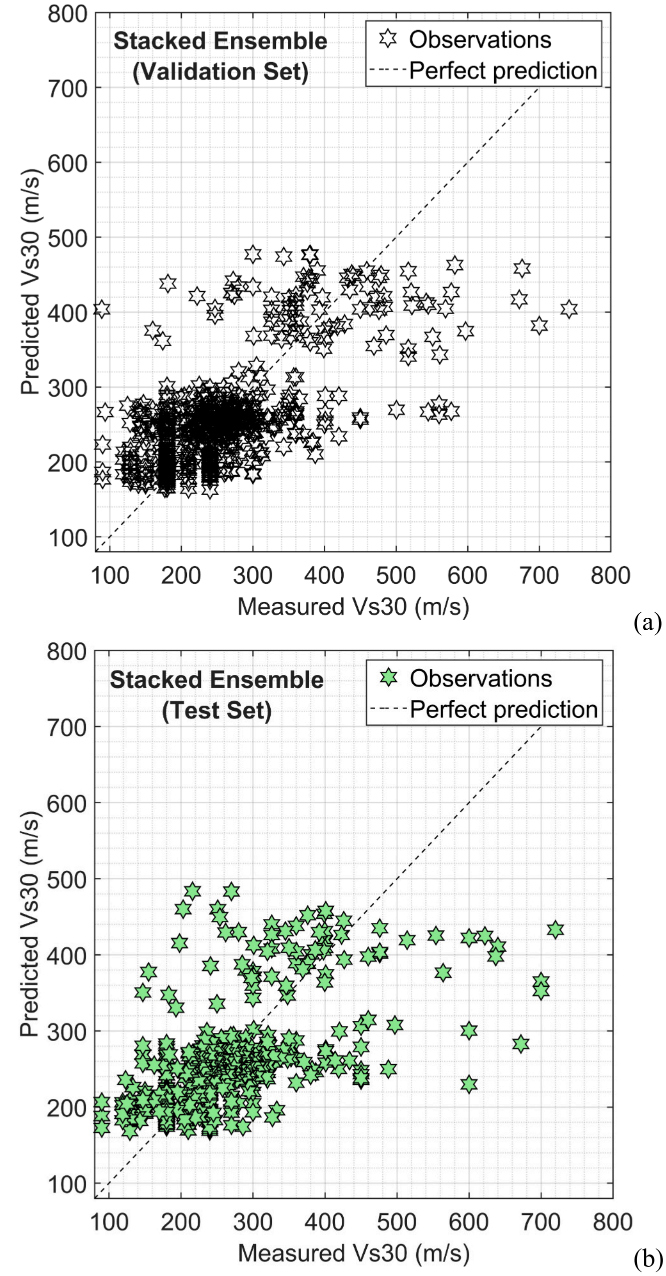


Fig. 13. Prediction results from the stacked ensemble model for - (a) Validation set (RMSE = 71.8 m/s, R^2 = 0.45, and MAE = 48.8 m/s), and (b) Test set (RMSE = 78.2 m/s, R^2 = 0.38, and MAE = 53.0 m/s).

contributors to the ensemble. Fig. 15(a) and (b) display the residual variations for the SEM model on the validation and test set observations, respectively.

5.1.1. *Comparison with Allen and Wald (2009) model*

Next, we compare the performance of the SEM to the Allen and Wald (2009) model. We have modified the predictions of the SEM model using the Allen and Wald (2009) criteria for the sites with slopes greater than 14 degrees. We made predictions using both models on the test dataset (Te) with 453 records for the comparison. Fig. 16 shows the comparison plots for the two models' performance using parameters MAE and RMSE. The SEM demonstrates substantial and consistent superiority over the AW09 model across all Vs30 categories, achieving an overall 64.6 % improvement in Mean Absolute Error and 52.6 % improvement in Root Mean Square Error across the test dataset (Te). The performance

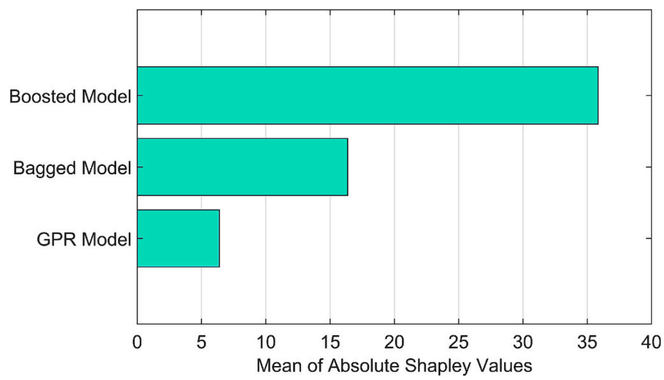


Fig. 14. Different machine learning models' importance ranking based on the mean of absolute Shapley values (representing average impact on model output magnitude) for the stacked ensemble model.

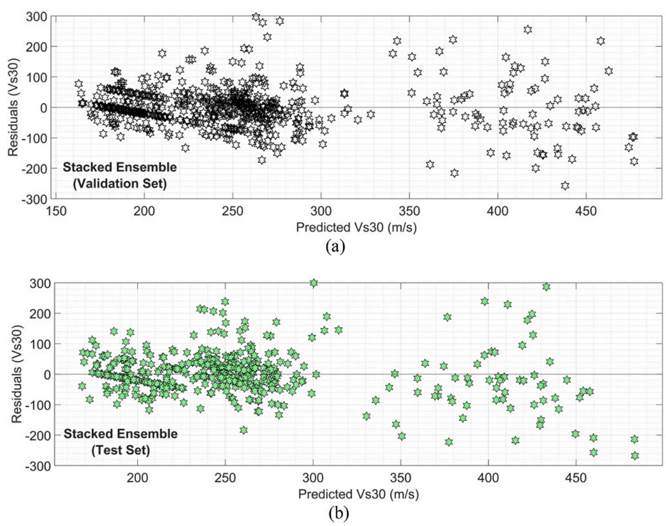


Fig. 15. Residuals from the stacked ensemble model for - (a) Validation set (RMSE = 71.8 m/s, R^2 = 0.445, and MAE = 48.8 m/s), and (b) Test set (RMSE = 78.3 m/s, R^2 = 0.38, and MAE = 53.0 m/s).

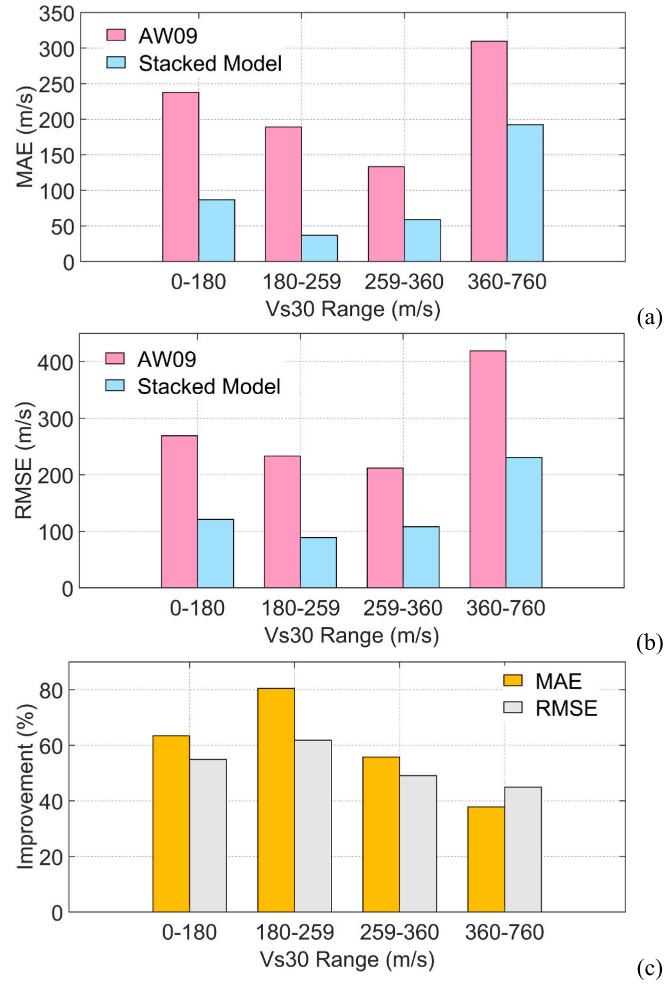


Fig. 16. Plots comparing the Stacked Model and Allen and Wald (2009) model's – (a) Mean Absolute Error (MAE), (b) Root Mean Squared Error (RMSE), (c) Performance Improvement (in percentage) on the test dataset (Te) across different Vs30 ranges (in m/s). For this analysis Stacked model's prediction has been modified using Allen and Wald's (2009) criteria for sites with slopes greater than 14 degrees.

advantages vary systematically across different Vs30 ranges, with the most pronounced improvements occurring in the 180 to 259 m/s category (Fig. 16a), which represents nearly half of all data points of the test dataset (Te) and encompasses typical urban soil conditions where the SEM achieves a remarkable 80.49 % error reduction (Fig. 16c). For very soft soils (0 to 180 m/s), the SEM reduces prediction errors by 63.44 % (Fig. 16a, c), demonstrating significant improvement in conditions where ground motion amplification effects are most critical while maintaining strong performance in medium-stiff conditions (259 to 360 m/s) with a 55.79 % error reduction across 130 sites (Fig. 16a, c). Even in the highest Vs30 category (360 to 760 m/s), where slope-based models typically perform better due to correlations between steep topography and rock outcrops, the SEM still achieves a 37.84 % improvement over AW09 (Fig. 16a, c).

Beyond average error reduction, the SEM shows improved error consistency through RMSE improvements that confirm enhanced performance without additional prediction volatility, making it particularly valuable for engineering applications requiring both accuracy and reliability (Fig. 16b, c). The comprehensive performance advantage across all Vs30 ranges, with the most significant improvements occurring in the ranges most relevant to urban seismic site characterization, positions the SEM as a better alternative to traditional slope-based Vs30 prediction methods for earthquake engineering design and seismic hazard assessment applications.

5.2. Site-classification maps of major cities

5.2.1. Delhi

The site classification map for Delhi (Fig. 17) shows that most of the urban area falls within NEHRP site class D, with smaller areas classified as C and E. This distribution is consistent with the geological setting of Delhi, which is primarily built on Quaternary alluvial deposits of the Indo-Gangetic Plains. The predicted values align well with the Vs30 contour map Satyam and Rao (2008) developed based on 117 MASW measurements, which reported values ranging from 230 to 350 m/s. However, the present map reveals finer spatial variability masked by the contour representation in previous studies. Compared to the NCS (2016) map, which broadly classified most of Delhi in the Vs30 range of 154 to 360 m/s, the present results provide more detailed differentiation within this range, which is critical for site-specific hazard assessment. A few regions on their map lie in the NEHRP site class C (Vs30 between 360 and 760 m/s), which is also the case in the present study.

5.2.2. Dhaka

The site classification map for Dhaka (Fig. 18) reveals a predominance of site classes D and E, reflecting the city's location on thick

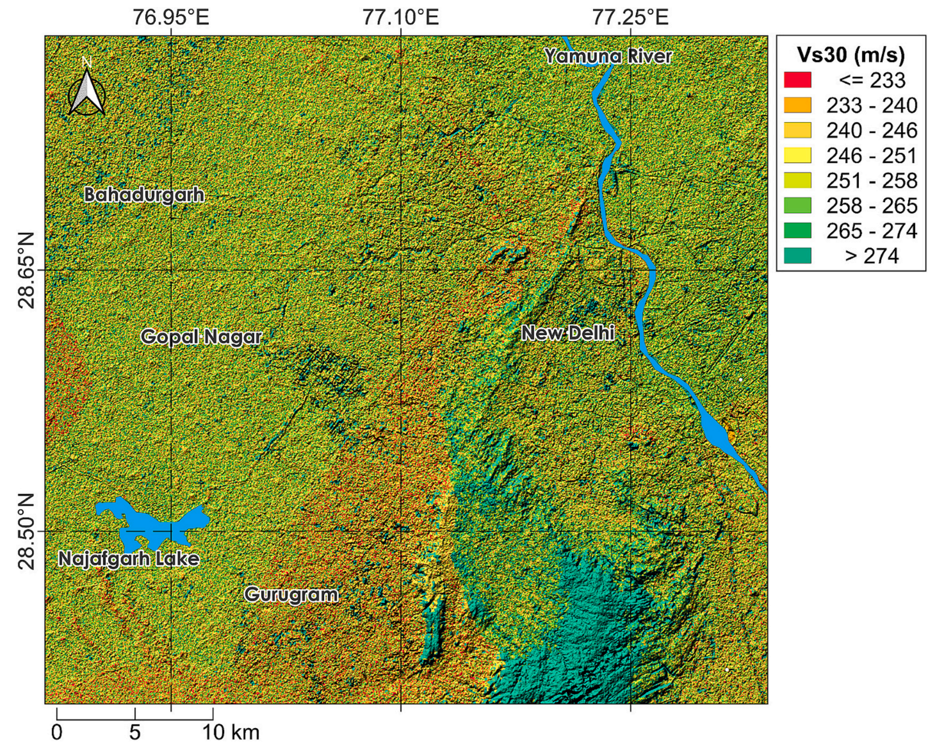


Fig. 17. Site class map of Delhi (India) region using a 50 m × 50 m grid based on the stacked ensemble model.

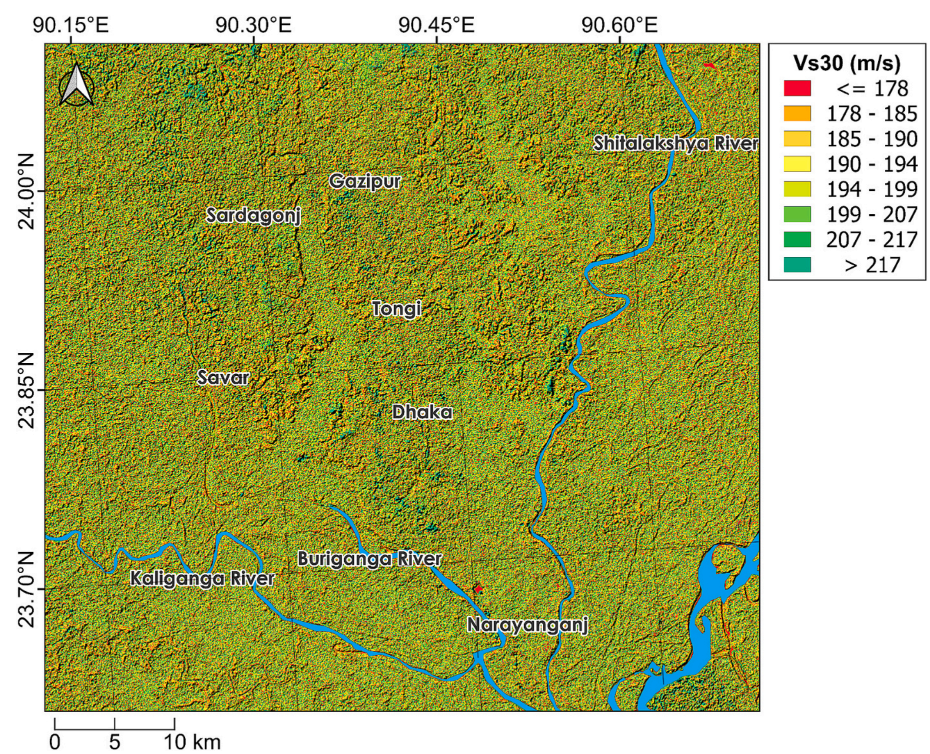
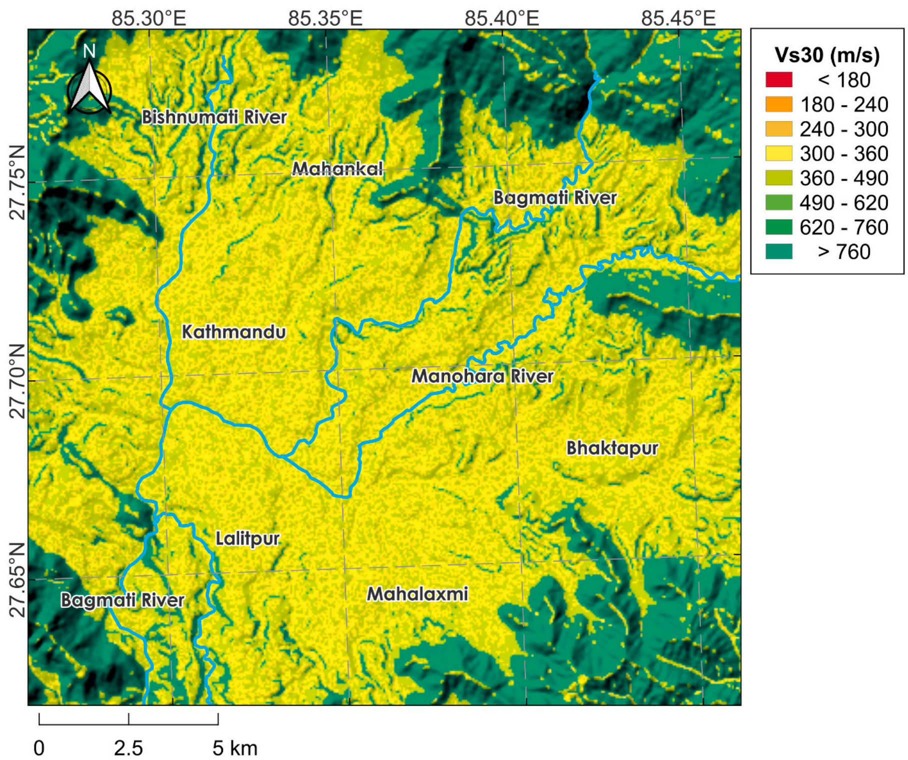


Fig. 18. Site class map of Dhaka (Bangladesh) region (50 m × 50 m raster grid) based on the modified stacked ensemble model.

Holocene deposits of the Bengal Basin. We developed a modified classification scheme subdividing the standard NEHRP classes to visualize spatial variations better. The predicted Vs30 values, ranging primarily between 145 and 260 m/s, agree with the map developed by Rahman et al. (2018a), who used Holocene soil thickness as a predictor. However, the present study's map covers a larger area and provides more detailed spatial resolution. The consistency between these independently developed maps validates our approach while offering improved spatial coverage.

5.2.3. Kathmandu

The Kathmandu Valley presents a more complex site classification pattern (Fig. 19a) due to its setting as an intermontane basin surrounded by steep mountains. This map differentiates rock sites (classes A and B) in the surrounding hills and soil sites (classes C and D) in the valley floor. To better visualize the variations within the valley, we created a second map (Fig. 19b) focusing only on on-site classes C and D. This reveals significant spatial heterogeneity in site conditions across the valley, with lower Vs30 values in the central and southern portions where lacustrine deposits are thickest. These prediction results are consistent with the Vs30 contour map developed by Gilder et al. (2022) using a kriging interpolation, with estimated values ranging from 0 to 900 m/s across the region. However, the current approach provides more detailed spatial variations than their contour representation, particularly within the valley where rapid lateral changes in subsurface conditions occur.



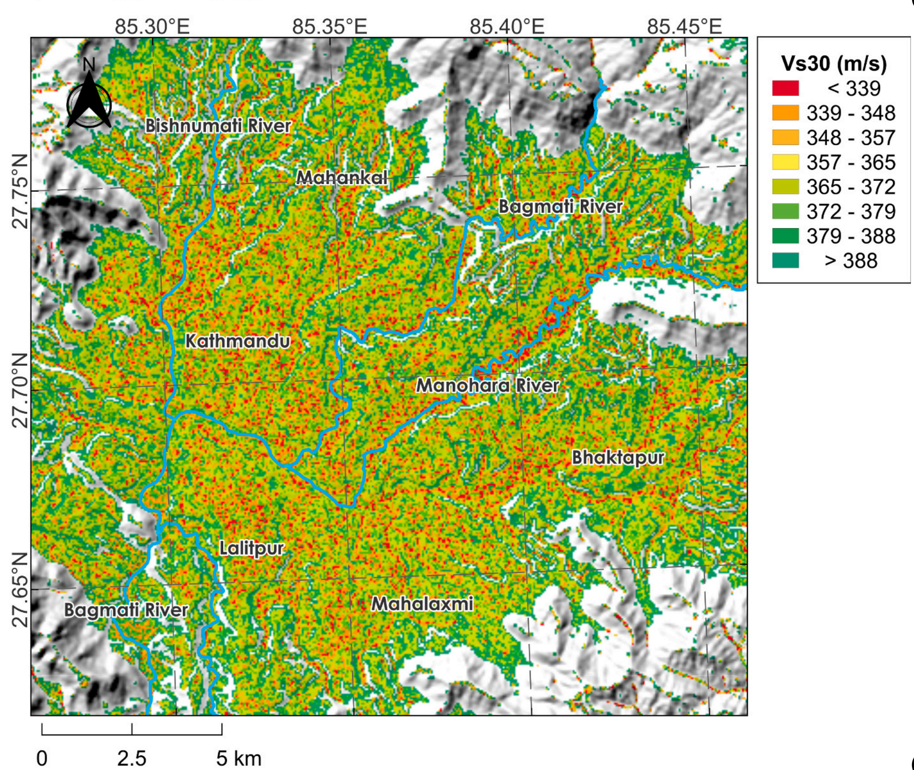


Fig. 19. Site class maps of Kathmandu (Nepal) based on a) NEHRP criteria, b) histogram-based (while discarding site classes A and B) using Vs30 values predicted by the modified stacked ensemble model (50 m × 50 m raster grid).

5.2.4. Guwahati

The site classification map for Guwahati (Fig. 20a) shows a clear distinction between rock sites (classes A and B) in the hills and soil sites (classes C and D) in the Brahmaputra River valley. Like Kathmandu, we created a second map (Fig. 20b) focusing only on the valley area better to visualize the variations in site classes C and D. The predicted Vs30 values in the valley range primarily from 200 to 360 m/s, which aligns well with the contour map developed by DST (2007) and subsequently used by Sharma and Rahman (2016). However, Fig. 20b map reveals more detailed spatial patterns not captured by the smooth contours in previous studies, reflecting the influence of local geological and geomorphological features on on-site conditions.

5.3. Site classification map for the Himalayan region

The regional site classification map for the entire Himalayan region (Fig. 21) at 1 km × 1 km resolution reveals the broad patterns of site conditions across this complex terrain. The map shows a clear correlation between physiographic provinces and site classes. The High Himalayan Zone is predominantly classified as site classes A and B, reflecting the presence of crystalline rocks and steep slopes. These areas typically exhibit Vs30 values exceeding 760 m/s. The Middle Himalayan Zone and Siwalik Hills display a complex pattern of site classes B and C, with occasional patches of class D in valley bottoms. This variability reflects the heterogeneous lithology and complex topography of these regions. The Indo-Gangetic Plains are classified as site classes C and D, with Vs30 values generally decreasing from north to south as sediment thickness increases from the mountainfront. The deltaic regions of

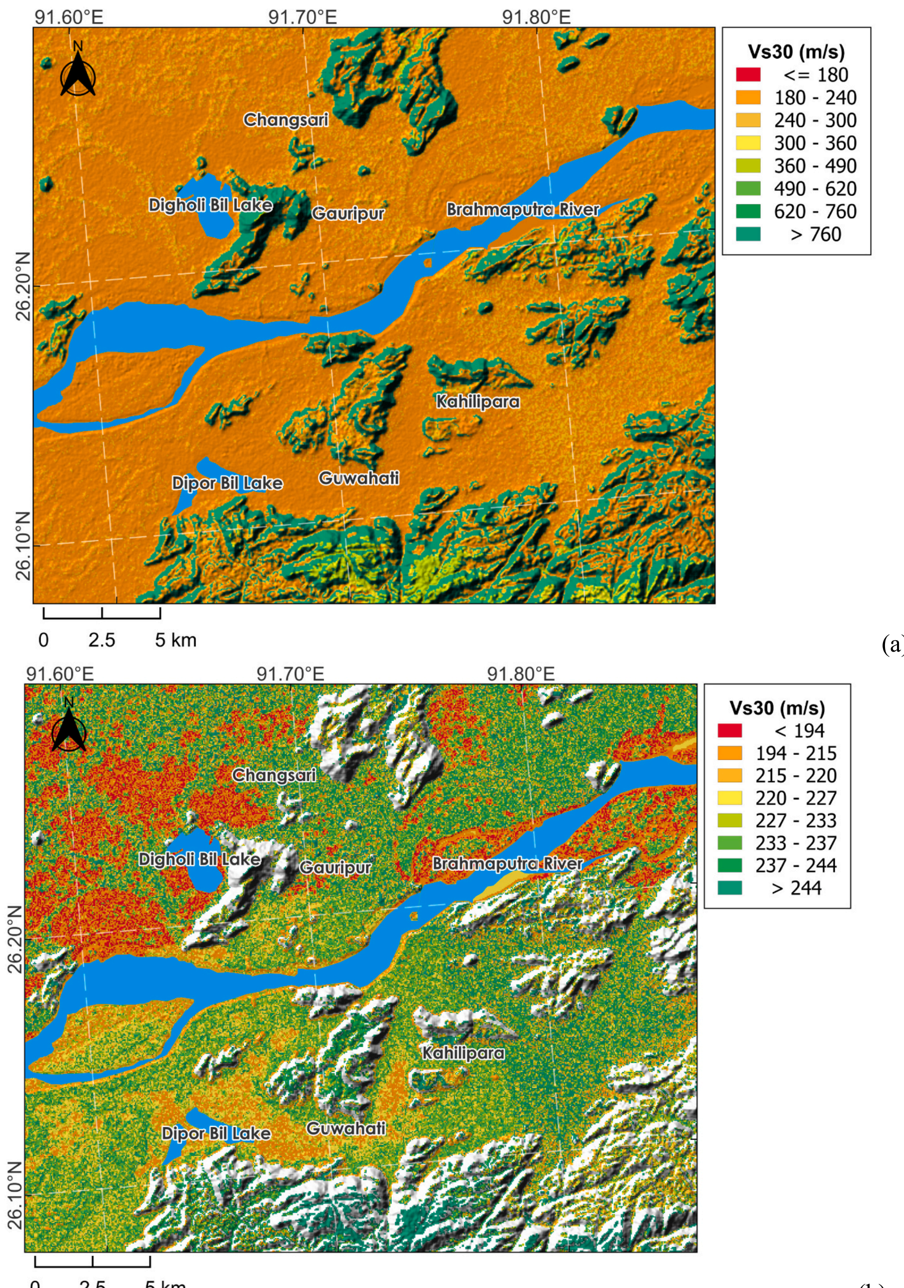


Fig. 20. Site class maps of Guwahati region (India) based on a) NEHRP criteria, b) histogram-based (discarding site class A and B) using Vs30 values predicted by the modified stacked ensemble model (50 m × 50 m raster grid).

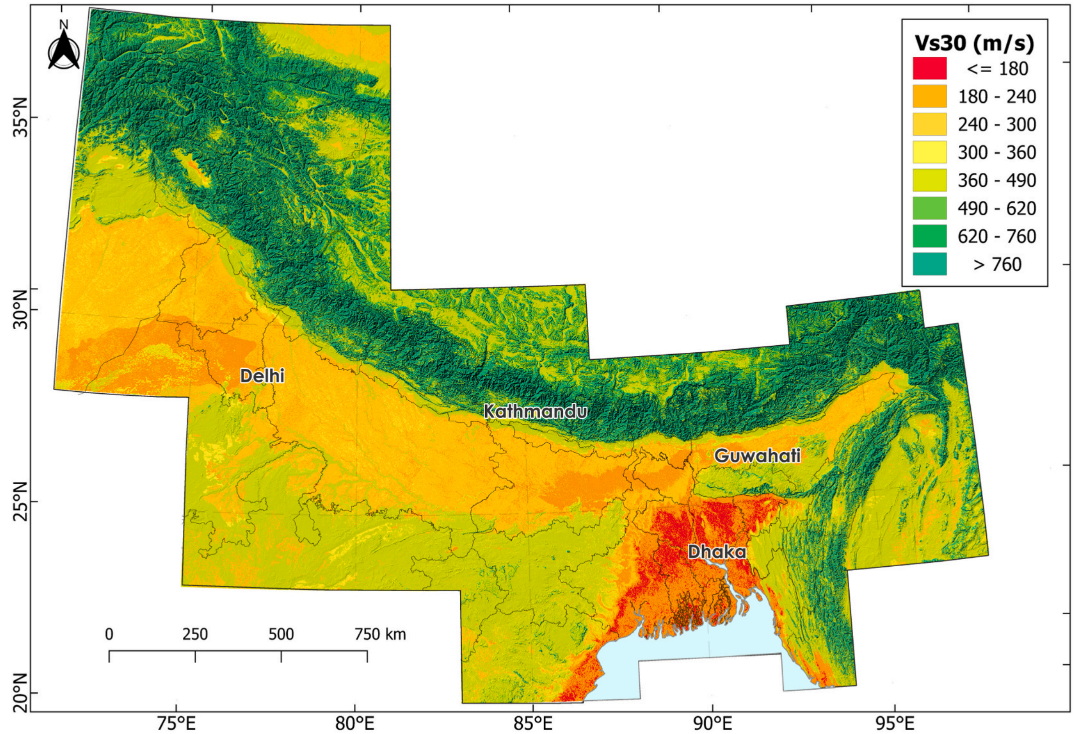


Fig. 21. Site classification map of the Himalayan region using stacked ensemble model-based predictions shown using a 1 km × 1 km raster grid.

Bangladesh and eastern India predominantly fall within site classes D and E, consistent with the presence of thick, soft sediments in these areas.

This regional classification represents a significant improvement over previous efforts by Sitharam et al. (2015), who used a slope-based approach at 10 km × 10 km resolution. While their map classified the Himalayan region into broad zones of site classes B, C, and D, Fig. 21 higher-resolution map reveals substantial local variations crucial for regional hazard assessment. Furthermore, the present study's integrated approach incorporating multiple predictors provides a more robust classification than single-parameter methods.

5.4. Implications of geological data resolution and model limitations

Using low-resolution geological maps (1:5,000,000 scale) represents a compromise between regional coverage and local detail. This limitation has following consequences for the model development and resulting site classification maps: (i) The coarse geological classification system (14 classes) may not capture important lithological variations that influence site response, particularly in areas with complex geology, such as intermontane valleys and tectonically active zones; (ii) Geological boundaries in the mapped data are necessarily generalized, potentially leading to misclassification of sites near these boundaries; and (iii) the inability to represent smaller geological features such as local fault zones, intrusions, or thin sedimentary layers may result in overlooking significant local amplification effects.

Despite these limitations, the current approach offers several advantages over previous studies. Integrating geological data, even at low resolution, with multiple geomorphological parameters provides a more comprehensive characterization of site conditions than purely topography-based methods. The machine learning framework enables identifying and weighing the most significant geological factors, maximizing the utility of available data. Furthermore, the consistent geological classification across the entire Himalayan region allows for a direct comparison of site conditions between different areas, facilitating regional hazard assessment.

Future work should explore integrating higher-resolution geological maps where available, potentially implementing a hierarchical approach that uses detailed geology for urban centres and generalized geology for remote areas. Developing geologically informed transfer learning techniques could also help address data gaps by adapting models trained in data-rich areas to geologically similar but data-poor regions, thereby improving prediction accuracy while maintaining the benefits of regional consistency.

6. Summary and conclusion

The seismic response of complex terrains such as the Himalayan region is influenced by a combination of geological, geomorphological, and geotechnical properties. However, most existing site classification approaches rely on a single parameter, limiting their effectiveness in capturing the region's complexity. This study presents a novel approach that integrates multiple proxies to develop comprehensive Vs30-based site classification maps for the Himalayan region using a modified stacked ensemble model.

We developed our approach through three key steps: (i) the creation of a comprehensive Vs30 database by combining direct measurements with data extracted through metric georeferencing of published maps; (ii) the identification and integration of twelve geological and geomorphological predictors; and (iii) development of a stacked ensemble model that combines the strengths of multiple ML algorithms, resulting in a final robust model. Our methodology addresses the challenge of underprediction for steeper slope sites by incorporating the AW09 criteria for slopes exceeding 14 degrees.

The final stacked ensemble model used for the Vs30 prediction and map development has RMSE values of 78.3 m/s and R^2 of 0.38 on the test dataset. The stacked ensemble model reveals significant spatial variability in Vs30 across the Himalayan region (Fig. 21). High Vs30 values (> 760 m/s, site classes A and B) predominate in mountainous areas with steep slopes and exposed bedrock. Moderate Vs30 values (from 360 to 760 m/s, site class C) characterize the middle mountains and foothills. Low Vs30 values (from 180 to 360 m/s, site class D) are prevalent in valley bottoms, river terraces, and the upper Indo-Gangetic Plains. Very low Vs30 values (< 180 m/s, site class E) are found primarily in deltaic regions and areas with thick, soft sediments.

The high-resolution urban-scale maps for Delhi, Dhaka, Kathmandu, and Guwahati capture local variations crucial for site-specific hazard assessment. The resulting site classification maps reveal significant spatial heterogeneity across the Himalayan region. Mountain areas are predominantly classified as site classes A and B, foothills as C and D, and deltaic regions as D and E. These patterns align with the known geological and geomorphological framework of the region. The high spatial resolution of our maps reveals detailed variations that coarser regional maps or interpolation-based approaches would miss.

This study advances seismic site classification methodology by demonstrating how integrating multiple data sources and machine learning techniques can overcome the limitations of previous approaches. The resulting maps provide essential information for earthquake hazard assessment, infrastructure planning, and building code implementation in this seismically vulnerable region. Future work should focus on expanding the dataset, particularly for steeper slopes, incorporating higher-resolution DEM models, and developing purely region-specific models as more data becomes available. This approach is particularly valuable for mountainous regions like the Himalayas, where comprehensive geotechnical data remains limited.

Authors contribution

Both authors contributed to the conceptualization of the study. HT contributed to the data collection, data analysis, model development, map preparation, writing, reviewing and editing of the manuscript. AP contributed to supervising, reviewing, and editing the manuscript.

CRediT authorship contribution statement

Harish Thakur: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **P. Anbazhagan:** Writing – review & editing, Supervision, Project administration, Conceptualization.

Declaration of competing interest

The authors declare that they have no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank M/s Secon Pvt. Ltd. for funding the project “Effect of shear wave velocity calibration on Amplification of shallow and deep soil sites” grant No: SECON/IISc/MoES/WO/07-18/0079.

Data availability

Data will be made available on a reasonable request.

References

Abbasnejadfard, M., Bastami, M., Jafari, M.K., Azadi, A., 2023. Spatial correlation models of VS30 values: a case study of the Tehran region. Eng. Geol. 325, 107300. https://doi.org/10.1016/j.enggeo.2023.107300.

Akin, M.K., Kramer, S.L., Topal, T., 2011. Empirical correlations of shear wave velocity (Vs) and penetration resistance (SPT-N) for different soils in an earthquake-prone area (Erbaa-Turkey). Eng. Geol. 119 (1–2), 1–17. https://doi.org/10.1016/j.enggeo.2011.01.007.

Allen, T.I., Wald, D.J., 2009. On the use of high-resolution topographic data as a proxy for seismic site conditions (VS 30). Bull. Seismol. Soc. Am. 99 (2A), 935–943. https://doi.org/10.1785/0120080255.

Anbazhagan, P., Kumar, A., Sitharam, T.G., 2013. Seismic site classification and correlation between standard penetration test N value and shear wave velocity for Lucknow City in Indo-Gangetic Basin. Pure Appl. Geophys. 170, 299–318. https://doi.org/10.1007/s00024-012-0525-1.

Anbazhagan, P., Srilakshmi, K.N., Bajaj, K., Moustafa, S.S., Al-Arifi, N.S., 2019. Determination of seismic site classification of seismic recording stations in the Himalayan region using HVSR method. Soil Dyn. Earthq. Eng. 116, 304–316. https://doi.org/10.1016/j.soildyn.2018.10.023.

Ansary, M., Tasmiah, A., Al Noman, M., 2023. Development of empirical correlations between shear wave velocity and in situ penetration test results for different types of soils in DMDP area, Bangladesh. Arab. J. Geosci. 16 (4), 280. https://doi.org/10.1007/s12517-023-11342-z.

Bajaj, K., Anbazhagan, P., 2019. Seismic site classification and correlation between VS and SPT-N for deep soil sites in Indo-Gangetic Basin. J. Appl. Geophys. 163, 55–72. https://doi.org/10.1016/j.jappgeo.2019.02.011.

Bandyopadhyay, S., Sengupta, A., Reddy, G., 2019. Development of correlation between SPT-N value and shear wave velocity and estimation of non-linear seismic site effects for soft deposits in Kolkata city. Geomech. Geoeng. 16 (1), 1–19. https://doi.org/10.1080/17486025.2019.1640898.

Bhutani, M., Naval, S., 2020. Assessment of seismic site response and liquefaction potential for some sites using borelog data. Civ. Eng. J. 6, 2103–2119. https://doi.org/10.28991/cej-2020-03091605.

Box, G.E., Cox, D.R., 1964. An analysis of transformations. J. R. Stat. Soc. Ser. B Stat Methodol. 26 (2), 211–243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x.

Brando, G., Pagliaroli, A., Cocco, G., Di Buccio, F., 2020. Site effects and damage scenarios: the case study of two historic centers following the 2016 Central Italy earthquake. Eng. Geol. 272, 105647. https://doi.org/10.1016/j.enggeo.2020.105647.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24 (2), 123–140. https://doi.org/10.1007/BF00058655.

BSSC, 2015. NEHRP Recommended Seismic Provisions for New Buildings and Other Structures. Washington, D.C.

Building Seismic Safety Council (BSSC), NEHRP Recommended Provisions for Seismic Regulations for New Buildings and Other Structures (FEMA 450), 2004. Building Seismic Safety Council, National Institute of Building Sciences, Washington, DC, USA.

Chatterjee, K., Choudhury, D., 2013. Variations in shear wave velocity and soil site class in Kolkata city using regression and sensitivity analysis. Nat. Hazards 69, 2057–2082. https://doi.org/10.1007/s11069-013-0795-7.

Crespo, M.J., Benjumea, B., Moratalla, J.M., Lacoma, L., Macau, A., González, Á., Stafford, P.J., 2022. A proxy-based model for estimating VS30 in the Iberian Peninsula. Soil Dyn. Earthq. Eng. 155, 107165. https://doi.org/10.1016/j.soildyn.2022.107165.

De Risi, R., De Luca, F., Gilder, C., Pokhrel, R., Vardanega, P., 2021. The SAFER geodatabase for the Kathmandu valley: Bayesian kriging for data-scarce regions. Earthquake Spectra 37 (2), 1108–1126. https://doi.org/10.1177/8755293020970977.

Di Martino, A., Sgattoni, G., Purri, F., Amorosi, A., 2024. Seismic amplification of late Quaternary paleovalley systems: 2D seismic response analysis of the Pescara paleovalley (Central Italy). Eng. Geol. 341, 107697. https://doi.org/10.1016/j.enggeo.2024.107697.

DST, 2007. Seismic Microzonation Atlas of Guwahati Region. Govt. of India Publication.

Džeroski, S., Ženko, B., 2004. Is combining classifiers with stacking better than selecting the best one? Mach. Learn. 54, 255–273. https://doi.org/10.1023/B:MACH.0000015881.36452.6e.

Falcone, G., Acunzo, G., Mendicelli, A., Mori, F., Naso, G., Peronace, E., Moscatelli, M., 2021. Seismic amplification maps of Italy based on site-specific microzonation dataset and one-dimensional numerical approach. Eng. Geol. 289, 106170. https://doi.org/10.1016/j.enggeo.2021.106170.

Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of online learning and an application to boosting. J. Comput. Syst. Sci. 55 (1), 119–139. https://doi.org/10.1006/jcss.1997.1504.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F., 2011. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. 42 (4), 463–484. https://doi.org/10.1109/TSMCC.2011.2161285.

Gangopadhyay, S., 2013. Engineering Geology. Oxford University Press.

Geyin, M., Maurer, B.W., 2023. U.S. national models and maps informed by remote sensing and machine learning. Seismol. Res. Lett. 94 (3), 1467–1477. https://doi.org/10.1785/0220220181.

Gilder, C.E., De Risi, R., De Luca, F., Mohan Pokhrel, R., Vardanega, P.J., 2022. Geostatistical framework for estimation of VS 30 in data-scarce regions. Bull. Seismol. Soc. Am. 112 (6), 2981–3000. https://doi.org/10.1785/0120210266.

Harinarayan, N., Kumar, A., 2018. Determination of NEHRP site class of seismic recording stations in the Northwest Himalayas and its adjoining area using HVSR method. Pure Appl. Geophys. 175, 89–107. https://doi.org/10.1007/s00024-017-1696-6.

Heath, D.C., Wald, D.J., Worden, C.B., Thompson, E.M., Smoczyk, G.M., 2020. A global hybrid VS 30 map with a topographic slope–based default and regional map insets. Earthquake Spectra 36 (3), 1570–1584. https://doi.org/10.1177/8755293020911137.

Ho, T.K., 1995. Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition, 1. IEEE, pp. 278–282. https://doi.org/10.1109/ICDAR.1995.598994.

Imam, A., Sharma, K., Kumar, V., 2023. Seismic site characterization using multichannel analysis of surface waves in the Singhbhum region of Jharkhand, India: a case study. Acta Geod. Geophys. 58 (2), 217–239. https://doi.org/10.1007/s40328-023-00408-9.

Jenness, J., 2006. Topographic Position Index (TPI), vol. v. 1.2. Jenness Enterprises, Flagstaff, AZ.
Kandpal, G., John, B., Joshi, K., 2009. Geotechnical studies in relation to seismic microzonation of Union Territory of Chandigarh. J. Indian Geophys. Union 13 (2), 75–83.
Kayal, J.R., 2008. Microearthquake Seismology and Seismotectonics of South Asia. Springer Science & Business Media. https://doi.org/10.1007/978-1-4020-8180-4.
Kim, H.S., Sun, C.G., Lee, M.G., Cho, H.I., 2021. Multivariate geotechnical zonation of seismic site effects with clustering-blended model for a city area, South Korea. Eng. Geol. 294, 106365. https://doi.org/10.1016/j.enggeo.2021.106365.
Kramer, S.L., Stewart, J.P., 2024. Geotechnical Earthquake Engineering. CRC Press. https://doi.org/10.1201/9781003512011.
Kuldeep Shekar, B., Mohan, G., Singh, S., 2022. Assessment of seismic vulnerability using the ambient noise recordings in Cachar Fold Belt, Assam. J. Geol. Soc. India 98 (6), 795–804. https://doi.org/10.1007/s12594-022-2070-6.
Kumar, P., Kumar, S., 2023. Development of synthetic acceleration time histories for seismic ground response studies for site classes C to E for Bihar region: a case study. Innov. Infrastruct. Solut. 8 (11), 292. https://doi.org/10.1007/s41062-023-01265-9.
Lehner, B., Grill, G., 2013. Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. Hydrol. Process. 27 (15), 2171–2186. https://doi.org/10.1002/hyp.9740.
Li, X., Jing, B., Liu, C., Yin, J., 2019. Site classification method based on geomorphological and geological characteristics and its application in China. Bull. Seismol. Soc. Am. 109 (5), 1843–1854. https://doi.org/10.1785/0120190058.
Lin, A., Wotherspoon, L., Bradley, B., Motha, J., 2021. Evaluation and modification of geospatial liquefaction models using land damage observational data from the 2010–2011 Canterbury Earthquake Sequence. Eng. Geol. 287, 106099. https://doi.org/10.1016/j.enggeo.2021.106099.
Liu, W., Chen, Q., Wang, C., Juang, C.H., Chen, G., 2017. Spatially correlated multiscale Vs30 mapping and a case study of the Suzhou site. Eng. Geol. 220, 110–122. https://doi.org/10.1016/j.enggeo.2017.01.026.
Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 4768–4777.
Mahajan, A.K., 2009. NEHRP soil classification and estimation of 1-D site effect of Dehradun fan deposits using shear wave velocity. Eng. Geol. 104 (3–4), 232–240. https://doi.org/10.1016/j.enggeo.2008.10.013.
Mahajan, A., Kumar, P., 2023. Near-surface characterization and delineation of 2-D shear wave velocity profile from geologically complex landscape of Kangra Valley, NW Himalaya, India. Bull. Eng. Geol. Environ. 82 (5), 184. https://doi.org/10.1007/s10064-023-03207-2.
Mahajan, A., Slob, S., Ranjan, R., Sporry, R., Champati Ray, P., Westen, C., 2007. Seismic microzonation of Dehradun City using geophysical and geotechnical characteristics in the upper 30 m of soil column. J. Seismol. 11, 355–370. https://doi.org/10.1007/s10950-007-9055-1.
Mahajan, A., Mundepi, A., Chauhan, N., Jasrotia, A., Rai, N., Gachhayat, T., 2012. Active seismic and passive microtremor HVSR for assessing site effects in Jammu city, NW Himalaya, India—a case study. J. Appl. Geophys. 77, 51–62. https://doi.org/10.1016/j.jappgeo.2011.11.005.
Martínez-Segura, M.A., García-Nieto, M.C., Navarro, M., Vásconez-Maza, M.D., Oda, Y., García-Jerez, A., Enomoto, T., 2024. Seismic characterization of the subsoil under a historic building: Cathedral Church of Saint Mary in Murcia case study. Eng. Geol. 335, 107529. https://doi.org/10.1016/j.enggeo.2024.107529.
MATLAB, 2023. Version R2023b. The MathWorks Inc., Natick, Massachusetts. URL. https://www.mathworks.com.
Matsuoka, M., Wakamatsu, K., Fujimoto, K., Midorikawa, S., 2005, June. Nationwide site amplification zoning using GIS-based Japan engineering geomorphologic classification map. In: Proc. 9th Int. Conf. on Struct. Safety and Reliability, pp. 239–246.
Matsuoka, M., Wakamatsu, K., Fujimoto, K., Midorikawa, S., 2006. Average shear-wave velocity mapping using Japan engineering geomorphologic classification map. Struct. Eng./Earthq. Eng. 23 (1), 57s–68s. https://doi.org/10.2208/jsceseee.23.57s.
Mihalić, S., Oštrić, M., Krkač, M., 2011. Seismic microzonation: a review of principles and practice. Geofizika 28 (1), 5–20.
Molnar, S., Assaf, J., Sirohey, A., Adhikari, S.R., 2020. Overview of local site effects and seismic microzonation mapping in Metropolitan Vancouver, British Columbia, Canada. Eng. Geol. 270, 105568. https://doi.org/10.1016/j.enggeo.2020.105568.
Mori, F., Mendicelli, A., Moscatelli, M., Romagnoli, G., Peronace, E., Naso, G., 2020. A new Vs30 map for Italy based on the seismic microzonation dataset. Eng. Geol. 275, 105745. https://doi.org/10.1016/j.enggeo.2020.105745.
Muthuganeisan, P., Raghukanth, S., 2016. Site-specific probabilistic seismic hazard map of Himachal Pradesh, India. Part I. Site-specific ground motion relations. Acta Geophys. 64, 336–361. https://doi.org/10.1515/acgeo-2016-0010.
Naik, S., Patra, N., Malik, J., 2014. Spatial distribution of shear wave velocity for late quaternary alluvial soil of Kanpur city, Northern India. Geotech. Geol. Eng. 32, 131–149. https://doi.org/10.1007/s10706-013-9698-3.
Nath, S., Adhikari, M., Maiti, S., Devaraj, N., Srivastava, N., Mohapatra, L., 2014. Earthquake scenario in West Bengal with emphasis on seismic hazard microzonation of the city of Kolkata, India. Nat. Hazards Earth Syst. Sci. 14 (9), 2549–2575. https://doi.org/10.5194/nhess-14-2549-2014.
Nath, S., Ghatak, C., Biswas, A., Srivastava, A., 2022. Liquefaction susceptibility mapping in West Bengal with emphasis on its Capital City Kolkata under the impact of a few great earthquakes triggered from the Himalaya and Northeast India. J. Rehabil. Civ. Eng. 10 (4), 56–96. https://doi.org/10.22075/jrce.2021.24034.1533.
Nazarian, S., Stokoe, K.H., 1984. In situ shear wave velocities from spectral analysis of surface wave. In: Proceedings of 8th Conference on Earthquake Engineering, San Francisco, 1984.
NCS, 2016. A report on Seismic Hazard Microzonation of NCT Delhi on 1:10,000 Scale. Govt. of India Publication.
Pandey, B., Jakka, R., Kumar, A., 2016. Influence of local site conditions on strong ground motion characteristics at Tarai region of Uttarakhand, India. Nat. Hazards 81, 1073–1089. https://doi.org/10.1007/s11069-015-2120-0.
Pandey, B., Jakka, R., Kumar, A., Sharma, M., 2021. Site characterization of strong-motion stations of Himalaya and adjoining plains. Arab. J. Geosci. 14 (10), 879. https://doi.org/10.1007/s12517-021-07231-y.
Panzera, F., Lombardo, G., Imposa, S., Grassi, S., Gresta, S., Catalano, S., Di Maio, E., 2018. Correlation between earthquake damage and seismic site effects: the study case of Lentini and Carlentini, Italy. Eng. Geol. 240, 149–162. https://doi.org/10.1016/j.enggeo.2018.04.014.
Park, C.B., Miller, R.D., Xia, J., 1999. Multichannel analysis of surface waves. Geophysics 64 (3), 800–808. https://doi.org/10.1190/1.1444590.
PESMOS, 2024. Department of Earthquake Engineering, IIT Roorkee. https://pesmos.org/location-of-instruments/.
Puri, N., Jain, A., 2021. Development of surface level seismic hazard maps considering local soil conditions for the State of Haryana, India. J. Geol. Soc. India 97, 1365–1378. https://doi.org/10.1007/s12594-021-1875-z.
QGIS Development Team, 2023. QGIS Geographic Information System. Open Source Geospatial Foundation. URL. http://qgis.org.
Rahman, M.Z., Siddiqua, S., Kamal, A., 2016. Shear wave velocity estimation of the near-surface materials of Chittagong City, Bangladesh for seismic site characterization. J. Appl. Geophys. 134, 210–225. https://doi.org/10.1016/j.jappgeo.2016.09.006.
Rahman, M.Z., Hossain, M.S., Kamal, A., Siddiqua, S., Mustahid, F., Farazi, A., 2018a. Seismic site characterization for Moulvibazar town, Bangladesh. Bull. Eng. Geol. Environ. 77, 1451–1471. https://doi.org/10.1007/s10064-017-1031-6.
Rahman, M.Z., Kamal, A., Siddiqua, S., 2018b. Near-surface shear wave velocity estimation and V s 30 mapping for Dhaka City, Bangladesh. Nat. Hazards 92, 1687–1715. https://doi.org/10.1007/s11069-018-3266-3.
Rahman, M.Z., Siddiqua, S., Kamal, A., 2019. Geology and topography based Vs 30 map for Sylhet City of Bangladesh. Bull. Eng. Geol. Environ. 78, 3069–3083. https://doi.org/10.1007/s10064-018-1331-5.
Rasmussen, C.E., Nickisch, H., 2010. Gaussian processes for machine learning (GPML) toolbox. J. Mach. Learn. Res. 11, 3011–3015.
Riley, S.J., DeGloria, S.D., Elliot, R., 1999. Index that quantifies topographic heterogeneity. Intermount. J. Sci. 5 (1–4), 23–27.
Riyaz, T., Singh, B., 2022. Site Characterization using MASW and Mapping of Average Shear Wave Velocity for Jammu, Jammu and Kashmir. Int. J. Innov. Res. Eng. Manag. 9 (2), 430–434. https://acspublisher.com/journals/index.php/ijirem/article/view/11083.
Sarkar, R., Kolathayar, S., Drukpa, D., Choki, K., Rai, S., Tshering, S., Yuden, K., 2021. Near-surface seismic refraction tomography and MASW for site characterization in Phuentsholing, Bhutan Himalaya. SN Appl. Sci. 3, 1–18. https://doi.org/10.1007/s42452-021-04347-4.
Satyam, D., Rao, K., 2008. Seismic site characterization in Delhi region using multi channel analysis of shear wave velocity (MASW) testing. Electron. J. Geotech. Eng. 13, 167–183.
Sharma, H., Mahajan, A.K., 2022. Near-Surface Site Characterization of Shimla city, Northwest Himalaya using Combination of Passive and Active Seismic Methods. https://doi.org/10.21203/rs.3.rs-1384465/v1.
Sharma, B., Rahman, S., 2016. Use of GIS based maps for preliminary assessment of subsoil of Guwahati city. J. Geosci. Environ. Prot. 4 (5), 106–116. https://doi.org/10.4236/gep.2016.45011.
Sharma, P., Sharma, M., Sawant, V., 2020. Estimation of seismic hazard and amplification of strong ground motions in Indo-Gangetic Plains. J. Seismol. Earthq. Eng. 22 (1).
Shukla, S., Sil, A., 2023. Stochastic finite fault modeling of Mw 6.4 earthquake along Indo-Burmese arc and site classification of its seismic recording station. Nat. Hazards Rev. 3 (1), 14–26. https://doi.org/10.1016/j.nhres.2022.12.005.
Sil, A., Sitharam, T., 2014. Dynamic site characterization and correlation of shear wave velocity with standard penetration test 'N'values for the city of Agartala, Tripura state, India. Pure Appl. Geophys. 171, 1859–1876. https://doi.org/10.1007/s00024-013-0754-y.
Sil, A., Sitharam, T., 2017. Detection of local site conditions in Tripura and Mizoram using the topographic gradient extracted from remote sensing data and GIS techniques. Nat. Hazards Rev. 18 (2), 04016009. https://doi.org/10.1061/(ASCE)NH.1527-6996.0000228.
Singh, M., Duggal, S.K., Pallav, K., Sharma, K.K., 2020. Correlation of shear wave velocity with standard penetration resistance value for Allahabad City. In: Advances in Computer Methods and Geomechanics: IACMAG Symposium 2019, 2. Springer, Singapore, pp. 467–481. https://doi.org/10.1007/978-981-15-0890-5_39.
Sitharam, T.G., Kolathayar, S., James, N., 2015. Probabilistic assessment of surface level seismic hazard in India using topographic gradient as a proxy for site condition. Geosci. Front. 6 (6), 847–859. https://doi.org/10.1016/j.gsf.2014.06.002.
Sörensen, R., Zinko, U., Seibert, J., 2006. On the calculation of the topographic wetness index: evaluation of different methods based on field observations. Hydrol. Earth Syst. Sci. 10 (1), 101–112. https://doi.org/10.5194/hess-10-101-2006.
Tempa, K., Sarkar, R., Dikshit, A., Pradhan, B., Simonelli, A., Acharya, S., Alamri, A., 2020. Parametric study of local site response for bedrock ground motion to earthquake in Phuentsholing, Bhutan. Sustainability 12 (13), 5273. https://doi.org/10.3390/su12135273.

Thompson, E.M., Wald, D.J., Worden, C.B., 2014. A VS30 map for California with geologic and topographic constraints. Bull. Seismol. Soc. Am. 104 (5), 2313–2321. https://doi.org/10.1785/0120130312.

Trifunac, M.D., 2016. Site conditions and earthquake ground motion–a review. Soil Dyn. Earthq. Eng. 90, 88–100. https://doi.org/10.1016/j.soildyn.2016.08.003.

Verdugo, R., 2019. Seismic site classification. Soil Dyn. Earthq. Eng. 124, 317–329. https://doi.org/10.1016/j.soildyn.2018.04.045.

Vilalta, R., Drissi, Y., 2002. A perspective view and survey of meta-learning. Artif. Intell. Rev. 18, 77–95. https://doi.org/10.1023/A:1019956318069.

Vilanova, S.P., Narciso, J., Carvalho, J.P., Lopes, I., Quinta-Ferreira, M., Pinto, C.C., Moura, R., Borges, J., Nemser, E.S., 2018. Developing a geologically based VS30 site-condition model for Portugal: Methodology and assessment of the performance of proxies. Bull. Seismol. Soc. Am. 108 (1), 322–337. https://doi.org/10.1785/0120170213.

Wald, D.J., Allen, T.I., 2007. Topographic slope as a proxy for seismic site conditions and amplification. Bull. Seismol. Soc. Am. 97 (5), 1379–1395. https://doi.org/10.1785/0120060267.

Walling, M.Y., Mohanty, W.K., Nath, S.K., Mitra, S., John, A., 2009. Microtremor survey in Talchir, India to ascertain its basin characteristics in terms of predominant frequency by Nakamura's ratio technique. Eng. Geol. 106 (3–4), 123–132. https://doi.org/10.1016/j.enggeo.2009.03.013.

Wandrey, C.J., 1998. Geologic Map of South Asia (geo8ag) [Data Set]. U.S. Geological Survey. https://doi.org/10.5066/P9YC1C8G.

Xie, J., Zimmaro, P., Li, X., Wen, Z., Song, Y., 2016. VS 30 empirical prediction relationships based on a new soil-profile database for the Beijing plain area, China. Bull. Seismol. Soc. Am. 106 (6), 2843–2854. https://doi.org/10.1785/0120160053.

Yao Xiaobai, A., 2020. Georeferencing and Geocoding. In: Kobayashi, Audrey (Ed.), International Encyclopedia of Human Geography. https://doi.org/10.1016/B978-0-08-102295-5.10548-7.

Yilar, E., Baise, L.G., Ebel, J.E., 2017. Using H/V measurements to determine depth to bedrock and Vs30 in Boston, Massachusetts. Eng. Geol. 217, 12–22. https://doi.org/10.1016/j.enggeo.2016.12.002.

Zahoor, F., Rao, K., Malla, S., Tariq, B., Bhat, W., 2021. Seismic site characterization using MASW of sites along Srinagar metro rail alignment, Jammu and Kashmir. In: Proceedings of the Indian Geotechnical Conference 2019: IGC-2019, IV, pp. 581–593. https://doi.org/10.1007/978-981-33-6564-3_49.

Zhang, Y., Ren, Y., Wen, R., Wang, H., Ji, K., 2023. Regional terrain-based VS 30 prediction models for China. Earth Planets Space 75 (1), 72. https://doi.org/10.1186/s40623-023-01826-3.