

Research Article

Cite this article: Dastres E, Shafiee Sarvestani G, Edalat M, Pourghasemi HR (2025). Habitat suitability modeling of dominant weed in canola (*Brassica napus*) fields using machine learning techniques. *Weed Sci.* **73**(e35), 1–15. doi: [10.1017/wsc.2025.5](https://doi.org/10.1017/wsc.2025.5)

Received: 14 June 2024
Revised: 11 November 2024
Accepted: 21 January 2025

Associate Editor:

Muthukumar V. Bagavathiannan, Texas A&M University

Keywords:

Ecological modeling; habitat suitability; machine learning; precision agriculture; weed management

Corresponding author:

Mohsen Edalat; Email: edalat@shirazu.ac.ir

Habitat suitability modeling of dominant weed in canola (*Brassica napus*) fields using machine learning techniques

Emran Dastres¹, Ghazal Shafiee Sarvestani¹, Mohsen Edalat²  and Hamid Reza Pourghasemi³

¹Ph.D, Department of Plant Production and Genetics, School of Agriculture, Shiraz University, Fars Province, Iran; ²Associate Professor, Department of Plant Production and Genetics, School of Agriculture, Shiraz University, Fars Province, Iran and ³Professor, Department of Soil Science, School of Agriculture, Shiraz University, Fars Province, Iran

Abstract

Weed infestations have been identified as a major cause of yield reductions in canola (*Brassica napus* L.), a vital oil crop that has gained significant prominence in Iran, especially within Fars Province. Weed management using machine learning algorithms has become a crucial approach within the framework of precision agriculture for enhancing the efficacy and efficiency of weed control strategies. The evolution of habitat suitability models for weeds represents a significant advancement in agricultural technology, offering the capability to predict weed occurrence and proliferation accurately and reliably. This study focuses on the issue of dominant weed infestation in canola cultivation, particularly emphasizing the prevalence and impact of wild oat (*Avena fatua* L.) as the dominant weed species in canola farming in 2023. We collected data on 12 environmental variables related to topography, climate, and soil properties to develop habitat suitability models. Three machine learning techniques, including random forest (RF), support vector machine (SVM), and boosted regression tree (BRT), were estimated based on the receiver operating characteristic (ROC) and area under the curve (AUC) to model the distribution of *A. fatua*. Model performance was quantified using the ROC curve and AUC metrics to identify the best predictive algorithm. The findings indicated that RF, BRT, and SVM models exhibited accuracies of 99%, 97%, and 96% for the habitat suitability of *A. fatua*, respectively. The Boruta feature selection method identified the slope variable as significantly influential in *A. fatua* habitat suitability modeling, followed by plan curvature, clay, temperature, and silt. This study serves as a case study that highlights the utility of machine learning for habitat suitability predictions when information on multiple environmental variables is available. This approach supports effective weed management strategies, potentially enhancing canola productivity and mitigating the ecological impacts associated with weed infestation.

Introduction

Canola (*Brassica napus* L.) has gained global significance as a valuable oilseed crop that is widely cultivated because of its high-quality oil and protein-rich by-products (Neik et al. 2020). Canola is the second major oilseed crop globally with increasing world demand and production, followed by soybean oil (Tu et al. 2024). Its versatility as a source of edible oil, animal feed, and biofuel contributes to canola's pivotal role in food security and renewable energy sectors (Tileuberdi et al. 2022). Canola exports have increased in recent decades, they are expected to expand by 40% by 2025 (Tiwarei et al. 2020).

Since 1996, canola production in Iran has grown consistently in the international oilseed marketplace (Spörl et al. 2022). The increasing demand for sustainable agriculture highlights the necessity of efficient canola cultivation practices, promoting its resilience to environmental stressors and optimizing yield (Majidian et al. 2024). A notable challenge in canola production is weed management, which can significantly reduce crop yield and quality by competing for resources such as nutrients, water, and sunlight (Hassan et al. 2023). This significant threat not only affects grain production and yield but may also compromise the quality of canola oil, showing the urgent need for the agricultural sector to explore innovative practices and technologies to mitigate this challenge (Walia and Kumar 2023). A critical component in the formulation of effective management plans is a comprehensive understanding of weed flora and its geographic distribution. Such knowledge facilitates the application of herbicides and development of other appropriate management techniques (Krämer et al. 2020; Nath et al. 2024).

Several weed species have been recognized for their significant effects on canola yield and cultivation. The management and control of these weeds are crucial for maintaining the

© The Author(s), 2025. Published by Cambridge University Press on behalf of Weed Science Society of America. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.



productivity and profitability of canola crops (Asaduzzaman et al. 2020). Wild oats (*Avena fatua* L.), belonging to the Gramineae family, is one of the most dominant weeds in canola and is currently found in approximately 50 countries globally (Matsushashi et al. 2021). Some studies have shown that *A. fatua* can significantly reduce crop yield, highlighting their severe impact on agricultural productivity (Tang et al. 2024). Moreover, *A. fatua* presents a significant challenge because of their substantial resistance to herbicides, increasing control efforts and causing an ongoing threat to canola cultivation (Onkokesung et al. 2022).

Geographic information systems (GIS) are one of the most effective and precise tools for producing weed distribution maps. These systems leverage advanced technologies to accurately identify areas infested by weeds, thereby facilitating targeted management approaches (Mohan and Giridhar 2022). Detailed species distribution and habitat suitability modeling (HSM) enabled by GIS technology plays a critical role in environmental management by providing in-depth assessments of the interactions between different species and their environments.

In recent years, machine learning algorithms have emerged as powerful tools for modeling the habitat suitability of weeds based on environmental variables (Rather et al. 2020). By analyzing data on soil composition, climate conditions, and other ecological factors, machine learning models can predict the likelihood of weed proliferation in specific areas (Bi et al. 2024). These insights can aid in preemptive weed management strategies tailored to environmental conditions, thereby enhancing the precision of crop management practices (Akhter et al. 2020).

The integration of machine learning techniques (MLTs) into HSM represents a cutting-edge approach that enhances the prediction and understanding of geographic distribution (Beery et al. 2021). By utilizing the power of algorithms and computational models, machine learning can analyze complex environmental and biological data to identify patterns and relationships that influence the presence or absence of species across different fields (Jeon et al. 2023). The use of habitat suitability as a measure for assessing the risk of weed infestation has increased globally (Hartl et al. 2024). HSM has been increasingly employed to identify areas that are potentially vulnerable to various weed species over extensive geographic areas (Qazi et al. 2023; Wang et al. 2023). Schartel et al. (2021) determined the habitat suitability of eight exotic species that were invasive in Baja California and assessed their distribution and invasion risk. Wan and Wang (2019) evaluated the compatibility of habitats for 10 dangerous weed species and proposed a strategy for mitigating the risks posed by these weed species by modifying prevention and control methods.

Several studies have used MLTs, such as support vector machine (SVM), random forest (RF), boosted regression trees (BRTs), classification and regression trees (CARTs), generalized additive models, and generalized linear models to predict species distribution and for HSM (Gholami et al. 2021; Mondal and Bhat 2021). RF is a group-learning method that uses multiple decision trees to improve prediction accuracy and is ideal for assessing habitat suitability by evaluating diverse environmental variables (Renjana et al. 2024). Environmental research widely employs the SVM framework, rooted in statistical learning theory. Although SVM demonstrates significant utility, its effectiveness in modeling habitats that favor the growth of specific plant species remains an area of ongoing investigation (Tazikheh et al. 2022). The BRT model combines the principles of boosting, an MLT, with regression trees, and creates a powerful predictive model (Salditt et al. 2023). Models such as RF, SVM, and BRT have gained prominence in

predicting natural events and hazard backgrounds because of their simplicity and efficacy (Berhane et al. 2021; Hasan et al. 2024; Hasannejadasl et al. 2023). However, the use of these models to assess habitat suitability for weed species in canola fields remains relatively underexplored in scientific literature.

This study set forth two primary aims to address the key challenges in canola farming within the Fars Province of Iran. First, we sought to identify and document the predominant weed species affecting canola cultivation across the region, thereby contributing essential data to local agronomic research. Second, we implemented and compared three advanced modeling approaches—RF, SVM, and BRT—to predict habitat suitability for the identified dominant weed species. The assessment of influential environmental factors facilitated by the Boruta algorithm further enhances the model interpretability and ecological insight. Additionally, the selection of the optimal model based on the receiver operating characteristic (ROC) curve and area under the curve (AUC) maximizes predictive accuracy, pioneering the application of these MLTs in weed habitat modeling. These aims collectively address a significant research gap, offering foundational knowledge that can improve precision in weed management strategies, reduce yield losses, and promote sustainable canola production.

Materials and Methods

Study Area

This investigation was performed in the southwestern region of the Fars Province, Iran, in 2023 (Figure 1). The research region is situated between 27.26°N and 30.41°N and between 51.49°E and 54.48°E. According to the FAO (2024), Iran has expanded its canola cultivation significantly, reaching a total of around 200,000 ha. This growth is part of the country's efforts to boost self-sufficiency in oilseed production, with regions like Fars Province playing a crucial role. Geographic analyses based on topographic maps show that Fars Province encompasses both mountainous terrain and plains. The province is also distinguished by its climatic diversity, with the four seasons exerting distinct effects on regional flora. This variation in climate is largely attributed to the varied elevation, ranging from 182 to 3,183 m above sea level. The Fars Province has an average annual rainfall of 315 mm and an average annual temperature of 15°C (Kheiri et al. 2024). The average slope of the Fars Province is 7°, which is particularly favorable for canola cultivation.

Methodology

This research followed a five-stage methodology: (1) data collection; (2) preparation of influential factors; (3) HSM using three models: RF, SVM, and BRT; (4) evaluation of models and selection of the best model; and (5) variable importance analysis, as illustrated in Figure 2.

Data Collection and Sampling

In the present study, sampling was conducted through 114 canola fields in 28 different counties of the Fars Province, based on the cultivation area of this crop. Some studies have demonstrated that the presence of weeds at the 6- to 8-leaf growth stage significantly reduces canola yield (Bečka et al. 2021). Chao et al. (2023) stated that the critical period for weeds in autumn canola growth can reduce plant performance by more than 10%, so canola should be maintained without weeds. Therefore, sampling was carried out

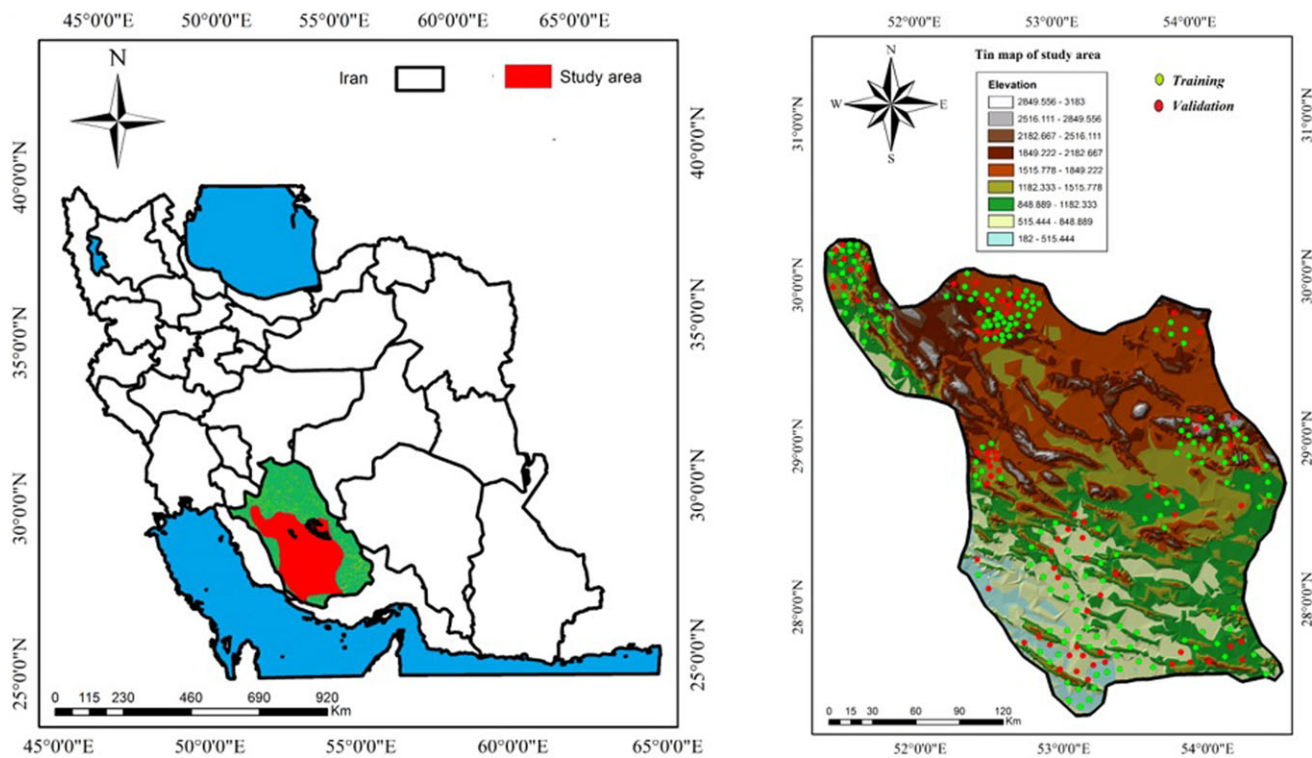


Figure 1. The research region as situated in Iran's Fars Province (left). Topographic map of research region showing locations for training and validation datasets (right).

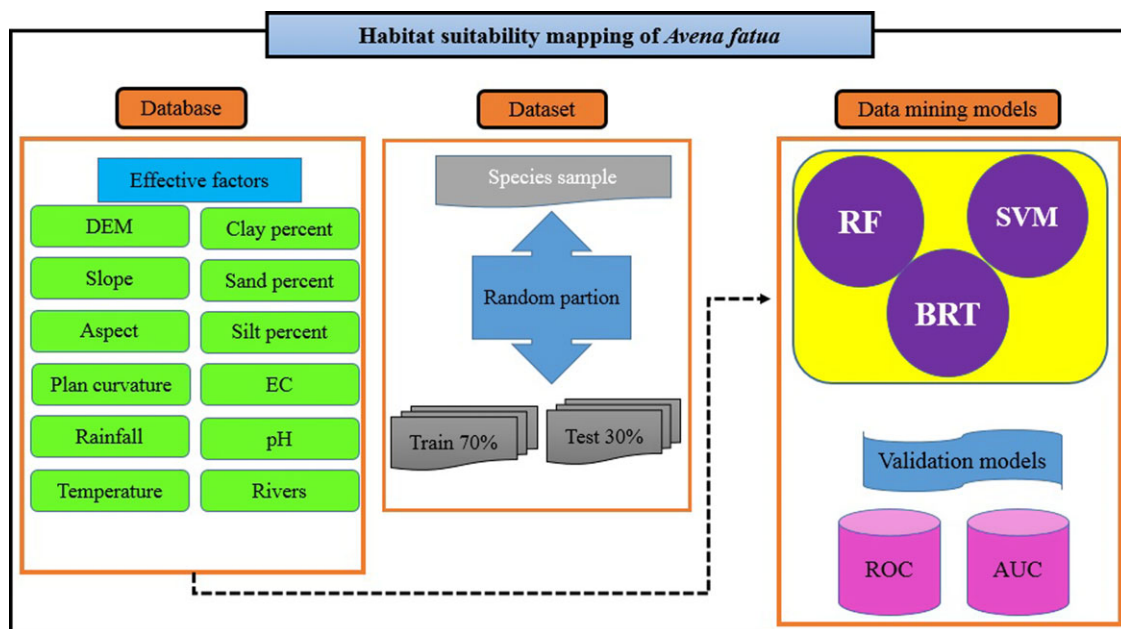


Figure 2. An *Avena fatua* habitat suitability mapping flowchart. AUC, area under the curve; BRT, boosted regression tree; DEM, digital elevation model; EC, electrical conductivity; RF, random forest; ROC, receiver operating characteristic; SVM, support vector machine.

during the winter season in 2023, when canola is in the 6- to 8-leaf growth stage. Sampling was conducted using a 0.25-m² quadrat in the form of a W-shaped field based on the cultivation area (Fried et al. 2022) in each country in Fars province (Table 1; Figure 3).

In addition to weed sampling, the geographic coordinates of each farm (latitude and longitude) were determined by using a GPS device. After weeds were collected from various canola fields,

they were accurately identified and counted based on genus and species. Soil samples from each point were transferred to the laboratory to determine the chemical and physical properties of the soil in each canola field. Based on Equations 1–5, the frequency %, uniformity %, mean field density (plant/m²), and abundance index of different species (Thomas 1985) were evaluated in Fars Province:

Table 1. Reviewing the fields of any county in Fars province

Area under canola cultivation (ha) in each county	Number of fields measured
<500	2
500 to 1,000	3
1,000 to 5,000	4
5,000 to 10,000	6
10,000 to 15,000	8
15,000 to 30,000	11
30,000 to 60,000	15
>60,000	1 field added to 15 for each 10,000 ha

$$F_k = \frac{\sum Y_i}{n} \times 100 \quad [1]$$

where n is the number of fields visited, Y_i is the presence or absence of species k in field i , and F_k is the frequency of species k across all the quadrats. The following formula was used to obtain the uniformity index for species k (U_k):

$$U_k = \frac{\sum_i^n \sum_j^m X_{ij}}{\sum_j^m m} \quad [2]$$

where X_{ij} indicates the presence or absence of species k in the i th quadrat and j th field, with n fields and m quadrats.

$$D_{ki} = \frac{\sum_1^n Z_j}{n} \times 4 \quad [3]$$

In Equation 3, m is the number of thrown quadrats and Z_j is the number of plants in the quadrat. D_{ki} is the density (number of plants per meter) of the k species at field number i .

$$\text{MDSK} = \frac{\sum_1^n D_{ki}}{n} \times 4 \quad [4]$$

In Equation 4, n is the number of fields visited, D_{ki} is the density (number of plants per meter) of k species on field number i , and MDSK is the mean density of species k .

$$A_{ik} = F_k + U_k + \text{MFDK} \quad [5]$$

Finally, Equation (5) was used to determine the dominance index of the weeds. Using this equation, the frequency (F_k), field uniformity (U_k), and mean density of species k (MDSK) were combined to determine the predominant weed species.

Important Factors

In general, for HSM, it is necessary to identify the factors that affect weed growth and development. For example, some studies have demonstrated that environmental factors, including topography, soil chemical and physical properties, road development, temperature, and rainfall, can affect weed distribution (Jehangir et al. 2024). Twelve layers were used as influencing factors: elevation, slope degree, slope aspect, plan curvature, distance from rivers, mean annual precipitation, mean annual temperature, pH, EC, and soil clay, silt, and sand percentages, which were considered to affect the growth and development of weed species. These 12 study layers were then converted to 30-m resolution for future analyses

(Kabiri et al. 2022) in ArcGIS v. 10.8.1 (<https://www.esri.com/en-us/arcgis/products/arcgis-desktop/overview>). The annual mean temperature and rainfall data were gathered from 29 meteorological organizations in the counties of Fars Province. The data were then converted to a point map using ArcGIS v. 10.8.1 software. The point map and study area were converted into temperature and rainfall maps using a 30-m resolution with the IDW algorithm (Figure 4A and 4B).

In total, 189 soil samples were collected at a depth of 30 cm. A hydrometer was used to determine the physical characteristics of the soil, such as the amounts of sand, silt, and clay (Feng et al. 2024). A pH meter and a conductivity meter were used to test the pH and EC of the soil, respectively. Sand, silt, clay, pH, and EC layers were also converted into a raster map with 30-m resolution (Figure 4C–G). A digital elevation model (DEM) of Fars Province was applied to assess elevation, slope degree, slope aspect, and plan curvature with a 30-m resolution (Figure 4H–K). Topographic maps at a resolution of 1:25,000 were used to create a raster map of the distance from the rivers to assess the impact of the rivers on habitat suitability (Figure 4L).

RF

RF is a supervised learning method developed by Breiman (2001) and consists of an ensemble of decision trees used for both classification and regression tasks. The RF model operates by constructing multiple trees during training and outputting the mode of the classes or mean prediction for classification or regression, respectively. This approach, which enhances model robustness and accuracy, is particularly effective for complex data, making it highly suitable for HSM (Talhimi et al. 2024).

In this study, the key parameters for RF, such as $n_{\text{estimators}}$ (number of trees in the forest), max_depth (maximum depth of each tree), and min_samples_split (minimum number of samples required to split a node), were optimized. We used grid search cross-validation to tune these parameters, with $n_{\text{estimators}}$ ranging from 100 to 500, max_depth from 10 to 50, and min_samples_split set to identify the optimal values. The performance of the model was evaluated using accuracy, F1 score, and AUC/ROC metrics, providing a comprehensive assessment of model accuracy and threshold-specific performance. The RF model was implemented using the RANDOM FOREST package in R (<https://cran.r-project.org/web/packages/randomForest/index.html>), which facilitates parameter tuning and cross-validation.

SVM

SVM, introduced by Vapnik (1997), is a nonparametric statistical method that does not assume any particular distribution of the dataset. SVM is effective for high-dimensional data with a relatively small number of samples, making it suitable for species distribution modeling (Kumar et al. 2024).

For our SVM model, the key parameters included C (penalty parameter) and the kernel type (linear, polynomial, sinusoidal, or radial basis function). The C parameter was tuned from a range of 0.1 to 10 on a log scale to balance the margin and misclassification tolerance, while kernel selection was optimized based on model performance. Accuracy, F1 score, and AUC/ROC metrics were used for model evaluation, emphasizing precision and recall owing to potential class imbalance. The SVM model was implemented using the $E1071$ package in R (<https://cran.r-project.org/web/packages/e1071/index.html>), which provides comprehensive support for parameter optimization and evaluation.



Figure 3. Spatial distribution of canola and weed sampling.

BRT

A BRT is an ensemble method that combines the predictions of several weak classifiers into a stronger overall model (Alnahit et al. 2022). It uses the CART framework to iteratively add trees that correct errors made by previous ones, optimizing both the learning_rate (learning speed) and n_estimators.

For this study, learning rate and estimators were optimized using a range of 0.01 to 0.1 for learning_rate and up to 500 trees for n_estimators. We also tuned max_depth to control tree complexity and prevent overfitting. The model was evaluated using accuracy, F1 score, and AUC/ROC metrics to provide a robust assessment of the predictive accuracy across thresholds.

We implemented BRT using the GBM package in R (<https://cran.r-project.org/web/packages/gbm/index.html>), which facilitates parameter tuning and cross-validation, including early stopping based on AUC/ROC performance.

Boruta Algorithm

A critical component of this research involves evaluating the importance of variables in spatial modeling for habitat suitability to guide optimal management strategies (López-Torres et al. 2023). The Boruta algorithm was chosen for this purpose because it effectively identifies influential variables by leveraging the RF model's capacity for variable selection (Li et al., 2023). The algorithm operates by iteratively comparing the importance of actual features to shadow features, which are randomized duplicates, thus distinguishing truly important predictors from noise (Xiao et al. 2024).

For the implementation, we used the BORUTA package in R (<https://cran.r-project.org/web/packages/Boruta/index.html>). The key parameters included maxRuns, set to 500 to ensure sufficient iterations for stable results, and doTrace, set to 2 for detailed output during the execution of the algorithm. The maxRuns parameter

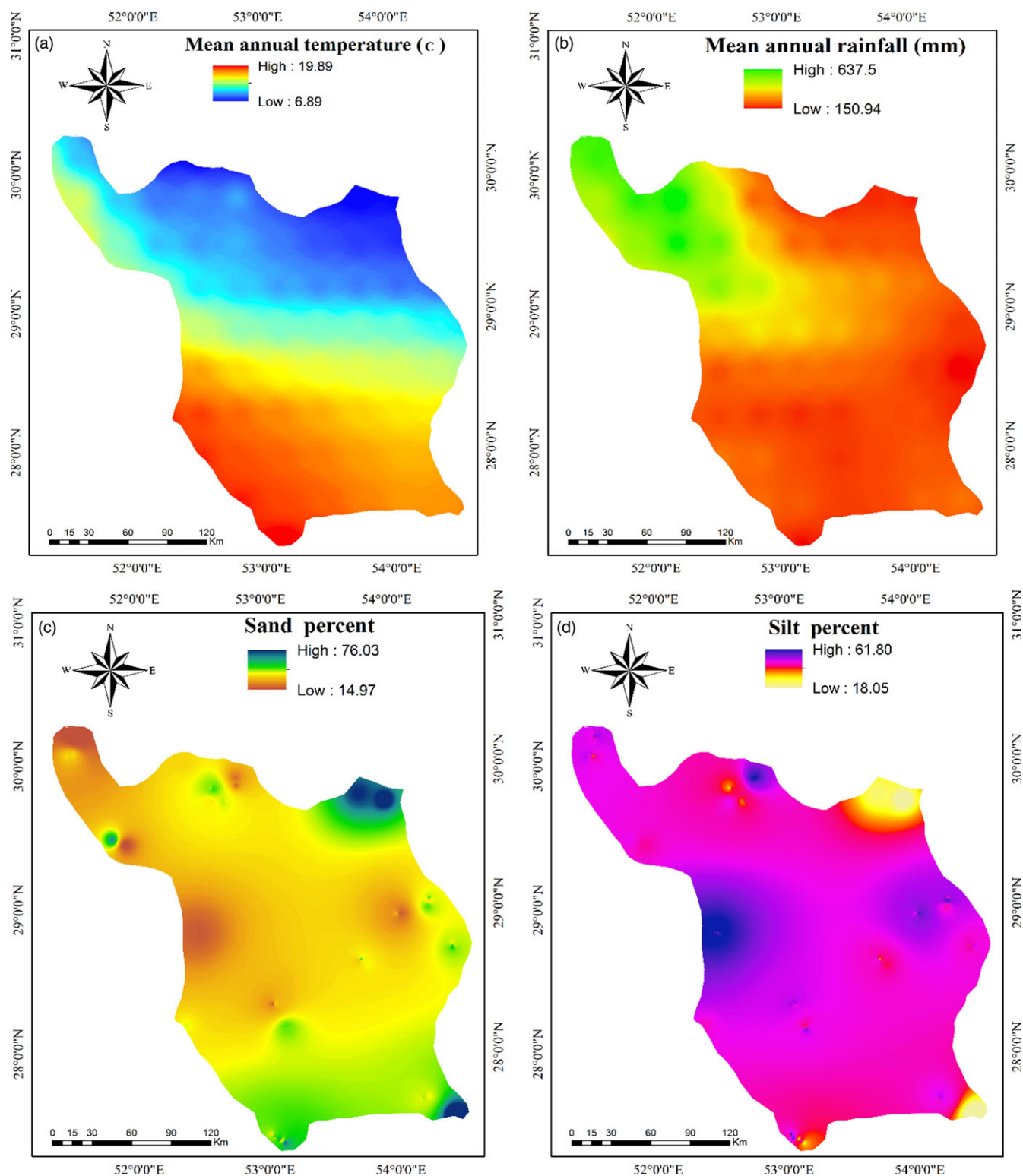
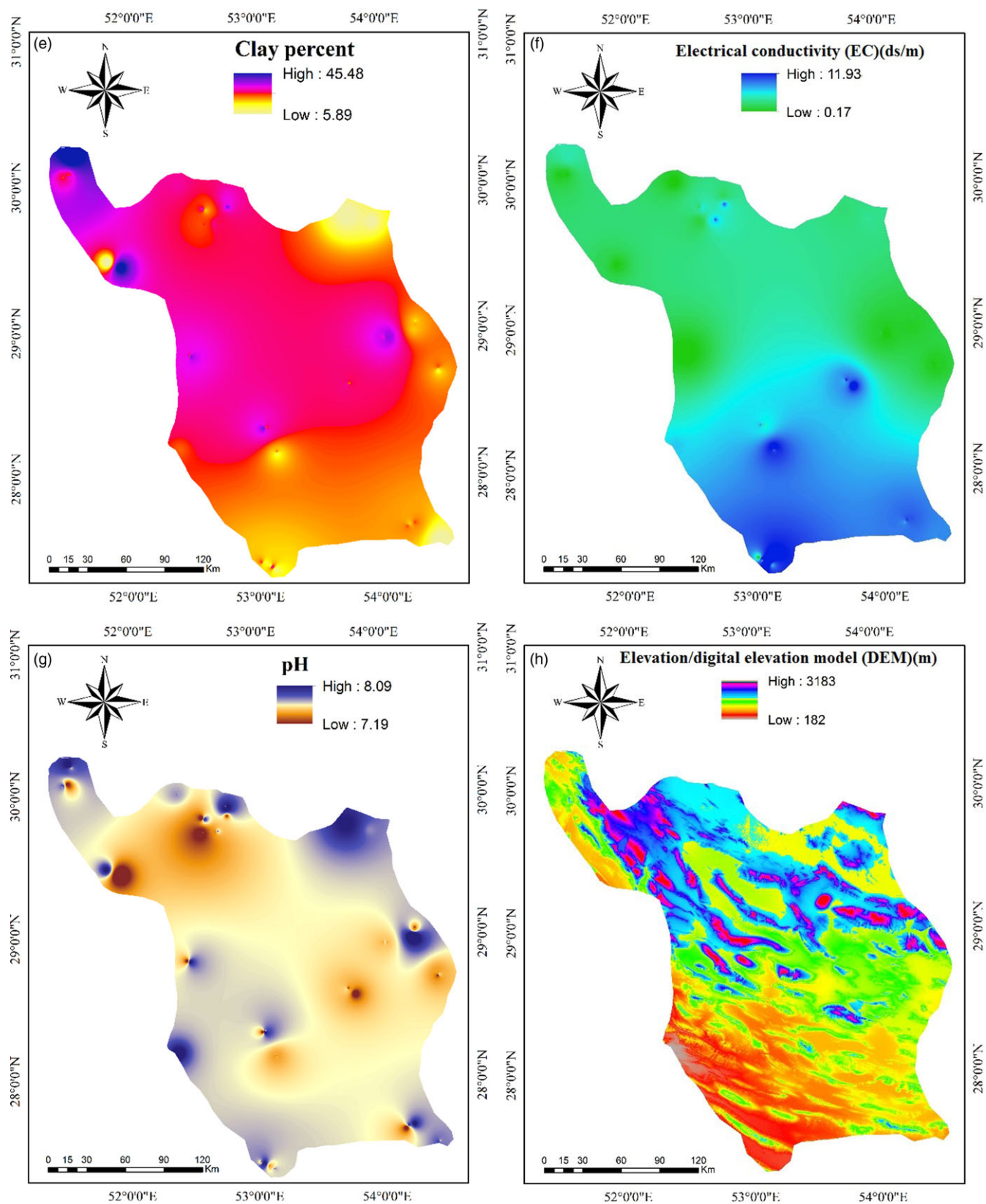
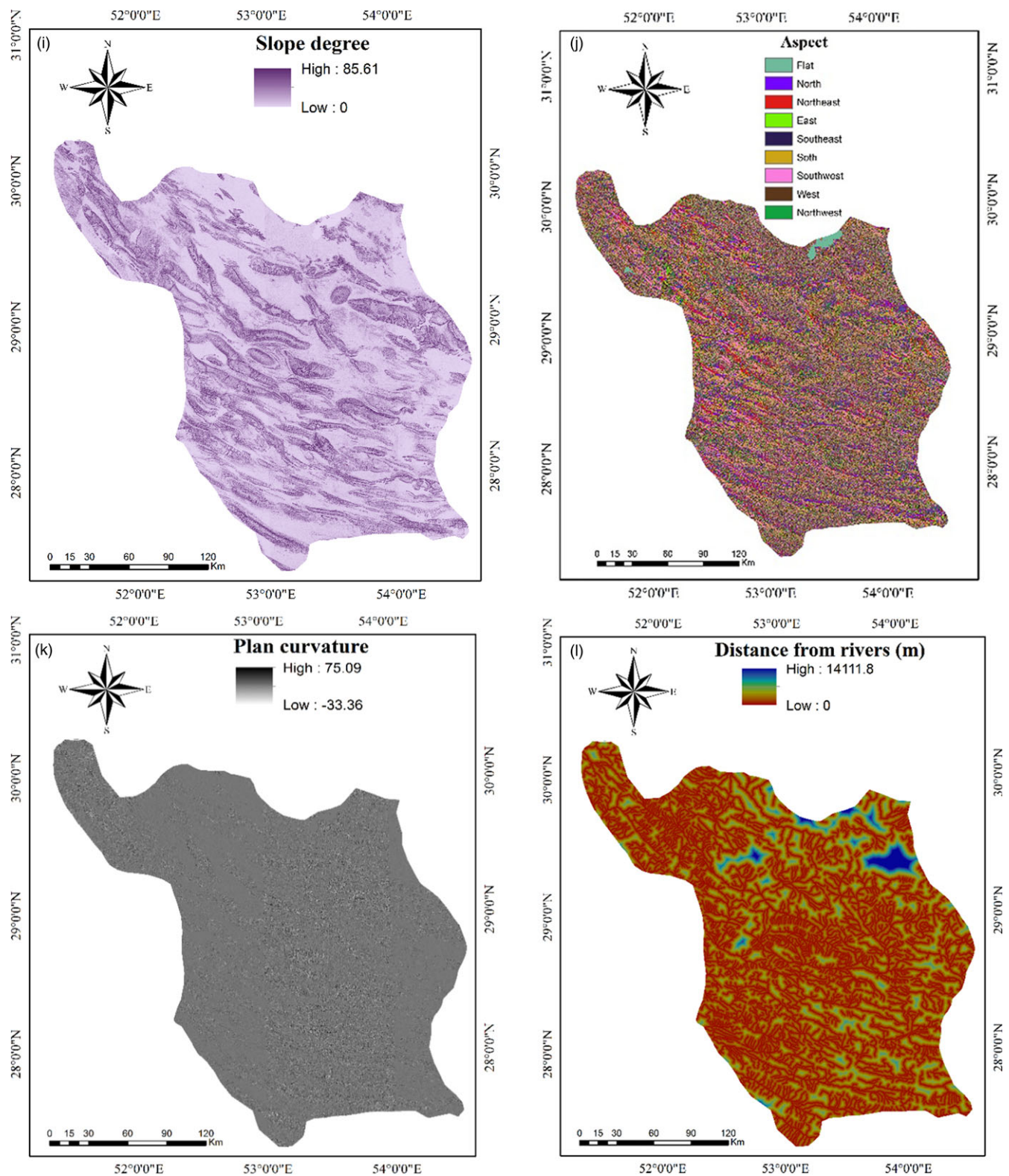


Figure 4. Important layers, including: (A) mean annual temperature, (B) mean annual precipitation, (C) sand percent, (D) silt percent, (E) clay percent, (F) electrical conductivity (EC), (G) pH, (H) elevation/digital elevation model, (I) slope degree, (J) slope aspect, (K) plan curvature, and (L) distance from rivers.

influences the stability and reliability of the variable importance ranking. Higher values provide more robust assessments by allowing more comparisons across iterations. Additionally, we used a P-value threshold of 0.05 to statistically identify significant variables.

The Boruta algorithm outputs three categories of variables: confirmed, tentative, and rejected (Han *et al.* 2022). This categorization helps refine the selection process by confirming variables with a statistically meaningful impact on habitat suitability while excluding non-informative features (Wang *et al.*





2022). The results of the Boruta algorithm provide a clearly ranked list of predictor variables crucial for understanding and managing habitat suitability patterns across different regions (Prasad et al. 2022). The variable importance derived from Boruta was instrumental in identifying which factors were most relevant in weed HSM, thereby guiding targeted management strategies.

Accuracy of Models

In HSM, where the goal is to forecast the presence or absence of a species in various locations, ROC and AUC metrics are essential tools for assessing model performance (Jamali et al. 2024). For this purpose, 70% of the presence data of the dominant weed were used

Table 2. The receiver operating characteristic (ROC) curve classification (Richardson et al. 2024).

Poor	Moderate	Good	Very good	Excellent
0.5–0.6	0.6–0.7	0.7–0.8	0.8–0.9	0.9–1

in the modeling process, while the remaining 30% of the data were utilized for validation and to evaluate the model's projected accuracy.

In this study, the 70:30 split between training and validation datasets was selected based on its established utility in predictive modeling and its practical alignment with the dataset size. This ratio is widely used in ecological and machine learning applications as a standard practice (Fielding and Bell 1997), balancing the competing requirements of sufficient data for model training and a reasonable subset for validation. The chosen split minimizes overfitting risk while allowing the evaluation of model performance on an independent dataset.

Given the dataset size, this split is particularly well suited to maximize the reliability of model parameter estimation and predictive accuracy. Despite the relatively modest dataset size, ecological modeling often operates with limited datasets due to challenges such as field collection constraints and environmental variability (Elith et al. 2006). While larger datasets are ideal, the 70:30 split effectively uses the available data to produce statistically sound results, consistent with studies in similar contexts (Hameed and Alamgir 2022).

The ROC curve and AUC are widely used metrics for assessing prediction models' accuracy. The ROC curve, a graphical representation, plots two parameters to show how well a classification model performs: the true-positive rate (TPR), or sensitivity, and the false positive rate (FPR), or 1-specificity, across different threshold values (Muschelli 2020). The TPR, represented on the y axis, indicates the proportion of real positives correctly identified by the model, while the FPR, shown on the x axis, represents the proportion of real negatives that are incorrectly classified as positives (Carrington et al. 2022). A single aggregate performance metric across all potential classification thresholds is provided by the ROC curve and AUC (Saha et al. 2023; Verbakel et al. 2020). The AUC value ranges from 0 to 1 and is classified into five performance categories: 0.5 to 0.6 (poor), 0.6 to 0.7 (moderate), 0.7 to 0.8 (good), 0.8 to 0.9 (very good), and 0.9 to 1.0 (excellent) (Table 2). In this study, the ROC-AUC was utilized to evaluate the RF, BRT, and SVM models using SPSS software v. 26 (<http://www.ibm.com>).

Collinearity Test of Effective Factors

The collinearity test of useful elements is a crucial technique in statistical analysis employed to diagnose the extent of multicollinearity among independent variables within a regression model (Barman et al. 2024). To quantitatively assess multicollinearity, the variance inflation factor (VIF) and tolerance indices were utilized. These metrics offer insights into the degree of linear association between an independent factor and the remaining independent variables in the model. A VIF value of 5 or 10 and above is generally regarded as demonstrating a problematic level of multicollinearity, indicating an exaggerated variance in an estimated regression coefficient by a factor of 5 or 10 because of its linear relationship with other variables (Cheng et al. 2022). The percentage of volatility of an independent variable that

Table 3. Frequency (%) of weeds in canola fields.

No.	Weed	Family	Frequency %
1	<i>Avena fatua</i> L.	Poaceae	58.48
2	<i>Sinapis arvensis</i> L.	Brassicaceae	45.68
3	<i>Malva neglecta</i> Wallr.	Malvaceae	36.62
4	<i>Chenopodium album</i> L.	Chenopodiaceae	34.46
5	<i>Convolvulus arvensis</i> L.	Convolvulaceae	31.64
6	<i>Centaurea depressa</i> M. Bieb.	Asteraceae	25.41
7	<i>Cerastium perfoliatum</i> L.	Caryophyllaceae	24.46
8	<i>Triticum aestivum</i> L.	Gramineae	23.83
9	<i>Daucus carota</i> L.	Umbelliferae	23.75
10	<i>Capsella bursa-pastoris</i> (L.) Medik.	Brassicaceae	23.81
11	<i>Descurainia sophia</i> (L.) Webb ex Prantl.	Brassicaceae	23.28
12	<i>Carthamus glaucus</i> M. Bieb.	Compositae	23.19
13	<i>Spergula arvensis</i> L.	Caryophyllaceae	23.01
14	<i>Trifolium pratense</i> L.	Leguminosae	22.70
15	<i>Tragopogon collinus</i> DC.	Asteraceae	22.67
16	<i>Amaranthus retroflexus</i> L.	Amaranthaceae	22.57
17	<i>Cynodon dactylon</i> (L.) Pers.	Poaceae	22.33
18	<i>Cardaria draba</i> (L.) Desv.	Brassicaceae	22.32
19	<i>Hordeum spontaneum</i> K. Koch	Poaceae	19.76
20	<i>Sonchus oleraceus</i> L.	Compositae	19.55
21	<i>Eruca sativa</i> Lam.	Brassicaceae	16.53
22	<i>Echinochloa crus-galli</i> (L.) P. Beauv.	Poaceae	15.50
23	<i>Portulaca oleracea</i> L.	Portulacaceae	15.44
24	<i>Eleusine indica</i> (L.) Gaertn.	Poaceae	14.24
25	<i>Suaeda aegyptiaca</i> (Hasselq.) Zohary	Amaranthaceae	14.23
26	<i>Galium aparine</i> L.	Labiaceae	10.99
27	<i>Carduus nutans</i> L.	Compositae	10.96
28	<i>Tribulus terrestris</i> L.	Zygophyllaceae	10.77
29	<i>Calendula arvensis</i> L.	Compositae	10.77

cannot be accounted for by other independent variables is called the tolerance. Hence, a lower tolerance value indicates a higher overlap of explanatory information among variables, signifying a potential multicollinearity issue. Typically, a tolerance value of less than 0.20 or 0.10 is considered indicative of significant multicollinearity (Negash and Alegn 2022).

Results and Discussion

Determining the Dominant Weed

Frequency percentages of genera and species were used to assess the dominant weeds. The initial findings from the sampling process indicated that *A. fatua* emerged as the most prevalent weed species, signifying its significant presence and impact in the sampled areas. Notably, 32 dominant weed species were identified, with *A. fatua* being the primary dominant species, with a frequency of 58.48% (Table 3). This indicated the critical importance of *A. fatua* in terms of its abundance and ecological influence on the studied environments.

Multicollinearity Test

Table 4 shows the collinearity between the factors affecting the species distribution modeling of *A. fatua* in the study area. Thus, based on the findings obtained, the tolerance coefficient is not less than 0.1 in any of the indices, and the VIF was not 5 or greater in any of the indices; therefore, there was no collinearity between the indices used. Otherwise, there will be multicollinearity between the

Table 4. Variance inflation factor (VIF).

Factors	Collinearity statistics	
	Tolerance	VIF
Slope aspect	0.259	1.796
Clay (%)	0.278	2.023
Elevation/DEM (m) ^a	0.632	1.528
Electrical conductivity (EC) (dS m ⁻¹)	0.566	3.000
pH	0.323	1.747
Plan curvature (100/m)	0.199	1.842
Mean annual rainfall (mm)	0.625	1.533
Distance from rivers (m)	0.627	1.531
Distance from roads (m)	0.289	2.002
Sand (%)	0.748	3.000
Silt (%)	0.600	1.549
Mean annual temperature (C)	0.657	1.512
Slope degree	0.829	2.068

^aDEM, digital elevation model.

Table 5. Habitat suitability classes areas for all applied models.

Models	Classes	Relative area %
Random forest (RF)	Low	66.56
	Moderate	16.35
	High	11.71
	Very high	5.38
Boosted regression tree (BRT)	Low	56.65
	Moderate	26.93
	High	11.87
	Very high	4.55
Support vector machine (SVM)	Low	35.57
	Moderate	37.89
	High	19.81
	Very high	6.74

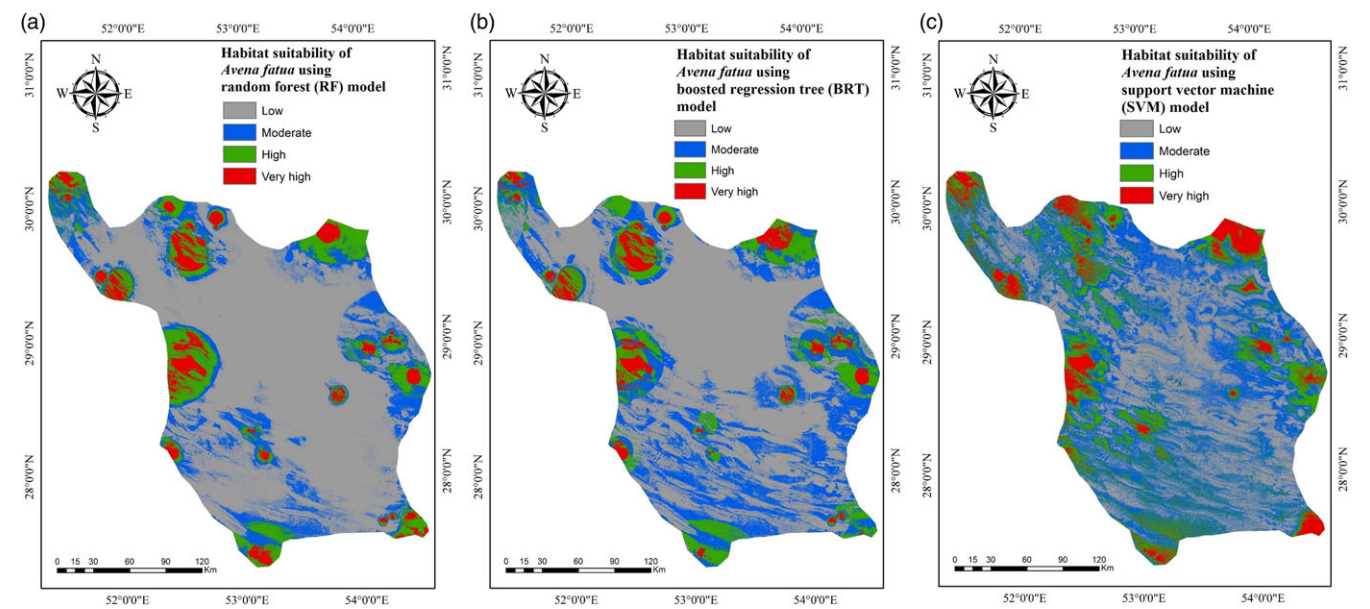


Figure 5. Habitat suitability maps of *Avena fatua* based on (A) random forest (RF), (B) boosted regression tree (BRT), and (C) support vector machine (SVM).

independent parameters and parameter estimates, and statistical significance standards will be targeted (Rovetta 2023). This leads to a lack of acceptable accuracy for spatial analysis, especially in RF, BRT, and SVM modeling.

MLTs

The final maps of the RF, SVM, and BRT models were divided into four classes to determine the suitability of the *A. fatua* habitats (Figure 5A).

RF Algorithm

According to the RF model, the low (66.56%), moderate (16.35%), high (11.71%), and very high (5.38%) classes had the largest relative areas (Table 5). In addition, the RF model map showed that the northern, northwestern, central, eastern, western, southeastern, and southwestern regions of the study area had the highest habitat suitability for *A. fatua*, although some centers had low habitat suitability for *A. fatua* (Figure 5A). However, the

northeast and parts of the center were not affected by this weed invasion.

BRT Algorithm

The habitat suitability map of *A. fatua* created using BRT showed that the low (56.65%), moderate (26.93%), high (11.71%), and very high (4.55%) classes had the largest relative areas (Table 5). The situation of the counties regarding the suitability of the habitat for *A. fatua* based on the BRT model was the same as that of the RF model (Figure 5B). This demonstrated that these models had the same performance in terms of predicting the habitat suitability of this weed.

SVM Algorithm

The SVM model had different classification conditions, such that the moderate (37.89%), low (37.89%), high (19.81%), and very high (6.74%) classes had the highest relative areas (Table 5). The suitability map of the SVM model showed that parts of the northern, northwestern, and southern study areas had greater

Table 6. Area under the curve (AUC).

Test result variable(s)	Area	SE	Asymptotic significant	Asymptotic 95% confidence interval	
				Lower bound	Upper bound
Boosted regression tree (BRT)	0.97	0.01	0.00	0.96	0.99
Random forest (RF)	0.99	0.01	0.00	0.98	1.000
Support vector machine (SVM)	0.96	0.01	0.00	0.94	0.98

habitat suitability for *A. fatua* (Figure 5C). In this model, small portions of the research area (west, northwest, southeast, east, and north) had the highest habitat suitability. According to the findings of the SVM model, it can be emphasized that the east had the highest habitat suitability for *A. fatua* (Figure 5C). In addition, counties in the southwest, southeast, and a large portion of the center of the research area had low habitat suitability.

Evaluation of Algorithms

In this study, the models were evaluated using the ROC curve and AUC. The most accurate models were the RF, BRT, and SVM models according to the ROC curve (Figure 6). Also, the AUC confirmed the accuracy of the RF (0.99%), BRT (0.97), and SVM (0.96) models (Table 6). Huang et al. (2021) have reported that the areas under the curve are 0.5 to 0.6 (poor), 0.6 to 0.7 (moderate), 0.7 to 0.8 (good), 0.8 to 0.9 (very good), and 0.9 to 1 (excellent). Therefore, the RF, BRT, and SVM models were excellent in this study.

Importance of Variables

In this study, the relevance of these variables is evaluated through the application of the Boruta algorithm. This method was used to determine the most influential factors in the analysis. The results of the Boruta algorithm demonstrated that the slope, plan curvature, clay, temperature, and silt factors had the greatest impact on the modeling of *A. fatua* habitat suitability (Table 7). Differences in the slope of the soil throughout the terrain may have affected the growth and expansion of *A. fatua*. This factor has a profound effect on vegetation dispersal patterns. One of the important effects of land slope is moisture absorption. For example, south-facing slopes subjected to higher solar irradiance typically exhibit reduced soil moisture levels, constraining plant growth.

Practical Implications and Conclusion

This study highlights the practical applications of machine learning algorithms, including RF, SVM, and BRT, for modeling the habitat suitability of *A. fatua* in canola fields. Each algorithm brings unique advantages to understanding weed distribution, which is crucial for devising sustainable and site-specific management strategies to mitigate the detrimental effects of *A. fatua* on crop productivity. By leveraging the strengths of these models, this research provides actionable insights that align with contemporary agricultural goals of improving efficiency while minimizing environmental impacts.

The RF model emerged as the most effective algorithm, achieving the highest accuracy (99%) in predicting habitat

suitability. This model was instrumental in identifying key environmental predictors, such as slope, soil texture, and plan curvature, that significantly influence *A. fatua* distribution. Its embedded feature selection capabilities not only enhanced interpretability but also allowed for the refinement of management practices in heterogeneous agricultural landscapes. Studies by Kang et al. (2022) and Melash et al. (2023) further validate the efficacy of RF in handling complex ecological datasets with numerous interacting variables. Additionally, RF's ensemble approach ensures model stability and robustness to outliers, making it particularly suitable for field-based ecological studies characterized by high variability in environmental conditions.

SVM also demonstrated its utility in analyzing high-dimensional datasets, with a classification accuracy of 96%. This algorithm excelled in differentiating between habitat suitability classes, providing detailed ecological niche maps that are indispensable for spatially targeted weed management. The ability of SVM to handle complex interactions among environmental variables has been documented in recent works, including those of Akhtar et al. (2024) and O'Neill et al. (2023). These studies emphasize the importance of SVM in addressing challenges posed by diverse agroecological conditions, where precision in habitat differentiation directly impacts the effectiveness of weed control measures.

The BRT model, with an accuracy of 97%, effectively captured nonlinear relationships between *A. fatua* occurrence and predictor variables. This capacity for addressing nonlinearity is particularly significant in weed science, where ecological interactions are rarely linear. The ensemble-based nature of BRT enhances its prediction precision, a feature corroborated by studies such as those of Montoya-Jiménez et al. (2022) and Kumari et al. (2024). By integrating BRT into HSM, this study adds to the growing body of evidence supporting its applicability in managing invasive species in agricultural systems.

Although the 70:30 training-validation split provides an efficient framework for ecological modeling, the dataset size remains a potential limitation of this study (Garcés et al. 2022). Smaller datasets inherently constrain the ability to capture rare patterns and subtle environmental interactions, which could impact model generalizability (Yu et al. 2024). However, this study operates within the boundaries of a case study approach, wherein the primary goal is to explore and demonstrate a method's applicability rather than achieve universal generalizability. To address this limitation, the dataset size and split were carefully chosen to balance robustness in model training and reliable validation. Previous studies have demonstrated that even smaller datasets can yield valuable insights when the modeling methodology is rigorous (Wisz et al. 2008). Additionally, the model's performance metrics, assessed using cross-validation, support the inference that the chosen split is sufficient for the study's aims. Future research could address this limitation by expanding the dataset through additional sampling or leveraging synthetic data-generation techniques to augment the dataset size. Nevertheless, for a case study framework, this approach aligns well with established methodologies, and the results provide meaningful insights into the ecological processes under investigation.

The complementary strengths of RF, SVM, and BRT underscore their collective utility in ecological modeling. RF and BRT were particularly effective in assessing feature importance, while SVM provided the highest resolution in classification tasks. This integrated approach offers a more comprehensive understanding of *A. fatua* habitat suitability and enables the creation of nuanced maps tailored to specific regional conditions. Such detailed

Table 7. Examining the significance of variables using the Boruta algorithm.

	Mean importance	Mean importance	Median importance	Maximum importance	Decision
Elevation/DEM ^a	−0.77	−0.93	−2.52	1.27	Confirmed
Aspect	13.28	13.20	11.79	14.60	Confirmed
Clay percent	19.61	19.69	18.42	20.93	Confirmed
Electrical conductivity (EC)	10.87	10.73	10.23	11.98	Confirmed
pH	9.02	9.25	9.14	9.87	Confirmed
Plan curvature	20.77	20.60	19.99	21.73	Confirmed
Annual mean rainfall	10.07	10.61	8.42	11.34	Confirmed
Distance from river	11.12	11.17	10.19	12.56	Confirmed
Sand percent	7.27	7.37	5.79	8.90	Confirmed
Silt percent	14.41	13.97	12.39	16.52	Confirmed
Slope degree	30.93	30.98	29.42	32.58	Confirmed
Annual mean temperature	14.75	14.89	13.50	16.25	Confirmed

^aDEM, digital elevation model.

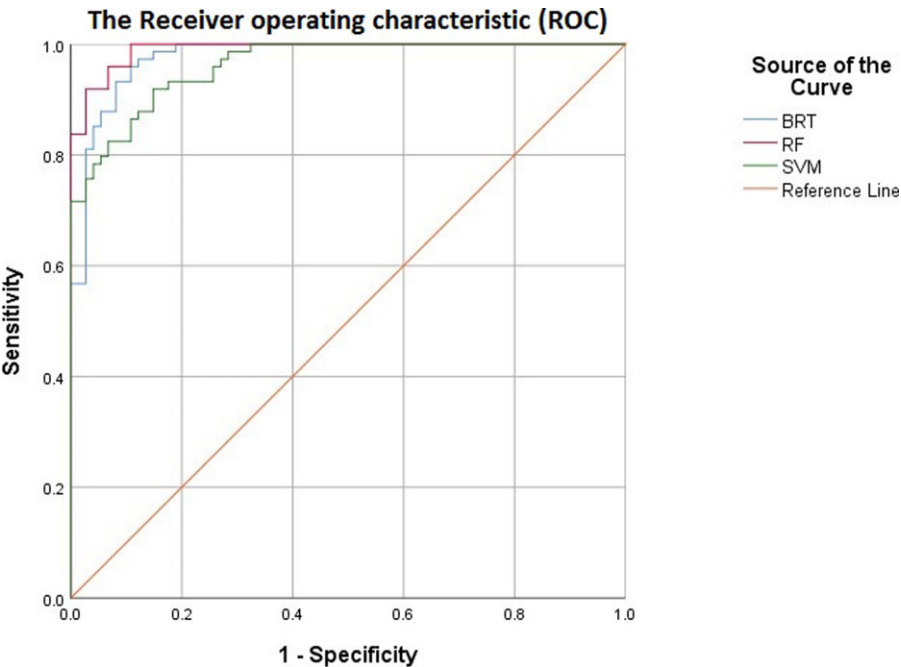


Figure 6. The receiver operating characteristic (ROC) curve for evaluating algorithms. BRT, boosted regression tree; RF, random forest; SVM, support vector machine.

mapping provides a critical basis for targeted interventions, ensuring that management resources are deployed efficiently and effectively in areas at high risk of weed invasion.

The practical implications of this study extend beyond theoretical modeling. By generating habitat suitability maps, this research equips agricultural practitioners with precise tools for implementing site-specific weed management strategies. This targeted approach not only minimizes herbicide usage but also supports environmentally conscious practices that align with the principles of sustainable agriculture. Topographic factors, such as slope and aspect, emerged as pivotal predictors, corroborating findings from Yang et al. (2023) and Vykydalová et al. (2024) that highlight the role of microclimatic conditions in shaping weed distribution. Similarly, the influence of soil texture and temperature on habitat suitability aligns with broader ecological studies, such as those by Dastres et al. (2023) and Yao (2023), emphasizing the adaptive strategies of *A. fatua* in diverse agroecological contexts.

While the study showcases the effectiveness of RF, SVM, and BRT, it also acknowledges limitations inherent to these models. The accuracy of predictions is influenced by data quality and representativeness, as highlighted in recent works by Hasan et al. (2024) and Xu et al. (2024). Algorithmic biases, environmental variability, and scalability challenges further underscore the need for continuous refinement of these methods. For instance, temporal and spatial changes in environmental conditions may reduce the reliability of predictions over time, necessitating the development of more adaptive and scalable modeling frameworks. Future research should focus on addressing these limitations to enhance the robustness and generalizability of machine learning applications in weed science.

In conclusion, this research advances the field of weed science by demonstrating the potential of machine learning models to improve habitat suitability predictions for dominant weeds like *A. fatua*. By integrating ecological, agronomic, and computational insights, the study lays a foundation for the development of

sustainable, data-driven weed management strategies. The findings not only highlight the efficacy of RF, BRT, and SVM in ecological modeling but also provide a road map for their broader application in addressing challenges associated with agricultural sustainability and biodiversity conservation.

Funding statement. This research was funded by Research Council of Shiraz University, grant/award no.: 98GCU1M75346.

Competing interests. No competing interests have been declared.

References

- Akhtar M, Tanveer M, Arshad M (2024) RoBoSS: a robust, bounded, sparse, and smooth loss function for supervised learning. *IEEE Trans Pattern Anal Mach Intell* 47:149–160
- Akhter MJ, Jensen PK, Mathiassen SK, Melander B, Kudsk P (2020) Biology and management of *Vulpia myuros*—an emerging weed problem in no-till cropping systems in Europe. *Plants* 9:715
- Alnahit AO, Mishra AK, Khan AA (2022) Stream water quality prediction using boosted regression tree and random forest models. *Stoch Environ Res Risk Assess* 36:2661–2680
- Asaduzzaman M, Pratley JE, Luckett D, Lemerle D, Wu H (2020) Weed management in canola (*Brassica napus* L): a review of current constraints and future strategies for Australia. *Arch Agron Soil Sci* 66:427–444
- Barman J, Biswas B, Rao KS (2024) A hybrid integration of analytical hierarchy process (AHP) and the multiobjective optimization on the basis of ratio analysis (MOORA) for landslide susceptibility zonation of Aizawl, India. *Nat Hazards* 120:8571–8596
- Bečka D, Bečková L, Kuchtová P, Cihlár P, Pazderů K, Mikšík V, Vašák J (2021) Growth and yield of winter oilseed rape under strip-tillage compared to conventional tillage. *Plant Soil Environ* 67:2
- Beery S, Cole E, Parker J, Perona P, Winner K (2021) Species distribution modeling for machine learning practitioners: a review. Pages 329–348 in *COMPASS '21: Proceedings of the 4th ACM SIGCAS Conference on Computing and Sustainable Societies*. New York: Association for Computing Machinery
- Berhane G, Kebede M, Alfarrah N (2021) Landslide susceptibility mapping and rock slope stability assessment using frequency ratio and kinematic analysis in the mountains of Mgulat area, Northern Ethiopia. *Bull Eng Geol Environ* 80:285–301
- Bi Z, Sun J, Xie Y, Gu Y, Zhang H, Zheng B, Ou R, Liu G, Li L, Peng X, Gao X, Wei N (2024) Machine learning-driven source identification and ecological risk prediction of heavy metal pollution in cultivated soils. *J Hazard Mater* 476:135109
- Breiman L (2001) Random forests. *Mach Learn* 45:5–32
- Carrington AM, Manuel DG, Fieguth PW, Ramsay T, Osmani V, Wernly B, Bennett C, Hawken S, Magwood O, Sheikh Y, McInnes M, Holzinger A (2022) Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation. *IEEE Trans Pattern Anal Mach Intell* 45:329–341
- Chao WS, Anderson JV, Li X, Gesch RW, Berti MT, Horvath DP (2023) Overwintering camelina and canola/rapeseed show promise for improving integrated weed management approaches in the Upper Midwestern US. *Plants* 12:1329
- Cheng J, Sun J, Yao K, Xu M, Cao Y (2022) A variable selection method based on mutual information and variance inflation factor. *Spectrochim Acta A Mol Biomol Spectrosc* 268:120652
- Dastres E, Jahangiri E, Edalat M, Zamani A, Amiri M, Pourghasemi HR (2023) Habitat suitability modeling of *Descurainia sophia* medicinal plant using three bivariate models. *Environ Monit Assess* 195:392
- Elith J, Graham H, Anderson R, Dudík M, Ferrier S, Guisan A, Zimmermann N (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29:129–151
- FAO (2024) Food and agriculture data. FAOSTAT. <https://www.fao.org/faostat/en/#home>
- Feng L, Khalil U, Aslam B, Ghaffar B, Tariq A, Jamil A, Farhan M, Aslam M, Soufan W (2024) Evaluation of soil texture classification from orthodox interpolation and machine learning techniques. *Environ Res* 246:118075
- Fielding AH, Bell JF (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environ Conserv* 24:38–49
- Fried G, Le Corre V, Rakotoson T, Buchmann J, Germain T, Gounon R, Chauvel B (2022) Impact of new management practices on arable and field margin plant communities in sunflower, with an emphasis on the abundance of *Ambrosia artemisiifolia* (Asteraceae). *Weed Res* 62:134–148
- Garcés P, Baumeister S, Mason L, Chatham CH, Holiga S, Dukart J, Jones EJH, Banaschewski T, Baron-Cohen S, Bölte S, Buitelaar JK, Durston S, Oranje B, Persico AM, Beckmann CF, et al. (2022) Resting state EEG power spectrum and functional connectivity in autism: a cross-sectional analysis. *Mol Autism* 13:22
- Gholami H, Mohamadifar A, Rahimi S, Kaskaoutis DG, Collins AL (2021) Predicting land susceptibility to atmospheric dust emissions in central Iran by combining integrated data mining and a regional climate model. *Atmos Pollut Res* 12:172–187
- Hameed MAB, Alamgir Z (2022) Improving mortality prediction in acute pancreatitis by machine learning and data augmentation. *Comput Biol Med* 150:106077
- Han X, Wang L, Wang Y, Yang J, Wan X, Liang T, Rinklebe J (2022) Mechanisms and influencing factors of yttrium sorption on paddy soil: experiments and modeling. *Chemosphere* 307:135688
- Hartl T, Srivastava V, Prager S, Wist T (2024) Evaluating climate change scenarios on global pea aphid habitat suitability using species distribution models. *Clim Chang Ecol* 7:100084
- Hasan MM, Roy SK, Talha MD, Ferdous MT, Nasher NMR (2024) Predictive landslide susceptibility modeling in the southeastern hilly region of Bangladesh: application of machine learning algorithms in Khagrachari district. *Environ Sci Pollut Res*, [10.1007/s11356-024-34949-5](https://doi.org/10.1007/s11356-024-34949-5)
- Hasannejadasl H, Osong B, Bermejo I, van der Poel H, Vanneste B, van Roermund J, Aben K, Zhang Z, Kiemeny L, Van Oort I, Verwey R, Hochstenbach L, Bloemen E, Dekker A, Fijten RRR (2023) A comparison of machine learning models for predicting urinary incontinence in men with localized prostate cancer. *Front Oncol* 13:1168219
- Hassan MS, Naz N, Ali H, Ali B, Akram M, Iqbal R, Ajmal S, Ali B, Ercisli S, Golokhvast KS, Hassan Z (2023) Ultra-responses of *Asphodelus tenuifolius* L. (wild onion) and *Convolvulus arvensis* L. (field bindweed) against shoot extract of *Trianthema portulacastrum* L. (horse purslane). *Plants* 12:458
- Huang W, Liu H, Zhang Y, Mi R, Tong C, Xiao W, Shuai B (2021) Railway dangerous goods transportation system risk identification: comparisons among SVM, PSO-SVM, GA-SVM and GS-SVM. *Appl Soft Comput* 109:107541
- Jamali F, Amininasab SM, Taleshi H, Madadi H (2024) Ensemble forecasting of Persian leopard (*Panthera pardus saxicolor*) distribution and habitat suitability in south-western Iran. *Wildl Res* 51(3), [10.1071/WR23010](https://doi.org/10.1071/WR23010)
- Jehangir S, Khan SM, Ahmad Z, Ejaz U, Ain QU, Lho LH, Han H, Raposo A (2024) Distribution of the *Cannabis sativa* L. in the western Himalayas: a tale of the ecological factors behind its continuous invasiveness. *Glob Ecol Conserv* 49: e02779
- Jeon J, Lee S, Oh C (2023) Age-specific risk factors for the prediction of obesity using a machine learning approach. *Front Public Health* 10:998782
- Kabiri S, Allen M, Okuonzia JT, Akello B, Ssabaganzi R, Mubiru D (2022) Detecting wetland encroachment and urban agriculture land classification in Uganda using hyper-temporal remote sensing. *AAS Open Res* 3:18
- Kang W, Kim G, Park Y (2022) Habitat suitability and connectivity modeling predict genetic population structure and priority control areas for invasive nutria (*Myocastor coypus*) in a temperate river basin. *PLoS ONE* 17:e0279082
- Kheiri M, Kambouzia J, Rahimi-Moghaddam S, Moghaddam SM, Vasa L, Azadi H (2024) Effects of agro-climatic indices on wheat yield in arid, semi-arid, and sub-humid regions of Iran. *Reg Environ Change* 24(1):10
- Krähmer H, Andreassen C, Economou-Antonaka G, Holec J, Kalivas D, Kolářová M, Novák R, Panozzo S, Pinke G, Salonen J, Sattin M (2020) Weed surveys and weed mapping in Europe: state of the art and future tasks. *Crop Prot* 129:105010

- Kumar A, Sinha S, Saurav S, Chauhan VB (2024) Prediction of unconfined compressive strength of cement–fly ash stabilized soil using support vector machines. *Asian J Civ Eng* 25:1149–1161
- Kumari G, Kotiyal PB, Singh H, Kumar M, Kumar N, Malik A, Singh S (2024) Predicting future climate change effects on biotic communities: a species distribution modeling approach. Pages 137–168 in Singh H, ed. *Forests and Climate Change: Biological Perspectives on Impact, Adaptation, and Mitigation Strategies*. Singapore: Springer Nature
- Li Q, Wei Y, Zhang T, Che F, Yao S, Wang C, Shi D, Tang H, Song B (2023) Predictive models and early postoperative recurrence evaluation for hepatocellular carcinoma based on gadoxetic acid-enhanced MR imaging. *Insights Imaging* 14:4
- López-Torres JF, Sánchez-García JY, Núñez-Ríos JE, López-Hernández C (2023) Prioritizing factors for effective strategy implementation in small and medium-size organizations. *Eur Bus Rev* 35:694–712
- Majidian P, Ghorbani HR, Farajpour M (2024) Achieving agricultural sustainability through soybean production in Iran: potential and challenges. *Heliyon* 10:e26389
- Matsushashi S, Asai M, Fukasawa K (2021) Estimations and projections of *Avena fatua* dynamics under multiple management scenarios in crop fields using simplified longitudinal monitoring. *PLoS ONE* 16:e0245217
- Melash AA, Bogale AA, Migbaru AT, Chakilu GG, Percze A, Ábrahám ÉB, Mengistu DK (2023) Indigenous agricultural knowledge: a neglected human-based resource for sustainable crop protection and production. *Heliyon* 9:e12978
- Mohan S, Giridhar MVSS (2022) A brief review of recent developments in the integration of deep learning with GIS. *Geomatics Environ Eng* 16(2):21–38
- Mondal R, Bhat A (2021) Comparison of regression-based and machine learning techniques to explain alpha diversity of fish communities in streams of central and eastern India. *Ecol Indic* 129:107922
- Montoya-Jiménez JC, Valdez-Lazalde JR, Ángeles-Perez G, De Los Santos-Posadas HM, Cruz-Cárdenas G (2022) Predictive capacity of nine algorithms and an ensemble model to determine the geographic distribution of tree species. *iForest* 15:363–371
- Muschelli J III (2020) ROC and AUC with a binary predictor: a potentially misleading metric. *J Classif* 37:696–708
- Nath CP, Singh RG, Choudhary VK, Datta D, Nandan R, Singh SS (2024) Challenges and alternatives of herbicide-based weed management. *Agronomy* 14:126
- Negash BT, Alelgn Y (2022) Proper partograph utilization among skilled birth attendants in Hawassa city public health facilities, Sidama region, Ethiopia, in 2021. *BMC Womens Health* 22:539
- Neik TX, Amas J, Barbeti M, Edwards D, Batley J (2020) Understanding host-pathogen interactions in *Brassica napus* in the omics era. *Plants* 9:1336
- O'Neill H, Khalid Y, Spink G, Thorpe P (2023) A one-class support vector machine for detecting valve stiction. *Digit Chem Eng* 8:100116
- Onkokesung N, Brazier-Hicks M, Tetard-Jones C, Benthall A, Edwards R (2022) Molecular diagnostics for real-time determination of herbicide resistance in wild grasses. *J Biotechnol* 358:64–66
- Prasad P, Loveson VJ, Das B, Kotha M (2022) Novel ensemble machine learning models in flood susceptibility mapping. *Geocarto Int* 37:4571–4593
- Qazi AW, Saqib Z, Zaman-ul-Haq M, Gardezi SMH, Khan AM, Khan I, Munir A, Ahmed I (2023) Modelling impacts of climate change on habitat suitability of three endemic plant species in Pakistan. *Pol J Environ Stud* 32:3281–3290
- Rather TA, Kumar S, Khan JA (2020) Multi-scale habitat modelling and predicting change in the distribution of tiger and leopard using random forest algorithm. *Sci Rep* 10:11473
- Renjana E, Firdiana ER, Angio MH, Ningrum LW, Lailaty IQ, Rahadianoro A, Martiansyah I, Zulkarnaen R, Rahayu A, Raharjo PD, Abywijaya IK, Usmadi D, Risna RA, Cropper WP Jr, Yudaputra A (2024) Spatial habitat suitability prediction of essential oil wild plants on Indonesia's degraded lands. *PeerJ* 12:e17210
- Richardson E, Trevizani R, Greenbaum JA, Carter H, Nielsen M, Peters B (2024) The receiver operating characteristic curve accurately assesses imbalanced datasets. *Patterns* 5:100994
- Rovetta A (2023) A framework to avoid significance fallacy. *Cureus* 15:6
- Saha S, Bera B, Shit PK, Bhattacharjee S, Sengupta N (2023) Prediction of forest fire susceptibility applying machine and deep learning algorithms for conservation priorities of forest resources. *Remote Sens Appl Soc Environ* 29:100917
- Salditt M, Humberg S, Nestler S (2023) Gradient tree boosting for hierarchical data. *Multivar Behav Res* 58:911–937
- Schartel TE, Cooper ML, May A, Daugherty MP (2021) Quantifying *Planococcus ficus* (Hemiptera: Pseudococcidae) invasion in Northern California vineyards to inform management strategy. *Environ Entomol* 50:138–148
- Spörl J, Speer K, Jira W (2022) Simultaneous mass spectrometric detection of proteins of ten oilseed species in meat products. *Foods* 11:2155
- Talhami M, Wakjira T, Alomar T, Fouladi S, Fezouni F, Ebead U, Altaee A, Al-Ejji M, Das P, Hawari AH (2024) Single and ensemble explainable machine learning-based prediction of membrane flux in the reverse osmosis process. *J Water Process Eng* 57:104633
- Tang W, Li Z, Guo H, Chen B, Wang T, Miao F, Yang C, Xiong W, Sun J (2024) Annual weeds suppression and oat forage yield responses to crop density management in an oat-cultivated grassland: a case study in eastern China. *Agronomy* 14:583
- Tazikheh S, Davoudi A, Shafiei A, Parsaei H, Atabaev TS, Ivakhnenko OP (2022) A comparison between the perturbed-chain statistical associating fluid theory equation of state and machine learning modeling approaches in asphaltene onset pressure and bubble point pressure prediction during gas injection. *ACS Omega* 7:30113–30124
- Thomas AG (1985) Weed survey system used in Saskatchewan for cereal and oilseed crops. *Weed Sci* 33:34–43
- Tileuberdi N, Turgumbayeva A, Yeskaliyeva B, Sarsenova L, Issayeva R (2022) Extraction, isolation of bioactive compounds and therapeutic potential of rapeseed (*Brassica napus* L.). *Molecules* 27:8824
- Tiwari AK, Nasreen S, Shahbaz M, Hammoudeh S (2020) Time-frequency causality and connectedness between international prices of energy, food, industry, agriculture and metals. *Energy Econ* 85:104529
- Tu M, Wang R, Guo W, Xu S, Zhu Y, Dong J, Yao X, Jiang L (2024) A CRISPR/Cas9-induced male-sterile line facilitating easy hybrid production in polyploid rapeseed (*Brassica napus*). *Horticult Res* 11:uhae139
- Vapnik VN (1997) The support vector method. Pages 261–271 in *International Conference on Artificial Neural Networks*. Berlin: Springer
- Verbakel JY, Steyerberg EW, Uno H, De Cock B, Wynants L, Collins GS, Van Calster B (2020) ROC curves for clinical prediction models part 1. ROC plots showed no added value above the AUC when evaluating the performance of clinical prediction models. *J Clin Epidemiol* 126:207–216
- Vykydalová L, Barroso PM, Děkanovský I, Neoralová M, Lumbantobing YR, Winkler J (2024) Interactions between weeds, pathogen symptoms and winter rapeseed stand structure. *Agronomy* 14:2273
- Walia S, Kumar R (2023) Wild marigold (*Tagetes minuta* L.) biomass and essential oil composition modulated by weed management techniques. *Ind Crops Prod* 161:113183
- Wan JZ, Wang CJ (2019) Contribution of environmental factors toward distribution of ten most dangerous weed species globally. *Appl Ecol Environ Res* 17:14835–14846
- Wang X, Liu X, Wang L, Yang J, Wan X, Liang T (2022) A holistic assessment of spatiotemporal variation, driving factors, and risks influencing river water quality in the northeastern Qinghai-Tibet Plateau. *Sci Total Environ* 851:157942
- Wang ZW, Yin J, Wang X, Chen Y, Mao ZK, Lin F, Wang XG (2023) Habitat suitability evaluation of invasive plant species *Datura stramonium* in Liaoning Province: based on Biomod2 combination model. *J Appl Ecol* 34:1272–1280
- Wisn MS, Hijmans RJ, Li J, Peterson AT, Graham CH, Guisan A, NCEAS Predicting Species Distributions Working Group (2008) Effects of sample size on the performance of species distribution models. *Divers Distrib* 14:763–773
- Xiao X, Ma H, Gan G, Li Q, Zhang B, Xia S (2024) Robust *k*-means-type clustering for noisy data. Pages 1–15 in *IEEE Transactions on Neural Networks and Learning Systems*. New York: Institute for Electrical and Electronics Engineers

- Xu P, Liang S, Hahn A, Zhao VT, Lo WT, Haller BC, Sobkowiak B, Chitwood MH, Colijn C, Cohen T, Rhee KY, Messer PW, Wells MT, Clark AG, Kim J (2024) e3SIM: epidemiological-ecological-evolutionary simulation framework for genomic epidemiology. *bioRxiv*, [10.1101/2024.06.29.601123](https://doi.org/10.1101/2024.06.29.601123)
- Yang X, Zhang X, Zhang P, Bidegain G, Dong J, Hu C, Li M, Zhang Z, Guo H (2023) Ensemble habitat suitability modeling for predicting optimal sites for eelgrass (*Zostera marina*) in the tidal lagoon ecosystem: implications for restoration and conservation. *J Environ Manag* 330:117108
- Yao W, Nan F, Li Y, Li Y, Liang P, Zhao C (2023) Effects of different afforestation years on soil properties and quality. *Forests* 14:329
- Yu K, Sun L, Chen J, Reynolds M, Chaudhary T, Batmanghelich K (2024) DrasCLR: a self-supervised framework of learning disease-related and anatomy-specific representation for 3D lung CT images. *Med Image Anal* 92:103062