Predicting ecological regulators of thymol and carvacrol biosynthesis in *Oliveria decumbens* Vent. using a hybrid ensemble machine learning model (RF+SVR-RBF) in arid regions of Iran

Emran DASTRES, Hassan ESMAEILI

2026-08-12

ChinaRxiv

Abstract

Dryland ecosystems, encompassing arid to semi-arid regions, impose strong climatic and edaphic constraints that profoundly shape plant functional traits and secondary metabolism. Understanding how environmental factors regulate phytochemical biosynthesis is essential for biodiversity conservation and sustainable resource management under increasing aridity. Oliveria decumbens Vent., an endemic medicinal species of the drylands of Fars Province, Iran, provides an excellent model for exploring the ecological determinants of metabolite variability in water-limited habitats. We integrated ecological predictors with machine learning to model the spatial variation of thymol and carvacrol concentrations across 59 georeferenced populations of O. decumbens. Three predictive models —Random Forest (RF), Support Vector Regression (SVR) with a Radial Basis Function (RBF) kernel (SVR-RBF), and a hybrid ensemble (RF+SVR-RBF)— were developed and evaluated. Model performance was quantified using root mean squared error (RMSE), mean absolute error (MAE), coefficient of determination (R2), and the concordance correlation coefficient (CCC). Generalized Linear Model (GLM) was applied to identify key environmental variables regulating metabolite biosynthesis. The hybrid ensemble consistently outperformed individual model, achieving the highest predictive accuracy (R2=0.82 for thymol and R2=0.80 for carvacrol). Spatial mapping revealed pronounced heterogeneity in metabolite distribution, with distinct functional hotspots in the northern and western semi-arid regions of Fars Province. GLM analysis indicated that mean annual temperature, slope aspect, slope degree, and sand content were strong positive predictors of thymol and carvacrol accumulation, whereas high soil potassium, clay percentage, and alkaline pH constrained metabolite production. This study shows that hybrid ensemble modeling effectively captures how environmental gradients regulate secondary metabolism in dryland plants. The proposed framework, combining RF with SVR-RBF, is transferable across arid environments. These findings support trait-based ecological predictions and offer practical insights for conservation and sustainable cultivation of high-value medicinal plants in water-limited ecosystems.

Full Text

J Arid Land (2026) 18(8): 1425–1445
doi: 10.1016/j.jaridl.2026.08.007; CSTR: 32276.14.JAL.20250553

Predicting ecological regulators of thymol and carvacrol biosynthesis in *Oliveria decumbens* Vent. using a hybrid ensemble machine learning model (RF+SVR-RBF) in arid regions of Iran

Emran DASTRES[1,2], Hassan ESMAEILI[1*]

[1] Department Agriculture, Medicinal Plants and Drugs Research Institute, Shahid Beheshti University, Tehran 1983969411, Iran;
[2] Department of Plant Production and Genetics, School of Agriculture, Shiraz University, Shiraz 7188637911, Iran

Abstract: Dryland ecosystems, encompassing arid to semi-arid regions, impose strong climatic and edaphic constraints that profoundly shape plant functional traits and secondary metabolism. Understanding how environmental factors regulate phytochemical biosynthesis is essential for biodiversity conservation and sustainable resource management under increasing aridity. *Oliveria decumbens* Vent., an endemic medicinal species of the drylands of Fars Province, Iran, provides an excellent model for exploring the ecological determinants of metabolite variability in water-limited habitats. We integrated ecological predictors with machine learning to model the spatial variation of thymol and carvacrol concentrations across 59 georeferenced populations of *O. decumbens*. Three predictive models—Random Forest (RF), Support Vector Regression (SVR) with a Radial Basis Function (RBF) kernel (SVR-RBF), and a hybrid ensemble (RF+SVR-RBF)—were developed and evaluated. Model performance was quantified using root mean squared error (RMSE), mean absolute error (MAE), coefficient of determination (R^2), and the concordance correlation coefficient (CCC). Generalized Linear Model (GLM) was applied to identify key environmental variables regulating metabolite biosynthesis. The hybrid ensemble consistently outperformed individual model, achieving the highest predictive accuracy (R^2=0.82 for thymol and R^2=0.80 for carvacrol). Spatial mapping revealed pronounced heterogeneity in metabolite distribution, with distinct functional hotspots in the northern and western semi-arid regions of Fars Province. GLM analysis indicated that mean annual temperature, slope aspect, slope degree, and sand content were strong positive predictors of thymol and carvacrol accumulation, whereas high soil potassium, clay percentage, and alkaline pH constrained metabolite production. This study shows that hybrid ensemble modeling effectively captures how environmental gradients regulate secondary metabolism in dryland plants. The proposed framework, combining RF with SVR-RBF, is transferable across arid environments. These findings support trait-based ecological predictions and offer practical insights for conservation and sustainable cultivation of high-value medicinal plants in water-limited ecosystems.

Keywords: dryland ecology; hybrid ensemble model; secondary metabolism; machine learning; conservation biology

Citation: Emran DASTRES, Hassan ESMAEILI. 2026. Predicting ecological regulators of thymol and carvacrol biosynthesis in *Oliveria decumbens* Vent. using a hybrid ensemble machine learning model (RF+SVR-RBF) in arid regions of Iran. Journal of Arid Land, 18(8): 1425–1445. https://doi.org/10.1016/j.jaridl.2026.08.007; https://cstr.cn/32276.14.JAL.20250553

---

*Corresponding author: Hassan ESMAEILI (E-mail: h_esmaeili@sbu.ac.ir)

Received 2025-11-07; revised 2026-05-08; accepted 2026-05-23

http://jal.xjegi.com; https://www.keaipublishing.com/en/journals/journal-of-arid-land/

1 Introduction

Medicinal and aromatic plants are integral components of terrestrial ecosystems, contributing to both ecological functioning and human well-being (Srivastava, 2018; Balkrishna et al., 2024). Their bioactive compounds mediate key ecological interactions, including defense against herbivores, attraction of pollinators, and adaptation to abiotic stress, while simultaneously providing therapeutic and economic value through their pharmacological properties (Park et al., 2025). Approximately 80.00% of the global population still depends on plant-based remedies for primary healthcare, highlighting the dual ecological and socio-economic significance of medicinal and aromatic plants (Majeed, 2017). However, increasing anthropogenic pressures, including overharvesting and habitat degradation, threaten their natural populations. This fact underscores the need for integrative approaches that link ecological understanding with conservation and sustainable cultivation strategies.

Dryland ecosystems, encompassing arid, semi-arid, dry, and sub-humid regions, represent critical reservoirs of medicinal and aromatic plants diversity (Rout et al., 2023). These landscapes are characterized by limited and unpredictable precipitation, intense evapotranspiration, and edaphic heterogeneity, all of which shape the adaptive and metabolic responses of plants (Forti et al., 2006). Secondary metabolites, particularly phenolic monoterpenes, play pivotal roles in mediating tolerance to these extreme conditions by protecting plants against oxidative stress, desiccation, and temperature fluctuations (Vilela et al., 2011). Understanding how environmental gradients regulate the biosynthesis and spatial variability of these compounds is essential for elucidating plant adaptation mechanisms and for developing sustainable resource management in water-limited ecosystems.

Oliveria decumbens Vent., an endemic annual herb of the Apiaceae family, provides an ideal model for such investigations. Native to the mountainous and semi-arid regions of Iran, this species is highly valued for its essential oils rich in thymol and carvacrol, compounds known for their antimicrobial, antioxidant, and anti-inflammatory properties (Boveiri Dehsheikh et al., 2024). Variability in the concentration of these metabolites is thought to reflect adaptive responses to climatic and edaphic gradients, yet the ecological mechanisms underlying this variation remain poorly understood (Esmaeili et al., 2018; Karami et al., 2020).

Linking metabolite production to environmental drivers offers an opportunity to clarify how semi-arid ecosystems shape plant chemical diversity and functional traits.

Recent advances in machine learning have opened new avenues for modeling complex ecological systems and trait–environment relationships (Hesami and Jones, 2020; Rather et al., 2020). Algorithms such as Random Forest (RF), Support Vector Regression (SVR), and Gradient Boosting Machines (GBM) have been widely applied to predict species distributions, habitat suitability, and trait variability (Catucci and Scardi, 2020; Dastres et al., 2023, 2025a). Hybrid ensemble frameworks that integrate the predictive strengths of multiple algorithms have demonstrated superior accuracy and stability, particularly in heterogeneous environmental settings (Dastres et al., 2025c).

The application of RF and SVR models, individually or in combination, has been increasingly documented across diverse plant species and ecological contexts. For instance, RF and SVR with Radial Basis Function (RBF) kernel (SVR-RBF) have been successfully employed to estimate chlorophyll content in winter wheat, with the RF+SVR hybrid demonstrating superior predictive performance for variety screening under late-sowing conditions (Wang et al., 2021). In medicinal plant research, machine learning approaches including RF and SVR have been integrated with species distribution models to evaluate habitat suitability and predict bioactive component accumulation in *Cynomorium songaricum* Rupr., identifying key environmental regulators such as precipitation, temperature, and soil properties (Ji et al., 2025). Similarly, recent work on *Bletilla striata* (Thunb. ex A. Murray) Rchb. f. has combined ensemble learning frameworks with phytochemical profiling to delineate high-quality cultivation zones, demonstrating the utility of such approaches for conservation and sustainable utilization of medicinal species (Mu et al.,

2025). In agronomic systems, stacking ensemble learning incorporating RF and SVR has been successfully applied to predict rapeseed flowering periods, achieving substantial improvements in accuracy over single model through the integration of meteorological variables and virtual sample generation techniques (Xie et al., 2024). Furthermore, ensemble modeling frameworks have demonstrated robust performance in species distribution modeling for invasive species management, with RF and SVM consistently outperforming traditional regression approaches in predicting habitat suitability across arid and semi-arid ecosystems (Ahmed et al., 2021).

These methods provide a robust computational basis for translating environmental data into biological insights, thereby enhancing the predictive understanding of plant responses to environmental stress (Singhal et al., 2025; Spake et al., 2025). The growing evidence across medicinal plants, crops, and native species underscores the versatility and transferability of RF, SVR-RBF, and hybrid ensemble approaches—precisely the methodological framework adopted in the present study on *O. decumbens*.

In this study, we combined spatially explicit phytochemical profiling with hybrid machine learning modeling to investigate the ecological regulation of thymol and carvacrol biosynthesis in *O. decumbens* across the semi-arid regions of Fars Province, Iran. A total of 18 environmental variables encompassing climatic, topographic, and edaphic factors were analyzed. We implemented RF, SVR-RBF, and a hybrid RF+SVR-RBF model to predict the spatial distribution of metabolite concentrations, while Generalized Linear Model (GLM) was employed to identify the most influential environmental determinants. The specific aims of this research were to: (1) quantify the relative contributions of climatic, edaphic, and topographic factors to thymol and carvacrol concentrations in *O. decumbens* and identify the most influential predictors using GLM analysis; (2) develop and compare predictive models (RF, SVR-RBF, and RF+SVR-RBF) for metabolite distribution under semi-arid conditions; (3) evaluate model performance across multiple statistical criteria to ensure ecological reliability; and (4) predict spatial distribution to guide conservation and sustainable cultivation of *O. decumbens* and related medicinal species in dryland ecosystems.

2 Materials and methods

2.1 Study area

The study was conducted in Fars Province, southwestern Iran, located between 27°–32°N and 50°–56°E (Fig. 1). The region lies at the intersection of the Zagros mountainous ranges and extensive plains, forming a heterogeneous landscape, where semi-arid to dry sub-humid climates predominate. The mean annual precipitation ranges from 150 to 350 mm and is highly variable, while the mean annual temperature varies between 18°C and 24°C, with high summer temperatures and elevated evapotranspiration rates imposing strong water limitation characteristic of dryland ecosystems (Hassanpouraghdam et al., 2022). These conditions, together with diverse topographic gradients, create strong ecological heterogeneity, ranging from relatively mesic mountain zones to expansive semi-arid plains and arid basins (Ibrahimi et al., 2024).

Fars Province supports a rich diversity of medicinal and aromatic plants that are adapted to water-limited conditions. Among them, *O. decumbens* is an endemic annual herb of the Apiaceae family that thrives across semi-arid habitats of the province. Its restricted distribution and strong dependence on climatic and edaphic gradients make it an ideal model for investigating how environmental drivers of drylands regulate secondary metabolite production. The combination of pronounced aridity, topographic variability, and soil heterogeneity renders Fars Province an ecologically significant region for exploring trait–environment interactions in dryland ecosystems (Azizi et al., 2022).

2.2 Modeling methodology

A comprehensive dataset of environmental variables and measured concentrations of thymol and carvacrol in *O. decumbens* was collected across different

Fig. 1. *Oliveria decumbens* Vent. distribution in Fars Province, Iran

Figure 1: Fig. 1. *Oliveria decumbens* Vent. distribution in Fars Province, Iran

Fig. 2 Flowchart illustrating the modeling process

Figure 2: Fig. 2 Flowchart illustrating the modeling process

regions of Fars Province. The

Fig. 1 *Oliveria decumbens* Vent. distribution in Fars Province, Iran

environmental predictors included topographic factors (elevation, slope aspect, slope degree, plan curvature, and profile curvature), hydrological variables (distance from rivers and distance from roads), climatic parameters (mean annual precipitation and mean annual temperature), and soil characteristics (clay, sand, silt, nitrogen, organic matter, potassium, phosphorus, pH, and electrical conductivity). To avoid redundancy and multi-collinearity, we preprocessed predictors using the variance inflation factor (VIF), ensuring suitability for regression-based machine learning models.

Three modeling approaches including RF, SVR-RBF, and hybrid RF+SVR-RBF model, were implemented to predict the continuous concentrations of thymol and carvacrol. In addition, a GLM with a Gaussian distribution and an identity link function was applied to assess the relative importance of environmental factors influencing metabolite concentrations.

Model training and validation followed a repeated nested 5-fold cross-validation framework to ensure robust generalization. Predictive performance was evaluated using root mean squared error (RMSE), mean absolute error (MAE), coefficient of determination (R^2), and the concordance correlation coefficient (CCC), offering a comprehensive assessment of both accuracy and agreement between observed and predicted concentrations (Fig. 2).

2.3 Essential oil extraction and analysis

In this study, *O. decumbens* samples were collected at the full flowering phenological stage during June 2024 from 59 georeferenced locations across Fars Province, Iran. The plant samples were botanically identified by the expert and a voucher specimen has been deposited at the Medicinal Plants and Drugs Research Institute Herbarium of Shahid Beheshti University under the accession number of MPH-3079. We performed essential oil extraction according to the British Pharmacopoeia (BP) standard extraction method using hydrodistillation with a Clevenger-type apparatus (hydrodistillation system according to BP method). For this purpose, the hydrodistillation was applied for essential oil extraction from 20 g of dried and chopped flowers, using Clevenger-type device for 3 h. The obtained essential oils were injected into gas chromatography for analysis.

Fig. 2 Flowchart illustrating the modeling process. GC, gas chromatography; GC-MS, gas chromatography-mass spectrometry; VIF, variance inflation factor; GLM, Generalized Linear Model; RMSE, root mean square error; MAE, mean absolute error; CCC, concordance correlation coefficient; RF, Random Forest; SVR-RBF, Support Vector Regression with Radial Basis Function kernel; EC, electrical conductivity; K, potassium; N, nitrogen; P, phosphorus; OM, organic matter.

The gas chromatography-mass spectrometry (GC-MS) analysis was conducted utilizing a Thermo Finnigan gas chromatograph equipped with a flame ionization detector (FID) featuring a fused silica capillary (5.00%-phenyl)-methylpolysiloxane (HP-5MS) column with dimensions of 60 m in length, 0.25 mm in inner diameter, and a film thickness of 0.25 μm. The instrument was coupled with a trace and quadrupole mass spectrometer operating in electron ionization mode at 70 electron volts. Helium was the carrier gas, and the ionization voltage was set at 70 electron volts. The ion source and interface temperatures were maintained at 200°C and 250°C, respectively. The mass spectrometer operated within a mass range spanning from 40 to 460 atomic mass units. The identification of essential oil constituents was conducted by determining their retention indices through temperature-programmed analysis with normal alkanes (C6–C24) and comparing them with the oil on a DB-5 column using identical chromatographic conditions. The compounds were identified by comparing their mass spectra with those in an internal reference mass spectra library by comparison with the essential oil mass database (https://essentialoilexperts.com/essential-oil-database/) and Wiley Science Solution (https://sciencesolutions.wiley.com/mass-spectral-databases/) libraries or with verified standards. Confirmation of identification was achieved by comparing retention indices with those of authenticated compounds or references in the literature. Quantification of the compounds was performed by determining relative area percentages from flame ionization detector signals, without the application of correction factors (Adams, 2017).

2.4 Environmental factors

A 30-m spatial-resolution digital elevation model (DEM), derived from Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) onboard National Aeronautics and Space Administration' s (NASA' s) Terra satellite imagery, served as the foundational dataset for topographic analysis (Saha and Agrawal, 2020). This DEM was meticulously processed to extract critical terrain attributes, including elevation, slope degree, slope aspect, plan curvature, and profile curvature (Fig. 3a–e). The high spatial resolution of the DEM ensured the accuracy and detail of these derived parameters, providing a robust basis for subsequent geospatial analyses and modeling.

endeavors. The DEM was geo-referenced to match the coordinate system used in the study area, and any anomalies or errors were corrected using standard preprocessing techniques.

The slope degree was calculated by applying a slope algorithm to each pixel in the DEM, utilizing the first derivative of elevation data. Plan curvature was derived from the second derivative of elevation data, indicating the curvature of the terrain surface in the horizontal plane (Amatulli et al., 2018). Profile curvature was extracted by assessing the second derivative of elevation data in the direction of the maximum slope (Amatulli et al., 2018). The slope aspect was determined by computing the compass direction of the hill for each pixel, aiding in understanding the orientation of the terrain (Capitani et al., 2013). Elevation values were directly obtained from the DEM (Li, 2018). The extracted parameters were validated against known control points and other high-resolution datasets, such as Global Positioning System (GPS)/leveling ground control points, to ensure their accuracy (Das et al., 2016).

We collected soil samples from 59 locations across Fars Province based on the presence of *O. decumbens*. Soil samples were collected from the 0–30 cm depth interval. This depth was selected because *O. decumbens* is an annual herb with a shallow fibrous root system concentrated in the upper soil horizon, as documented in the botanical description of the species (Mousavi et al., 2023). The 0–30 cm depth represents the primary zone of nutrient and water uptake for this plant and is the standard sampling depth for assessing edaphic factors influencing secondary metabolite accumulation in herbaceous medicinal species. The physical properties of the soil, specifically the proportions of sand, silt, and clay, were analyzed using the hydrometer method (Mwendwa, 2022). Soil pH was determined by mixing samples with distilled water in a 1:1 ratio and measuring with a pH meter (Regasa et al., 2024). Electrical conductivity (EC) was assessed by mixing soil with distilled water in a 1:2 ratio and measuring the conductivity with an EC meter (Khan et al., 2023). Organic matter (OM) content was quantified via the Walkley-Black Wet Digestion Method using spectrophotometer, involving the oxidation of soil organic carbon with potassium dichromate and sulfuric acid (de Santana et al., 2019). Nitrogen (N) content was determined using the Kjeldahl method, where soil samples are digested with sulfuric acid to convert organic N to ammonium sulfate (Wang and Cai, 2024). Phosphorus (P) was measured using the Olsen method, extracting it with a sodium bicarbonate solution and measuring it colorimetrically (Li et al., 2025). Potassium (K) was extracted using ammonium acetate and measured with a flame photometer (Selassie et al., 2020). Following analysis at the Soil Science Laboratory of Shiraz University, data on clay, silt, sand, K, P, OM, N, EC, and pH were used to create raster maps with a 30-m resolution through the inverse distance weighting algorithm using Geographic Information System (GIS) software (ArcMap) (Fig. 3f–n) (Dastres et al., 2022).

Mean annual temperature and mean annual precipitation data were obtained from 24 synoptic weather stations distributed across Fars Province. Data spanning the most recent 5-a period (2020–2024) were utilized to calculate mean values. The inverse distance weighting algorithm was applied to convert the mean annual temperature and mean annual precipitation data into 30-m-resolution raster maps (Fig. 3o and p) (Dastres et al., 2025d), effectively creating continuous raster surfaces from discrete point data.

A 1:25,000 scale topographic map of Fars Province was used to extract information on the terrain and hydrography of the region (Chu et al., 2019). Using GIS software, the Euclidean distance algorithm was employed to generate raster maps that indicate the distances from rivers and roads (Fig. 3q and r). The Euclidean distance algorithm calculates the straight-line distance from each cell in the raster to the nearest river or road, thus creating a continuous surface that represents proximity to river or road networks throughout Fars Province (AlHamaydeh et al., 2021).

2.5 Evaluation of factors with VIF collinearity test

To assess multi-collinearity among independent variables, we employed the tolerance (TOL) and VIF collinearity tests. These measures are essential in regression analysis for detecting multi-collinearity. The TOL index, defined as $1-R^2$, measures the variance explained by other

predictors, with values below 0.1 indicating significant multi-collinearity (Cheng et al., 2022). VIF, calculated as 1/TOL, indicates how much a regression coefficient' s variance is inflated due to multi-collinearity, with values above 5 suggesting high multi-collinearity. SPSS v.25.0 software was used to compute the TOL and VIF values.

2.6 Machine learning algorithms

To predict continuous concentrations of secondary metabolites (thymol and carvacrol), we employed two base learners, i.e., RF regression and SVR-RBF kernel, and a hybrid ensemble combining their predictions.

2.6.1 RF

RF regression is an ensemble method that fits multiple decision trees on bootstrap samples of the data and averages their predictions to generate continuous outputs (Liu et al., 2024). At each split, a random subset of predictor variables is considered, which reduces correlation among trees and enhances generalization. RF is robust against multi-collinearity, noise, and outliers, common in metabolomics datasets (Shahjaman et al., 2021). In this study, "ntree" value in RF regression was set between 500 and 1000, and "mtry" value was set to the square root of the total number of features.

2.6.2 SVR-RBF kernel

SVR identifies a function approximating the relationship between input variables and continuous responses within a margin of error (Keshtegar and Yaseen, 2022). The RBF kernel allows the model to capture complex nonlinear patterns between metabolite concentrations and predictor features. The penalty parameter and kernel coefficient were optimized via nested cross-validation to prevent overfitting and ensure robust generalization (Chen et al., 2022).

2.6.3 Hybrid model

The hybrid ensemble leverages complementary strengths of the two base learners: RF reduces prediction variance via tree aggregation, while SVR captures non-linear relationships. Predictions from both models were combined using weighted soft averaging, with weights determined according to cross-validated performance metrics (He et al., 2023). This approach enhances overall prediction accuracy and stability, especially for metabolites exhibiting complex concentration patterns.

2.7 Modeling framework

Because the target variables (thymol and carvacrol) were measured as continuous percentages, prediction was treated as a regression problem rather than classification. Two base learners, including RF and SVR-RBF kernel, were implemented and integrated through a weighted soft-averaging ensemble to leverage complementary bias–variance characteristics (Yoon and Kang, 2023). Sensitivity analyses were additionally conducted using a stacking meta-learner to evaluate model robustness.

Given that the response variables were bounded between 0.00% and 100.00%, all concentrations were first rescaled to the [0, 1] interval. To handle boundary values (concentrations close to 0.00% or 100.00%), we applied a small offset prior to transformation following the empirical logit approach recommended for proportional data with extreme values (Warton and Hui, 2011). Specifically, for observed proportion (p), adjusted value p' was calculated as follows:

$$p' = (p \times (n-1) + 0.5)/n, \qquad (1)$$

where n is the sample size ($n = 59$). This adjustment ensures that values of 0 and 1 are mapped to non-extreme logit values while preserving the original data distribution. The p' was then transformed to the logit scale:

$$\text{logit}(p) = \ln(p/(1-p)). \qquad (2)$$

This transformation linearizes the bounded response and stabilizes variance, making the data suitable for regression modeling. After model prediction on the logit scale, values were back-transformed to the original percentage scale for reporting using the inverse logit transformation:

Fig. 3 Classified raster map depicting 18 environmental influencing factors. (a), elevation; (b), slope degree; (c), slope aspect; (d), plan curvature; (e), profile

Fig. 3. Classified raster map depicting 18 environmental influencing factors.

Figure 3: Fig. 3. Classified raster map depicting 18 environmental influencing factors.

curvature; (f), EC (electrical conductivity); (g), pH; (h), K (potassium); (i), N (nitrogen); (j), OM (organic matter); (k), P (phosphorus); (l), clay; (m), sand; (n), silt; (o), mean annual precipitation; (p), mean annual temperature; (q), distance from rivers; (r), distance from roads.

$$\text{P\_pred} = \exp(\text{logit\_pred})/(1 + \exp(\text{logit\_pred})), \qquad (3)$$

where P_pred is the predicted probability; and logit_pred is the linear logit output. P_pred was derived from the logit_pred using the standard logistic transformation (sigmoid function), ensuring that the predicted values range between 0 and 1.

All analyses were conducted in R v.4.3.2 software (R Core Team, 2023) using the CRAN packages for RF regression (https://cran.r-project.org/package=randomForest), e1071 for SVR (https://cran.r-project.org/package=e1071), and caretEnsemble for model ensembling (https://cran.r-project.org/package=caretEnsemble).

Given the limited sample size (n=59), we applied nested repeated k-fold cross-validation for both hyperparameter tuning and model performance estimation. Each iteration used 10 folds with 10 repetitions to ensure that every observation contributed to both model training and out-of-fold validation, reducing optimism and variance in predictive estimates (Coley et al., 2023).

2.8 Model evaluation, uncertainty, and statistical comparison

Model performance was evaluated using four complementary metrics: RMSE, MAE, R^2, and CCC. These metrics jointly capture the magnitude, directionality, and overall agreement between predicted and observed metabolite concentrations (Karim et al., 2024). The CCC was specifically included as an evaluation metric because it assesses both precision and accuracy simultaneously. Unlike R^2, which only measures the proportion of variance explained by the model and is sensitive to the range of data, CCC evaluates the degree to which pairs of predicted and observed values fall on the 45° line through the origin (Elshaboury and Elmousalami, 2025).

To quantify uncertainty, we summarized each performance metric using bias-corrected 95.00% confidence intervals obtained by bootstrapping over cross-validation folds (Pintar et al., 2023). This resampling approach provides a robust estimate of precision and accounts for variability arising from both sampling and model stochasticity.

For the weighted soft-averaging ensemble, we determined optimal weights based on cross-validation performance using a grid search optimization approach. Specifically, weights of RF (w_RF) and SVR-RBF kernel (w_SVR) were assigned to the predictions of these two models, respectively, satisfying the constraint w_RF+w_SVR=1. The weight combination that minimized the RMSE across cross-validation folds was selected as optimal. A systematic search was performed over the interval [0.00, 1.00] with increments of 0.05, evaluating ensemble performance for each weight combination. The final ensemble prediction (Pre_ensemble) for each observation was calculated as follows:

$$\text{Pre}_{\text{ensemble}} = \text{w\_RF} \times \text{Pre\_RF} + \text{w\_SVR} \times \text{Pre\_SVR}. \quad (4)$$

where Pre_RF and Pre_SVR are the predicted values of FR and SVR, respectively. This approach ensures that the model with superior predictive accuracy receives proportionally higher weight in the final ensemble, while maintaining the complementary bias-variance characteristics of both algorithms. The optimal weights were determined separately for thymol and carvacrol predictions, as model performance varied slightly between the two metabolites.

To assess whether observed differences among models were statistically significant, we performed a non-parametric Friedman test on the average ranks of RMSE and MAE across repeated folds, followed by a Nemenyi post-hoc (Critical Difference; CD) test for pairwise comparisons. These tests are appropriate for comparing multiple algorithms over repeated resampling procedures without assuming normality or homoscedasticity.

Model ranks were computed per iteration, with lower ranks corresponding to better performance. When the Friedman test detected a significant overall difference (α=0.05), the Nemenyi test was used to identify which pairs of models differed significantly. The results were visualized using CD diagrams, providing an intuitive representation of statistical separations in model accuracy.

All statistical analyses were performed in R v.4.3.2 software using the CRAN packages PMCMRplus (https://cran.r-project.org/package=PMCMRplus) and scmamp (https://cran.r-project.org/package=scmamp). The complete workflow in R v.4.3.2 software, including scripts for cross-

validation and significance testing, is publicly available at the website of https://github.com/dastres-lab/hybrid-metabolite-modeling.

2.9 GLM for factor importance

To assess the contribution of predictor variables to the concentrations of thymol and carvacrol in *O. decumbens*, we employed GLM. GLM extends traditional linear regression by allowing flexible assumptions about the response variable,

Fig. 4. Medicinal plant *O. decumbens* (a) and chemical structure of its main compounds, including thymol and carvacrol (b)

Figure 4: Fig. 4. Medicinal plant *O. decumbens* (a) and chemical structure of its main compounds, including thymol and carvacrol (b)

accommodating potential deviations from constant variance or normality in metabolite data (McCoy et al., 2023).

Given the continuous nature of the target variables, we applied a Gaussian distribution with an identity link function. This setup is mathematically equivalent to standard linear regression but provides a formal GLM framework for statistical inference. The model was implemented in R v.4.3.2 software using the glm package (https://cran.r-project.org/web/packages/glm.predict/index.html).

Maximum likelihood estimation was used to estimate model parameters. The significance of each predictor was assessed to identify which factor most strongly influence thymol and carvacrol concentrations. Variables with statistically significant effects were interpreted as the main drivers of metabolite variation in *O. decumbens*. This GLM-based analysis served as an independent evaluation of variable importance, complementing the machine learning models (RF, SVR, and hybrid of RF and SVR) and providing a conventional statistical perspective on the relationship between predictors and metabolite levels.

3 Results

3.1 Essential oil compositions

The essential oil analysis led to the identification of 23 chemical components. Among these, thymol and carvacrol were identified as the principal bioactive compounds. The concentration of thymol ranged from 20.10% to 38.80%, while carvacrol varied from 18.70% to 51.90% across the samples. These variations reflect the environmental heterogeneity and adaptive responses of *O. decumbens* populations in different ecological zones. The chemical structures of thymol and carvacrol are illustrated in Figure 4.

Fig. 4 Medicinal plant *O. decumbens* (a) and chemical structure of its main compounds, including thymol and carvacrol (b)

3.2 Assessment of environmental factors

The results of the multi-collinearity analysis are summarized in Table 1. None of the 18 environmental variables exceeded commonly accepted thresholds for the VIF or TOL values, indicating the absence of problematic multi-collinearity. This step was conducted to ensure that the predictive modeling process would not be adversely affected by redundant or highly correlated input variables.

Although RF is generally robust to multi-collinearity, removing highly collinear variables can improve the interpretability of variable importance rankings.

Table 1 Collinearity test of independent variables

Variable	TOL	VIF	Variable	TOL	VIF
Slope aspect	0.67	1.49	EC	0.59	1.78
Slope degree	0.76	1.12	N	0.33	3.74
Elevation	0.56	2.02	K	0.48	3.01
Profile curvature	0.97	1.02	P	0.21	4.78
Plan curvature	0.89	1.09	OM	0.26	3.92
Silt	0.28	4.05	Mean annual temperature	0.29	4.12
Sand	0.37	3.56	Mean annual precipitation	0.32	4.01
Clay	0.45	3.09	Distance from roads	0.43	3.19
pH	0.29	4.01	Distance from rivers	0.52	2.87

Note: EC, electrical conductivity; N, nitrogen; P, phosphorus; K, potassium; OM, organic matter; TOL, tolerance; VIF, variance inflation factor.

3.3 Spatial patterns in habitat suitability

To evaluate habitat suitability for *O. decumbens* with respect to thymol and carvacrol biosynthesis, we applied three machine learning approaches: RF, SVR-RBF, and the hybrid RF+SVR-RBF. The continuous outputs of these models were reclassified into four suitability categories—low, moderate, high, and very high—using the Jenks Natural Breaks optimization method. This approach minimizes within-class variance and maximizes between-class heterogeneity, thus providing an objective and ecologically meaningful framework for visualizing suitability patterns. Classification thresholds were independently determined for each model output to ensure consistency in interpretation while retaining the distinct spatial characteristics inherent to different algorithms.

At the landscape scale, the highest suitability zones for thymol and carvacrol production were predominantly concentrated in the northern, northwestern, and —though to a lesser extent—the western and southern parts of the study area. In contrast, the central and eastern parts of Fars Province were largely unsuitable, reflecting suboptimal environmental conditions for secondary metabolite accumulation (Fig. 5a–f).

The performance of the three algorithms revealed distinct spatial patterns in predicting habitat suitability for thymol and carvacrol. The hybrid RF+SVR-RBF model consistently provided the most balanced and spatially explicit predictions. For thymol, this model classified 22.24% of the study area (8947.78 km^2) as very high suitability and 31.19% (12,551.09 km^2) as low suitability, with moderate and high classes accounting for 29.96% (12,055.18 km^2) and 16.62% (6687.24 km^2), respectively (Fig. 6a; Table 2).

Fig. 5. Spatial prediction of thymol (a–c) and carvacrol (d–f) contents in *O. decumbens* using RF, SVR-RBF, and hybrid RF+SVR-RBF models based on current environmental conditions

Figure 5: Fig. 5. Spatial prediction of thymol (a–c) and carvacrol (d–f) contents in *O. decumbens* using RF, SVR-RBF, and hybrid RF+SVR-RBF models based on current environmental conditions

The RF model produced a more evenly distributed classification, with 24.03% (9668.88 km^2) in the very high and 29.42% (11,837.64 km^2) in the low category (Fig. 6a; Table 2). In contrast, SVR-RBF displayed a tendency to overestimate unsuitable or marginal zones, assigning more than half of the study area (52.19%; 21,001.44 km^2) to the low class, while reducing the high and very high categories to 10.66% (4288.34 km^2) and 20.69% (8327.37 km^2), respectively (Fig. 6a; Table 2).

For carvacrol, the hybrid model again demonstrated superior spatial differentiation, allocating 20.00% of the study area (8047.96 km^2) to very high suitability and 15.23% (6128.43 km^2) to high suitability. Low and moderate classes comprised 38.36% (15,438.29 km^2) and 26.41% (10,626.60 km^2), respectively (Fig. 6b; Table 2). The RF model predicted a relatively even spread across classes, with 20.70% (8330.73 km^2) of the area in the very high and 30.40% (12,232.28 km^2) in the low suitability category (Fig. 6b; Table 2). By contrast, SVR-RBF was strongly biased toward the low class, predicting 52.63% (21,179.25 km^2) of the area as low suitability, while the very high suitability class was limited to 14.37% (5782.08 km^2) (Fig. 6b; Table 2).

Fig. 5 Spatial prediction of thymol (a–c) and carvacrol (d–f) contents in *O. decumbens* using RF, SVR-RBF, and hybrid RF+SVR-RBF models based on current environmental conditions

3.4 Predictive accuracy of individual and hybrid models

The predictive performance of RF, SVR-RBF, and their hybrid RF+SVR-RBF model in predicting thymol and carvacrol concentrations is summarized in Table 3. The hybrid model consistently outperformed both individual model across all evaluation metrics, including RMSE, MAE, R^2, and CCC. For thymol, the hybrid RF+SVR-RBF model achieved the lowest prediction errors (RMSE=0.015; MAE=0.012) and the highest concordance with observed values (CCC=0.88; R^2=0.82). The RF model showed intermediate accuracy (RMSE=0.019; R^2=0.74), while SVR-RBF exhibited slightly lower predictive power (RMSE=0.021; R^2=0.68). A similar pattern was observed for carvacrol, where the hybrid ensemble again achieved superior performance (RMSE=0.016; CCC=0.86; R^2=0.80), followed by RF (RMSE=0.020; R^2=0.72) and SVR-RBF (RMSE=0.022; R^2=0.66). These results highlight the ability of hybrid approach to integrate the complementary strengths of RF (robustness and feature interpretability) and SVR (nonlinear generalization), leading to improved

Fig. 6. Percentage of habitat suitability of thymol (a) and carvacrol (b) for each suitability class

Figure 6: Fig. 6. Percentage of habitat suitability of thymol (a) and carvacrol (b) for each suitability class

predictive reliability for both target compounds.

3.5 Statistical significance of model performance differences

While numerical differences in predictive metrics indicated that the hybrid ensemble achieved better predictive accuracy, the magnitude of improvement required statistical verification. Therefore, a non-parametric Friedman test followed by the CD test was conducted to assess whether the observed performance differences among the three models were statistically significant.

The Friedman test revealed a significant overall difference among models ($P < 0.05$). The subsequent CD test results (Table 4) confirmed that the hybrid RF+SVR-RBF model performed significantly better than SVR-RBF ($P < 0.05$) for both ΔRMSE and ΔMAE. The difference between RF+SVR-RBF and RF was also statistically significant for ΔRMSE, but only marginally important for ΔMAE, suggesting comparable robustness between these two models. In contrast, the RF and SVR-RBF models did not differ significantly in either metric, indicating that their

Fig. 6 Percentage of habitat suitability of thymol (a) and carvacrol (b) for each suitability class

Table 2 Area optimal habitat of *O. decumbens* for the maximum production of thymol and carvacrol

Model	Classification	Thymol Area (km^2)	Carvacrol Area (km^2)
RF	Low	11,837.64	12,232.28
RF	Moderate	10,638.40	12,265.54
RF	High	8096.37	7412.74
RF	Very high	9668.88	8330.73
SVR-RBF	Low	21,001.44	21,179.25
SVR-RBF	Moderate	6624.14	6551.71
SVR-RBF	High	4288.34	6728.24
SVR-RBF	Very high	8327.37	5782.08
RF+SVR-RBF	Low	12,551.09	15,438.29
RF+SVR-RBF	Moderate	12,055.18	10,626.60
RF+SVR-RBF	High	6687.24	6128.43

Model	Classification	Thymol Area (km^2)	Carvacrol Area (km^2)
RF+SVR-RBF	Very high	8947.78	8047.96

Note: RF, Random Forest; SVR-RBF, Support Vector Regression (SVR) with Radial Basis Function (RBF) kernel; RF+SVR-RBF, hybrid ensemble.

Table 3 Predictive performance of RF, SVR-RBF, and their hybrid RF+SVR-RBF model for thymol and carvacrol concentrations

Essential oil	Model	RMSE	MAE	R^2	CCC
Thymol	RF	0.019	0.015	0.74	0.81
Thymol	SVR-RBF	0.021	0.017	0.68	0.76
Thymol	RF+SVR-RBF	0.015	0.012	0.82	0.88
Carvacrol	RF	0.020	0.016	0.72	0.79
Carvacrol	SVR-RBF	0.022	0.018	0.66	0.74
Carvacrol	RF+SVR-RBF	0.016	0.013	0.80	0.86

Note: RMSE, root mean squared error; MAE, mean absolute error; R^2, coefficient of determination; CCC, concordance correlation coefficient.

Table 4 Statistical comparison of model performance based on the Critical Difference (CD) test

Model comparison	ΔRMSE	Significant	ΔMAE	Significant
RF+SVR-RBF vs. RF	0.004	Significant	0.003	Marginal
RF+SVR-RBF vs. SVR-RBF	0.006	Significant	0.005	Significant
RF vs. SVR-RBF	0.002	Not significant	0.002	Not significant

performance levels were relatively similar. These statistical analyses demonstrate that, despite the close numerical values, the hybridization of RF and SVR-RBF leads to a measurable and statistically validated improvement in predictive accuracy and reliability.

Bar chart showing average rank by model: RF+SVR-RBF = 1.00, RF = 2.00, SVR-RBF = 3.00. The y-axis is "Average rank" and the x-axis is "Model".

Figure 7: Bar chart showing average rank by model: RF+SVR-RBF = 1.00, RF = 2.00, SVR-RBF = 3.00. The y-axis is "Average rank" and the x-axis is "Model".

Figure 7 visualizes the relative performance ranking of the three models based on the CD test results. Lower average ranks indicate better predictive performance. The hybrid RF+SVR-RBF model achieved the lowest average rank (1.00), confirming its superior performance compared with both individual models. The RF model obtained an intermediate rank (2.00), while SVR-RBF ranked the third (3.00), consistent with the quantitative evaluation presented in Table 3. This graphical representation further emphasizes that, while the differences between models were not numerically large, the hybrid model achieved a consistently better rank distribution.

Fig. 7 Statistical comparison of model performance based on the Critical Difference (CD) test

3.6 Determining the importance of environmental factors using GLM

To identify the key ecological drivers shaping both habitat suitability and the biosynthesis of secondary metabolites (thymol and carvacrol) in *O. decumbens*, we employed GLM. This approach enabled us to disentangle the relative contributions of 18 environmental variables to the distribution of metabolites across semi-arid landscapes.

Among the predictors, mean annual temperature emerged as the most influential factor, exhibiting a strong and positive effect on metabolite concentrations. This result suggests that moderately warm conditions enhance both the occurrence of *O. decumbens* and the biosynthesis of its dominant compounds. Aspect was the second most important driver, highlighting the role of

slope orientation in regulating solar radiation, microclimatic conditions, and water availability. Slope ranked the third, with steeper topographies favoring metabolite accumulation, likely due to improved drainage and reduced soil compaction. Sand content was the fourth major contributor, exerting a positive influence, indicating that coarser soil textures promote better aeration and root function.

In contrast, several edaphic and topographic variables constrained metabolite concentrations. Higher soil K content, increased clay percentage, alkaline soil pH, and micro topographic concavities (plan curvature) were negatively associated with thymol and carvacrol, reflecting the adverse effects of nutrient imbalances, finer soil textures, and poorly drained or alkaline microsites on secondary metabolite production (Fig. 8).

Fig. 8 Importance of environmental variables based on generalized linear model (GLM)

Figure 8: Fig. 8 Importance of environmental variables based on generalized linear model (GLM)

Fig. 8 Importance of environmental variables based on generalized linear model (GLM)

4 Discussion

This study established an integrative and spatially explicit framework that linked environmental gradients to the biosynthesis of thymol and carvacrol in *O. decumbens*, an endemic medicinal herb of Iranian drylands. By coupling georeferenced phytochemical data with advanced machine learning algorithms, we disentangled the complex relationships between habitat characteristics and metabolite accumulation across environmentally heterogeneous semi-arid landscapes. The approach illustrated how combining ecological datasets with hybrid computational models can reveal the mechanisms underlying chemical adaptation to environmental stress. Our findings provided both mechanistic insight into secondary metabolism and a predictive basis for managing medicinal plant resources in water-limited ecosystems.

A major outcome of this work was the superior performance of the hybrid RF+SVR-RBF model compared with single model. The ensemble consistently produced lower predictive errors (RMSE and MAE) and higher agreement metrics (R^2 and CCC), confirming that hybridization between decision tree-based and kernel-based methods can synergistically enhance model stability and generalizability. RF captured hierarchical and interaction effects among predictors, while SVR-RBF effectively modeled continuous gradients, and their integration minimized both bias and variance. Similar benefits of hybrid ensembles have been demonstrated for other medicinal species such as *Nepeta persica* Boiss. and *Salvia leriifolia* L. (Dastres et al., 2025b, e), reinforcing the value of multi-algorithmic fusion in ecological modeling under environmental complexity.

Hybrid models typically entail increased computational costs due to the need for training multiple algorithms, hyperparameter tuning across different architectures, and cross-validation for weight optimization (Toderean et al., 2025). In this study, the ensemble framework required approximately 2.3 times longer computation time compared with individual RF or SVR-RBF model. Additionally, ensemble method can reduce model interpretability compared with simpler single-algorithm approach, as the combination of predictions from different underlying mechanisms obscures direct tracing of prediction outcomes to specific input variables (Zhong et al., 2022). To address this trade-off between predictive performance and interpretability, we deliberately balanced the modeling framework by: (1) using GLM analysis as a complementary interpretable approach to identify the direction and significance of individual environmental drivers; (2) reporting variable importance measures from RF, which remain accessible even within the ensemble context; and (3) limiting the ensemble to only two base models (rather than many) to maintain transparency in the weighting scheme. This multi-pronged strategy allowed us to leverage the superior predictive accuracy of the hybrid model while preserving ecological interpretability, a critical consideration for studies aimed at informing conservation and management decisions.

While the hybrid ensemble modeling framework employed in this study demonstrated robust predictive performance for thymol and carvacrol concentrations in *O. decumbens*, it is important to point out the limitations imposed by sample size relative to model complexity. With 59 sampling points and 18 environmental predictors, our sample-to-predictor ratio of approximately 3.3:1.0 fall within what is commonly considered a "small data" scenario in ecological machine learning. Previous study has suggested that sample-to-feature ratios below 20.0:1.0 may increase the risk of overfitting and challenge model generalizability to unsampled regions (Wang et al., 2025).

Despite these theoretical concerns, several factors in our study design mitigated the associated risks and support the reliability of our findings. First, the implementation of nested repeated k-fold cross-validation (10 folds with 10 repetitions) ensured that every observation contributed to both training and validation, reducing optimism in performance estimates (Wang et al., 2024). Second, the use of ensemble method (RF+SVR-RBF) with weighted soft averaging leveraged complementary bias-variance characteristics, which has been shown to enhance model stability under data-limited conditions. Third, our logit transformation of bounded response variables improved data distribution properties and model convergence. Comparable studies in medicinal plant research have successfully employed similar sample sizes. For instance, Wang et al. (2024) developed robust classification models for Chinese herbal medicine origin identification using 568 samples with 3448 spectral features, demonstrating that appropriate methodological choices can overcome sample size limitations in high-dimensional settings.

From a biogeographical perspective, the northern and northwestern parts of Fars Province emerged as metabolite-rich zones, with additional hotspots in the western uplands. These regions are characterized by moderate thermal regimes, heterogeneous topography, and sandy soils conditions identified by the GLM as strong positive predictors. Conversely, central and eastern lowlands exhibited limited suitability, likely due to unfavorable soil texture and microtopographic features. Specifically, areas with concave surface curvature (negative values) in these lowlands tend to accumulate water during precipitation events, creating transient waterlogged conditions that can restrict root aeration and reduce oxygen availability for root metabolism, factors known to negatively impact terpenoid biosynthesis in aromatic plants (Sun and Fernie, 2024). Additionally, these micro-depressions often serve as cold air convergence zones during spring nights, increasing frost risk and potentially damaging young tissues at critical phenological stages. In contrast, convex microtopographic positions (positive curvature) on slopes promote rapid drainage and cold air shedding, creating more favorable microclimatic conditions for secondary metabolite accumulation (Troise et al., 2025). Such spatial clustering of high-thymol and high-carvacrol populations may reflect localized adaptation to stress gradients, a common phenomenon in aromatic taxa inhabiting arid and semi-arid ecosystems (Najar et al., 2022; Singh et al., 2023).

Among environmental drivers, mean annual temperature emerged as the dominant and positive regulator of secondary metabolite accumulation. Moderate heat levels may stimulate enzymatic activities in the monoterpene biosynthetic pathway, particularly through enhanced expression of geranyl diphosphate synthase and terpene synthase genes (Yan et al., 2017). Temperature-driven

modulation of volatile terpenoid metabolism has been reported in several Lamiaceae and Apiaceae species, suggesting that thermal conditions fine-tune carbon allocation between primary and secondary metabolism (Willi and Van Buskirk, 2022). Aspect, the second most influential variable, reflects the role of solar radiation exposure and evapotranspiration in shaping plant microclimates. South- and southwest-facing slopes, which receive higher insolation, provide warmer and drier microsites that can increase metabolic fluxes through stress-induced signaling (e.g., abscisic acid (ABA) and jasmonates), enhancing terpenoid biosynthesis (Walton et al., 2005; Deák et al., 2024).

Slope degree also showed a positive effect, likely due to improved aeration, lower soil compaction, and reduced pathogen pressure on steeper terrains (Wen et al., 2024). Sandy soils further favored secondary metabolism by increasing oxygen diffusion and reducing mechanical impedance to root growth—conditions that facilitate precursor availability for terpenoid formation (Hill et al., 2023). Similar findings were observed for *N. persica*, where sandy and well-drained soils were associated with elevated nepetalactone content (Dastres et al., 2025b).

Conversely, inhibitory effects were associated with soil K, clay content, and pH. Elevated K concentrations may disturb osmotic equilibrium and nutrient uptake, constraining the flow of carbon skeletons into secondary pathways (Benito et al., 2023). Excessive clay reduces drainage and limits microbial diversity, which in turn affects rhizosphere-metabolite feedbacks (Tang et al., 2025). Alkaline pH may also restrict the availability of micronutrients such as ferrum (Fe), zinc (Zn), and manganese (Mn)—essential cofactors for enzymes involved in monoterpene synthesis (Moreno-Jiménez et al., 2019). Microtopographic depressions (negative plan curvature) were negatively related to metabolite levels, likely reflecting the effects of waterlogging and hypoxia on root respiration and metabolic activity (Wan et al., 2024).

Collectively, these results reveal a suite of ecological trade-offs controlling secondary metabolism in *O. decumbens*. Optimal metabolite accumulation occurs under moderately warm, well-drained, and aerated conditions, whereas environments characterized by high clay content, alkalinity, or ionic imbalance inhibit biosynthetic efficiency. This pattern aligns with broader evidence that both climatic and edaphic factors jointly control phytochemical diversity in dryland ecosystems (Najar et al., 2022; Singh et al., 2023). Furthermore, these results extend chemical ecological understanding by highlighting that subtle spatial shifts in soil texture, slope, and temperature gradients can generate disproportionate biochemical variation within a single species—reflecting the high sensitivity of dryland flora to microhabitat heterogeneity.

Beyond ecological implications, the generated spatial predictions hold strong applied relevance. The ability to map high-metabolite zones provides a scientific basis for *in situ* conservation of chemically valuable populations, while informing *ex situ* cultivation strategies that reduce collection pressure on wild stands. The integration of ecological modeling and metabolite mapping thus offers a dual framework for both biodiversity conservation and sustainable resource utilization. Similar spatially guided conservation strategies have proven effective for optimizing cultivation of other aromatic plants in arid regions (Måren et al., 2015; Dastres et al., 2025e). Ultimately, this study bridges machine-learning advances with biological interpretation, demonstrating how data-driven modeling can elucidate environmental regulation of plant secondary metabolism and guide adaptive management under global climate acidification.

5 Conclusions

This study highlighted the potential of integrating spatially explicit metabolite profiling with ensemble machine learning to elucidate the ecological regulation of secondary metabolism in *O. decumbens*. The hybrid RF+SVR-RBF model achieved the highest predictive accuracy for thymol and carvacrol concentrations, demonstrating the advantage of ensemble learning in capturing complex environmental interactions within heterogeneous dryland ecosystems. Our results indicated that mean annual temperature, slope aspect, slope degree, and sand content act as strong positive drivers of metabolite biosynthesis, while high soil K, clay content, alkaline pH, and

concave topographies exert negative effects. These findings suggest that *O. decumbens* exhibits adaptive metabolic plasticity to cope with semi-arid stressors, with metabolite production optimized under warm, well-drained, and aerated conditions. The combined influence of climatic, edaphic, and topographic gradients underscores the integrative nature of environmental control over chemical diversity in dryland medicinal and aromatic plants. By establishing a predictive ecological framework that links metabolite variation to environmental heterogeneity, this study contributes to both the theoretical understanding of plant chemical ecology and the practical goals of conservation and sustainable utilization. The spatial predictions generated here provide valuable tools for identifying and prioritizing ecologically and chemically valuable populations, supporting site-specific strategies for *in situ* conservation and sustainable cultivation. Future research should expand upon this framework by incorporating population genomic data to investigate the interactions between genetic variation and environmental factors in shaping metabolite phenotypes. Additionally, integrating functional traits and long-term climate scenarios will be essential for understanding the adaptive evolution of phytochemical pathways under accelerating climate change impacts. Such interdisciplinary approaches will be crucial for safeguarding the ecological and chemical resilience of medicinal species in arid ecosystems.

Conflict of interest

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

Acknowledgements

The authors gratefully acknowledge the support and resources provided by the Shahid Beheshti University, Iran.

Author contributions

Emran DASTRES: conceptualization, methodology, investigation, formal analysis, writing - original draft preparation; Hassan ESMAEILI: supervision, validation, investigation, writing - review and editing: All authors approved the manuscript.

References

Adams R P. 2017. Identification of Essential Oil Components by Gas Chromatography/Mass Spectrometry (4^{th} ed.). Carol Stream: Allured Publishing.

Ahmed N, Atzberger C, Zewdie W. 2021. Species distribution modelling performance and its implication for Sentinel-2-based prediction of invasive *Prosopis juliflora* in lower Awash River basin, Ethiopia. Ecological Processes, 10(1): 8, doi: 10.1186/s13717-021-00285-6.

AlHamaydeh M, Al-Shamsi G, Aly N, et al. 2021. Seismic risk quantification and GIS-based seismic risk maps for Dubai-UAE_Dataset. Data in Brief, 39: 107566, doi: 10.1016/j.dib.2021.107566.

Amatulli G, Domisch S, Tuanmu M N, et al. 2018. A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. Scientific Data, 5: 180040, doi: 10.1038/sdata.2018.40.

Azizi K, Soltani Z, Aliakbarpour M, et al. 2022. Bionomics of phlebotomine sand flies in different climates of leishmaniasis in Fars Province, southern Iran. Journal of Arthropod-Borne Diseases, 16(2): 148–161.

Balkrishna A, Sharma N, Srivastava D, et al. 2024. Exploring the safety, efficacy, and bioactivity of herbal medicines: Bridging traditional wisdom and modern science in healthcare. Future Integrative Medicine, 3(1): 35–49.

Benito P, Bellón J, Porcel R, et al. 2023. The biostimulant, potassium humate ameliorates abiotic stress in *Arabidopsis thaliana* by increasing starch availability. International Journal of Molecular Sciences, 24(15): 12140, doi: 10.3390/ijms241512140.

Boveiri Dehsheikh A, Mahmoodi Sourestani M, Enayatizamir N, et al. 2024. *Oliveria decumbens*, a long-neglected plant with promising phytochemical and biological properties. Chemistry & Biodiversity, 21: e202400810, doi: 10.1002/cbdv.202400810.

Capitani M, Ribolini A, Bini M. 2013. The slope aspect: A predisposing factor for landsliding? Comptes Rendus Geoscience, 345(11–12): 427–438.

Catucci E, Scardi M. 2020. A machine learning approach to the assessment of the vulnerability of *Posidonia oceanica* meadows. Ecological Indicators, 108: 105744, doi: 10.1016/j.ecolind.2019.105744.

Chen X Y, Jiang Z H, Tai Q L, et al. 2022. Construction of a photosynthetic rate prediction model for greenhouse strawberries with distributed regulation of light environment. Mathematical Biosciences and Engineering, 19(12): 12774–12791.

Cheng J H, Sun J, Yao K S, et al. 2022. A variable selection method based on mutual information and variance inflation factor. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 268: 120652, doi: 10.1016/j.saa.2021.120652.

Chu H J, Chen Y C, Ali M Z, et al. 2019. Multi-parameter relief map from high-resolution DEMs: A case study of mudstone badland. International Journal of Environmental Research and Public Health, 16(7): 1109, doi: 10.3390/ijerph16071109.

Coley R Y, Liao Q Q, Simon N, et al. 2023. Empirical evaluation of internal validation methods for prediction in large-scale clinical data with rare-event outcomes: A case study in suicide risk prediction. BMC Medical Research Methodology, 23(1): 33, doi: 10.1186/s12874-023-01844-5.

Das S, Patel P P, Sengupta S. 2016. Evaluation of different digital elevation models for analyzing drainage morphometric parameters in a mountainous terrain: A case study of the Supin–Upper Tons Basin, Indian Himalayas. SpringerPlus, 5: 1544, doi: 10.1186/s40064-016-3207-0.

Dastres E, Edalat M, Moayedi G, et al. 2022. An evaluation of alternative statistical models for predicting habitat suitability for weeds. Weed Research, 62(6): 404–421.

Dastres E, Jahangiri E, Edalat M, et al. 2023. Habitat suitability modeling of

Descurainia sophia medicinal plant using three bivariate models. Environmental Monitoring and Assessment, 195(3): 392, doi: 10.1007/s10661-023-10996-2.

Dastres E, Esmaeili H, Edalat M. 2025a. Species distribution modeling of *Malva neglecta* Wallr. weed using ten different machine learning algorithms: An approach to site-specific weed management (SSWM). European Journal of Agronomy, 167: 127579, doi: 10.1016/j.eja.2025.127579.

Dastres E, Esmaeili H, Sonboli A, et al. 2025b. Optimizing cultivation areas for *Salvia leriifolia* using advanced spatial prediction and hybrid machine learning algorithms for maximum bioactive compound yield. Smart Agricultural Technology, 12: 101482, doi: 10.1016/j.atech.2025.101482.

Dastres E, Rabiei-Dastjerdi H, Esmaeili H, et al. 2025c. Modeling habitat suitability for endangered herb (*Salvia leriifolia* Benth) using innovative hybrid machine learning algorithms. Environmental and Sustainability Indicators, 26: 100694, doi: 10.1016/j.indic.2025.100694.

Dastres E, Sarvestani G S, Edalat M, et al. 2025d. Habitat suitability modeling of dominant weed in rapeseed (*Brassica napus*) fields using machine learning techniques. Weed Science, 73: e35, doi: 10.1017/wsc.2025.5.

Dastres E, Sonboli A, Esmaeili H, et al. 2025e. Predicting nepetalactone accumulation in *Nepeta persica* using machine learning algorithms and geospatial analysis. Scientific Reports, 15: 31535, doi: 10.1038/s41598-025-17039-5.

de Santana F B, de Souza A M, Poppi R J. 2019. Green methodology for soil organic matter analysis using a national near infrared spectral library in tandem with learning machine. Science of The Total Environment, 658: 895–900.

Deák B, Botta-Dukát Z, Rádai Z, et al. 2024. Meso-scale environmental heterogeneity drives plant trait distributions in fragmented dry grasslands. Science of The Total Environment, 947: 174355, doi: 10.1016/j.scitotenv.2024.174355.

Elshaboury N, Elmousalami H. 2025. Wind speed and power forecasting using Bayesian optimized machine learning models in Gabal Al-Zayt, Egypt. Scientific Reports, 15(1): 28500, doi: 10.1038/s41598-025-13140-x.

Esmaeili H, Karami A, Maggi F. 2018. Essential oil composition, total phenolic and flavonoids contents, and antioxidant activity of *Oliveria decumbens* Vent. (Apiaceae) at different phenological stages. Journal of Cleaner Production, 198: 91–95.

Forti M, Lavie Y, Ben-Dov Y, et al. 2006. Long-term plant survival and development under dryland conditions in an experimental site in the semi-arid Negev of Israel. Journal of Arid Environments, 65(1): 1–28.

Hassanpouraghdam M B, Ghorbani H, Esmaeilpour M, et al. 2022. Diversity and distribution patterns of endemic medicinal and aromatic plants of Iran: Implications for conservation and habitat management. International

Journal of Environmental Research and Public Health, 19(3): 1552, doi: 10.3390/ijerph19031552.

He B H, Ma S G, Zhang X Y, et al. 2023. Rank-based greedy model averaging for high-dimensional survival data. Journal of the American Statistical Association, 118(544): 2658–2670.

Hesami M, Jones A M P. 2020. Application of artificial intelligence models and optimization algorithms in plant cell and tissue culture. Applied Microbiology and Biotechnology, 104(23): 9449–9485.

Hill R D, de Castro J, Mira M M, et al. 2023. Over-expression of the barley Phytoglobin 1 (HvPgb1) evokes leaf-specific transcriptional responses during root waterlogging. Journal of Plant Physiology, 283: 153944, doi: 10.1016/j.jplph.2023.153944.

Ibrahimi H, Mohammadi H, Ghaderi E, et al. 2024. A new Drusinae (Trichoptera: Limnephilidae) species from the Zagros Mountains in the Kurdistan Province, Iran. Zootaxa, 5406(3): 451–460.

Ji J C, Wei X X, Guan H, et al. 2025. Evaluation of the impact of ecological factors on the habitat suitability and bioactive components accumulation of the medicinal holoparasitic plant *Cynomorium songaricum* using machine learning models. Frontiers in Plant Science, 16: 1586682, doi: 10.3389/fpls.2025.1586682.

Karami A, Khoshbakht T, Esmaeili H, et al. 2020. Essential oil chemical variability in *Oliveria decumbens* (Apiaceae) from different regions of Iran and its relationship with environmental factors. Plants, 9(5): 680, doi: 10.3390/plants9050680.

Karim M R, Ahmed S, Reza M N, et al. 2024. Geometric feature characterization of apple trees from 3D LiDAR point cloud data. Journal of Imaging, 11(1): 5, doi: 10.3390/jimaging11010005.

Keshtegar B, Yaseen Z M. 2022. Reinforcing bar development length modeling using integrative support vector regression model with response surface method: New approach. ISA Transactions, 128: 423–434.

Khan M O, Klamerus-Iwan A, Kupka D, et al. 2023. Short-term impact of different doses of spent coffee grounds, salt, and sand on soil chemical and hydrological properties in an urban soil. Environmental Science and Pollution Research, 30(36): 86218–86231.

Li Z Y, Zheng L, ZHENG Z Z, et al. 2025. A new method for turbidity removal and decolorization in the determination of available phosphorus using sodium bicarbonate leaching. Rock and Mineral Analysis, 44(6): 1192–1202. (in Chinese)

Li Y K. 2018. Determining topographic shielding from digital elevation models for cosmogenic nuclide analysis: A GIS model for discrete sample sites. Journal of Mountain Science, 15(5): 939–947.

Liu F G, Chen Z R, Xu J, et al. 2024. Interpretable machine learning-based influence factor identification for 3D printing process-structure linkages. Polymers, 16(18): 2680, doi: 10.3390/polym16182680.

Majeed M. 2017. Evidence-based medicinal plant products for the health care of world population. Annals of Phytomedicine, 6(1): 1–4.

Måren I E, Karki S, Prajapati C, et al. 2015. Facing north or south: Does slope aspect impact forest stand characteristics and soil properties in a semiarid trans-Himalayan valley? Journal of Arid Environments, 121: 112–123.

McCoy D, Hubbard A, Van der Laan M. 2023. CVtreeMLE: Efficient estimation of mixed exposures using data adaptive decision trees and cross-validated targeted maximum likelihood estimation in R. Journal of Open Source Software, 8(82): 4181, doi: 10.21105/joss.04181.

Moreno-Jiménez E, Plaza C, Saiz H, et al. 2019. Aridity and reduced soil micronutrient availability in global drylands. Nature Sustainability, 2(5): 371–377.

Mousavi S S, Karami A, Haghighi T M, et al. 2023. Phytochemistry of *Oliveria decumbens* Vent. (Apiaceae) and its therapeutic potential: A systematic review. Fitoterapia, 170: 105647, doi: 10.1016/j.fitote.2023.105647.

Mu K L, Ran F, Feng Y Q, et al. 2025. Integrated evaluation of habitat suitability and spatial prediction of medicinal quality for *Bletilla striata* using species distribution models and machine learning approaches. Industrial Crops and Products, 236: 121937, doi: 10.1016/j.indcrop.2025.121937.

Mwendwa S. 2022. Revisiting soil texture analysis: Practices towards a more accurate Bouyoucos method. Heliyon, 8(5): e09395, doi: 10.1016/j.heliyon.2022.e09395.

Najar B, Pistelli L, Fratini F. 2022. Exploitation of marginal hilly land in Tuscany through the cultivation of *Lavandula angustifolia* Mill.: Characterization of its essential oil and antibacterial activity. Molecules, 27(10): 3216, doi: 10.3390/molecules27103216.

Park S H, Jang J Y, Lim H, et al. 2025. Ecological dynamics of volatile organic compound-mediated interactions in *Aristolochia contorta* with parasitoids and herbivores. Ecology and Evolution, 15(1): e71175, doi: 10.1002/ece2.71175.

Pintar A L, Levine Z H, Yoon H W, et al. 2023. Linearity characterization and uncertainty quantification of spectroradiometers via maximum likelihood and the non-parametric bootstrap. Metrologia, 60(1): 015002, doi: 10.1088/1681-7575/acb5f8.

Rather T A, Kumar S, Khan J A. 2020. Multi-scale habitat modelling and predicting change in the distribution of tiger and leopard using random forest algorithm. Scientific Reports, 10(1): 11473, doi: 10.1038/s41598-020-68167-z.

Regasa A, Haile W, Abera G. 2024. Assessment of soil acidity and fertility status under different land uses types in Sayo District of Western Ethiopia. Russian Agricultural Sciences, 50(2): 172–184.

Rout S, Kumar R V, Sachan K. 2023. Flora diversity and conservation status of medicinal plants: Insights from a semi-arid to sub-humid subtropical region. Plant Science Archives, 8(4): 23–25.

Saha A K, Agrawal S. 2020. Mapping and assessment of flood risk in Prayagraj district, India: A GIS and remote sensing study. Nanotechnology for Environmental Engineering, 5(1): 11, doi: 10.1007/s41204-020-00073-1.

Selassie Y G, Molla E, Muhabie D, et al. 2020. Response of crops to fertilizer application in volcanic soils. Heliyon, 6(12): e05629, doi: 10.1016/j.heliyon.2020.e05629.

Shahjaman M, Rahman M R, Islam T, et al. 2021. rMisbeta: A robust missing value imputation approach in transcriptomics and metabolomics data. Computers in Biology and Medicine, 138: 104911, doi: 10.1016/j.compbiomed.2021.104911.

Singh D K, Singh A, Gacem A, et al. 2023. Multiple site dissimilarities of herbaceous species due to coal fly ash dumping based on soil heavy metal toxication. Toxics, 11(2): 90, doi: 10.3390/toxics11020090.

Singhal R, Izquierdo P, Ranaweera T, et al. 2025. Using supervised machine-learning approaches to understand abiotic stress tolerance and design resilient crops. Philosophical Transactions of the Royal Society B: Biological Sciences, 380(1927): 20240252, doi: 10.1098/rstb.2024.0252.

Spake R, Jackson E E, Bullock J M, et al. 2025. Precision ecology for targeted conservation action. Nature Ecology & Evolution, 9(7): 1102–1111.

Srivastava A K. 2018. Significance of medicinal plants in human life. In: Tewari A, Tiwari S. Synthesis of Medicinal Agents from Plants. Amsterdam: Elsevier, 1–24.

Sun Y M, Fernie A R. 2024. Plant secondary metabolism in a fluctuating world: Climate change perspectives. Trends in Plant Science, 29(5): 560–571.

Tang J Q, Wu H Y, Li Y Q, et al. 2025. The response microbial of the cucumber rhizosphere network keystone taxa of the cucumber rhizosphere to continuous fertilization. Microorganisms, 13(2): 451, doi: 10.3390/microorganisms13020451.

Toderean L, Daian M, Cioara T, et al. 2025. Heuristic based federated learning with adaptive hyperparameter tuning for households energy prediction. Scientific Reports, 15(1): 12564, doi: 10.1038/s41598-025-96443-3.

Troise A D, De Pascale S, Piluzza G, et al. 2025. Metabolite variations in Sulla shoots induced by drought stress. Journal of the Science of Food and Agriculture, 105(2): 1069–1079.

Vilela A E, González-Paleo L, Ravetta D A. 2011. Secondary metabolites of woody plants from arid lands: Mechanisms involved, functions and uses. Ecología Austral, 21(3): 317-327.

Walton J C, Martinez-Gonzalez F, Worthington R. 2005. Desert vegetation and timing of solar radiation. Journal of Arid Environments, 60(4): 697-707.

Wan Y L, Yang Z X, Sun K L, et al. 2024. Micrometer groove array structures inspired by desert grasses on curved substrates for enhanced surface condensation efficiency. Langmuir, 40(38): 20035-20044.

Wang J F, Cai Q L. 2024. Nitrogen migration and transformation characteristics of the soil in Karst areas under the combined application of oxalic acid and urea inhibitors. Frontiers in Plant Science, 15: 1386912, doi: 10.3389/fpls.2024.1386912.

Wang J J, Zhou Q, Shang J L, et al. 2021. UAV- and machine learning-based retrieval of wheat SPAD values at the overwintering stage for variety screening. Remote Sensing, 13(24): 5166, doi: 10.3390/rs13245166.

Wang Y J, Jin C L, Ma L, et al. 2024. A robust TabNet-based multi-classification algorithm for infrared spectral data of Chinese herbal medicine with high-dimensional small samples. Journal of Pharmaceutical and Biomedical Analysis, 242: 116031, doi: 10.1016/j.jpba.2024.116031.

Wang Y, Dong J C, Zhou Y C, et al. 2025. Addressing the data scarcity problem in ecotoxicology via small data machine learning methods. Environmental Science & Technology, 59(12): 5867-5871.

Warton D I, Hui F K. 2011. The arcsine is asinine: The analysis of proportions in ecology. Ecology, 92(1): 3-10.

Wen D N, Yang L, Ni K, et al. 2024. Topography-driven differences in soil N transformation constrain N availability in Karst ecosystems. Science of The Total Environment, 908: 168363, doi: 10.1016/j.scitotenv.2023.168363.

Willi Y, Van Buskirk J. 2022. A review on trade-offs at the warm and cold ends of geographical distributions. Philosophical Transactions of the Royal Society B: Biological Sciences, 377(1848): 20210022, doi: 10.1098/rstb.2021.0022.

Xie Q W, Xue F C, Chen J F. 2024. Ensemble learning prediction model for rapeseed flowering periods incorporating virtual sample generation. Transactions of the Chinese Society of Agricultural Engineering, 40(19): 159-167.

Yan N, Liu Y H, Zhang H B, et al. 2017. Solanesol biosynthesis in plants. Molecules, 22(4): 510, doi: 10.3390/molecules22040510.

Yoon T, Kang D. 2023. Multi-modal stacking ensemble for the diagnosis of cardiovascular diseases. Journal of Personalized Medicine, 13(2): 373, doi: 10.3390/jpm13020373.

Zhong H Y, Brandeau M L, Yazdi G E, et al. 2022. Metamodeling for policy simulations with multivariate outcomes. Medical Decision Making, 42(7): 872-884.

Note: Figure translations are in progress. See original paper for figures.