Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Comparative Assessment of Feature Selection Methods for Machine Learning-Based Soil pH Prediction

Mir Nasser Navidi

n.navidi@areeo.ac.ir

Soil and Water Research Institute, Agricultural Research, Education and Extension Organization (AREEO), Karaj

Shahrokh Fatehi

Soil and Water Research Department, Kermanshah Agricultural and Natural Resources Research and Education Center, AREEO, Kermanshah, Iran

Ayeh Lotfollahi

Soil and Water Research Institute, Agricultural Research, Education and Extension Organization (AREEO), Karaj

Ahmad Aliyari Boroujeni

Department of Mathematics and Computer Science, Faculty of Sciences, University of Zanjan, Zanjan, Iran

Article

Keywords: XGBoost, Random Forest, Simulated annealing, Genetic algorithm, Soil pH

Posted Date: July 14th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10187873/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Soil pH is a key indicator of soil quality, particularly in arid and semi-arid environments. Accurate prediction of its spatial variability is essential for sustainable land management and digital soil mapping (DSM). Feature selection (FS) is a critical step for improving model efficiency and interpretability; however, its effectiveness depends on the machine learning (ML) algorithm. This study evaluated six FS methods (VIF, RFE, SA-FS1, SA-FS2, GA-FS1, GA-FS2) combined with four ML algorithms (SVM, RF, Cubist, XGBoost) to predict soil pH. Results demonstrated that FS performance is strongly algorithm-dependent, significantly impacting prediction accuracy and model stability. The selected predictor subsets varied markedly, ranging from 4 to 24 variables. Simulated annealing and genetic algorithm variants generally outperformed deterministic methods when coupled with complex models like RF and XGBoost. Notably, XGBoost combined with SA-FS1 achieved the highest validation performance and transferability, offering the optimal balance between accuracy and generalization. Conversely, SVM and Cubist showed greater sensitivity to variable selection, often exhibiting overfitting. These findings confirm that no single FS method is universally optimal; rather, performance depends on the model–feature interaction. This study highlights the importance of tailored FS strategies to enhance the reliability, interpretability, and transferability of soil pH predictions in environmental modeling.

1. Introduction

Soil is one of the most fundamental but increasingly limited natural resources, and it is under growing pressure from degradation processes worldwide (FAO, 2015). These pressures significantly reduce its capacity to deliver ecosystem services and support human well-being (Montanarella et al., 2016; Dominati et al., 2010). As a key environmental resource, soil is also essential for achieving the United Nations Sustainable Development Goals (Keesstra et al., 2016; Bouma, 2014). Soil systems exhibit strong spatial heterogeneity, encompassing a wide range of physical, chemical, and biological properties. While this variability supports ecosystem functions, it also increases vulnerability, making the assessment of soil variability essential for sustainable land use planning and soil management (Ibáñez et al., 1998). According to the FAO, about 33% of global soils are already degraded, and this may exceed 90% by 2050 (FAO, 2015; Sediqi et al., 2025).

Among soil quality indicators, soil pH is one of the most important, as it directly regulates key biological, chemical, and physical processes in the soil system. This role is especially critical in arid and semi-arid regions, where pH strongly affects agricultural productivity, ecosystem functioning, and environmental quality (Brevik et al., 2015). Soil pH is influenced by both natural and human factors, including climate, parent material, land use, vegetation, topography, and management practices. Therefore, understanding its spatial variability is essential for sustainable soil management and land use planning (Guo et al., 2022). In digital soil mapping, selecting appropriate environmental covariates is a crucial step for improving prediction accuracy, typically guided by frameworks such as CLORPT and SCORPAN, which link soil formation to environmental and landscape factors (McBratney et al., 2003; Minasny & McBratney, 2016).

The selection of auxiliary variables plays a key role in determining the performance of machine learning models. Removing redundant or irrelevant variables can improve accuracy, reduce multicollinearity, and speed up model calibration, but excessive elimination may also lead to information loss and reduced predictive performance (Machado, 2017). Feature selection (FS) methods are widely used to identify relevant predictors

and improve model efficiency and interpretability (Atef et al., 2025). These methods are generally grouped into four categories, filter, wrapper, embedded, and ensemble approaches, each with different strengths and limitations in terms of computational cost, robustness, and sensitivity to model structure (Hu et al., 2023; Chen et al., 2022).

Although feature selection plays a critical role in improving model accuracy, reducing redundancy, and enhancing computational efficiency, a wide range of feature selection methods is currently available, and their relative performance remains unclear across different modeling frameworks. Because each method relies on different assumptions and selection mechanisms, their effectiveness may vary depending on the structure of the dataset and the machine learning algorithm employed. In this study, soil pH was used as a representative test case to evaluate and compare the performance of different feature selection strategies in digital soil mapping. Feature selection methods are generally classified into four categories based on their methodology and interaction with predictive models.

a) Filter methods:

Filter approaches use statistical criteria independent of the predictive model to evaluate variable relevance. Common techniques include correlation analysis, multicollinearity diagnostics, and information-theoretic measures such as Information Gain (Reddy et al., 2024). These methods are computationally efficient and suitable for high-dimensional datasets, but because variables are evaluated independently, they often fail to capture nonlinear relationships and predictor interactions (Nzekwe et al., 2024).

b) Wrapper methods:

Wrapper techniques evaluate subsets of variables based on the predictive performance of a machine learning model. Examples include RFE (Lin et al., 2012), genetic algorithms (Theodoridis and Gkikas., 2020), and simulated annealing (Lin et al., 2008). By repeatedly training models with different predictor subsets, wrapper methods can identify nonlinear relationships and interactions. However, this iterative process is computationally expensive, especially for large datasets (Seijo-Pardo et al., 2017).

c) Embedded methods:

Embedded approaches integrate feature selection directly into model training. Representative methods include Boruta (Atef et al., 2025), and RF-based variable importance measures (Nicodemus et al., 2010). These methods are generally more computationally efficient than wrapper techniques, although their effectiveness depends on the underlying learning algorithm.

d) Ensemble methods:

Ensemble feature selection combines multiple strategies, such as filter, wrapper, and embedded methods, to improve the robustness and stability of selected predictors (Bolón-Canedo et al., 2015). By aggregating results from different approaches, ensemble methods reduce the bias of individual techniques and improve reliability in complex environmental and high-dimensional datasets (Chandrashekar & Sahin, 2014). In digital soil mapping (DSM), two main strategies are commonly used for variable selection (McBratney et al., 2003). In the first, variable reduction is performed before model development using methods such as correlation analysis,

multicollinearity diagnostics, or dimensionality reduction techniques like PCA (Hengl et al., 2007). Although computationally efficient, these methods mainly detect linear relationships and may overlook important nonlinear interactions among environmental covariates (Chandrashekar & Sahin, 2014).

2- Materials and Methods

2.1. Study Area

The study area covers approximately 57,850 hectares and is located about 60 km northwest of Kermanshah Province, Iran. It extends between 46°08' to 46°26' E longitude and 34°33' to 34°51' N latitude (Fig. 1). The geology of the region is dominated by calcareous formations, with localized occurrences of marl, conglomerate, and sandstone. The soils of the area are mainly classified into four soil orders: Inceptisols, Vertisols, Aridisols, and Mollisols. Land use in the region is diverse and includes forested areas, rangelands, rainfed agricultural lands, and irrigated croplands, reflecting the heterogeneous environmental and management conditions of the landscape.

2.2. Soil sampling and testing

Soil sampling locations were selected using the conditional Latin Hypercube Sampling (cLHS) approach (Minasny & McBratney, 2006). This stratified sampling method efficiently captures the multivariate distribution of environmental covariates by incorporating both quantitative and categorical auxiliary variables. Owing to its ability to represent environmental variability, cLHS has been widely used in digital soil mapping studies. In the present study, the cLHS algorithm implemented in the R environment was combined with environmental covariates to determine the optimal locations of 120 sampling sites across the study area. Soil samples were collected from the 0–30 cm depth at each site and analyzed under laboratory conditions. Descriptive statistical analyses of soil pH were then conducted in R, and the resulting summary statistics are presented in the following table.

2.3. Environmental variables

In this study, 106 environmental covariates were generated from multiple data sources based on the SCORPAN framework. Topographic conditions were represented using a range of terrain attributes, vegetation conditions through spectral vegetation indices, and lithological characteristics through parent material–related variables. Vegetation indices were derived from Landsat imagery acquired from the USGS Earth Explorer platform (https://earthexplorer.usgs.gov/). The satellite images had a spatial resolution of 30 × 30 m and cloud coverage of less than 5%.

Topography-related covariates were extracted from a Digital Elevation Model (DEM) obtained from the same platform and processed in SAGA GIS using standard geomorphometric procedures. To ensure consistency, all 106 covariates were standardized to a spatial resolution of 30 × 30 m. Radiometric and atmospheric corrections for both the satellite imagery and DEM were carried out in R using the RStoolbox package. Due to the relatively small spatial extent of the study area and its limited climatic variability, climate-related variables were not included in the modeling framework (Chen et al., 2022).

2.4. Categorization and description of environmental covariates

The environmental covariates used in this study were classified into several categories according to their environmental characteristics and their relevance within the SCORPAN framework (Table 1). These categories are described as follows:

Table 1
Classification some of the environmental variables used in this study.

Landsat OLI	Variables	Abbreviation	Variables	Abbreviation
	Landsat-8 OLI six bands and thermal infrared band	**B2:B7**		
Topographic variables	Channel Network Base Level	CNBL	Total Curvature	TC
	Average Vertical Distance	AVD	Morphometric Protection Index	MPI
	Analytical Hillshading	AH	(Melton Ruggedness Number)	MRRTF
	Mid.Slope Position	MSP	Terrain Ruggedness Index	TRI
	Gradient	G	Terrain Surface Convexity	TSC
	Catchment Slope	CS	Terrain Surface Texture	TST
	Slope Length	SL	Ridge Level	RL
	Slope	S	Maximum Height	MH
	Aspect	A	Normalized Height	NH
	Convexity	C	Standardized Height	SH
	Downslope Curvature	DC	Relative Slope Position	RSP
	Local Downslope Curvature	LDC	Modified Catchment Area	MCA
	Local Upslope Curvature	LUC	Specific Catchment Area	SCA
	Longitudinal Curvature	LC	Total Catchment Area	TCA
	Plan Curvature	PC	Surface Area	SA
	Tangential Curvature	TgC	Real Surface Area	RSA
	Compare to Flat Terrain	CFT	LS Factor	LSF
Variables related to the network and water flow	Channel Initiation Threshold	CIT	CIT Index (Convergence Index Threshold)	CITI
	Channel Network Distance	CND	Stream Power Index	SPI

Landsat OLI	Variables	Abbreviation	Variables	Abbreviation
	Landsat-8 OLI six bands and thermal infrared band	**B2:B7**		
	Overland Flow Distance	OFD	Cell Balance	CB
	Horizontal Overland Flow Distance	HOFD		
Radiation, energy and climate indicators	Hillshade	HS	Diurnal Anisotropic Heating	DAH
	Sky View Factor	SVF	Terrain View Factor	TVF
	Total Insolation	TI	Mass Balance Index	MBI
	Direct Insolation	DI	Northness	N
	Direct to Diffuse Ratio	DDR		
Vegetation cover indices and spectral indices	Normalized Difference Vegetation Index	NDVI	Perpendicular Vegetation Index	PVI
	Transformed Soil Adjusted Vegetation Index	TSAVI	Ratio Vegetation Index	RVI
	Transformed Vegetation Index	TVI	B2, B5, B7	
	Corrected Transformed Vegetation Index	CTVI	Thiam.s Transformed Vegetation Index	TTVI
	Enhanced Vegetation Index	EVI	Grain Size Index	GSI
	Difference Vegetation Index	DVI		
Chemical and geological variables	Carbonate Normalized Ratio	CNR	Rock Outcrop Normalized Ratio	RNR

2.5. Environmental Variables

Topographic variables play a fundamental role in controlling hydrological and erosional processes by reflecting spatial variations in land surface roughness and geomorphometric structure (Table 1). These factors influence water accumulation, energy redistribution, sediment transport, and the initiation or suppression of erosion (Senanayake et al., 2020). Terrain attributes such as slope, SL, curvature, elevation, and terrain ruggedness directly affect overland flow velocity, sediment detachment, mechanical weathering, and the spatial variability of soil chemical properties, including pH, through their effects on moisture availability and leaching intensity (Poggio et al., 2021).

Network and flow variables further characterize the pathways and timing of water movement within watersheds (Deemy et al., 2022). Indices such as channel network distance, overland flow distance, and CIT

influence sediment routing, dissolved material transport, and erosion dynamics, thereby improving the understanding of spatial erosion patterns (Table 1).

Radiation, energy, and climate indicators also contribute significantly to hydrological and erosional dynamics (Petrov et al., 2025). These variables describe the amount and distribution of surface energy, influencing evapotranspiration, vegetation growth, soil moisture regimes, and the soil carbon cycle. Changes in energy balance alter runoff generation and erosion intensity by modifying vegetation density and moisture conditions (Table 1) (Poggio et al., 2021).

Vegetation indices provide indirect but important information on soil erosion processes by representing vegetation density, vigor, and spatial distribution. Spectral indicators such as NDVI are commonly associated with reduced erosion risk because dense vegetation cover protects the soil surface from raindrop impact, reinforces soil structure through root systems, and decreases runoff energy (Ayalew et al., 2020). Remote sensing data therefore offer an effective approach for quantifying vegetation cover and assessing its influence on erosion processes (Table 1) (Veneros et al., 2020).

Chemical and geological variables represent the mineralogical and chemical properties of soils and parent materials, influencing both soil stability and erosion susceptibility. Indices such as CNR and ROR describe carbonate content, rock exposure, and buffering capacity, which regulate soil acidity, aggregate stability, sediment retention, and pH dynamics (Deemy et al., 2022). These properties play a key role in controlling soil resistance to erosion and related biogeochemical processes (Table 1) (Zhou., 2024).

2.6. Feature selection based on optimization algorithms

Dimensionality reduction techniques play a crucial role in simplifying modeling processes, particularly when many input variables are involved (Nikraftar et al., 2019). The main objective of variable selection is to identify an optimal subset of predictors that enhances model performance and robustness. The quality and relevance of training data strongly influence overall model accuracy. As data dimensionality increases, supervised and unsupervised algorithms may face challenges such as higher computational complexity and reduced efficiency (Guyon & Elisseeff., 2003). Redundant or highly correlated variables can negatively affect generalization and decrease predictive accuracy, while irrelevant predictors unnecessarily increase model complexity. Therefore, applying variable selection techniques helps reduce overfitting, increase accuracy, and lower computational cost and training time (Cheng., 2024). In general, these algorithms are divided into supervised and unsupervised approaches. Supervised methods use response-variable information to guide selection and are typically grouped into filter, wrapper, and embedded methods.

In filter-based approaches, variables are selected based on their statistical relationship with the target variable, using measures such as correlation or mutual dependence. These methods evaluate each predictor independently of the learning algorithm, making them computationally efficient and well suited for high-dimensional data. In contrast, wrapper methods assess subsets of variables by repeatedly training and validating a predictive model. The dataset is partitioned, and the model is iteratively trained while predictors are added or removed according to their contribution to performance. Wrapper techniques often use search strategies such as forward selection and backward elimination to explore variable combinations (Cheng., 2024). In this study, commonly used filter- and wrapper-based variable selection strategies were applied. To determine the most informative predictors for soil pH estimation, five optimization-based methods were

implemented. A brief description of their implementation and characteristics is provided in the following sections.

2.6.1. Variance Inflation Factor (VIF)

The VIF is a widely used filter-based method for detecting multicollinearity among predictor variables. A high VIF value indicates that a predictor has a strong linear relationship with other variables in the dataset. This method evaluates only statistical relationships among explanatory variables and operates independently of machine learning algorithms. Strong intercorrelations among predictors can make variable importance estimates and model coefficients unstable or biased, potentially leading to misleading interpretations. Therefore, assessing multicollinearity is an essential step in model development. The VIF quantifies how much the variance of a regression coefficient is inflated due to multicollinearity and is calculated as follows (Dormann et al., 2013; O'Brien, 2007):

$$\mathrm{VIF}_{\mathrm{X}_1} = \frac{1}{1 - \mathrm{R}^2_{\mathrm{X}_1}}$$

where $R^2$is the coefficient of determination obtained from a regression model in which $X_1$is treated as the response variable and the remaining predictors serve as explanatory variables. Generally, a VIF value greater than five indicates substantial correlation among predictors, suggesting that the variable should be removed to ensure reliable inference. Multicollinearity can reduce parameter stability and impair predictive performance. To address this issue, VIF analysis was performed so that correlated variables were excluded from the predictor set (Dormann et al., 2013; O'Brien, 2007). To further reduce uncertainty associated with multicollinearity, the collinear package was used. The adjustment criteria included a maximum pairwise correlation of 0.5 and a maximum VIF threshold of 2.5. This automated procedure removed highly correlated predictors, resulting in more stable, robust, and interpretable models.

2.6.2. Recursive Feature Elimination (RFE)

RFE is a widely used wrapper-based feature selection method designed to identify the most relevant predictors in machine learning models (Chen et al., 2022). The algorithm follows a recursive backward elimination strategy in which model performance and variable importance are repeatedly evaluated while progressively reducing the number of predictors. Initially, a model is trained using the full set of variables, and its performance is assessed using RMSE and k-fold cross-validation. Variable importance scores are then calculated, the least important variable is removed, and the model is retrained using the reduced predictor set (Poggio et al., 2021). This process continues until only one predictor remains. The optimal subset is selected based on the model achieving the lowest RMSE (Kuhn, 2008).

In this study, RFE was implemented in R using the caret package with an RF model (Poggio et al., 2021; Sediqi et al., 2025). A RF regression model was first constructed using all environmental variables, and RF-derived importance measures were used to rank predictors. Variables with the lowest importance were iteratively removed, while the model was retrained and evaluated through cross-validation at each step. This approach enabled identification of the predictor subset that maximized RF predictive accuracy while minimizing variable redundancy (Chen et al., 2022; Guyon et al., 2002).

2.6.3. Simulated Annealing Feature Selection (SA-FS)

This method uses the nature-inspired optimization algorithm Simulated Annealing (SA) to identify an optimal subset of predictor variables. SA is a stochastic optimization technique inspired by the physical annealing process in metallurgy, where a system is gradually cooled to reach a stable low-energy state (Kirkpatrick et al., 1983). In feature selection, the algorithm starts with an initial subset of variables and iteratively generates new candidate subsets through random modifications. Each subset is evaluated based on model performance, and subsets that improve predictive accuracy are accepted as the current solution. SA can also occasionally accept inferior solutions, especially during early search stages, allowing the algorithm to escape local optima and explore the predictor space more effectively (Chandrashekar & Sahin, 2014).

In this study, SA-based variable selection was implemented in R using the caret package. The algorithm was configured through the safsControl function to optimize RF model performance. Model evaluation employed repeated cross-validation with ten replicates to improve robustness and reduce uncertainty. The number of optimization iterations was set to 250, during which the algorithm tracked model performance and candidate variable subsets. The final output was the predictor subset with the best performance. This approach enables identification of a parsimonious set of informative and interrelated environmental variables that most effectively improve model prediction (Blum & Roli, 2003; Kuhn & Johnson, 2013).

2.6.4. Genetic Algorithm Feature Selection (GA-FS)

The Genetic Algorithm (GA) is an adaptive optimization technique inspired by natural selection and evolutionary genetics (Mukhopadhyay., 2009). In variable selection, GA optimizes feature subsets through crossover, mutation, population control, and fitness evaluation, while integrating effectively with other modeling techniques to improve predictive performance (Ghodmare et al., 2018). Its main advantages include efficient exploration of large predictor spaces, support for parallel computation, and flexibility across different data structures (Waysi et al., 2025).

GA iteratively recombines candidate subsets to generate improved solutions. Each subset is evaluated using a fitness function, commonly based on RMSE, and high-performing subsets are retained through successive generations (Theodoridis and Gkikas., 2020). In this study, GA was implemented using ten-fold cross-validation over 100 iterations to optimize environmental variable selection for soil pH prediction (Moser et al., 2000). The iterative search identified informative predictors, reduced unnecessary computations, and improved model accuracy and efficiency (Waysi et al., 2025). Similar to Simulated Annealing, GA functions as a wrapper-based stochastic search approach that evaluates candidate subsets based on predictive model performance. GA- and SA-based variable selection were implemented using the SVM-RFE package in R (Guo et al., 2022; Kuhn., 2008).

2.7. Data modeling with machine learning

Machine learning (ML) techniques have been increasingly applied to map soil properties using environmental covariates, due to their ability to capture complex and nonlinear relationships between soil attributes and environmental predictors (Naimi et al., 2021; Theodoridis and Gkikas., 2020). Among these approaches, RF and Extreme Gradient Boosting (XGBoost) are two of the most widely used ensembles learning algorithms and have demonstrated strong performance in digital soil mapping (DSM) applications (Chen & Guestrin, 2016).

Compared with traditional regression-based approaches, such as decision trees, multiple linear regression, and support vector machines, RF and XGBoost are generally more robust to noisy or sparse data and exhibit superior predictive capability due to their ability to model complex nonlinear interactions among predictors (Fatima et al., 2023). In this study, RF and XGBoost were applied both for identifying optimized predictor variables and for predicting and mapping soil properties across the study area.

2.7.1. Random Forest

RF, developed by Breiman (2001), is a decision tree–based ensemble algorithm that constructs multiple trees using bootstrap sampling and random variable selection. By aggregating predictions from individual trees, RF reduces overfitting and shows strong robustness to noise and outliers, making it well suited for high-dimensional datasets and variable selection tasks (Li et al., 2025). Model calibration depends on several user-defined parameters, including the number of trees (ntree), the number of variables selected at each split (mtry), and the minimum terminal node size (nodesize). In this study, ntree was set to 500 because prediction error stabilized at this level. The mtry parameter was set to its default value, equal to the square root of the total number of predictors, while nodesize was fixed at 5. The final model was trained in the R environment with the ranger package, which provides an efficient implementation of RF for high-dimensional data (Dunkl & Ließ, 2022).

2.7.2. XGBoost

The XGBoost algorithm is an advanced gradient boosting framework based on decision trees, designed to enhance prediction accuracy and minimize errors in hybrid models (Chen and Guestrin, 2016; Li et al., 2022). The learning process is performed by a specific function that measures the error between the model prediction and the actual data. At each iteration, the model residuals, representing prediction errors, are calculated, and a new regression tree is fitted to reduce these residuals. This procedure is repeated iteratively, with each new tree added to the ensemble of previously fitted trees, progressively improving the overall prediction. Finally, the predictions from all trees are weighted and combined to produce the final output (Arif Ali et al., 2023; Chen & Guestrin, 2016). This methodology enables the modeling of complex nonlinear relationships and multivariate interactions among environmental variables, resulting in enhanced predictive accuracy for soil organic carbon content (Hengl et al., 2018).

2.7 3. Cubist

Cubist is a rule-based regression algorithm derived from Quinlan's M5 model tree approach (Quinlan, 1992; Kuhn., 2020). The method first partitions the predictor space using a decision tree and then fits linear regression models at the terminal nodes based on variables involved in the splitting process. During pruning, the algorithm evaluates both terminal and internal node models to simplify the tree and generate optimized rule-based models. Each rule represents a path from the root to a terminal node and is refined through pruning and simplification procedures (Quinlan, 1992; Kuhn & Johnson, 2013). Cubist can also use a boosting-like strategy through ensembles called committees, where successive models focus on correcting errors from previous ones, improving predictive performance (Kuhn & Johnson, 2013). In this study, Cubist was implemented in R using the Cubist package, and model optimization was performed with the caret package. Two parameters were tuned: the number of committees (5–50, with increments of 5) and the number of neighbors (1, 5, and 9). Unlike conventional decision trees such as CART, which assign constant values to

terminal nodes, Cubist fits linear regression models at each leaf, allowing it to capture complex relationships through localized linear modeling. The model also provides variable importance measures similar to RF, facilitating interpretation of predictor contributions (Kuhn & Johnson, 2013; Kuhn., 2020).

2.7.4. Support vector machine (SVM)

SVM is a supervised machine learning algorithm based on statistical learning theory. It is widely used for both classification and regression tasks and has been extensively applied in digital soil mapping due to its strong generalization capability and ability to model nonlinear relationships between environmental variables and soil properties (Guo et al., 2022; Cortes & Vapnik, 1995). The algorithm maps input variables into a higher dimensional feature space using kernel functions, allowing complex nonlinear relationships to be modeled effectively. The optimal regression function is determined by maximizing the margin while minimizing prediction errors using a regularization parameter C and an ε insensitive loss function (Jakkula, 2006).

Kernel functions such as the radial basis function (RBF), polynomial, and linear kernels enable SVM to capture nonlinear patterns in soil–environment relationships, with the RBF kernel being the most widely used in digital soil mapping due to its flexibility and strong predictive performance (Hengl et al., 2018). However, the performance of SVM depends on proper tuning of hyperparameters, including C, γ, and ε, which are typically optimized using cross validation techniques (Kuhn & Johnson, 2013; Guo et al., 2022).

2.8. Model performance evaluation

2.8.1. Evaluation of Model Prediction Accuracy

To evaluate the predictive accuracy of the models for soil pH mapping, the variables selected by six feature selection approaches, VIF, RFE, SA-FS1, SA-FS2, GA-FS1, and GA-FS2, were used as inputs for model training. Model performance was assessed using repeated cross-validation with ten repetitions (Picard & Cook, 1984). This approach repeatedly partitions the dataset into training and testing subsets, enabling a more reliable evaluation of model robustness and generalization ability while reducing uncertainty in performance estimation (Kuhn & Johnson, 2013).

Several statistical metrics were used to evaluate soil pH prediction accuracy. The coefficient of determination (R^2) measured the proportion of variance in observed values explained by the model, while the root means square error (RMSE) quantified prediction error. In addition, the mean absolute error (MAE) was calculated as a complementary indicator representing the average absolute difference between observed and predicted values (Chai and Draxler., 2014). Generally, lower RMSE and MAE values indicate higher model accuracy, whereas higher R^2 values reflect stronger explanatory power. The dataset was divided into calibration (80%) and validation (20%) subsets. The calibration dataset was used to train the models, while the validation dataset was used to assess final predictive performance. Accordingly, model errors for the calibration and validation datasets were reported as RMSEc and RMSEv, respectively (Sediqi et al., 2025).

Furthermore, in the genetic algorithm approach, a desirability function was applied to simultaneously optimize prediction accuracy and model complexity. This method balances model performance with the number of selected variables, resulting in models with adequate predictive power and simpler structures. These evaluation criteria allow comparison of different feature selection methods and assessment of their effects on

uncertainty and prediction quality, which is essential for analyzing the stability of environmental prediction models (Chen et al., 2022). Finally, the evaluation metrics RMSE, R², and MAE were calculated to compare model performance, and their mathematical formulations are defined as follows.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left[Z_{x_i}^{*} - z_{x_i}\right]^2}{n}}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|z_{x_i}^{*} - z_{x_i}\right|$$

$$R^2 = \frac{\left(n(\sum_{i=1}^{n} z_{x_i}^{*} z_{x_i} - \sum_{i=1}^{n} z_{x_i}^{*} \sum_{i=1}^{n} z_{x_i}\right)^2}{\left[n\sum_{i=1}^{n} {z_{x_i}^{*}}^2 - \left(\sum_{i=1}^{n} z_{x_i}^{*}\right)^2\right]\left[n\sum_{i=1}^{n} {z_{x_i}}^2 - \left(\sum_{i=1}^{n} z_{x_i}\right)^2\right]}$$

RMSE = root mean square error

$Z_{x_i}^{*}$ = the defined variable

Z_{x_i} = the measured variable

R^2 = the coefficient of determination

3. Results and discussion

3.1. Descriptive statistics

Soil pH values ranged from 6.7 to 8.3, with a coefficient of variation of 5.2%, indicating relatively low variability across the study area. The distribution showed slight negative skewness (− 0.65), suggesting a modest tendency toward lower pH values relative to the mean. The kurtosis value was approximately zero, indicating that the shape of the distribution was close to normal distribution in terms of peakedness. However, according to the normality test (p-value < 0.05), the soil pH data significantly deviated from a normal (Table 2).

Table 2
Descriptive statistics of soil pH.

Soil properties	Min	Max	Median	Mean	Std.d	CV (%)	SKEW	KURT	*p*-value
pH	6.65	8.3	7.7	7.65	0.397	5.2	-0.65	-0.34	0.00011

3.2. Selected methods for variable reduction

To evaluate uncertainty and examine the impact of different variable reduction methods on model performance, six feature-selection techniques were applied. These methods were chosen to compare their effectiveness in reducing multicollinearity, identifying optimal predictor subsets, and improving model reliability.

The variable-reduction techniques used in this study included:

a) VIF,

b) SA-FS – which, in a single run, produced two types of outputs: the set of top variables and the set of optimal variables (SA-FS1 and SA-FS2),

c) GA-FS – which, in a single run, produced two types of outputs: the set of top variables and the set of optimal variables (GA-FS1 and GA-FS2),

d) RFE.

These methods represent different categories of feature-selection strategies, including filter-based and wrapper-based approaches, enabling a comprehensive comparison of their ability to reduce predictor redundancy and enhance the robustness of the modeling framework.

3.3. Selected Variables and Performance of Variable Reduction Methods

A review of the variables used in soil pH prediction models indicates that different variable reduction methods emphasize distinct dataset characteristics, leading to different subsets of key predictors. The number of selected variables varied considerably among methods, ranging from 4 to 24 (Table 3). These differences reflect variations in algorithm sensitivity, search strategy, and criteria used to evaluate variable importance. For example, the VIF method produced a compact subset of 10 variables, whereas SA-FS2 retained 24 variables, capturing broader environmental variability. Despite these differences, several predictors repeatedly appeared across methods, highlighting their importance in representing topographic, hydrological, geological, and vegetation controls on soil pH variability.

Topographic variables were consistently identified as influential predictors. This finding emphasized the role of microtopography in controlling water redistribution, soil moisture, sediment transport, and nutrient accumulation, all of which directly influence soil chemical properties (Moore et al., 1993; Hengl et al., 2017).

The results also show that the effectiveness of feature selection methods strongly depends on the modeling algorithm. The present study revealed clear interactions between variable reduction methods and model structure. Tree-based algorithms such as RF and Cubist generally benefited from larger and more diverse predictor subsets, whereas SVM was more sensitive to feature dimensionality and parameter tuning. These findings confirm that there is no universally optimal variable selection strategy; rather, its suitability depends on the learning behavior and complexity of the predictive algorithm (Guyon & Elisseeff, 2003; Kuhn & Johnson, 2013; Guo et al., 2022).

Methods selecting a moderate number of variables, particularly VIF and GA-FS2 (10–12 variables), provided a better balance between model complexity, predictive accuracy, and generalizability (Cheng et al., 2022; Nikraftar et al., 2019). Systematic evaluation of variable reduction approaches together with model-specific behavior provides a robust framework for identifying key environmental controls on soil pH and developing reliable predictive models for DSM (Samuel-Rosa et al., 2015; Poggio et al., 2021).

Table 3
The number and types of selected variables used in the feature selection methods under study.

Variable reduction method	Number of variables	Sample variables
VIF	10	MPI, AVD, CFT, CNR, …
RFE	4	CNR, LUC, PVI, SL.
SA-FS1	5	LSF, TgC, MH, N, B7.
SA-FS2	24	AH, CS, Dem, C, CTVI, …
GA-FS1	5	SL, DDR, LC, PVI, …
GA-FS2	10	CIT, B5, EVI, MRRTF, PVI, …

3.3.1. Variance Inflation Factor (VIF)

In this approach, variables such as the Morphometric Index, AVD, and CNR were identified as influential predictors. The primary objective of the VIF method is to reduce multicollinearity among auxiliary variables, thereby improving the stability, interpretability, and statistical reliability of predictive models (Cheng et al., 2022).

The predictors retained through this procedure mainly represent topographic and hydrological controls, which are known to play a key role in regulating soil chemical dynamics, particularly soil pH. Topography influences soil pH distribution by controlling environmental processes such as water flow pathways, soil moisture accumulation, surface drainage, sediment transport, and erosion intensity, which collectively affect soil chemical reactions and nutrient redistribution (Moore et al., 1993; Hengl et al., 2017). Compared with many other feature selection techniques, the VIF method typically results in a smaller subset of predictors because variables with high levels of multicollinearity are systematically removed (Hengl et al., 2017; Kuhn & Johnson, 2013). While this reduction can simplify model structure and improve interpretability, its effectiveness depends on the modeling algorithm used. Some ML algorithms, particularly tree-based ensemble models such as RF and extreme gradient boosting (XGBoost), are relatively robust to correlated predictors and can retain predictive performance even when multicollinearity is present (Breiman, 2001; Hengl et al., 2017). Consequently, strict removal of correlated variables may sometimes lead to the exclusion of informative predictors that contribute to model performance.

In the present study, models developed using the variables selected by the VIF method showed moderately good performance during the calibration phase, with R^2 values ranging from 0.55 to 0.68 and a concordance coefficient of approximately 0.80. However, model performance declined considerably during the validation phase, where R^2 values decreased to 0.15–0.37 and concordance dropped below 0.45. This reduction in predictive accuracy likely reflects a limitation of the VIF approach, which focuses primarily on reducing multicollinearity among predictors and does not explicitly consider the strength of the relationship between auxiliary variables and the target variable (soil pH). As a result, some variables with meaningful explanatory power may have been excluded, ultimately reducing the model's ability to generalize to independent datasets (Guyon & Elisseeff, 2003; Kuhn & Johnson, 2013).

3.3.2. Recursive Feature Elimination (RFE)

RFE is a widely used method for identifying the most influential variables in predictive modeling. In this study, variables such as the CNR and LUC were identified as key factors affecting soil pH prediction. RFE typically selects a smaller but more targeted set of variables compared to random search-based approaches. Similar observations were reported by Guo et al. (2022), who demonstrated that choosing a compact and optimal subset of variables, when aligned with the spatial structure of the data, can maintain model accuracy while reducing computational complexity. In the present study, RFE performed comparably to VIF during the calibration phase, achieving R^2 values up to 0.68 and concordance around 0.80. However, performance declined markedly during the validation phase, with R^2 values dropping below 0.05 in some models. This discrepancy highlights RFE's strong dependence on the training dataset and its sensitivity to the correlation structure among variables (Heuvelink et al., 2018; Jiang et al., 2022).

The iterative process of variable elimination in RFE, although effective in reducing multicollinearity, may inadvertently remove variables that exhibit nonlinear or interaction effects. Consequently, models developed using RFE may achieve high accuracy on training data but demonstrate limited generalizability when applied to independent datasets (Ramezan., 2022). In nonlinear algorithms such as RF or XGBoost, this limitation often manifests as overfitting, since RFE eliminates variables based on their local importance within the training data rather than their stability across the entire predictor space (Zerouali et al., 2024). Overall, the findings of this study indicate that while RFE can enhance model simplicity and interpretability by reducing the number of predictors, its application without stability assessments or ensemble feature selection frameworks may result in unstable predictions and increased uncertainty (Szatmári et al., 2019). To achieve a more appropriate balance among accuracy, stability, and generalizability, it is recommended to combine RFE with multi-stage validation strategies or to integrate it within hybrid frameworks (e.g., RFE–SA or RFE–GA).

3.3.3. Simulated Annealing Feature Selection (SA-FS1/2)

In this study, two modes of the SA-FS algorithm were evaluated. The first mode, SA-FS1, selected a limited subset of variables, choosing only five predictors during the modeling process. In contrast, the second mode, SA-FS2, selected a broader range of variables, including AH, TVI, CS, and TRI (Table 3). This variation reflects the multidimensional nature of the processes governing soil pH and highlights the interactions among topographic, hydrological, and vegetation factors in shaping the spatial distribution of soil properties (Park and Burt., 2002).

The inclusion of variables with diverse characteristics in SA-FS2 demonstrates the algorithm's ability to capture the heterogeneous structure of the dataset, although it also introduces a relative increase in uncertainty regarding the optimal selection of predictors. The broader variable set selected in SA-FS2 may provide a more comprehensive representation of the environmental factors governing the spatial variability of soil pH. In the performance evaluation, the SA-FS1 method demonstrated the best results during the validation phase (Validation R^2 = 0.62, Concordance = 0.74), achieving a well-balanced trade-off between calibration accuracy (R^2 up to 0.69, Concordance up to 0.74) and validation stability.

The SA-FS2 version also performed similarly, with the difference that a better balance between validation and calibration results was observed, although its generalization power was slightly lower than SA-FS1. These

findings suggest that SA-FS based algorithms possess greater structural stability compared to methods such as VIF or RFE and are more effective at mitigating fluctuations in model performance when variable combinations are altered (Belkadi and Drias., 2022). Leveraging stochastic optimization and incremental search mechanisms, SA-FS2 algorithms can achieve a balance between prediction accuracy and model stability, thereby reducing uncertainty associated with variable selection (Chang et al., 2025). Notably, the SA-FS1 method emerged as the most efficient among the approaches evaluated, offering an optimal compromise between accuracy and generalizability. This makes it a robust and effective option for digital soil modeling.

3.3.4. Genetic Algorithm Feature Selection (GA-FS 1/2)

In this study, the GA-FS method was evaluated. In a single execution, GA-FS1 identified five variables, whereas GA-FS2 identified ten variables. This variation in selected variable sets highlights the inherent uncertainty associated with algorithmic variable selection. In both runs, variables such as MRRTF, PVI, and SL were consistently selected, underscoring their key role in the spatial prediction of soil pH (Tables 3 and 4). This consistency indicates that indicators related to land surface roughness and vegetation are fundamental drivers of spatial variability in soil pH (Baltensweiler et al., 2020).

Due to its stochastic nature and reliance on a random search process, the genetic algorithm may produce different variable combinations in each execution. While this property demonstrates the algorithm's search capability, it can also introduce instability and increase uncertainty in variable selection (Chen et al., 2022). In the performance evaluation, GA-FS1 demonstrated the best balance between accuracy and stability (Validation R^2 = 0.45; Concordance = 0.57). In contrast, GA-FS2, despite a lower R^2 value (0.31), exhibited more favorable Bias and RMSE indices, indicating greater relative stability and reduced error fluctuations in independent datasets.

These results suggest that increasing the number of selected variables does not necessarily improve model accuracy and can, in fact, increase variance and reduce generalizability (Seifi et al., 2020). Overall, the findings indicate that the genetic algorithm performs better during model calibration than validation, reflecting a tendency to overfit the training data. Nevertheless, it remains effective in identifying nonlinear relationships among variables. These results underscore the importance of integrating genetic algorithms with stability control strategies, such as multi-step cross-validation, to minimize uncertainty and enhance model robustness (Sadik and Samui., 2024; Al-Shabeeb et al., 2022).

Table 4
Number of variables and Sample variables in properties selection method.

Properties Selection Method	Number of variables	Sample variables
VIF	10	MPI, AVD, CFT, CNR, …
RFE	5	MRVBF, MRRTF, CND, TST, …
SA-FS1	5	LSF, TgC, MH, N, B7.
SA-FS2	24	AH, CS, Dem, C, CTVI, …
GA-FS1	5	SL, DDR, LC, PVI, …
GA-FS2	10	CIT, B5, EVI, MRRTF, PVI, …

3.4. Examining the overlap of variables in variable reduction methods

The results of this study indicate that, although there is partial overlap among the variables selected by different algorithms, notable differences exist in the final composition of each subset. This partial overlap suggests that certain predictors, such as SL and LUC, possess structural and stable importance in soil pH prediction and have been consistently identified across most variable selection methods (Wadoux, 2019). Additionally, indices such as the CNR have been recognized as key determinants in regulating soil chemical processes (Guo et al., 2022). The repeated inclusion of these variables across different selection techniques reflects the structural stability of the models and enhances their spatial robustness and generalizability to new datasets (Ferhatoglu and Miller, 2022). However, certain variables were selected or excluded depending on the specific algorithm and optimization criteria applied (Table 5).

These discrepancies likely arise from the inherent differences among dimensionality reduction methods and their varying sensitivity to multicollinearity, nonlinear interactions, and interdependencies among environmental predictors (Schmidt et al., 2024). For example, methods based on mutual information or ordinal ranking (e.g., RFE) typically favor the elimination of correlated variables, whereas stochastic optimization approaches such as Genetic Algorithms (GA) or Simulated Annealing (SA) are more inclined to retain complementary variables to model more complex interactions in the data (Ferhatoglu and Miller., 2022). Therefore, adopting a hybrid feature selection strategy or implementing multi-algorithm cross-validation is recommended to enhance confidence in the stability of the selected variable set and to mitigate fluctuations associated with reliance on a single selection algorithm (Schmidinger and Heuvelink, 2023; Chen et al., 2022).

The analysis of selected variables indicates that soil pH modeling requires an integrated framework that simultaneously incorporates topographic, hydrological, vegetation, radiation-related, and soil chemical composition variables (Hosseini et al., 2023; Dindaroglu et al., 2022). The findings of the present study show that the concurrent application of multiple variable reduction (feature selection) methods facilitates the identification of a stable and influential subset of predictors that play a decisive role in soil pH modeling. By leveraging the complementary strengths of different algorithms, this combined approach improves predictive accuracy while substantially reducing uncertainty associated with variable selection (Ferhatoglu and Miller, 2022).

Analysis of the results indicates that, among all examined variables, ten predictors were consistently selected by multiple variable selection methods, highlighting their importance and high stability in predicting soil pH variability (Guo et al., 2022). In contrast, variables identified by only a single method likely reflect differences in algorithm sensitivity to multicollinearity and nonlinear interactions among predictors (Bogaert et al., 2023). The primary contribution to pH modeling comes from topographic indices, emphasizing the central role of geomorphological dynamics in regulating soil chemical processes (Wadoux, 2019). These indices, which capture land surface shape, slope, and roughness, directly influence surface water flow, moisture retention, and erosion intensity, thereby affecting dissolution, leaching, and ion exchange processes that govern soil pH (Hosseini et al., 2023; Dindaroglu et al., 2022). For instance, the LUCI, a key morphometric parameter, identifies areas of concentrated flow and mineral deposition, locations where chemical reactions are more intense and pH variations are typically greater (Ferhatoglu and Miller, 2022).

The MRRTF, which quantifies the degree of topographic roughness and discontinuity, indirectly influences soil pH by affecting soil permeability, subsurface flow, and effective drainage (Hoffmann et al., 2014; Ren et al., 2019). Collectively, these findings underscore the importance of multi-objective approaches in variable selection and the necessity of integrating topographic, vegetation, and hydrological indices for more accurate modeling of soil chemical properties (Ferhatoglu and Miller, 2022; Isong et al., 2022). Among the variable groups, hydrological and water flow indices rank second in importance for predicting soil pH variability. These indices capture the processes of horizontal material transport and accumulation across the land surface and reflect the hydrogeomorphological dynamics of landscape (Sylvain et al., 2021). Water flow pathways, by redistributing chemicals and dissolved ions, facilitate the movement of H^+ and OH^- ions across spatial scales, thereby exerting a significant influence on the spatial patterns of soil pH (Sediqi et al., 2025).

Among these indicators, the OFD, which represents potential water movement and surface solute transport pathways, and the SCA, which reflects the effective area for runoff and dissolved substance accumulation, have demonstrated consistent importance and a pivotal role in the spatial regulation of soil pH (Poggio et al., 2021; Wadoux, 2019). These variables are critical for maintaining soil chemical balance by regulating soil moisture regimes, influencing redox processes, and controlling the accumulation or leaching of alkaline and acidic compounds (Ferhatoglu and Miller, 2022; Isong et al., 2022). However, differences in the selection of other hydrological variables across algorithms highlight their sensitivity to optimization criteria, multicollinearity, and complex interactions among environmental factors (Schmidt et al., 2024). consequently, variable selection approaches that analyze the overlap of variables identified by multiple independent methods can produce a stable, interpretable, and environmentally meaningful set of predictors (Ferhatoglu and Miller, 2022). This multi-algorithm strategy not only improves prediction accuracy but also enhances model generalizability in heterogeneous landscapes (Wadoux, 2019).

Table 5
The number of repetitions variable in methods.

Variable	The number of repetitions	Variable type
SL	4	Topo-related variables
LUC	3	Topo-related variables
MRRTF	3	Topo-related variables
DDR	3	Hydrology-related variables
PVI	2	Vegetation-related variables
SCA	2	Topo-related variables
CNR	2	Geological
TST	2	Topo-related variables
N	2	Topo-related variables
OFD	2	Hydrology-related variables

3.5. Evaluation of Variable Selection Methods in MLA

The assessment of variable selection methods across the four MLA revealed notable differences in calibration accuracy, generalization capability, and sensitivity to uncertainty.

3.5.1. Support Vector Machine algorithm

In the Support Vector Machine (SVM) model, calibration results were moderate for most variable selection methods. The coefficient of determination (R^2) during calibration ranged from 0.44 for GA-FS1 to 0.61 for SA-FS2, with the concordance index showing a similar trend. This pattern indicates a tendency of the SVM model to overfit training data and demonstrates that existing variable selection strategies are insufficient for maintaining model stability on independent datasets (Poggio et al., 2021; Guo et al., 2022). The substantial increases in mean square error (MSE) and root mean square error (RMSE), along with observed negative bias, confirm that the SVM model systematically underestimates out-of-sample predictions (Table 6).

This behavior is attributed to the strong of model dependence on training data and its sensitivity to kernel choice, as well as the C and γ hyperparameters (Rainio et al., 2024). Recent studies have shown that using a genetic algorithm (GA) for variable selection can improve the performance of the SVM model to some extent, because GA is able to adaptively navigate the search space and identify a more optimal combination of variables (Wutzl et al., 2019). However, in the absence of fine-tuning of parameters and appropriate cross-validation, this combination is also prone to overfitting (Guo et al., 2022; Correa et al., 2025). Overall, the results suggest that although SVM can achieve high accuracy during the training phase, its performance in validation is often poor unless variables are optimally selected and parameters finely tuned. Therefore, integrated strategies combining intelligent variable selection with multi-stage model calibration are recommended to improve reliability in predicting soil properties such as pH.

3.5.2. Random Forest (RF) Algorithm

In the RF model, the highest calibration accuracy among all algorithms was observed. Using the GA-FS1 method, the coefficient of determination (R^2) reached 0.71 and the concordance index approached 0.80, highlighting the strong capability of RF to fit training data. However, in the validation phase, a slight decrease in performance was noted, with R^2 and concordance values falling to moderate ranges, indicating mild overfitting and relative sensitivity to the structure of the training dataset (Nenkam et al., 2024). Compared to other algorithms, RF demonstrated more stable performance on independent data, particularly when using the VIF and SA-FS2 variable selection methods, both achieving $R^2 \approx 0.37$ and establishing a reasonable balance between accuracy and generalizability. In contrast, GA-FS1 and GA-FS2 performed less effectively on validation data, with R^2 values of 0.27 and 0.26, respectively (Table 6).

These results suggest that while RF models data effectively during calibration, careful tuning of key parameters, such as the number of trees (ntree) and the number of variables per split (mtry), is necessary to maintain stability on independent datasets (Poggio et al., 2021). Multiple studies have emphasized the superiority of RF over SVM in predicting soil properties. Emadi et al. (2020) demonstrated that in northern Iran, RF achieved higher accuracy in SOC prediction than SVM by effectively incorporating topographic and climatic factors. The RF model has a high ability to model nonlinear relationships and complex interactions between soil and environmental variables (Ferhatoglu and Miller., 2022). Overall, these findings indicate that the RF model, when paired with appropriate variable selection methods, especially VIF and SA-FS2, provides a

favorable balance between accuracy, stability, and generalizability, making it a reliable option for modeling soil properties.

3.5.3. Cubist Model Algorithm

The Cubist model exhibited well performance during the calibration phase, with R^2 values ranging from 0.67 to 0.68 and concordance indices between 0.80 and 0.82. However, performance decreased markedly the validation phase. For certain variable selection methods, such as GA-FS2, R^2 dropped to 0.08, while the VIF and SA-FS2 methods reached only $R^2 \approx 0.26$. Concordance indices ranged from 0.37 to 0.46, indicating only partial agreement between predictions and observed data, and highlighting the model's sensitivity to new data (Table 6). This pattern suggests that, despite its high calibration accuracy, Cubist exhibits overfitting in the training data (Reddy et al., 2024). The results indicate that Cubist tends to be less stable on validation datasets, even when its calibration RMSE is lower (Adeniyi et al., 2023). Overall, the literature emphasizes that Cubist's strengths lie in its extrapolation capability and high training accuracy, whereas its primary limitations are instability on independent data and sensitivity to model parameters (Correa et al., 2025). Therefore, in practical applications, it is recommended to implement multi-stage cross-validation, optimize pruning and committee parameters, and combine Cubist with robust variable selection methods to enhance model stability and generalizability.

3.5.4. XGBoost model Algorithm

The XGBoost model performed well in the calibration phase, with R^2 values between 0.50 (SA-FS1) and 0.68 (VIF). Compared to other models, the generalization power of XGBoost was more favorable in the validation phase. With the set of variables selected by the SA-FS1 algorithm, this model showed the best performance among all models and variable selection methods, which reached $R^2 = 0.62$ and Concordance = 0.74 (Table 6). These findings indicate that random search-based variable selection can identify the main patterns of the data well while maintaining the generalizability of the model to new data. In contrast, methods such as RFE and VIF, despite high calibration accuracy, performed poorly in validation, indicating their sensitivity to overfitting and limitations in generalization (Chen et al., 2022).

Table 6
Validation of machine learning models using different feature selection methods.

Variable reduction method	Statistical parameters	Calibration data				Validation data			
		Models				Models			
		SVM	RF	Cubist	XGBoost	SVM	RF	Cubist	XGBoost
VIF	R^2	0.55	0.65	0.67	0.68	0.15	0.37	0.22	0.26
	concordance	0.71	0.74	0.80	0.80	0.34	0.43	0.42	0.45
	MSE	0.07	0.04	0.06	0.04	0.13	0.15	0.16	0.15
	RMSE	0.27	0.21	0.24	0.21	0.36	0.38	0.39	0.38
	bias	0.01	0	0	0.01	-0.06	0.08	-0.09	-0.09
RFE	R^2	0.51	0.66	0.68	0.65	0.19	0.32	0.24	0.02
	concordance	0.68	0.76	0.80	0.79	0.39	0.42	0.43	0.36
	MSE	0.08	0.04	0.05	0.044	0.11	0.15	0.17	0.199
	RMSE	0.29	0.21	0.24	0.21	0.33	0.39	0.41	0.43
	bias	0.02	0.003	0.006	0.005	0.02	0.08	-0.105	0.08
SA-FS1	R^2	0.56	0.69	0.68	0.50	0.11	0.32	0.17	0.62
	concordance	0.72	0.77	0.81	0.68	0.27	0.39	0.37	0.74
	MSE	0.07	0.04	0.06	0.072	0.14	0.14	0.16	0.075
	RMSE	0.27	0.20	0.24	0.27	0.38	0.38	0.39	0.27
	bias	0.002	-0.004	0.005	-0.05	-0.004	0.09	-0.08	0.033
SA-FS2	R^2	0.61	0.67	0.70	0.65	0.18	0.37	0.26	0.21
	concordance	0.75	0.76	0.82	0.78	0.37	0.43	0.46	0.38
	MSE	0.06	0.04	0.05	0.045	0.13	0.14	0.13	0.18
	RMSE	0.25	0.20	0.22	0.21	0.35	0.38	0.36	0.43
	bias	-0.00	0.00	-0.00	-0.002	-0.00	0.08	-0.10	0.07
GA-FS1	R^2	0.44	0.71	0.44	0.63	0.18	0.27	0.18	0.45
	concordance	0.62	0.80	0.62	0.77	0.37	0.44	0.37	0.57
	MSE	0.08	0.04	0.08	0.047	0.14	0.10	0.14	0.12
	RMSE	0.29	0.21	0.29	0.22	0.38	0.32	0.38	0.35
	bias	0.04	-0.00	0.04	-0.001	0.03	-0.09	0.03	0.06

GA-FS2	R^2	0.51	0.64	0.68	0.58	0.06	0.26	0.08	0.31
	concordance	0.68	0.74	0.81	0.74	0.25	0.41	0.29	0.44
	MSE	0.07	0.05	0.05	0.052	0.19	0.11	0.21	0.15
	RMSE	0.27	0.24	0.22	0.23	0.44	0.33	0.46	0.39
	bias	0.022	0	0	-0.001	0.06	-0.08	-0.05	0.062

XGBoost ability to handle unbalanced data, represent complex and nonlinear relationships, and employ iterative pruning of trees improves performance and reduces overfitting Alonso-Sarria et al., 2025). In contrast, RF makes predictions based on the sum of the outputs of all trees, where each tree has its own unique conditions and the probability of random prediction is higher (Chen et al., 2022).

Some studies have reported that XGBoost is sometimes less accurate than RF in SOM prediction; these differences are mainly due to different study designs, sample sizes, validation strategies, and model parameter configurations (Theodoridis and Gkikas., 2020). The results of this study emphasize that variable selection plays a critical role in model performance. Using various variable reduction techniques to optimize high-dimensional variables greatly increases the prediction accuracy. These findings highlight the importance of choosing the appropriate variable reduction method in achieving reliable predictions with machine learning models and high-dimensional data, and provide a valuable reference for large-scale digital soil mapping.

3.6. Examining the differences between the results of algorithms

In this section, a key question is raised: are the observed differences driven by the algorithms themselves, or are they merely an artifact of the small study area? More specifically, this study investigates how variable selection is influenced by the learning algorithm and what this reveal about model behavior.

The results indicate that differences among models are not solely due to the limited spatial extent, but are largely related to algorithm characteristics and their interaction with feature selection methods. Although a small study area can increase variance and sensitivity, the consistent patterns across all feature selection methods suggest that each algorithm responds differently to the structure of the feature space. In other words, models react differently to variable removal, which cannot be explained only by spatial limitations.

The findings show that feature selection effects are strongly algorithm-dependent. SVM was the most sensitive model to variable selection. Its performance, especially under deterministic methods such as RFE and VIF, was unstable, with a clear gap between calibration and validation. This indicates strong dependence on feature space structure, kernel choice, and parameter tuning, where removing variables may eliminate important nonlinear relationships. In contrast, stochastic methods such as SA-FS2 improved SVM stability by preserving more complex interactions.

RF showed a different behavior, being more robust to changes in predictor sets and maintaining a reasonable balance between accuracy and generalization, even when some feature selection methods performed poorly. This stability is likely due to its ensemble structure and random sampling, which reduce dependence on

specific variables. However, RF still showed performance degradation with some inappropriate subsets (e.g., GA-FS combinations), particularly in validation.

Cubist exhibited the largest gap between calibration and validation, indicating high sensitivity to training data structure. Its strong training performance and weak generalization suggest overfitting to local patterns, especially when feature selection reduces environmental heterogeneity.

In contrast, XGBoost showed the most consistent behavior and achieved the best generalization when combined with SA-FS1. This suggests that boosting methods are sensitive to feature space composition but can better reconstruct environmental relationships when nonlinear interactions are preserved. The weaker performance of deterministic methods like RFE in XGBoost further indicates that stepwise elimination may remove complementary information needed for boosting.

Overall, there is no universally “best” feature selection method independent of the algorithm. Its effectiveness depends on model architecture, nonlinear modeling capacity, and sensitivity to feature space structure. Stochastic and global search-based methods (SAFS and GAFS) are generally more compatible with complex models, especially XGBoost, while deterministic methods may oversimplify the feature space and reduce generalization. This is a key implication for digital soil mapping and high-dimensional environmental modeling.

4. Conclusion

This study evaluated feature selection methods and machine learning algorithms for predicting soil pH within a digital soil mapping (DSM) framework. The findings indicate that feature selection is not an algorithm-independent process; rather, it is a model-dependent mechanism that strongly influences prediction accuracy, model stability, and uncertainty. A key strength of this study is the comprehensive comparison of six feature selection strategies, including filter-, wrapper-, and stochastic-based approaches, combined with four machine learning algorithms. This multi-algorithm framework enabled a robust assessment of how different models respond to variations in the feature space and helped identify a stable subset of environmental predictors consistently controlling soil pH variability.

The results showed that stochastic and global-search-based methods, particularly SA and GA variants, generally performed better with complex models such as XGBoost and RF. In contrast, SVM and Cubist were highly sensitive to feature selection and frequently exhibited overfitting and weaker generalization. Among all combinations, XGBoost coupled with SA-FS1 demonstrated the most stable and transferable performance. These findings confirm that no single feature selection method is universally optimal; instead, predictive performance depends on the interaction between the learning algorithm and the structure of the feature space.

Several limitations should be acknowledged. The relatively small spatial extent of the study area may have increased model variance and limited the transferability of the findings to larger or more heterogeneous regions. Climatic variables were also excluded because of their limited spatial variability, which may have reduced the completeness of environmental representation. In addition, the stochastic nature of some feature selection methods, such as GA and SA, may lead to variation in the predictors selected across repeated runs.

Future research directions are substantial. Extending this framework to larger and more climatically diverse regions would help test the scalability of the findings. Integrating hybrid feature selection strategies, such as SA–GA–RFE combinations or group-based frameworks, may further enhance robustness. Explicitly incorporating uncertainty quantification into feature selection could also provide deeper insight into model reliability. Overall, effective DSM modeling depends not only on accurate machine learning algorithms but also on carefully designed feature selection strategies, whose interaction is essential for achieving reliable, generalizable, and interpretable soil pH predictions in high-dimensional environmental datasets.

Declarations

Declaration of interests

Ayeh Lotfollahi reports that financial support for this work was provided by the Iranian National Science Foundation. All other authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Ethical approval

This is not applicable.

Consent to participate

This is not applicable.

Consent to publish

All authors have given their permission for publishing this work.

Competing interests

The authors declare no competing interests.

Permissions Statement

No specific permission was required for conducting this study and collecting samples.

Funding

This work was financially supported by the Iran National Science Foundation (INSF), the Ministry of Science, Research and Technology, and the Agricultural Research, Education and Extension Organization (AREEO), and was conducted under project No. 4045530.

Author Contribution

M.N.N Writing original draft, Edited and optimized the manuscript.Sh.F Data curation, Software, writing original draft, Edited and optimized the manuscript.A.L Conceptualization, Methodology, Investigation, Software, writing original draft, Edited, Restructured, and optimized the manuscript.A.A.B Writing original draft, Edited and optimized the manuscript.

Acknowledgement

The authors sincerely thank Dr. David G. Rossiter and Dr. Gerard BM Heuvelink for their valuable guidance in conducting the analyses and for their assistance in editing and improving the manuscript.

Data Availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

References

1. Adeniyi, O. D., Brenning, A., Bernini, A., Brenna, S. & Maerker, M. Digital mapping of soil properties using ensemble machine learning approaches in an agricultural lowland area of Lombardy, Italy. *Land* **12** (2), 494. https://doi.org/10.3390/land12020494 (2023).
2. Alonso-Sarria, F., Blanco-Bernardeau, A., Gomariz-Castillo, F., Jimenez-Bastida, H. & Romero-Diaz, A. Estimation of soil properties using machine learning techniques to improve hydrological modeling in a semiarid environment: Campo de Cartagena (Spain). *Earth Sci. Inf.* **18** (3), 1–24. https://doi.org/10.1007/s12145-025-01833-w (2025).
3. Al-Shabeeb, A. R., Al-Fugara, A. K., Khedher, K. M., Mabdeh, A. N. & Al-Adamat, R. Spatial mapping of landslide susceptibility in Jerash governorate of Jordan using genetic algorithm-based wrapper feature selection and bagging-based ensemble model. *Geomatics Nat. Hazards Risk*. **13** (1), 2252–2282 (2022).
4. Arif Ali, Z., Abduljabbar, Z. H., Tahir, H., Bibo Sallow, A. & Almufti, S. M. eXtreme gradient boosting algorithm with machine learning: a review. *Acad. J. Nawroz Univ.* **12** (2), 320–334 (2023).
5. Atef, M., Gabr, M. I., Seoud, W. & Ouf, S. Feature Selection and Classification in Machine Learning: Methods and Models for Peer-to-Peer Lending. *MSA-Management Sci. J.* **4** (4), 96–113. https://doi.org/10.21608/msamsj.2025.405941.1121 (2025).
6. Ayalew, D. A., Deumlich, D., Šarapatka, B. & Doktor, D. Quantifying the sensitivity of NDVI-based C factor estimation and potential soil erosion prediction using Spaceborne earth observation data. *Remote Sens.* **12** (7), 1136. https://doi.org/10.3390/rs12071136 (2020).
7. Baltensweiler, A., Heuvelink, G. B., Hanewinkel, M. & Walthert, L. Microtopography shapes soil pH in flysch regions across Switzerland. *Geoderma* **380**, 114663. https://doi.org/10.1016/j.geoderma.2020.114663 (2020).
8. Belkadi, W. H. & Drias, Y. September. Advancements in digital soil mapping: From data acquisition to uncertainty estimation-a comprehensive review. In *Artificial intelligence doctoral symposium* (pp. 162–177). Singapore: Springer Nature Singapore. (2022). https://doi.org/10.1007/978-981-99-4484-2_13

9. Blum, C. & Roli, A. Metaheuristics in combinatorial optimization: Overview and conceptual comparison. *ACM Comput. Surveys*. **35** (3), 268–308. https://doi.org/10.1145/937503.937505 (2003).
10. Bogaert, P., Taghizadeh-Mehrjardi, R. & Hamzehpour, N. Model averaging of machine learning algorithms for digital soil mapping: A minimum variance framework. *Geoderma* **437**, 116604. 10.4067/S0718-34292020000400099 (2023).
11. Bolón-Canedo, V., Sánchez-Maroño, N. & Alonso-Betanzos, A. Feature selection for high-dimensional data. *Progress Artif. Intell.* **4**, 65–75. https://doi.org/10.1007/s13748-015-0065-0 (2015).
12. Bouma, J. Soil science contributions towards sustainable development goals and their implementation. *Geoderma Reg.* **16**, e00201 (2014).
13. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32. https://doi.org/10.1023/A:1010933404324 (2001).
14. Brereton, R. G. & Lloyd, G. R. Support vector machines for classification and regression. *Analyst* **135** (2), 230–267 (2010).
15. Brevik, E. C., Pereg, L., Steffan, J. J. & Burgess, L. C. Soil ecosystem services and human health. *Curr. Opin. Environ. Sci. Health*. **5**, 87–92. https://doi.org/10.1016/j.coesh.2015.04.003 (2015).
16. Chai, T. & Draxler, R. R. Root means square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.* **7** (3), 1247–1250 (2014).
17. Chandrashekar, G. & Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **40** (1), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024 (2014).
18. Chang, J. G. et al. Hydro-Topographic Contribution to In-Field Crop Yield Variation Using High-Resolution Surface and GPR-Derived Subsurface DEMs. *Remote Sensing*, *17*(17), p.3061. (2025).
19. Chen, T. & Guestrin, C. August. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794). (2016).
20. Chen, Y. et al. Comparison of feature selection methods for mapping soil organic matter in subtropical restored forests. *Ecol. Ind.* **135**, 108545. https://doi.org/10.1016/j.ecolind.2022.108545 (2022).
21. Cheng, J., Sun, J., Yao, K., Xu, M. & Cao, Y. A variable selection method based on mutual information and variance inflation factor. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *268*, p.120652. (2022).
22. Cheng, X. A comprehensive study of feature selection techniques in machine learning models. *Available at SSRN 5154947*. (2024).
23. Correa, J., Tavakoli, H., Vogel, S. & Gebbers, R. Overfitting due to data leakage in soil sensor calibration: Examples from lab-based and in-situ soil NIR spectroscopy. *Computers and Electronics in Agriculture*, *239*, p.110920. (2025).
24. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20** (3), 273–297 (1995).
25. Deemy, J. B. et al. Hydrology, geomorphology, and soils: An overview. *Fundamentals of tropical freshwater wetlands*, pp.43–86. (2022).
26. Dindaroglu, T. et al. The use of remote sensing to characterise geomorphometry and soil properties at watershed scale. *Int. J. Global Warming*. **27** (4), 402–421 (2022).
27. Dominati, E., Patterson, M. & Mackay, A. A framework for classifying and quantifying the natural capital and ecosystem services of soils. *Ecol. Econ.* **69**, 1858–1868 (2010).

28. Dunkl, I. & Ließ, M. On the benefits of clustering approaches in digital soil mapping: an application example concerning soil texture regionalization. *Soil* **8** (2), 541–558 (2022).
29. Emadi, M. et al. Predicting and mapping of soil organic carbon using machine learning algorithms in Northern Iran. *Remote Sens.* **12** (14), 2234. https://doi.org/10.3390/rs12142234 (2020).
30. FAO & ITPS. *Status of the World's Soil Resources (SWSR) – Main Report* (Food and Agriculture Organization of the United Nations, 2015).
31. Fatima, S. et al. Xgboost and random forest algorithms: an indepth analysis. *Pak. J. Sci. Res.* **3** (1), 26–31 (2023).
32. Food and Agriculture Organization of the United Nations. *Status of the World's Soil Resources (SWSR) – Main Report* (FAO Intergovernmental Technical Panel on Soils (ITPS), 2015).
33. Ghodmare, P., Jibhakate, M., Meshram, N., Bele, V. & Ghire, P. A review paper on brief introduction of genetic algorithm. *Int. J. Sci. Technol. Eng.* **4** (8), 42–44 (2018).
34. Guo, J., Wang, K. & Jin, S. Mapping of soil pH based on SVM-RFE feature selection algorithm. *Agronomy*, *12*(11), p.2742. (2022).
35. Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3** (Mar), 1157–1182 (2003).
36. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46**, 389–422. https://doi.org/10.1023/A:1012487302797 (2002).
37. Hengl, T., Heuvelink, G. B. M. & Rossiter, D. G. About regression-kriging: From equations to case studies. *Comput. Geosci.* **33** (10), 1301–1315. https://doi.org/10.1016/j.cageo.2007.05.001 (2007).
38. Hengl, T. et al. SoilGrids250m: Global gridded soil information based on machine learning. *PLoS one*. **12** (2), e0169748 (2017).
39. Heuvelink, G. B. Uncertainty and uncertainty propagation in soil mapping and modelling. In Pedometrics (439–461). Cham: Springer International Publishing. (2018).
40. Hoffmann, U., Hoffmann, T., Johnson, E. A. & Kuhn, N. J. Assessment of variability and uncertainty of soil organic carbon in a mountainous boreal forest (Canadian Rocky Mountains, Alberta). *Catena* **113**, 107–121 (2014).
41. Hosseini, F. S. et al. Geospatial artificial intelligence (GeoAI) and satellite imagery fusion for soil physical property predicting. *Sustainability*, *15*(19), p.14125. (2023).
42. Hu, B. et al. Mapping soil organic matter and identifying potential controls in the farmland of Southern China: Integration of multi-source data, machine learning and geostatistics. *Land. Degrad. Dev.* **34** (17), 5468–5485 (2023).
43. Isong, I. A., John, K., Okon, P. B., Ogban, P. I. & Afu, S. M. Soil quality estimation using environmental covariates and predictive models: An example from tropical soils of Nigeria. *Ecological Processes*, *11*(1), p.66. (2022).
44. Jakkula, V. Tutorial on support vector machine (svm). *School EECS Wash. State Univ.* **37** (2.5), 3 (2006).
45. Jiang, S. H., Huang, J., Griffiths, D. V. & Deng, Z. P. Advances in reliability and risk analyses of slopes in spatially variable soils: A state-of-the-art review. *Computers and Geotechnics*, *141*, p.104498. (2022).

46. Keesstra, S. D. et al. The significance of soils and soil science towards realization of the United Nations Sustainable Development Goals. *SOIL* **2**, 111–128 (2016).
47. Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. Optimization by simulated annealing. *Science* **220** (4598), 671–680. https://doi.org/10.1126/science.220.4598.671 (1983).
48. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **28** (5), 1–26. https://doi.org/10.18637/jss.v028.i05 (2008).
49. Kuhn, M. & Johnson, K. *Applied Predictive Modeling* (Springer, 2013).
50. Kuhn, M. Modern Rule-Based Models. (2020).
51. Li, J. et al. Application of XGBoost algorithm in the optimization of pollutant concentration. *Atmospheric Research*, *276*, p.106238. (2022).
52. Lin, S. W., Tseng, T. Y., Chou, S. Y. & Chen, S. C. A simulated-annealing-based approach for simultaneous parameter optimization and feature selection of back-propagation networks. *Expert Syst. Appl.* **34** (2), 1491–1499 (2008).
53. Machado, D. F. T. Accuracy and uncertainty of digital soil mapping approaches to extract and transfer soil information from reference area. (2017).
54. McBratney, A. B., Mendonça Santos, M. L. & Minasny, B. On digital soil mapping. *Geoderma* **117**, 3–52. https://doi.org/10.1016/S0016-7061(03)00223-4 (2003).
55. Minasny, B. & McBratney, A. B. A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Comput. Geosci.* **32** (9), 1378–1388 (2006).
56. Minasny, B. & McBratney, A. B. Digital soil mapping: A brief history and some lessons. *Geoderma* **264**, 301–311. https://doi.org/10.1016/j.geoderma.2015.07.017 (2016).
57. Montanarella, L. et al. World's soils are under threat. *SOIL* **2**, 79–82. https://doi.org/10.5194/soil-2-79-2016 (2016).
58. Moore, I. D., Grayson, R. B. & Ladson, A. R. Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. *Hydrol. Process.* **5** (1), 3–30 (1991).
59. Moser, A. & Narasimha Murty, M. April. On the scalability of genetic algorithms to very large-scale feature selection. In Workshops on Real-World Applications of Evolutionary Computation (77–86). Berlin, Heidelberg: Springer Berlin Heidelberg. (2000).
60. Mukhopadhyay, D. M., Balitanas, M. O., Farkhod, A., Jeon, S. H. & Bhattacharyya, D. Genetic algorithm: A tutorial review. *Int. J. grid Distrib. Comput.* **2** (3), 25–32 (2009).
61. Naimi, S., Ayoubi, S., Zeraatpisheh, M. & Dematte, J. A. M. Ground observations and environmental covariates integration for mapping of soil salinity: a machine learning-based approach. *Remote Sensing*, *13*(23), p.4825. (2021).
62. Nenkam Mentho, A. et al. Y. and Applications and Challenges of Digital Soil Mapping in Africa. (2024).
63. Nicodemus, K. K., Malley, J. D., Strobl, C. & Ziegler, A. The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC bioinformatics*, *11*(1), p.110. (2010).
64. Nikraftar, Z., Rajabi-Kiasari, S. & Seydi, S. T. Genetic algorithm-based feature selection for landslide susceptibility mapping in northern Iran. *Int. Archives Photogrammetry Remote Sens. Spat. Inform. Sci.* **42**, 821–825 (2019).

65. Nzekwe, C. J., Kim, S. & Mostafa, S. A. Interaction selection and prediction performance in high-dimensional data: a comparative study of statistical and tree-based methods. *J. Data Sci.* https://doi.org/10.6339/24-jds1127 (2024).
66. O'Brien, R. M. A caution regarding rules of thumb for variance inflation factors. *Qual. Quantity*. **41**, 673–690. https://doi.org/10.1007/s11135-006-9018-6 (2007).
67. Park, S. J. & Burt, T. P. Identification and characterization of pedogeomorphological processes on a hillslope. *Soil Sci. Soc. Am. J.* **66** (6), 1897–1910 (2002).
68. Petrov, G., Chanev, M. & Filchev, L. September. Satellite-derived vegetation indices used in archaeological research. In *Eleventh International Conference on Remote Sensing and Geoinformation of the Environment (RSCy2025)* (Vol. 13816, pp. 685–693). SPIE. (2025).
69. Quinlan, J. R. November. Learning with continuous classes. In *5th Australian joint conference on artificial intelligence* (Vol. 92, pp. 343–348). (1992).
70. Rainio, O., Teuho, J. & Klén, R. Evaluation metrics and statistical tests for machine learning. *Scientific reports*, *14*(1), p.6086. (2024).
71. Ramezan, C. A. Transferability of recursive feature elimination (RFE)-derived feature sets for support vector machine land cover classification. *Remote Sensing*, *14*(24), p.6218. (2022).
72. Reddy, B. B. et al. Predictive soil mapping using random forest models: applications in pH and soil organic matter assessment. (2024).
73. Ren, H. et al. Determining landscape-level drivers of variability for over fifty soil chemical elements. *Sci. Total Environ.* **657**, 279–286 (2019).
74. Sadik, L. & Samui, P. Uncertainty-aware prediction of bearing capacity of shallow foundations resting on cohesionless soils using Bayesian regression. *Geotech. Geol. Eng.* **42** (5), 3919–3937 (2024).
75. Samuel-Rosa, A., Heuvelink, G. B. M., Vasques, G. M. & Anjos, L. H. C. Do more detailed environmental covariates deliver more accurate soil maps? Geoderma, **243–244**, 214–227. (2015). https://doi.org/10.1016/j.geoderma.2014.12.017
76. Schmidinger, J. & Heuvelink, G. B. *Validation of uncertainty predictions in digital soil mapping* Vol. 437, p.116585 (Geoderma, 2023).
77. Schmidt, M. et al. Filling the gaps in soil data: A multi-model framework for addressing data gaps using pedotransfer functions and machine-learning with uncertainty estimates to estimate bulk density. *Catena*, *245*, p.108310. (2024).
78. Sediqi, A. S., Naseri, A. K. & Haidari, M. D. Digital soil mapping: Challenges and future prospects. *ESRJ* **63** (2), 61–67 (2025).
79. Seifi, A., Ehteram, M., Singh, V. P. & Mosavi, A. Modeling and uncertainty analysis of groundwater level using six evolutionary optimization algorithms hybridized with ANFIS, SVM, and ANN. *Sustainability*, **12**(10), p.4023. (2020). https://doi.org/10.3390/su12104023
80. Seijo-Pardo, B., Porto-Díaz, I., Bolón-Canedo, V. & Alonso-Betanzos, A. Ensemble feature selection: Homogeneous and heterogeneous approaches. *Knowl. Based Syst.* **118**, 124–139 (2017).
81. Senanayake, S., Pradhan, B., Huete, A. & Brennan, J. A review on assessing and mapping soil erosion hazard using geo-informatics technology for farming system management. *Remote sensing*, *12*(24), p.4063. (2020).

82. Sylvain, J. D., Anctil, F. & Thiffault, É. Using bias correction and ensemble modelling for predictive mapping and related uncertainty: A case study in digital soil mapping. *Geoderma*, **403**, p.115153. (2021). https://doi.org/10.1016/j.geoderma.2021.115153
83. Szatmári, G. & Pásztor, L. Comparison of various uncertainty modelling approaches based on geostatistics and machine learning algorithms. *Geoderma*, *337*, pp.1329–1340. (2019). https://doi.org/10.1016/j.geoderma.2018.09.008
84. Theodoridis, P. K. & Gkikas, D. C. Optimal feature selection for decision trees induction using a genetic algorithm wrapper-a model approach. In Strategic Innovative Marketing and Tourism: 8th ICSIMAT, Northern Aegean, Greece, 2019 (583–591). Cham: Springer International Publishing. (2020).
85. Veneros, J. et al. Application of remote sensors for the analysis of vegetation cover and water bodies. (2020).
86. Wadoux, A. M. C. Using deep learning for multivariate mapping of soil with quantified uncertainty. *Geoderma* **351**, 59–70 (2019).
87. Wadoux, A. M. C., Minasny, B. & McBratney, A. B. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. *Earth-Science Reviews*, *210*, p.103359. (2020).
88. Waysi, D., Ahmed, B. T. & Ibrahim, I. M. Optimization by nature: A review of genetic algorithm techniques. *The Indonesian J. Comput. Science*, **14**(1). (2025).
89. Wutzl, B. et al. Genetic algorithms for feature selection when classifying severe chronic disorders of consciousness. *PloS one*, **14**(7), e0219683. (2019).
90. Zerouali, B. et al. Enhancing deep learning-based slope stability classification using a novel metaheuristic optimization algorithm for feature selection. Scientific Reports, 14(1), p.21812. (2024).
91. Zhang, P., Yin, Z. Y. & Jin, Y. F. Machine learning-based modelling of soil properties for geotechnical design: review, tool development and comparison. *Arch. Comput. Methods Eng.* **29** (2), 1229–1245 (2022).
92. Zhou, J. M. The relationship between soil pH and geochemical components. *Environmental Earth Sciences*, *83*(13), p.402. (2024).

Figures

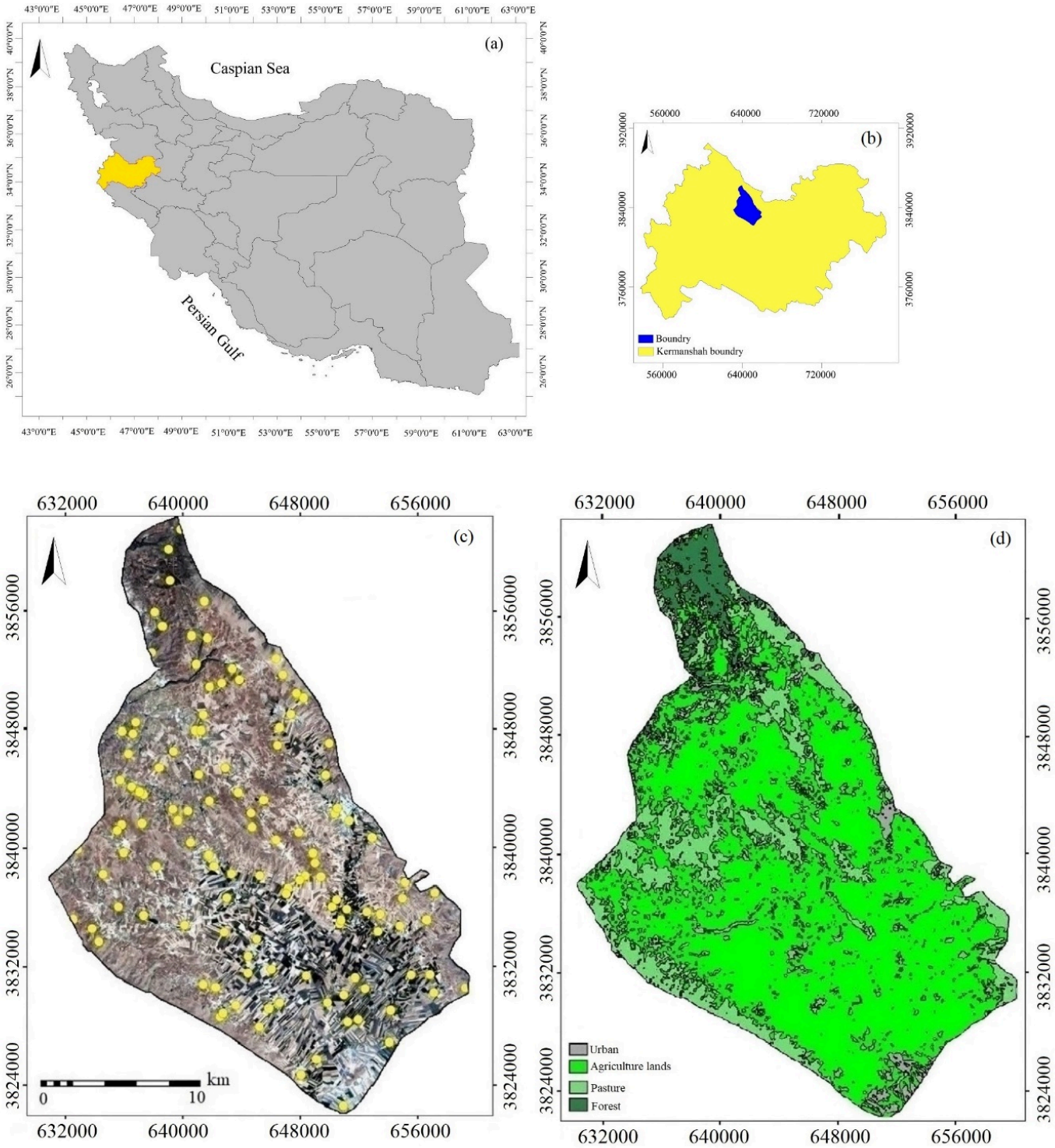


Figure 1

(a) Location of Kermanshah Province within Iran; (b) location of the study area within Kermanshah Province; (c) spatial distribution of soil sampling points; and (d) land-use map of the study area (UTM projection, and Zone 38).

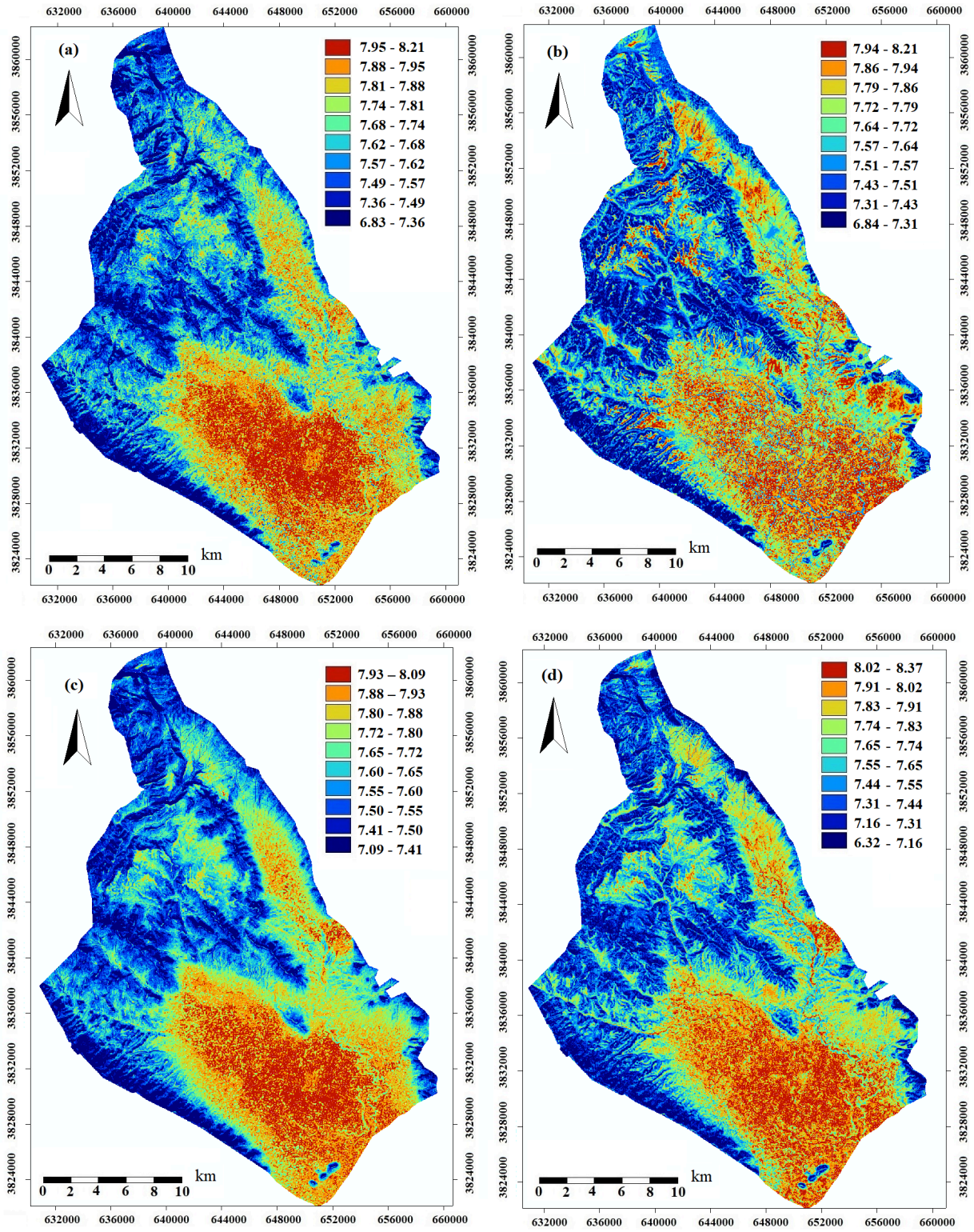


Figure 2

Spatial prediction maps of soil pH produced by (a) XGBoost, (b) SVM, (c) RF, and (d) Cubist models using the VIF method.

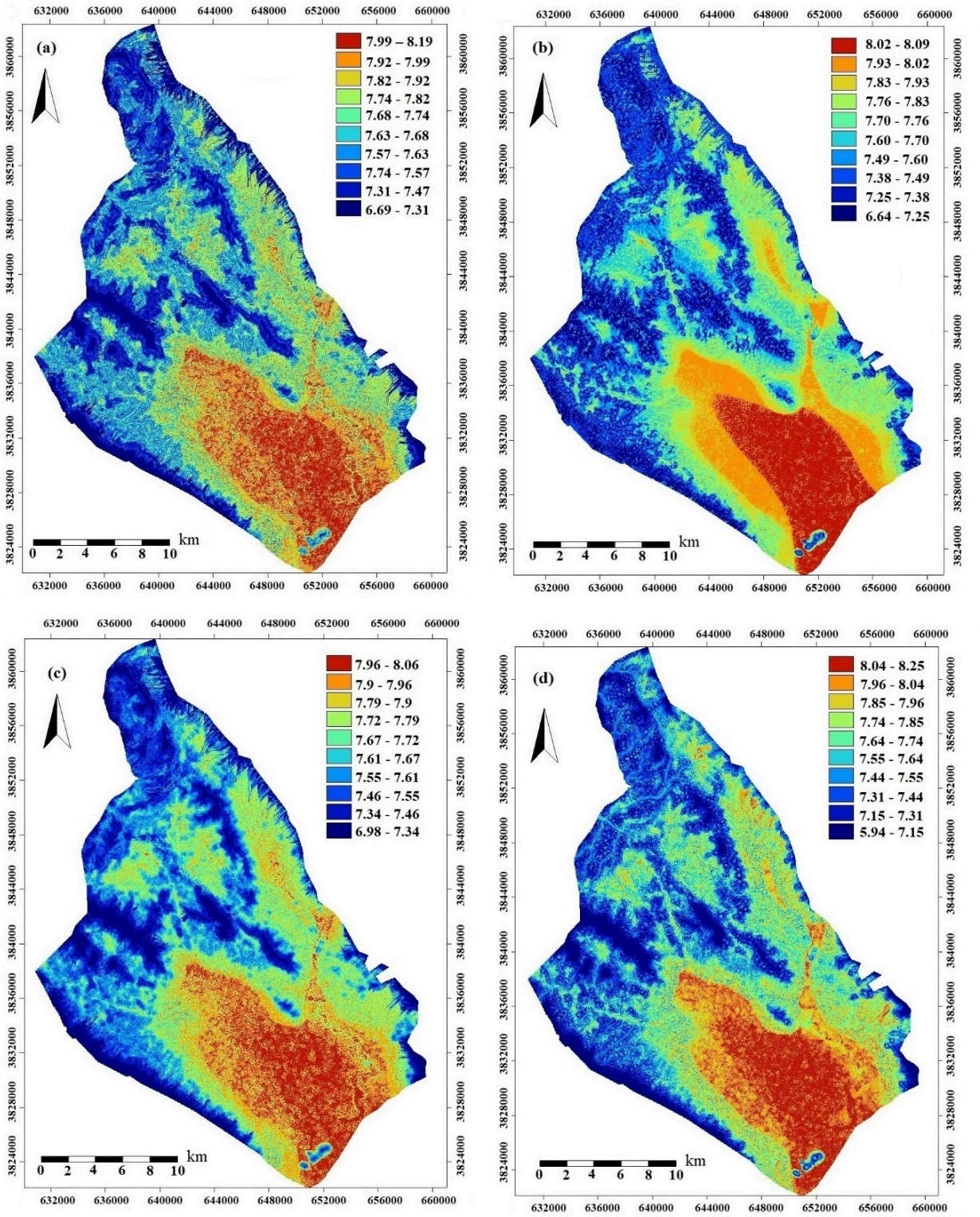


Figure 3

Spatial prediction maps of soil pH produced by (a) XGBoost, (b) SVM, (c) RF, and (d) Cubist models using the RFE method.

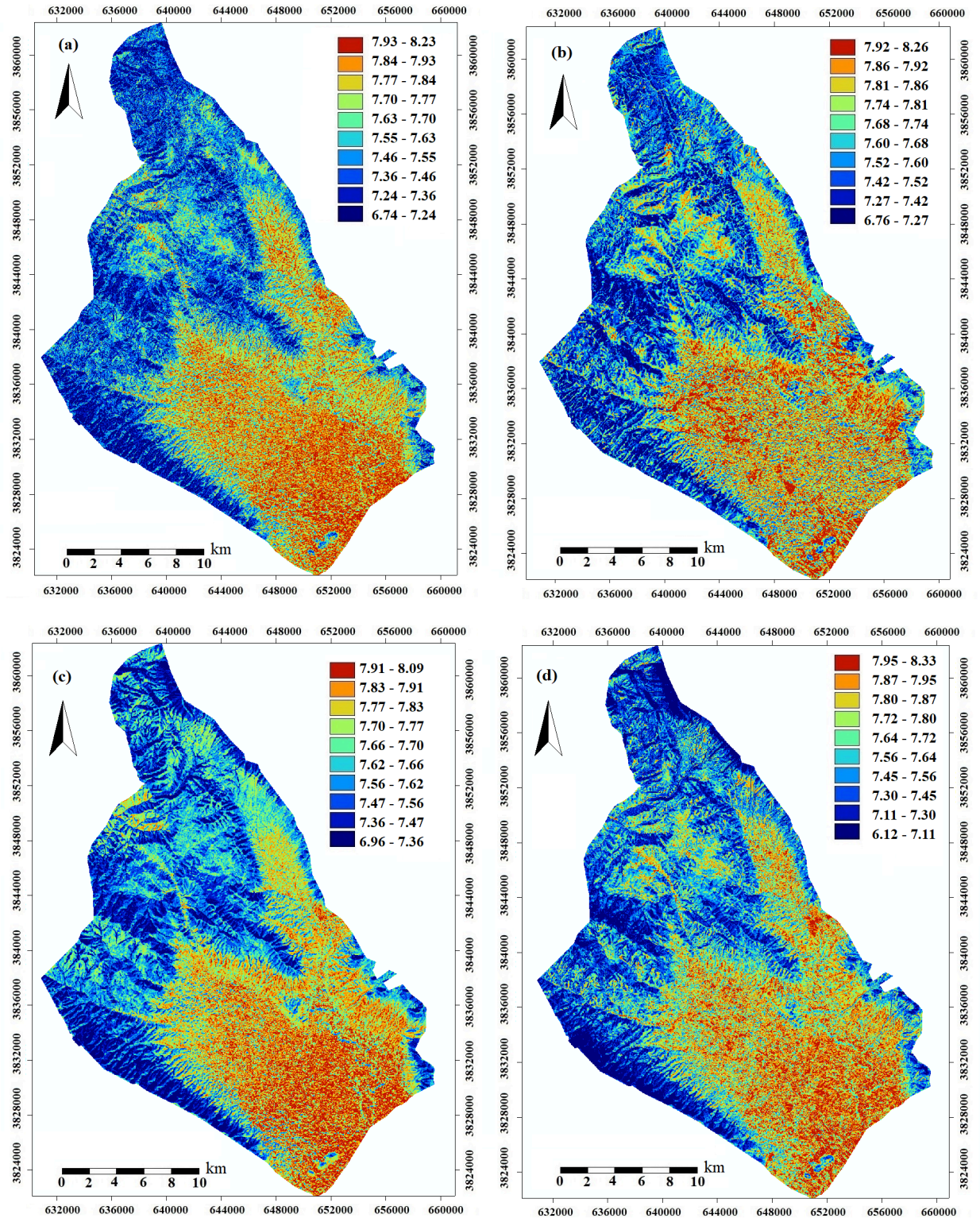


Figure 4

Spatial prediction maps of soil pH produced by (a) XGBoost, (b) SVM, (c) RF, and (d) Cubist models using the SA-FS1 method.

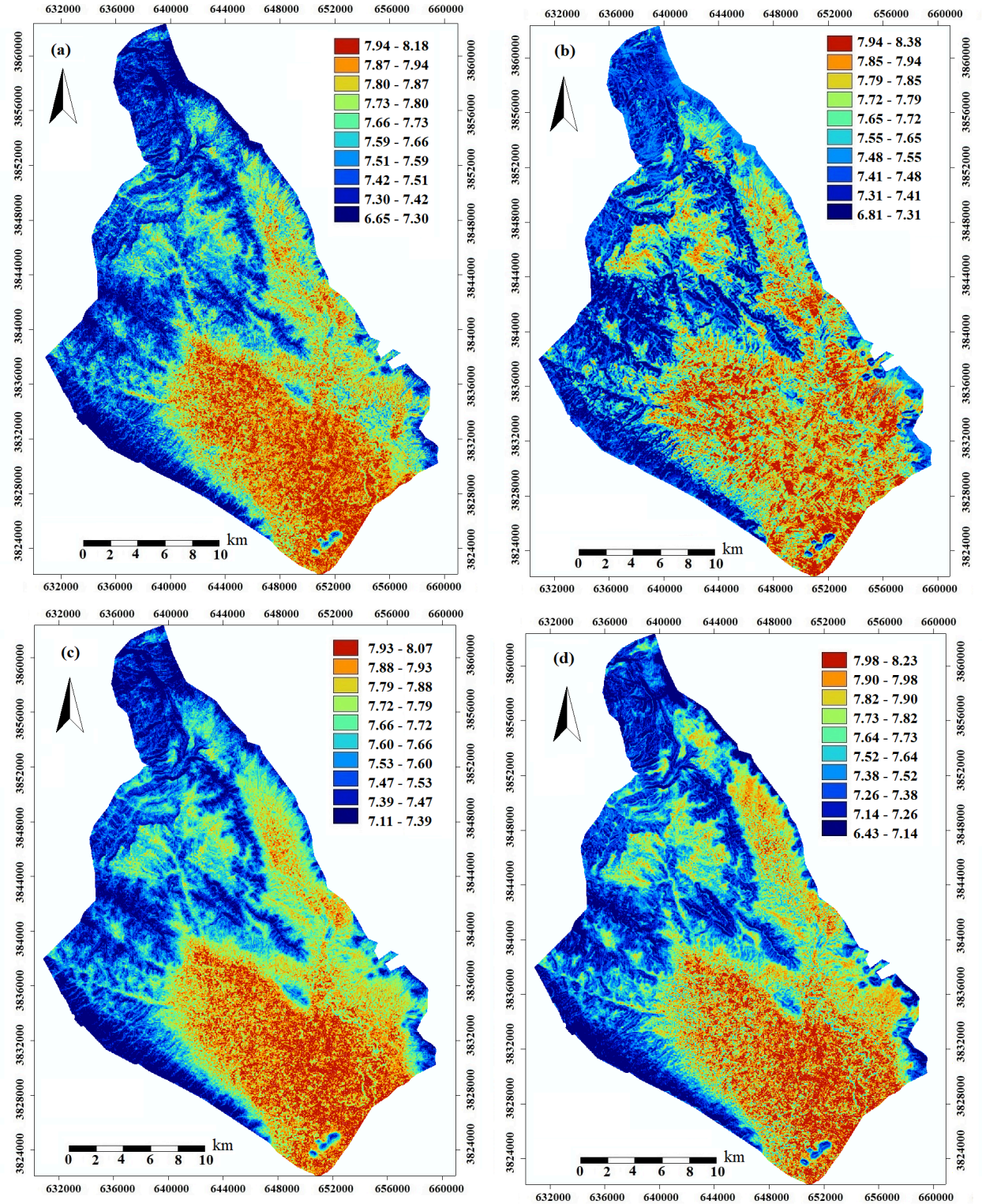


Figure 5

Spatial prediction maps of soil pH produced by (a) XGBoost, (b) SVM, (c) RF, and (d) Cubist models using the SA-FS2 method.

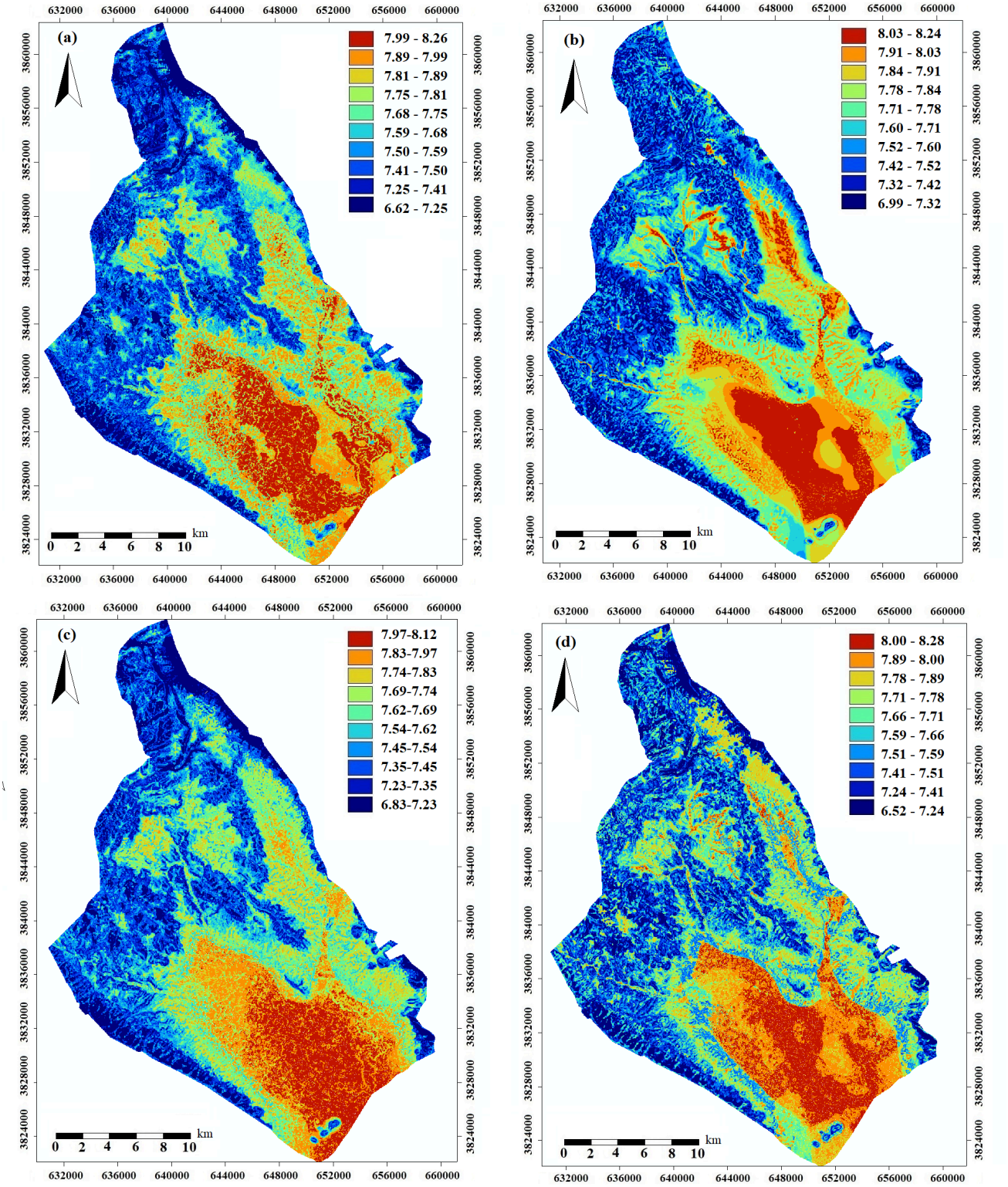


Figure 6

Spatial prediction maps of soil pH produced by (a) XGBoost, (b) SVM, (c) RF, and (d) Cubist models using the GA-FS1 method.

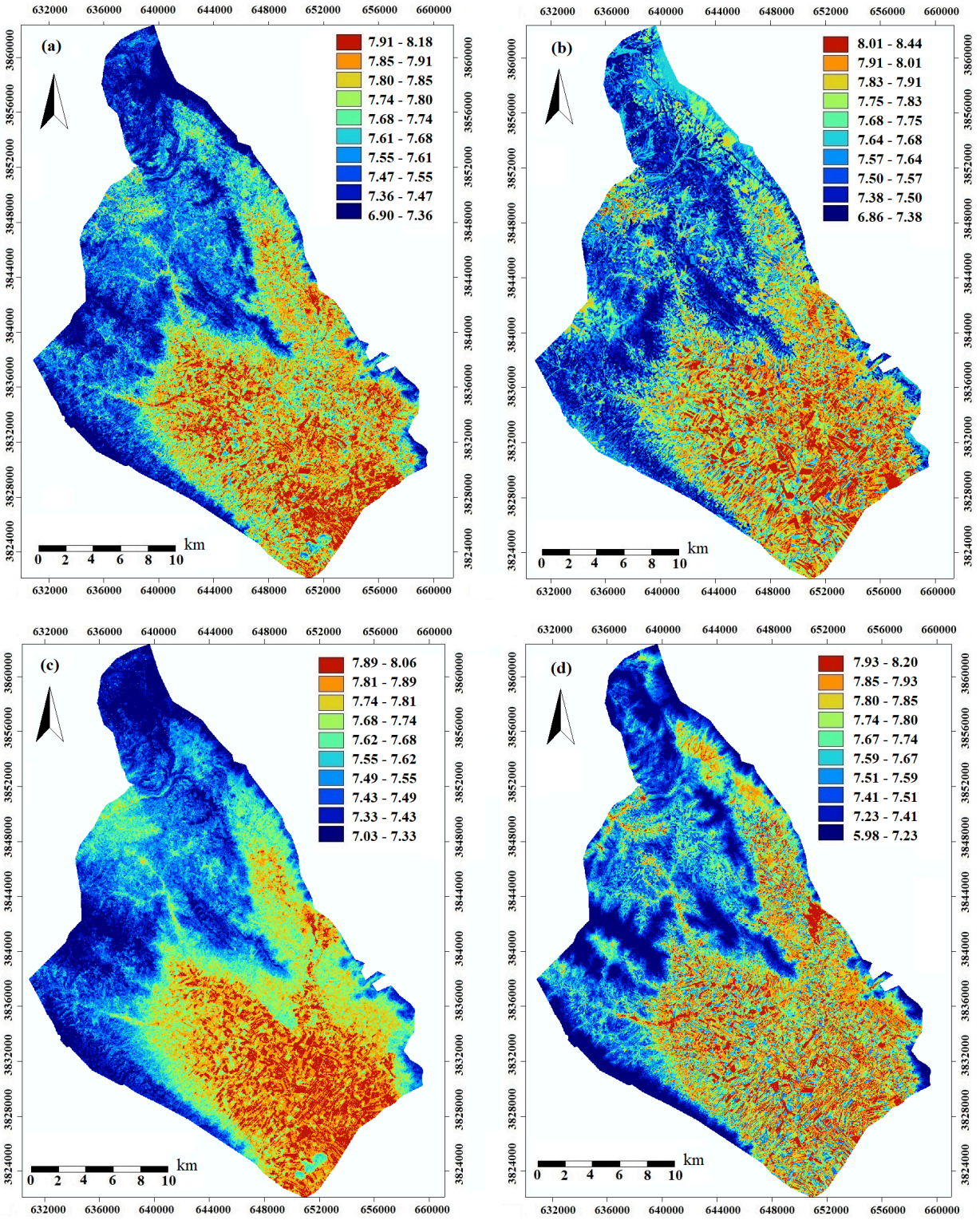


Figure 7

Spatial prediction maps of soil pH produced by (a) XGBoost, (b) SVM, (c) RF, and (d) Cubist models using the GA-FS2 method.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterialTable.docx
- SupplementaryMaterialFig.docx