Evaluating the habitat suitability modeling of Aceria alhagi and Alhagi maurorum in their native range using machine learning techniques

Emran Dastres
Shiraz University
Farzad Bijani
Rafsanjan University
Ruhollah Naderi (naderi.ruhollah@gmail.com)
Shiraz University
Afshin Zamani
Shiraz University
Mohsen Edalat
Shiraz University

Article

Keywords: Algorithm, data mining, machine learning, special modeling

Posted Date: January 12th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2441475/v1

Additional Declarations: No competing interests reported.

Abstract

Spatial locational modeling techniques are increasingly used in species distribution modeling. However, the implemented techniques differ in their modeling performance. In this study, we tested the predictive accuracy of three algorithms, namely "random forest (RF)," "support vector machine (SVM)," and "boosted regression trees (BRT)" to prepare habitat suitability mapping of an invasive species, *Alhagi maurorum*, and its potential biological control agent, *Aceria alhagi*. Location of this study was in Fars Province, southwest of Iran. The spatial distributions of the species were forecasted using GPS devices and GIS software. The probability values of occurrence were then checked using three algorithms. The predictive accuracy of the machine learning (ML) techniques was assessed by computing the "area under the curve (AUC)" of the "receiver-operating characteristic" plot. When the *Aceria alhagi* was modeled, the AUC values of RF, BRT and SVM were 0.89, 0.81, and 0.79, respectively. However, in habitat suitability models (HSMs) of *Alhagi maurorum* the AUC values of RF, BRT and SVM were 0.89, 0.80, and 0.73, respectively. The RF model provided significantly more accurate predictions than other algorithms. The importance of factors on the growth and development of *Alhagi maurorum* and *Aceria alhagi* was also determined using the partial least squares (PLS) algorithm, and the most crucial factors were the road and slope. Habitat suitability modeling based on algorithms may significantly increase the accuracy of species distribution forecasts, and thus it shows considerable promise for different conservation biological and biogeographical applications.

Introduction

Camelthorn, *Alhagi maurorum, is a 60–100 cm tall and glabrous undershrub*[1]*. It is native to Iran, Pakistan,* Afghanistan*, Russia, Turkey, Iraq, Syria, Palestine, Cyprus, Egypt, India and China*[1]*.* It has been accidentally introduced in several regions of the world: South Africa has classified it as a category 1 invasive species; Australia has declared it a state-prohibited weed in Victoria; and the USA has categorized it as a poisonous plant in seven states[2]. *Alhagi maurorum* is currently recognized as a harmful species in Iranian agricultural ecosystems. Around the world, invasive weeds have had an influence on both controlled and field regions[3]. An instance of an invasive plant that prominently has caused issues in a variety of ecosystems in North America's dry climate is *Alhagi maurorum*[4]. Because the invasive species is not native to the new environment, it might drive out other species that have been naturally occurring in that region and hinder their development and reproduction[5]. In reality, invasive species are those that grow quickly outside of their natural habitat. Through resource competition, hunting, and parasitic illnesses, they may have reduced the number of native species[5]. On the other hand, as vegetation has grown, native species now make up a significant portion of the pasture and field vegetation[6]. Native vegetation produces certain natural disturbances in the living environments of people, plants, and animals in addition to not being beneficial to the nearby living things[7]. Invasive plants can have significant negative economic effects, decrease biodiversity, and change how ecosystems work[8]. Therefore, the central emphasis of rehabilitation study in invader-dominated ecosystems is on invasive plant management. Invasive species are a major threat to biodiversity[9]. The most prevalent and prolific invasive species in invading ecology are plants[10]. Studying and controlling species that have been introduced outside of their natural habitats is a part of invasion science[11]. There is no denying that some of these imported species have an influence on the existence and abundance of other species, as well as the potential for significant economic repercussions[12]. Mechanical and chemical approaches for controlling invasive species are costly[13]. For instance, it is estimated that pesticides for use against exotic weeds in farming alone cost £90 million in the United Kingdom each year[14]. More significantly, since control measures are frequently required for an extended length of time, chemical and mechanical approaches are not considered as sustainable over the longer period[15]. Classical biological control, a deliberate application of natural enemies from the native range to control the weed in its introduced range, may offer a long-term, self-replicating, and successful control, particularly in more delicate situations[16].

Aceria alhagi Vidović & Kamali, an important species of eriophyid mite, has recently been reported as a promising candidate for the biological control of *Alhagi maurorum*[17]. *Aceria alhagi* was able to impose significant reductions in plant biomass (26%), seed production (95%) and photosynthesis (53%) of camelthorn and thus it would diminish the weed's competitive ability and long-distance spread via seeds[17]. However, little particular research has been conducted on *Aceria alhagi*, despite the fact that it is well known as a crucial species for the management of camelthorn.

For creating habitat suitability maps for invasive species, machine learning approaches including "Random forest (RF)", "Support vector machine (SVM)", and "Boosted regression trees (BRT)" have been widely utilized recently[18–20]. It enhances the use of computer models with geographic information systems (GIS)[21]. Additionally, the use of GIS technology offers appropriate substitutes for the efficient implementation of vast and intricate geographical information[22]. Several investigations into weed and invasive plant assessment utilizing GIS and RS methods have been carried out[23–25]. Recently, the study of biological invasion has shown a lot of interest in remote sensing technologies. Many features of remote sensing technologies are advantageous for identifying, mapping, and keeping an eye on invaders[26]. The study of annual and long-term patterns in biological invasion is complicated by spatial heterogeneity. However, because of its wide scope, remote sensing has the ability to provide the necessary data[27]. Maps of the distribution of different plant and animal species, as well as of their ecosystems, landscapes, bio-climatic conditions, and invasion-promoting factors, have already been created using GIS and remote sensing[13,28]. The control of invasive species can now be performed much easier by machine learning technologies and precision agriculture (PA)[29]. An emphasis on strategies to ensure the credibility of present and future practices has emerged in response to the increased focus on the potential ramifications of machine learning and artificial intelligence[30]. The present focus in this direction reflects the understanding that preserving machine learning's credibility may be essential for assuring the acceptance and effective acceptance of AI-driven solutions and services[31]. This makes it possible for researchers to explore fresh and highly productive ideas. Farmers may also obtain knowledge and data to make the best choice for controlling invasive species by using smart farming and machine learning in information systems[32]. Modern technology, such as machine learning, provides advanced methods for controlling invasive species[33]. Agriculture can experience economic prosperity by utilizing these cutting-edge technologies. Due to the fact that agriculture has contributed significantly to global economic growth, experts are now able to search for new, precise, high-productivity technologies[34]. Farmers may gather information and data through the use of precise agriculture, particularly machine learning, to make the best choice for high farm productivity. Modern technology known as precision agriculture provides advanced methods for increasing farm productivity[35]. Agriculture can experience economic prosperity by utilizing this cutting-edge technology. Precision agriculture has a wide range of uses, including the detection of plant diseases, weed identification, agricultural yield production, and plant pest identification[36]. The ML has a wide range of uses, including plant disease detection, weed identification, agricultural yield production, and plant pest identification[37,38].

Models of invasion risk based on comparisons between biodiversity and environmental characteristics have become more prevalent in an effort to predict future invasions. These models are frequently introduced as a method for managers to identify treatments for preventing invasions in their early stages. The majority of habitat suitability models for invasive plants, however, are based on occurrence data, which has no known correlation with the effects of invasive species. As a result, developing HSMs based on regions with high invasive abundance and diversity may be a more management-relevant method of predicting invasion risk[39]. In southwest Iran, we contrast the presence and abundance of two distinct species. In order to assess how well HSMs predict the risk of *Alhagi maurorum* as a native species in Iran and as an invading species in the introduced range, we ascertain the degree to which occurrence records indicate regions of high species abundance and compare suitability model findings. We also produce *Aceria alhagi* habitat suitability maps to analyze the biological control mechanisms developed by HSMs. This study examines the RF, BRT, and SVM algorithms in order to create maps of the habitat suitability for *Alhagi maurorum* and *Aceria alhagi* in Fars Province, Iran. This study's major goal is to use maps of habitat suitability to aid *Aceria alhagi* in its biological control of *Alhagi maurorum*. Additionally, it demonstrates how the *Aceria alhagi* may be utilized as a biological control agent in the new regions by comparing the parameters used in the research area (the native range) with the parameters of the areas where the *Alhagi maurorum* is recognized as an invasive species. In order to detect where *Alhagi maurorum* and *Aceria alhagi* are established in the study area and to better manage *Alhagi maurorum*, maps of the suitability of *Alhagi maurorum and Aceria alhagi* need to be studied simultaneously.

Material And Methods

2.1 study area

The Fars Province in Iran's southwest served as the study's location (Fig. 1). The study area is located between latitudes 28°00' 29" and 31°00' 36" N, and 52°11' 32" and 54°11' 49" E. Iran has 90 million ha of pastureland, and Fars province is accounting for 15% of that total[40]. Fars Province is one of the largest and most densely populated. A variety of climates exists in this province, making it feasible to cultivate crops all year round[41]. There is around 1,200,000 ha of the land used for agriculture in Fars province[42]. Almost every sort of weather and climate, including cold, mountainous, temperate, semi-arid, desert, forest, and semi-tropical climates, can be found in this province[43]. This allows for the cultivation of a variety of agricultural crops in this province under various climatic conditions[43]. The variation of topography in Fars has had an impact on the types of flora that grow there[44]. Therefore, the elevation range of the study area is 270 to 3491 m. The mean annual temperature ranges from 3 to 19°C, and the mean annual rainfall ranges from 141 to 637 mm. Since the geography and climatic diversity of the studied region clearly demonstrate the importance of determining the habitat's suitability, the current study intends to throw additional light on this topic.

2.2. Methodology

In this study, 13 variables that impact the growth and development of *Aceria alhagi* and *Alhagi maurorum* were taken into account. Elevation, slope degree, slope aspect, plan curvature, distance from rivers, annual mean temperature and rainfall, pH, EC, clay%, silt%, and sand% were among the parameters. Additionally, a global positioning system (GPS) device was used to track the existence of *Alhagi maurorum* and *Aceria alhagi* throughout the province of Fars (Fig. 2). At each location where the species was present, soil samples were also collected in the same manner. The *Alhagi mauorum* data was split into two portions, 70% and 30%, which were utilized for the modeling process and the models' assessment, respectively. For data on *Aceria alhagi* presence, the same process was repeated. In this study, habitat suitability maps were created using RF, BRT, and SVM machine models. The accuracy of the models was assessed using the "receiver operating characteristic (ROC)" curve, and the best model was chosen. A flowchart has been created for this study, to put it briefly (Fig. 3).

2.3. Creating Study Layers

A digital elevation model (DEM) from the "Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER)" with a resolution of 30 m was used to extract the topographic parameters, such as slope degree, aspect, and elevation (Fig. 4A, B and C) (Chen et al., 2022). Additionally, a plan curvature map with a spatial resolution of 30 meters was created using the DEM as a supplementary characteristic (Fig. 4D).

Due to the presence of *Aceria alhagi* and *Alhagi maurorum* at various locations, the samples of soil were analyzed for a thorough investigation of the chemical and physical characteristics of the soil. A hydrometer was used to test the soil's physical characteristics (its percentages of sand, silt, and clay)[45]. A pH meter was used to measure the pH of the soil. A conductivity meter was used to analyze the EC[46]. By using the IDW ("inverse distance weight") method, the values for each physical and chemical characteristic (sand, silt, clay, pH, and EC) were calculated for each soil sample. The data points were then assembled in Arc-GIS 10.8 to plot and analyze each variable. The soil layers were finally constructed using the IDW algorithm, point maps, and study area (Fig. 4E, F, G, H and I). For assessing quantified climatic and ecological data, "inverse distance weighting (IDW)" is among the best methods[47]. When the IDW optimizes the size across locations, the values for each atypical site are based on the weighted sum of the nearby weather stations[48].

Additionally, the mean annual rainfall and temperature were calculated from meteorological stations gathered throughout the Fars province. Using IDW, the representative value for each was determined (Fig. 4J and K). Using a topographic map of the Fars province, a raster map showing the "distance from rivers" and "distance from roads" was produced by ArcGIS 10.8's Euclidean distance algorithm. (Fig. 4L and M).

2.4. Machine Learning Algorithms

2.4.1. Random forest (RF)

The RF is a supervised learning technique that classifies data using multiple trees[49]. A large portion of the decision trees are produced by the RF algorithm as a result of the replacement and ongoing modification of the target's influencing elements. Decisions are then made by combining all of these trees[50]. Three user-defined parameters make up the RF: the number of variables utilized in each tree's creation, which represents the capacity of each independent tree; the number of trees inside the RF; and the minimum number of sensor nodes[51]. With stronger independent trees and a lessening correlation between them, RF prediction power rises[52]. In this approach, a bootstrap tree is grown using 66% of the data, and then a regression analysis is injected at random as the tree grows to split a node. Additionally, the fitted tree is assessed using the other 33% of the dataset[53]. The program makes numerous iterations of this procedure before using the mean of all anticipated values as its final conclusion. The mean decline in precision and the average drop in Gini (a measurement of how inputs are distributed throughout a population) are two parameters in this model that are used to rank each of the important components. When deciding the relative importance of effective factors, the average decrease accuracy (ADA) is more useful than the average decrease Gini index, particularly when considering how environmental factors interact[12,54].

2.4.2. Boosted Regression Trees (Brt)

By integrating multiple models, BRT is one of the several methods that can help a single model to perform better by integrating multiple models[55]. Regression and boosting are the two modeling techniques used by BRT[56]. Boosting is a technique for improving model accuracy, and consequently building, combining, and averaging several models is more efficient and precise than using a single model alone[57]. The single decision tree's biggest flaw is that its comparatively poor data processing is remedied by BRT. For BRT, only the initial tree from the datasets is formed; subsequent trees are developed using the leftover training data from the tree that came before them[58]. Trees are not generated on all the training data; rather, they use a subset of it. The fundamental principle behind this approach is to combine a number of weak prediction models with high predictive error to get strong predictions with low predictive error[59].

2.4.3. Support Vector Machine (Svm)

SVMs were created for both problems involving regression and classification[21]. Originally introduced by Noble[60], SVMs for classification seek to identify the hyperplane that maximizes the margin dividing the hyperplane among two types of data. The nearest point from either class is as far away from the hyperplane as possible. A soft margin approach was given for the situation when the categories are not distinguishable by a linear border, and it permits some locations to be on the incorrect end of a margin[61]. The user-defined variable C controls how often these incorrect classifications occur. SVMs' high level of popularity is largely due to their capacity to represent intricate, non-linear connections[62]. Therefore, the dataset's items are projected onto a high-dimensional space using a nonlinear feature value (often a kernel function). In this expanded feature space, the best separating hyperplane is built, which results in a non-linear decision border in the original space[63]. As a result, numerous techniques to enhance SVMs for multi-class categorization have been suggested[64]. The so-called "one-against-one" strategy was used in this instance. For a wide variety of classification problems in the current world, SVMs produce great prediction accuracy. For categorization, no distributional presumptions are necessary[65].

2.5. Partial Least Squares (Pls)

PLS streamlines and combines information from multiple regression and principal component analysis[66]. Finding Y (reaction) from X (latent) and defining the common organization of the two variables are the goals of PLS[67]. The best way to assess the relative relevance of variables in a model is through PLS regression[68]. Multiple regression is frequently replaced by this regression technique because it might be problematic or because there is a strong connection between the variables[69]. PLS may also provide better accuracy than other techniques. PLS is often regarded as the gold standard for parameter evaluation[70]. This algorithm's framework presupposes that the function of the independent variables at a place explains the value of the dependent variable there. They are also capable of accurately predicting the value of the dependent variable at subsequent times[71].

2.6. Algorithms Evaluation

To validate algorithms, 30% of species presence sites that were not part of the modeling process were utilized. The accuracy of the final habitat suitability maps created by three machine learning algorithms was assessed by utilizing the "receiver operating characteristic (ROC)" curve and calculating the "area under the curve (AUC) " Thus, the best model has the highest AUC, as well as the values (AUC) range from 0.5 to 1[72]. In general, AUC values of 0.9-1, 0.8–0.9, 0.7–0.8, 0.6–0.7, and 0.5–0.6, respectively, imply excellent, very good, good, moderate, and poor classes[73].

Results And Discussion

3.1. Effective factors collinearity test

In general, the multi-collinearity test of parameters is crucial for a habitat suitability map[74]. Multi-collinearity denotes the existence of a linear connection between parameters. In this study, the “tolerance (TOL)” and “variance inflation factor (VIF)” indices are employed to check for multi-collinearity when the values of TOL and VIF are 0.1 and 5 or 10, suggesting multi-collinearity between independent variables, respectively[75,76]. Table 1 displays the findings of the multi-collinearity analysis performed on the 13 habitat suitability parameters employed in this study. This research used 13 variables, including elevation, slope, aspect, plan curvature, silt%, clay%, sand%, distance from rivers, distance from roads, EC, pH, mean annual rainfall, and mean annual temperature. As a consequence, three models could utilize these factors to create the final habitat suitability map.

Table 1
Collinearity Test of Effective Factors

Factors	Tolerance	VIF
Elevation (m)	0.23	4.35
Distance from rivers (m)	0.78	1.39
Distance from roads (m)	0.37	2.68
% Sand	0.25	4.97
% Silt	0.28	4.96
Slope degree	0.80	1.25
Mean annual temperature (^{o}C)	0.20	4.98
% Clay	0.19	4.99
Plan curvature (100/m)	0.93	1.07
EC (dS/m)	0.72	1.38
pH	0.80	1.24
Mean annual rainfall (mm)	0.45	2.22
Aspect	0.95	1.05

3.2. Algorithms For Determining Habitat Suitability

3.2.1. Aceria alhagi

Three models—RF, BRT, and SVM—were utilized to create the *Aceria alhagi* habitat's suitability map in the province of Fars (Fig. 5). Figure 5A shows the outcomes of RF algorithm modeling based on *Aceria alhagi* presence and non-presence locations and 13 influencing factors. The output of this model demonstrates that *Aceria alhagi* prefers the northern portion of Fars Province which is characterized by mean annual temperature of 3.82–7.70°C and, elevation of 1740–3491m. In other words, the region with the best habitat for *Aceria alhagi* also has the coldest climate. The RF model has been assigned the classifications of low (30.40%), moderate (30.09%), high (21.86%), and very high (17.06%) as the final output of the models is separated into four groups (Fig. 6A).

The BRT model is another algorithm that was applied to assess the habitat suitability of *Aceria alhagi*. This model's final outputs are essentially identical to RF (Fig. 5B). According to the BRT model, *Aceria alhagi's* habitat is not as suitable in the southern part. In fact, there are a few remote areas in the southwest with ideal habitats. According to the categorization used to assess the habitat suitability of *Aceria alhagi*, the low, moderate, high, and very high classes received values of 35.63%, 30.02%, 21.28%, and 13.07%, respectively (Fig. 6B).

The third model that was utilized in this study to assess the compatibility of the *Aceria alhagi* habitat was the SVM model. Figure 5C displays the SVM model's findings for four classes. The result of this model demonstrates that the low, medium, high, and low classes are essentially identical across the various areas of Fars Province. As a result, habitat suitability is present in the low 36.63%, moderate 24.65%, high 17.94%, and very high 20.78% classifications (Fig. 6C).

3.2.2. Alhagi maurorum

Three models of RF, BRT, and SVM were also employed to calculate the attractiveness maps of the *Alhagi maurorum* habitat. Figure 5 shows the outcomes of the models. The RF algorithm's findings revealed that the southeast, southwest, northeast, and northwest parts of Fars Province are the species' preferred habitats (Fig. 5D). Additionally, the RF model's final map was separated into four classes: low 25.19%, moderate 33.58%, high 28.70%, and very high 14.56%. (Fig. 6D).

In order to more accurately assess the findings, three comparable models for *Alhagi maurorum* were utilized in this work. The BRT model was utilized for this purpose as well, to determine if the habitat of *Alhagi maurorum* was suitable, and its findings were given in four groups (Fig. 5E). The outcomes of the BRT model used to identify the places in Fars Province that have a high habitat suitability for *Alhagi maurorum* are essentially identical to those of the RF model, with the exception of a portion of the northern areas. Additionally, the BRT algorithm's classification result revealed that the low 30.11%, moderate 27.46%, high 27.87%, and very high 14.56% classes have habitat suitability (Fig. 6E).

In Fig. 5F, the SVM model's findings are presented. It is evident that the *Alhagi maurorum* habitat suitability areas in the Fars Province based on the SVM model are comparable to the BRT model, with the exception being that the northeastern areas do not exhibit high habitat suitability. This model's output was similarly split into four groups, with low, moderate, high, and very high classes being given the values 24.05%, 27.90%, 26.43%, and 21.62%, respectively (Fig. 6F).

Researchers have recently turned to habitat suitability modeling as a reliable and practical method for managing the habitats of various pests, insects, and plant species[77]. Researchers compared the effectiveness of seven data analysis strategies for forecasting the spread of China berry (*Melia azedarach* L.) in a study using three standard metrics for evaluating model accuracy. The RF model offered the maximum degree of accuracy in creating a climate niche model because of its considerable durability and stability. According to the RF forecast findings, *M. azedarach* would profit from future changes in climate by expanding its range, which has a propensity to move north and west of where?[78]. In a different investigation, it was proven that RF and BRT performed better than decision trees, MaxLike, and Lasso overall[79]. Variable significance and complicated variations in reaction to the resolution play a key role in how well models work (REF). Wunderlich et al.[79] encourage researchers to regularly investigate a variety of algorithms, parameters, and frequencies because RF and BRT are strongly advised but could necessitate bias correction techniques.

3.3. General Discussion

Except for Bioclim, all "machine learning and regression" models produced accurate predictions, according to the present results. "Random Forest (RF)" outperformed the other investigated models with 99% AUC and 93% TSS, followed in decreasing order by "Boosted Regression Trees (BRT)", "ensemble," "Generalized Additive Model (GAM)," "Support Vector Machine (SVM)," and "Generalized Linear Model (GLM)"[76]. Our findings also showed that RF and BRT models are better able to simulate the dispersal of the *Aceria alhagi* and *Alhagi maurorum* species. Additionally, for remote sensing-based intrusive SDM, the application of machine learning techniques like the RF and BRT algorithms is absolutely crucial. Similarly, it had been shown in other studies that BRT, Maxent, MLP, RF, and SVM showed excellent performance, with RF being the best at predicting the distribution of *Bombus formosellus*[80].

Prosopis juliflora is anticipated to spread to more regions in Ethiopia, according to Sintayehu et al.[81] who used a variety of algorithms including RF, BRT, SVM, and GLM. They stressed that *P. juliflora* is expected to spread rapidly to numerous drylands in Ethiopia, including major areas in "Afar", "Oromia", "Southern", "Dire Dawa", "Somalia", "Amhara", "Tigray", and "Gambella". This will reduce agricultural output and pose a danger to the region's biodiversity. The invasive species' ongoing range expansion has already had a negative impact on ecosystem services, the economy, and biodiversity. Many pastoralists across the world, in particular, rely on natural resources and other natural ecosystems for their livelihood to survive[82]. We need coordinated and extensive actions due to the existing situation and probable future increases in the range and abundance of invasive species worldwide. The study's findings will also aid in the early discovery and control of invasive species in prospective habitat-friendly niches. Based on our research, we recommend cooperation amongst various stakeholders, research institutes and authorities for early detection and eradication efforts at the national level to create and apply comprehensive biological management of *Alhagi maurorum* by *Aceria alhagi* that would minimize the adverse effects by reducing camelthorn' size and seed production[17]. Although *Alhagi maurorum* is native to Iran, the research findings are extremely helpful for areas where this species is invasive. According to the findings of another study, the RF model outperformed other methods, and it is useful for mapping the proportionate covering of species distribution in agro climatic settings like those of the Afar Region (The Afar Region, previously designated as Region 2, is the home of the Afar people and a local state in northeastern Ethiopia). The GLM, the GBM-BRT, and the DNN performed poorly when considering specificity, precision, kappa, and the AUC, although the GBM and the SVM only slightly less accurately predicted outcomes[83]. However, if a substantially greater quantity of data (i.e., the response variable) is utilized, if there is a lack of training data, or if the research is carried out in a different agroecological environment, MLTs' performances may change (REF).

The results of research by Mudereri et al. (2020)[84] show that RF, CART, SVM, BRT, GLM, and FDA have been used to predict the likelihood of Striga (*Striga asiatica*) incidence in Zimbabwe using multi-source bio - climatic and remotely sensed data. It has been determined that RF, CART, SVM, and the wide range of communication processes yield the most accurate Striga incidence prediction results in Zimbabwe. Additionally, several SVM kernels were utilized to generate GPMs with satisfactory performance. Their performance, however, lags below RF performance. In order to create the habitat suitability model, Pourghasemi and Rahmati (2018)[85] used a variety of models, including the "generalized linear model (GLM)", "generalized additive model (GAM)", "classification and regression trees (CART)", "boosted regression trees (BRT)", "multivariate adaptive regression spline (MARS)", "random forests (RF)", "support vector machines (SVM)", "artificial neural networks (ANN)", "maximum entropy (Maxent)", "penalized maximum likelihood GLM (GLMNET)", "domain, and radial basis function network (RBF)". Their distribution model identified basins as having the highest likelihood of harboring invasive Fallopia species. The Southern Slovak Basin and the Koice Basin have the greatest potential for the propagation of this species.

3.4. Choosing The Optimal Algorithm

As was mentioned in the preceding section, three algorithms—RF, BRT, and SVM—were applied in this work to predict habitat suitability of *Aceria alhagi*. Based on ROC-AUC, machine learning algorithms were assessed. The results demonstrate that RF (89%), BRT (81%), and SVM (79%), respectively, were more accurate at predicting the events when the algorithms were applied to create the map of suitable *Aceria alhagi* habitats (Fig. 7 and Table 2). In other words, the SVM model had good accuracy, whereas the RF and BRT models had very good accuracy.

Table 2
Evaluating algorithms and selecting the best algorithm for *Aceria alhagi* based on the AUC

Test Result Variable(s)	Area	Std. Error^{a}	Asymptotic Sig.^{b}	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
BRT	0.816	0.053	0.000	0.712	0.921
RF	0.890	0.044	0.000	0.803	0.977
SVM	0.790	0.059	0.000	0.674	0.906

A habitat suitability map of *Alhagi maurorum* was also created using machine learning techniques, and the outcomes were quite similar. The ROC curve and area under the curve (AUC) findings show that the RF, BRT, and SVM algorithms have accuracy rates of 89%, 80%, and 73%, respectively (Fig. 7 and Table 3). As a result, the RF and BRT models had very good accuracy, while the SVM model had good accuracy. In general, a key tactic in the process model is the assessment of estimated outcomes[86]. As a standard procedure, the ROC curve is used to evaluate the accuracy of diagnostic tests[87]. Area under the curve (AUC) values for the ROC technique range from 0.5 to 1.0[88]. If the constructed model is unable to forecast the existence of species more correctly than probability, the AUC is equal to 0.5. In comparison, the prediction has an AUC value of 1, which is ideal[65]. When training the habitat suitability models, the AUC value takes the species pixels into account[89]. Using existing species in the training phase, this approach was utilized to assess the accuracy of habitat suitability maps. However, in order to calculate accuracy in the validation stage, we employed species that weren't used in the training stage[90]. What is evident is that in recent years, ROC-AUC has been widely employed to assess habitat suitability maps[91].

Table 3
Evaluating algorithms and selecting the best algorithm for *Alhagi maurorum* based on the AUC

Test Result Variable(s)	Area	Std. Error^{a}	Asymptotic Sig.^{b}	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
BRT	0.800	0.043	0.000	0.716	0.884
RF	0.894	0.031	0.000	0.834	0.955
SVM	0.733	0.048	0.000	0.640	0.826

3.5. Importance Of Factors By PIs

Alhagi maurorum and *Aceria alhagi* are threshold-dependent processes influenced by a wide array of useful parameters[92]. Therefore, in order to conduct a habitat suitability evaluation, it is required to determine the parameters that are effective for *Aceria alhagi* and *Alhagi maurorum*, as well as their significance among the conditioning factors[69]. A greater understanding of the impact that each influencing factor has on the overall evaluation of habitat suitability was achieved by developing the PLS approach after training data selection. For example, Fig. 8A and B show the 13 variables for *Aceria alhagi* and *Alhagi maurorum* habitat suitability models in the correct order of significance[93,94]. The findings show that, in that order, roads, slope, clay, and temperature are the most important variables for *Aceria alhagi*. However, plan, aspect, rain, and elevation, were of the least consequence (Fig. 8A).

The PLS algorithm also looked at the parameters that were important in the *Alhagi maurorum* habitat suitability modeling process. The findings revealed that the three most important variables were road, slope, and EC. On the other hand, the suitability of the *Alhagi maurorum*'s habitat was not significantly impacted by rain, silt, or aspect, respectively (Fig. 8B).

In the current study, the abundance of *Aceria alhagi* and *Alhagi maurorum* was substantially greater close to the roads. The findings of Delgado et al. (2017)[95], who discovered early indications of relatively high *Aceria alhagi* abundance near roads and in the area of road underpasses, are consistent with our study. This outcome was connected to the vegetation around the road and the presence of ticks. Another study established that closer to road borders than farther away, increased tick abundance was seen[96]. Along remote road edges with little traffic, adult ticks were seen acting aggressively. Ticks may have a better chance of finding hosts if they spend a lot of time on roadside vegetation. Our findings also suggest that roads may contribute to an increase in tick development and transmission. Since roads act as a barrier to stopping tick movement, Hornok et al. (2017)[97] show that roads may influence disparities in tick species composition and tick-borne pathogen frequency along their two sides. The slope is a significant factor that determines whether a certain tick habitat is suitable. A study indicated that younger ticks are sparser on lower slopes, while older ticks are more numerous on higher slopes[98].

One of the main factors contributing to the degradation of plant ecosystems is human disturbance. The amount of the *Alhagi maurorum* increased as the distance from the highways shrank in this investigation as well. According to Jahantigh and Pessarakl)2021([99] *Alhagi maurorum* distribution expanded as the distance from a road decreased. Furthermore, it is crucial to consider how the slope component affects the distribution of *Alhagi maurorum*. Water runoff and the spread of invasive plant seeds are both a result of the land's slope[100]. As a result, in this study, it is also determined that *Alhagi maurorum* is more abundant on low slopes.

3.6. The Perspective Of Hsms And Mlts

In general, it is evident that *Aceria alhagi* has been shown by Bijani et al. (2021)[17] to act as a potential biological control by preventing the growth and development of *Alhagi maurorum*. The main goal of this research was to find a way to extend the control of *Alhagi maurorum* such that even the threat of its appearance could be used in areas where it is known to be an invasive species. When it comes to managing invasive plants, habitat suitability models (HSMs) and species distribution models (SDMs) are often utilized nowadays[101,102]. As a result, a novel approach in this area was to apply habitat suitability modeling.

By identifying the environmental factors limiting a species' distribution, HSMs seek to define the "envelope" that best captures the species' geographic range boundaries[103]. They are created by connecting the distributions of extant species to their current surroundings. By extrapolating these associations to certain environmental change scenarios, future species' natural geographical ranges are projected[104]. Measures of climate (such as temperature and rainfall), landscape structure (such as connectivity indices), vegetation heterogeneity (such as ecotone cover), resources (such as insect availability), soil characteristics (such as physical and chemical properties), the topography of an area (such as elevation, slope, aspect, and so on), and biotic information are frequently used as variables for habitat suitability modeling of plants[105].

Environmental variables can exert direct or indirect effects on species and are optimally chosen to reflect the three main types of influences on the species: (1) limiting factors, defined as factors controlling species’ eco-physiology (e.g., minimum winter temperature or high summer temperatures) or appearance (e.g., competition and facilitation); (2) disturbances, defined as all types of perturbations affecting environmental systems (e.g., fire frequency); and (3) resources, defined as all materials that can be assimilated by organisms (e.g., availability of seeds or insects). The environmental data related to these three main types of influence depict the environmental niche of the species[106]. The environmental information pertaining to these three primary categories of effect shows the species' environmental niche[103]. The ecological niche is often multidimensional, and different aspects may be significant at various geographical scales. In the patterns of habitat utilization, these scale-dependent interactions between niche traits and plant species distributions frequently produce hierarchical structures[107].

SDMs are very important, although the field of computer science has paid them very little attention. Although mapping habitat appropriateness using HSMs is our main objective, our other objective with this effort, we hope to do two things: first, provide computer scientists with the knowledge they need to understand the SDM literature and, second, create ML-based SDM algorithms that are beneficial to the environment. These characteristics could be extremely useful in ecology and agriculture, with potential future uses in plant management and conservation. The method may be used, for instance, to model distribution changes brought on by climate change. Additionally, it represents a novel strategy in relation to the many models mentioned in the literature.

Machine learning technology has recently been created, particularly for SDMs[108]. Numerous studies attest to the remarkable accuracy of algorithmically generated habitat suitability maps[109–111]. From our perspective, the main issue with the majority of these comparisons is that they only validate model performance (defined as the match up among both predicted and observed species' distributions) against the needs under current conditions, despite the fact that most models are approximately accurate in trying to project distributions under present environmental conditions. However, highly diverse model structures may be the origin of what appear to be minor variations in estimates of present distributions, leading to unsettlingly divergent projections for novel conditions.

The overall conclusion is that *Alhagi maurorum* can be biologically controlled in both its native and invasive ranges by introducing habitat suitability maps. In fact, the findings of this study add to those of Bijani et al. (2021)[17]. They found that the *Alhagi maurorum* was controlled by the *Aceria alhagi*. Now that we have created maps of habitat suitability, we can extend the reach of this biological control. By annihilating the inflorescences and branches of *Alhagi maurorum, Aceria alhagi* has the ability to stop its growth. We may now considerably more successfully accomplish our aim of controlling *Alhagi maurorum* by taking into account the habitat suitability maps of both *Alhagi maurorum* and *Aceria alhagi*. We can steer *Aceria alhagi* in that direction by using maps that show the favorable and vulnerable locations of the *Alhagi maurorum* habitat. Since *Aceria alhagi* can control *Alhagi maurorum*, it is predicted that *Alhagi maurorum* would be more controlled in regions with greater *Aceria alhagi* habitat. The *Aceria alhagi* habitat suitability map also conveys the idea that by taking crucial aspects into account, we may expand the *Aceria alhagi* range and in order to control *Alhagi maurorum*. For instance, in this study, roads, slope, clay, and temperature were the most significant elements; thus, the *Aceria alhagi* may be produced by taking these aspects into account. Furthermore, this approach could be extensively explored in regions where *Alhagi maurorum* is regarded as an invasive species. In other words, the regions that need to be managed are identified by creating a map of the habitats of *Alhagi maurorum* and *Aceria alhagi.*

Conclusion

In this work, we validate the claim that adopting HSMs approaches, particularly machine learning technologies, can result in considerable increases in the accuracy of species distribution forecasts. Our findings may have significant ramifications for area protection and management planning studies, in which incomplete or biased field data should be appropriately supplemented by species distribution modeling. However, our findings also demonstrated that algorithms enhance the habitat suitability prediction's accuracy. Even though the *Alhagi maurorum* is native to Iran, it can damage a lot of crops. According to our research, the distribution of these two species can be impacted by the slope of the land and roads, since they have the biggest impact on the habitat suitability of the *Alhagi maurorum* and the *Aceria alhagi*. Furthermore, our research has revealed that *Aceria alhagi* are more common in colder climates, implying that *Aceria alhagi* could be used as a biological control in other colder climate areas where *Alhagi maurorum* is recognized as an invasive species. Therefore, it is suggested to carefully monitor the *Aceria alhagi* habitats in these areas to enable early detection and stop the invasion of *Alhagi maurorum*. The outcomes of our models may assist in the development of management strategies to postpone or stop invasions as well as help identify the environmental factors that encourage *Alhagi maurorum*'s propensity for invasion. Combination of habitat suitability models can offer insightful information about the threat presented by invasive species, but as our research demonstrates, care should be taken in choosing the environmental factors that are used to predict species dispersal.

Declarations

Data availability statement:

All data used for analyses are available from the corresponding author upon request.

Funding:

This research was partly funded by the department of Plant Production and Genetics, School of Agriculture, Shiraz University, Shiraz, Iran.(Number of fund: KM:32145)

Author contributions:

ED, FB, RN, AZ and ME designed the experiments, ran models, analyzed the results, and wrote and reviewed the manuscript. All authors reviewed the final manuscript.

Competing interests:

The authors declare that they have no competing interests.

Author information

Affiliations

Department of Plant Production and Genetics, School of Agriculture, Shiraz University, Shiraz, Iran

Emran Dastres, Ruhollah Naderi, Afshin Zamani & Mohsen Edalat

Department of Plant Production and Genetics, Rafsanjan University, Rafsanjan, Iran

Farzad bijani

Research involving plants

The department of Plant Production and Genetics (PPG) has granted us permission to undertake research on 170 *Alhagi maurorum* ecotypes chosen from the province of Fars. International, national, and/or institutional rules and regulations are followed in the current study's usage of plants and experimental research procedures. There is no research by any of the authors in this study that involved using either human subjects or animals.

References

1. El-Zahar, H. *et al.* UPLC-PDA-MS/MS profiling and healing activity of polyphenol-rich fraction of Alhagi maurorum against oral ulcer in rats. *Plants* **11**, 455 (2022).

2. Yuan, C. *et al.* Anti-human ovarian cancer and cytotoxicity effects of nickel nanoparticles green-synthesized by Alhagi maurorum leaf aqueous extract. *J. Exp. Nanosci.* **17**, 113–125 (2022).
3. Kazemi, M. & Ghasemi Bezdi, K. An investigation of the nutritional value of camelthorn (Alhagi maurorum) at three growth stages and its substitution with part of the forage in Afshari ewes' diets. *Anim. Feed Sci. Technol.* **271**, 114762 (2021).
4. Ebrahimi, A., Ehteshami, M. & Dahrazma, B. Isotherm and kinetic studies for the biosorption of cadmium from aqueous solution by Alhaji maurorum seed. *Process Saf. Environ. Prot.* **98**, 374–382 (2015).
5. Heikal, Y. M., El-Esawi, M. A., Naidu, R. & Elshamy, M. M. Eco-biochemical responses, phytoremediation potential and molecular genetic analysis of Alhagi maurorum grown in metal-contaminated soils. *BMC Plant Biol.* **22**, 383 (2022).
6. Lambdon, P. W. & Hulme, P. E. How strongly do interactions with closely-related native species influence plant invasions? Darwin's naturalization hypothesis assessed on Mediterranean islands. *J. Biogeogr.* **33**, 1116–1125 (2006).
7. Schlaepeer, M. A., Sax, D. O. V. F. & Olden, J. The potential conservation value of non-native species. *Conserv. Biol.* **25**, 428–437 (2011).
8. Cuthbert, R. N., Diagne, C., Haubrock, P. J., Turbelin, A. J. & Courchamp, F. Are the "100 of the world's worst" invasive species also the costliest? *Biol. Invasions* **24**, 1895–1904 (2022).
9. Seastedt, T. R. Biological control of invasive plant species: a reassessment for the Anthropocene. *New Phytol.* **205**, 490–502 (2015).
10. Moran, P. J. *et al.* Chapter 6 - Mass-production of arthropods for biological control of weeds: a global perspective. in (eds. Morales-Ramos, J. A., Rojas, M. G. & Shapiro-Ilan, D. I. B. T.-M. P. of B. O. (Second E.) 157–194 (Academic Press, 2023). doi:https://doi.org/10.1016/B978-0-12-822106-8.00014-2.
11. Morin, L. *et al.* Decline of the invasive plant Asparagus asparagoides within the first seven years after release of biological control agents in Australia. *Biol. Control* **165**, 104795 (2022).
12. Li, J. *et al.* Interactions between invasive plants and heavy metal stresses: a review. *J. Plant Ecol.* **15**, 429–436 (2022).
13. McCulloch, G. A., Makinson, J. R., Purcell, M. F., Raghu, S. & Walter, G. H. A prospective and iterative approach to finding safe weed biological control agents – testing ecological and evolutionary hypotheses with molecular evidence. *Biol. Control* **169**, 104887 (2022).
14. Johnson, A. C., Williams, R. J. & Matthiessen, P. The potential steroid hormone contribution of farm animals to freshwaters, the United Kingdom as a case study. *Sci. Total Environ.* **362**, 166–178 (2006).
15. Canavan, K., Magengelele, N. L., Paterson, I. D., Williams, D. A. & Martin, G. D. Uncovering the phylogeography of Schinus terebinthifolia in South Africa to guide biological control. *AoB Plants* **14**, plab078 (2022).
16. Strong, D. R. & Pemberton, R. W. Biological Control of Invading Species–Risk and Reform. *Science (80-.).* **288**, 1969–1970 (2000).

17. Bijani, F. *et al.* Impact of Aceria alhagi (Acari: Eriophyidae) as a potential biological control agent on the invasive weed Alhagi maurorum (Fabaceae) in its native range. *Exp. Appl. Acarol.* **83**, 229–239 (2021).
18. Chlingaryan, A., Sukkarieh, S. & Whelan, B. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Comput. Electron. Agric.* **151**, 61–69 (2018).
19. Kiangala, S. K. & Wang, Z. An effective adaptive customization framework for small manufacturing plants using extreme gradient boosting-XGBoost and random forest ensemble learning algorithms in an Industry 4.0 environment. *Mach. Learn. with Appl.* **4**, 100024 (2021).
20. G C, S. *et al.* Weed and crop species classification using computer vision and deep learning technologies in greenhouse conditions. *J. Agric. Food Res.* **9**, 100325 (2022).
21. Xia, J. *et al.* Performance optimization of support vector machine with oppositional grasshopper optimization for acute appendicitis diagnosis. *Comput. Biol. Med.* **143**, 105206 (2022).
22. Talukdar, S. *et al.* Coupling geographic information system integrated fuzzy logic-analytical hierarchy process with global and machine learning based sensitivity analysis for agricultural suitability mapping. *Agric. Syst.* **196**, 103343 (2022).
23. Ismail, R., Mutanga, O. & Peerbhay, K. The identification and remote detection of alien invasive plants in commercial forests: An Overview. *South African J. Geomatics* **5**, 49 (2016).
24. Rodríguez-Garlito, E. C., Paz-Gallardo, A. & Plaza, A. Automatic Detection of Aquatic Weeds: A Case Study in the Guadiana River, Spain. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **15**, 8567–8585 (2022).
25. Lake, T. A., Briscoe Runquist, R. D. & Moeller, D. A. Deep learning detects invasive plant species across complex landscapes using Worldview-2 and Planetscope satellite imagery. *Remote Sens. Ecol. Conserv.* **8**, 875–889 (2022).
26. Mahabir, R., Croitoru, A., Crooks, A. T., Agouris, P. & Stefanidis, A. A critical review of high and very high-resolution remote sensing approaches for detecting and mapping slums: trends, challenges and emerging opportunities. *Urban Sci.* **2**, (2018).
27. Guo, Y., Zhao, Y., Rothfus, T. A. & Avalos, A. S. A novel invasive plant detection approach using time series images from unmanned aerial systems based on convolutional and recurrent neural networks. *Neural Comput. Appl.* **34**, 20135–20147 (2022).
28. Haltuch, M. A., Berkman, P. A. & Garton, D. W. Geographic information system (GIS) analysis of ecosystem invasion: Exotic mussels in Lake Erie. *Limnol. Oceanogr.* **45**, 1778–1787 (2000).
29. Racelis, A. E. *et al.* Facilitative Ecological Interactions Between Invasive Species: Arundo donax Stands as Favorable Habitat for Cattle Ticks (Acari: Ixodidae) Along the U.S.–Mexico Border. *J. Med. Entomol.* **49**, 410–417 (2012).
30. Guo, Q. *et al.* Urban tree classification based on object-oriented approach and random forest algorithm using unmanned aerial vehicle (uav) multispectral imagery. *Remote Sensing* vol. 14 (2022).

31. Flores, A. M., Demsas, F., Leeper, N. J. & Ross, E. G. Leveraging machine learning and artificial intelligence to improve peripheral artery disease detection, treatment, and outcomes. *Circ. Res.* **128**, 1833–1850 (2021).
32. Taniguchi, M. *et al.* Combining machine learning and nanopore construction creates an artificial intelligence nanopore for coronavirus detection. *Nat. Commun.* **12**, 3726 (2021).
33. Roshan Kumar & Purabi Saha. A review on artificial intelligence and machine learning to improve cancer management and drug discovery. *Int. J. Res. Appl. Sci. Biotechnol.* **9**, 149–156 (2022).
34. Jin, S., Huang, J., Hu, R. & Rozelle, S. The Creation and Spread of Technology and Total Factor Productivity in China's Agriculture. *Am. J. Agric. Econ.* **84**, 916–930 (2002).
35. Akhter, R. & Sofi, S. A. Precision agriculture using IoT data analytics and machine learning. *J. King Saud Univ. - Comput. Inf. Sci.* **34**, 5602–5618 (2022).
36. Aslan, M. F., Durdu, A., Sabanci, K., Ropelewska, E. & Gültekin, S. S. A Comprehensive Survey of the Recent Studies with UAV for Precision Agriculture in Open Fields and Greenhouses. *Appl. Sci.* **12**, (2022).
37. Lin, Z. & Chou, W.-C. Machine Learning and Artificial Intelligence in Toxicological Sciences. *Toxicol. Sci.* **189**, 7–19 (2022).
38. Kler, R. *et al.* Machine Learning and Artificial Intelligence in the Food Industry: A Sustainable Approach. *J. Food Qual.* **2022**, 8521236 (2022).
39. Bradley, B. A. Distribution models of invasive plants over-estimate potential impact. *Biol. Invasions* **15**, 1417–1429 (2013).
40. Raiesi, F. The conversion of overgrazed pastures to almond orchards and alfalfa cropping systems may favor microbial indicators of soil quality in Central Iran. *Agric. Ecosyst. Environ.* **121**, 309–318 (2007).
41. Heidari, A., Mahmoodi, S., Roozitalab, M. H. & Mermut, A. R. Diversity of clay minerals in the vertisols of three different climatic regions in Western Iran. *J. Agric. Sci. Technol.* **10**, 269–284 (2008).
42. Khormali, F., Abtahi, A. & Stoops, G. Micromorphology of calcitic features in highly calcareous soils of Fars Province, Southern Iran. *Geoderma* **132**, 31–46 (2006).
43. Hashempour Motlagh Shirazi, S., Khalili, D., Zand-Parsa, S. & Shirvani, A. Spatio-temporal variability of extreme precipitation characteristics under different climatic conditions in Fars province, Iran. *Environ. Dev. Sustain.* **24**, 11348–11368 (2022).
44. Ghaemi, Z. & Noshadi, M. Surface water quality analysis using multivariate statistical techniques: a case study of Fars Province rivers, Iran. *Environ. Monit. Assess.* **194**, 178 (2022).
45. Orhan, U., Kilinc, E., Albayrak, F., Aydin, A. & Torun, A. Ultrasound Penetration-Based Digital Soil Texture Analyzer. *Arab. J. Sci. Eng.* **47**, 10751–10767 (2022).
46. Shang, X., Yu, J., Wang, C. & Du, Y. An aptasensing strategy using the phosphatase-mimic nanozyme and pH meter as signal readout. *Electroanalysis* **34**, 535–541 (2022).

47. Qu, R. *et al.* Prediction on the combined toxicities of stimulation-only and inhibition-only contaminants using improved inverse distance weighted interpolation. *Chemosphere* **287**, 132045 (2022).
48. Tahama, K., Baride, A., Gupta, G., Erram, V. C. & Baride, M. V. Spatial variation of sub-surface heterogenieties within the dyke swarm of Nandurbar region, Maharashtra, India, for groundwater exploration using Inverse Distance Weighted technique. *HydroResearch* **5**, 1–12 (2022).
49. Biau, G. & Scornet, E. A random forest guided tour. *TEST* **25**, 197–227 (2016).
50. Lin, W., Wu, Z., Lin, L., Wen, A. & Li, J. An Ensemble Random Forest Algorithm for Insurance Big Data Analysis. *IEEE Access* **5**, 16568–16575 (2017).
51. Sarica, A., Cerasa, A. & Quattrone, A. Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review. *Front. Aging Neurosci.* **9**, (2017).
52. Daneshvar, D. & Behnood, A. Estimation of the dynamic modulus of asphalt concretes using random forests algorithm. *Int. J. Pavement Eng.* **23**, 250–260 (2022).
53. Schonlau, M. & Zou, R. Y. The random forest algorithm for statistical learning. *Stata J.* **20**, 3–29 (2020).
54. Guo, Y. A new paradigm of "Real-Time" stroke risk prediction and integrated care management in the digital health era: innovations using machine learning and artificial intelligence approaches. *Thromb. Haemost.* **122**, 005–007 (2022).
55. Abedi, R., Costache, R., Shafizadeh-Moghadam, H. & Pham, Q. B. Flash-flood susceptibility mapping based on XGBoost, random forest and boosted regression trees. *Geocarto Int.* **37**, 5479–5496 (2022).
56. Hao, M. *et al.* Global potential distribution of Oryctes rhinoceros, as predicted by Boosted Regression Tree model. *Glob. Ecol. Conserv.* **37**, e02175 (2022).
57. Alnahit, A. O., Mishra, A. K. & Khan, A. A. Stream water quality prediction using boosted regression tree and random forest models. *Stoch. Environ. Res. Risk Assess.* **36**, 2661–2680 (2022).
58. Gong, E. S. *et al.* Identification of key phenolic compounds responsible for antioxidant activities of free and bound fractions of blackberry varieties' extracts by boosted regression trees. *J. Sci. Food Agric.* **102**, 984–994 (2022).
59. Sharma, P. & Sahoo, B. B. Precise prediction of performance and emission of a waste derived Biogas–Biodiesel powered Dual–Fuel engine using modern ensemble Boosted regression Tree: A critique to Artificial neural network. *Fuel* **321**, 124131 (2022).
60. Noble, W. S. What is a support vector machine? *Nat. Biotechnol.* **24**, 1565–1567 (2006).
61. Zhou, J. *et al.* Predicting tunnel squeezing using support vector machine optimized by whale optimization algorithm. *Acta Geotech.* **17**, 1343–1366 (2022).
62. Essam, Y. *et al.* Predicting streamflow in Peninsular Malaysia using support vector machine and deep learning algorithms. *Sci. Rep.* **12**, 3883 (2022).
63. Pant, M. & Kumar, S. Fuzzy time series forecasting based on hesitant fuzzy sets, particle swarm optimization and support vector machine-based hybrid method. *Granul. Comput.* **7**, 861–879 (2022).

64. Syah, R., Towfighi Naeem, M. H., Daneshfar, R., Dehdar, H. & Soulgani, B. S. On the prediction of methane adsorption in shale using grey wolf optimizer support vector machine approach. *Petroleum* **8**, 264–269 (2022).

65. Khan, A. M. *et al.* MaxEnt modelling and impact of climate change on habitat suitability variations of economically important Chilgoza Pine (Pinus gerardiana Wall.) in South Asia. *Forests* **13**, (2022).

66. Franzoi, M., Costa, A., Goi, A., Penasa, M. & De Marchi, M. Effectiveness of visible – Near infrared spectroscopy coupled with simulated annealing partial least squares analysis to predict immunoglobulins G, A, and M concentration in bovine colostrum. *Food Chem.* **371**, 131189 (2022).

67. Pan, H. *et al.* Label-free Au NRs-based SERS coupled with chemometrics for rapid quantitative detection of thiabendazole residues in citrus. *Food Chem.* **375**, 131681 (2022).

68. Menduni, G. *et al.* High-concentration methane and ethane QEPAS detection employing partial least squares regression to filter out energy relaxation dependence on gas matrix composition. *Photoacoustics* **26**, 100349 (2022).

69. Lou, Z., Wang, Y., Si, Y. & Lu, S. A novel multivariate statistical process monitoring algorithm: Orthonormal subspace analysis. *Automatica* **138**, 110148 (2022).

70. Fernández-Habas, J. *et al.* Estimating pasture quality of Mediterranean grasslands using hyperspectral narrow bands from field spectroscopy by Random Forest and PLS regressions. *Comput. Electron. Agric.* **192**, 106614 (2022).

71. Zhu, P., Chen, X., Meng, F., He, X. & Chen, W. Thermal risk modeling and safety optimization of an arylamine diazo reaction based on PLS algorithm. *Process Saf. Environ. Prot.* **166**, 108–112 (2022).

72. Gneiting, T. & Walz, E.-M. Receiver operating characteristic (ROC) movies, universal ROC (UROC) curves, and coefficient of predictive ability (CPA). *Mach. Learn.* **111**, 2769–2797 (2022).

73. Mallick, J. *et al.* Proposing receiver operating characteristic-based sensitivity analysis with introducing swarm optimized ensemble learning algorithms for groundwater potentiality modelling in Asir region, Saudi Arabia. *Geocarto Int.* **37**, 4361–4389 (2022).

74. Hirzel, A. & Guisan, A. Which is the optimal sampling strategy for habitat suitability modelling. *Ecol. Modell.* **157**, 331–341 (2002).

75. Akinwande, M. O., Dikko, H. G. & Samson, A. Variance inflation factor: As a condition for the inclusion of suppressor variable(s) in regression analysis. *Open J. Stat.* **5**, 754–767 (2015).

76. O'brien, R. M. A caution regarding rules of thumb for variance inflation factors. *Qual. Quant.* **41**, 673–690 (2007).

77. Mohammady, M. *et al.* Modeling and Prediction of Habitat Suitability for Ferula gummosa Medicinal Plant in a Mountainous Area. *Nat. Resour. Res.* **30**, 4861–4884 (2021).

78. Feng, L. *et al.* Predicting suitable habitats of Melia azedarach L. in China using data mining. *Sci. Rep.* **12**, 12617 (2022).

79. Wunderlich, R. F., Mukhtar, H. & Lin, Y.-P. Comprehensively evaluating the performance of species distribution models across clades and resolutions: choosing the right tool for the job. *Landsc. Ecol.*

37, 2045–2063 (2022).

80. Lu, M.-L. & Huang, J.-Y. Predicting negative Effects of Climate Change on Taiwan's endemic Bumblebee Bombus formosellus. *J. Insect Conserv.* **26**, 21–32 (2022).
81. Sintayehu, D. W., Dalle, G. & Bobasa, A. F. Impacts of climate change on current and future invasion of Prosopis juliflora in Ethiopia: environmental and socio-economic implications. *Heliyon* **6**, e04596 (2020).
82. Yan, X., Zhenyu, L., Gregg, W. P. & Dianmo, L. Invasive species in China — an overview. *Biodivers. Conserv.* **10**, 1317–1341 (2001).
83. Shiferaw, H., Bewket, W. & Eckert, S. Performances of machine learning algorithms for mapping fractional cover of an invasive plant species in a dryland ecosystem. *Ecol. Evol.* **9**, 2562–2574 (2019).
84. Mudereri, B. T. *et al.* Multi-source spatial data-based invasion risk modeling of Striga (Striga asiatica) in Zimbabwe. *GIScience Remote Sens.* **57**, 553–571 (2020).
85. Pourghasemi, H. R. & Rahmati, O. Prediction of the landslide susceptibility: Which algorithm, which precision? *CATENA* **162**, 177–192 (2018).
86. Pinsky, P. F. Scaling of True and Apparent ROC AUC with Number of Observations and Number of Variables. *Commun. Stat. - Simul. Comput.* **34**, 771–781 (2005).
87. Carrington, A. M. *et al.* 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 (2023).
88. Zhuang, H. *et al.* A flagship species-based approach to efficient, cost-effective biodiversity conservation in the Qinling Mountains, China. *J. Environ. Manage.* **305**, 114388 (2022).
89. Malik, K., Saranya, K. R. L., Reddy, C. S. & Varghese, A. O. Predicting the habitat suitability of Dipterocarpus indicus: an endemic and endangered species in the Western Ghats, India. *Spat. Inf. Res.* **30**, 729–738 (2022).
90. Mallett, S., Halligan, S., Collins, G. S. & Altman, D. G. Exploration of Analysis Methods for Diagnostic Imaging Tests: Problems with ROC AUC and Confidence Scores in CT Colonography. *PLoS One* **9**, e107633 (2014).
91. Yousaf, A. *et al.* Identification of suitable habitat for Taxus wallichiana and Abies pindrow in moist temperate forest using maxent modelling technique. *Saudi J. Biol. Sci.* **29**, 103459 (2022).
92. Marini, F. *et al.* Eriophyid mites in classical biological control of weeds: progress and challenges. *Insects* **12**, 513 (2021).
93. Song, Z. & Yu, C. Investor sentiment indices based on k-step PLS algorithm: A group of powerful predictors of stock market returns. *Int. Rev. Financ. Anal.* **83**, 102321 (2022).
94. Sarstedt, M., Radomir, L., Moisescu, O. I. & Ringle, C. M. Latent class analysis in PLS-SEM: A review and recommendations for future applications. *J. Bus. Res.* **138**, 398–407 (2022).

95. Delgado, J. D., Abreu-Yanes, E., Abreu-Acosta, N., Flor, M. D. & Foronda, P. Vertebrate ticks distribution and their role as vectors in relation to road edges and underpasses. *Vector-Borne Zoonotic Dis.* **17**, 376–383 (2017).
96. Bellato, A. *et al.* Risk of tick-borne zoonoses in urban green areas: A case study from Turin, northwestern Italy. *Urban For. Urban Green.* **64**, 127297 (2021).
97. Hornok, S. *et al.* Impact of a freeway on the dispersal of ticks and Ixodes ricinus-borne pathogens: forested resting areas may become Lyme disease hotspots. *Acta Vet. Hung.* **65**, 242–252 (2017).
98. Cadenas, F. M. *et al.* Phenology of Ixodes ricinus and Infection with Borrelia burgdorferi sensu lato Along a North- and South-Facing Altitudinal Gradient on Chaumont Mountain, Switzerland. *J. Med. Entomol.* **44**, 683–693 (2007).
99. Jahantigh, M. & Pessarakli, M. Causes and Effects of Gully Erosion on Agricultural Lands and the Environment. *Commun. Soil Sci. Plant Anal.* **42**, 2250–2255 (2011).
100. Agarwal, T., Gupta, A. K., Patel, A. K. & Shekhawat, N. S. Micropropagation and validation of genetic homogeneity of Alhagi maurorum using SCoT, ISSR and RAPD markers. *Plant Cell, Tissue Organ Cult.* **120**, 313–323 (2015).
101. Mainali, K. P. *et al.* Projecting future expansion of invasive species: comparing and improving methodologies for species distribution modeling. *Glob. Chang. Biol.* **21**, 4464–4480 (2015).
102. Chapman, D., Pescott, O. L., Roy, H. E. & Tanner, R. Improving species distribution models for invasive non-native species with biologically informed pseudo-absence selection. *J. Biogeogr.* **46**, 1029–1040 (2019).
103. Fragnière, Y. *et al.* Mapping habitats sensitive to overgrazing in the Swiss Northern Alps using habitat suitability modeling. *Biol. Conserv.* **274**, 109742 (2022).
104. Store, R. & Jokimäki, J. A GIS-based multi-scale approach to habitat suitability modeling. *Ecol. Modell.* **169**, 1–15 (2003).
105. Hirzel, A. H., Le Lay, G., Helfer, V., Randin, C. & Guisan, A. Evaluating the ability of habitat suitability models to predict species presences. *Ecol. Modell.* **199**, 142–152 (2006).
106. Broennimann, O. *et al.* Measuring ecological niche overlap from occurrence and spatial environmental data. *Glob. Ecol. Biogeogr.* **21**, 481–497 (2012).
107. Nzei, J. M. *et al.* Ecological niche modeling of water lily (Nymphaea L.) species in Australia under climate change to ascertain habitat suitability for conservation measures. *Plants* **11**, 1874 (2022).
108. Kellenberger, B., Cole, E., Marcos, D. & Tuia, D. Training techniques for presence-only habitat suitability mapping with deep learning. in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium* 5085–5088 (2022).
109. Sarma, K. *et al.* Habitat suitability of Gymnocladus assamicus - A critically endangered plant of Arunachal Pradesh, India using machine learning and statistical modeling. *Acta Ecol. Sin.* **42**, 398–406 (2022).

110. Edalat, M. *et al.* Spatial mapping Zataria multiflora using different machine-learning algorithms. *CATENA* **212**, 106007 (2022).

111. Kopsco, H. L., Smith, R. L. & Halsey, S. J. A scoping review of species distribution modeling methods for tick vectors. *Front. Ecol. Evol.* **10**, 893016 (2022).

Figures

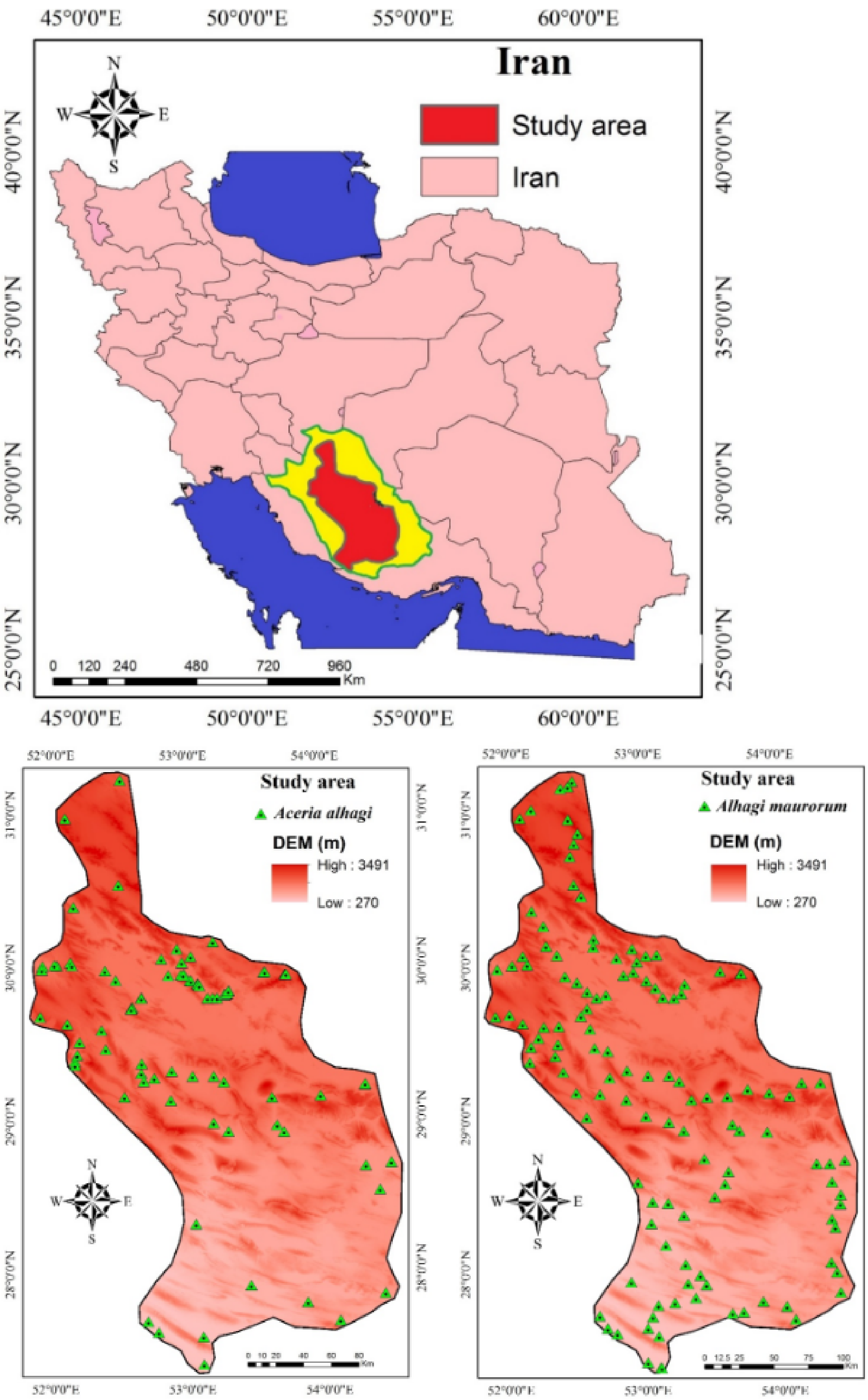

Figure 1

The study area in Fars province, southwest of Iran

Figure 2

Identification and sampling of *Alhagi maurorum* and *Aceria alhagi* (photos by Frazad Bijani)

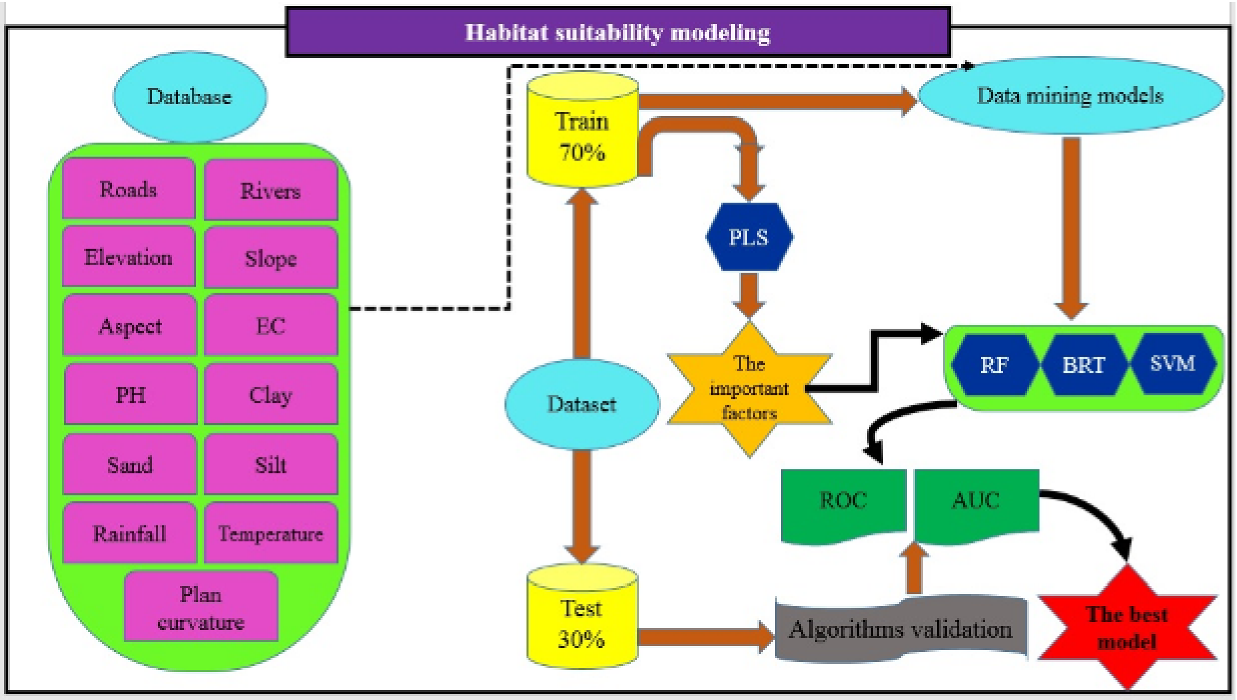

Figure 3

A flowchart of the habitat suitability modeling process

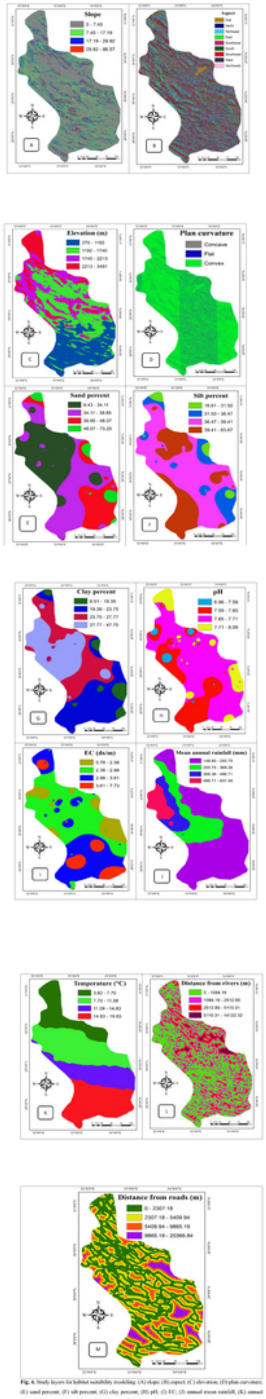

Figure 4

Study layers for habitat suitability modeling: (A) slope; (B) aspect; (C) elevation; (D) plan curvature; (E) sand percent; (F) silt percent; (G) clay percent; (H) pH; (I) EC; (J) annual mean rainfall; (K) annual mean temperature; (L) distance from rivers; (M) distance to roads

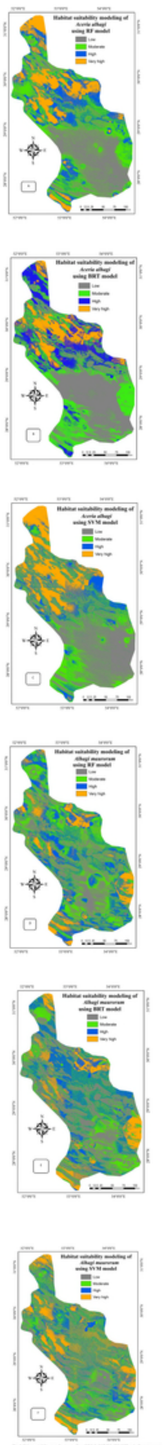

Figure 5

Habitat suitability maps for *Aceria alhagi* (RF "A", BRT "B", SVM "C") and *Alhagi maurorum* (RF "D", BRT "E", SVM "F")

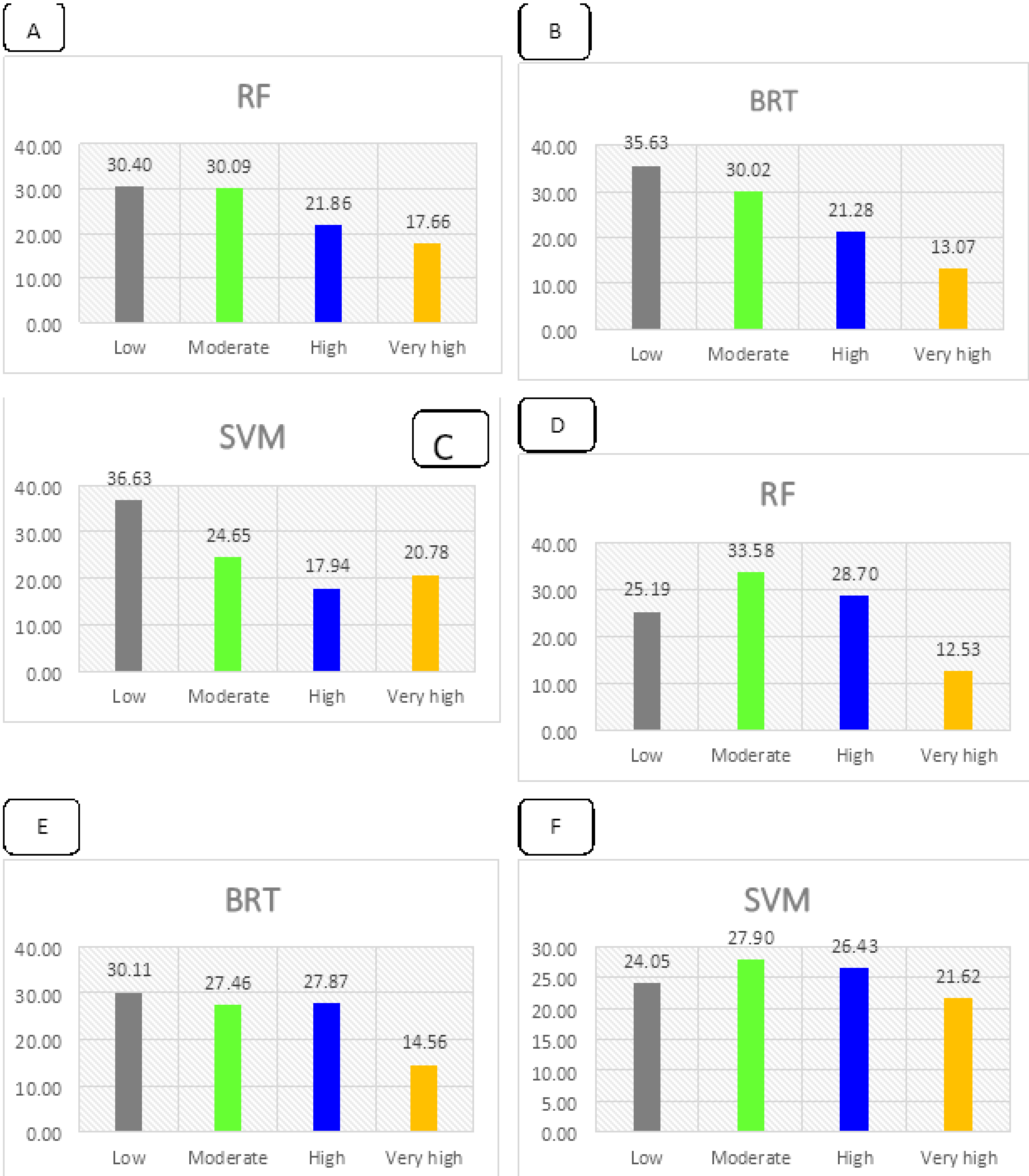

Figure 6

Evaluation of the accuracy of the three algorithms based on the percentage value of each class: *Aceria alhagi* (A): RF model; B: BRT model; C: SVM model); and *Alhagi maurorum* (D): RF model; E: BRT model; F: SVM model)

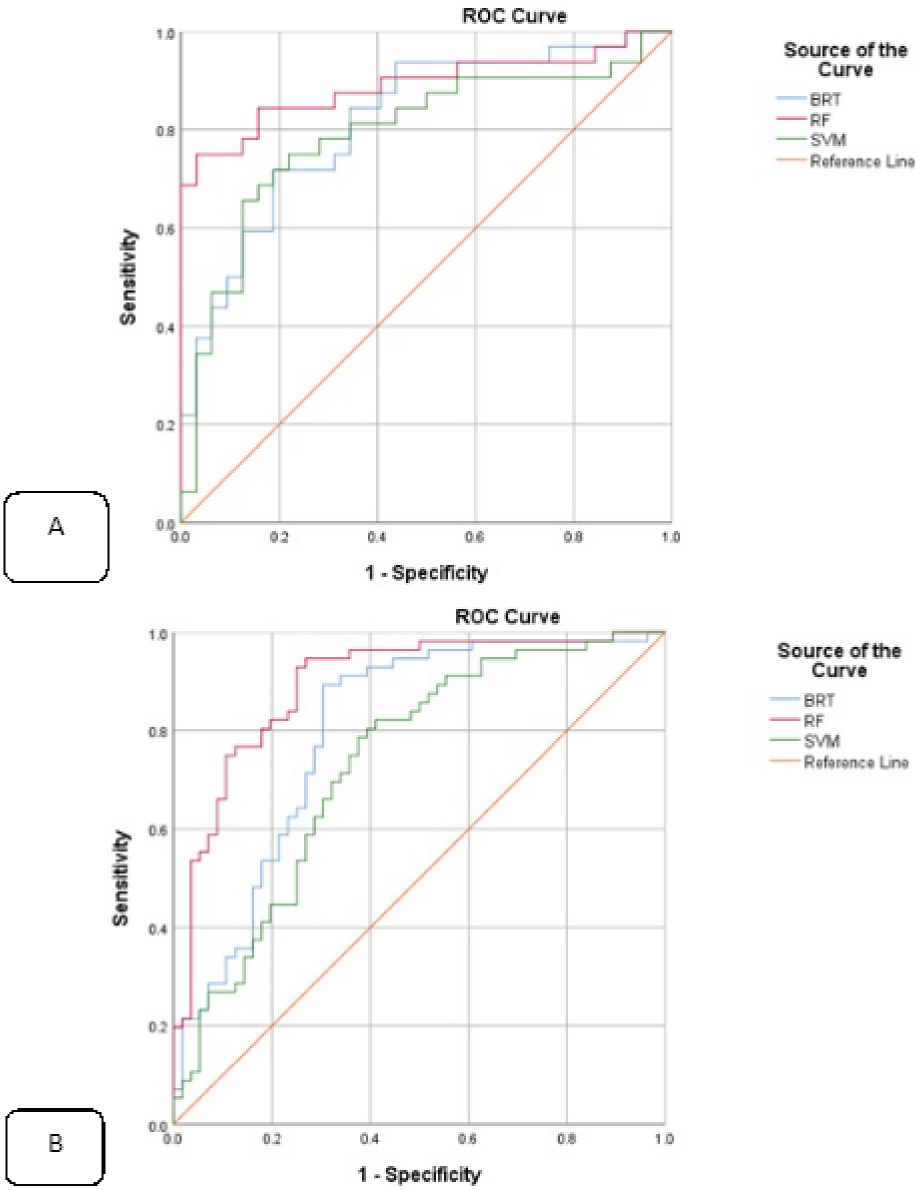

Figure 7

Evaluating algorithms and selecting the best algorithm based on the ROC curve: (A) *Aceria alhagi*; (B) *Alhagi maurorum*

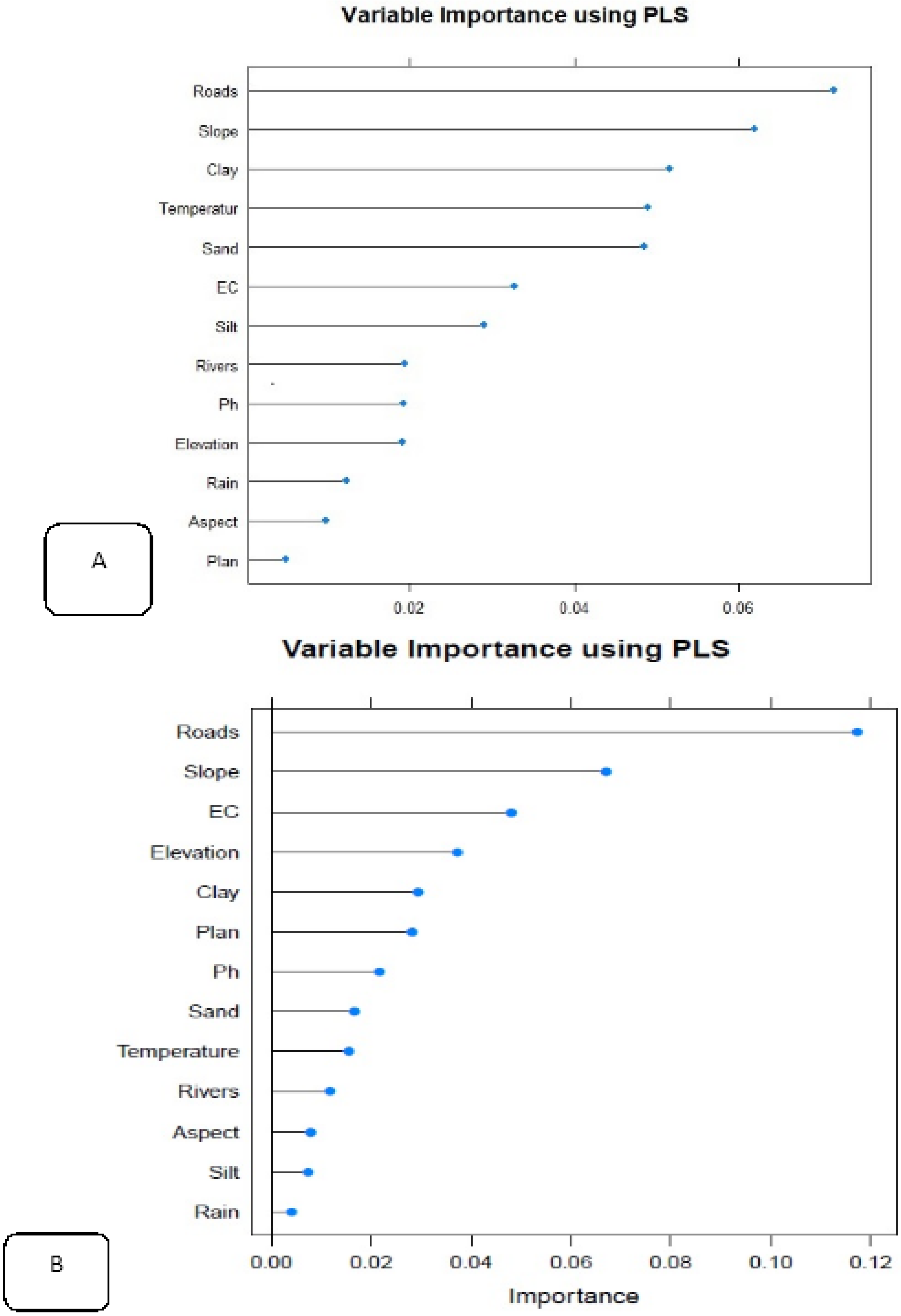

Figure 8

Determining the most important factor based on the PLS algorithm: (A) *Aceria alhagi*; (B) *Alhagi maurorum*

Supplementary Files

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

- Conflictofinterest.doc
- GraphicalAbstract.docx
- Highlights.docx