



Synthetic seed production in *Crataegus monogyna* L. and prediction of regeneration of synthetic seeds with machine learning algorithms

Metin Koçak¹ · Mustafa Can Yılmaz² · Cansu Kuzğun¹ · Sibel Turan Sirke¹ · Mehtap Yildiz¹

Received: 20 December 2024 / Accepted: 26 March 2025 / Published online: 8 April 2025
© The Author(s) 2025

Abstract

Crataegus monogyna is a complex species that is essential in both ecological and therapeutic domains. Its versatility across several settings, along with its extensive phytochemical composition, renders it a significant focus of research in both botany and medicine. One of the research areas that could be focused on is the propagation of *C. monogyna* under in vitro conditions. In this study, we encapsulated nodal segments of sterile shoots from *C. monogyna* plants growing naturally in Van, Türkiye. Encapsulated propagules were cultured in hormone-free Murashige and Skoog medium or Murashige and Skoog medium supplemented with 2 mg/L indole-3-acetic acid (IAA) after being stored for 30, 60, or 90 days at −20, 4, or 24 °C. The regeneration of synthetic seeds under the effects of hormone (IAA), storage temperature, and storage period was predicted using five machine learning algorithms: Decision Tree (DT), Gaussian Process (GP), Multi-Layer Perceptron (MLP), Random Forest (RF), and XGBoost (Extreme Gradient Boosting). Feature importance analysis was conducted to identify the key factors influencing regeneration outcomes. The DT, MLP, RF, and XGBoost models achieved high prediction accuracy (97.3%). Furthermore, while the DT and XGBoost models identified temperature as the most influential factor, the MLP and RF models found hormone to be the most significant. Surface and contour plot analyses were also employed to assess the relationships visually between key features of the regeneration process.

Key message

The synthetic seeds of *Crataegus monogyna* were regenerated after being stored at 4 °C and 24 °C for 1, 2, and 3 months, and the regenerations of the synthetic seeds were predicted with five different machine learning algorithms with high accuracy.

Keywords *Crataegus monogyna* · Synthetic seed · Encapsulation · Machine learning

Introduction

Hawthorn (*Crataegus* spp.), belonging to the Rosaceae family, is estimated to include 150–200 species from different sources worldwide. Türkiye is the most important gene center of naturally growing hawthorn species and is reported to contain 22 species of the hawthorn genus. Hawthorn grows naturally in Türkiye's mountainous areas, bushes, and rocky terrains, and no cultural practice is applied to these natural plants. Hawthorn species are found wild in the Central Anatolia, Aegean, Eastern Anatolia, Mediterranean, and Northern Anatolia regions. In these regions, different hawthorn species show intense morphological diversity in terms of plant habit and form, especially fruit color and shape. The wide diversity of hawthorn species is attributed to the

Communicated by Maurizio Lambardi.

✉ Metin Koçak
metinkocak@yyu.edu.tr

✉ Mehtap Yildiz
mehtapyildiz@gmail.com

¹ Faculty of Agriculture, Agricultural Biotechnology Department, Van Yuzuncu Yil University, Van, Turkey

² Faculty of Agriculture, Animal Science Department, Van Yuzuncu Yil University, Van, Turkey

scattering of hawthorn seeds by some hawthorn-eating birds and mammals (Özcan et al. 2005; Dönmez 2007).

C. monogyna is a deciduous shrub or small tree belonging to the Rosaceae family. It typically reaches 2 to 8 m and is characterized by its thorny branches and white flowers, which bloom in spring. The shrub is widely distributed across Europe, particularly in hedgerows, scrubs, and woodlands, and it thrives in various of soil types, demonstrating adaptability to medium conditions regarding pH, nutrients, and water supply (Fichtner and Wissemann 2021; Martinelli et al. 2021). The fruit of *C. monogyna* is fleshy, red, and aromatic, making it ecologically significant and valuable for human consumption and wildlife (Martinelli et al. 2021).

The ecological role of *C. monogyna* extends beyond its physical presence; it serves as a vital food source for various frugivores due to its bright red berries, which are attractive to birds and other animals (Lediuk et al. 2014). This species is particularly noted for its drought tolerance, allowing it to flourish in sun-exposed areas and habitat corridors, contributing to biodiversity and ecosystem stability (Urban et al. 2023). Furthermore, *C. monogyna* is known for its ability to hybridize with other hawthorn species, which can enhance its genetic diversity and adaptability (Christensen et al. 2014; Kuhn et al. 2021).

From a medicinal perspective, *C. monogyna* has been utilized in traditional medicine for centuries, particularly for its cardiovascular benefits. Extracts from its fruits, leaves, and flowers are commonly employed in herbal remedies to treat conditions such as hypertension, arrhythmias, and heart failure (Bernatoniene et al. 2009). The pharmacological effects of hawthorn are largely attributed to its rich composition of bioactive compounds, including flavonoids, tannins, and vitamin C, which exhibit antioxidant properties and contribute to its therapeutic efficacy (Martinelli et al. 2021; Bahorun et al. 2003; Yanakieva 2023). Recent studies have highlighted the potential of *C. monogyna* extracts in reducing oxidative stress and improving mitochondrial function in cardiac tissues, underscoring its relevance in modern phytotherapy (Bernatoniene et al. 2009).

Synthetic seeds are coated somatic embryos or other vegetative in vitro-derived propagules with special substances containing nutrients and protective compounds. This new strategy is especially beneficial for plants with restricted seed production, by mimicking zygotic seeds using plant tissues and/or organs other than reproductive organs. Sodium alginate is widely used for coating somatic embryos in synthetic seed production due to its low cost, low toxicity, and gel stability (Ghosh and Haque 2019). The hardening of the insoluble calcium alginate surrounding the propagule occurs through an ion exchange reaction between the Na^+ in the sodium alginate solution and the calcium in the Ca^{2+} chloride solution. (Daud et al. 2008). The encapsulation of the

propagule in this way provides significant protection for the propagule and nutrients needed by the propagule during the seed regeneration stage.. The biodegradable matrix that makes up the synthetic seed effectively protects the explants from physical damage and external environmental factors. It also serves as an endosperm for the somatic embryo, containing nutrients necessary for early growth (Jang et al. 2020). This technique is mainly recommended for conserving and propagating endangered species, medicinal plants, and commercial plants (Jang et al. 2020).

In recent years, machine learning (ML) techniques have become increasingly valuable in agricultural research, particularly in the field of plant biotechnology (Yoosefzadeh Najafabadi et al. 2023). ML approaches have been employed to predict outcomes, optimize processes, and gain deeper insights into complex biological systems (Dirk et al. 2021; Hesami et al. 2022; Niazian and Niedbala 2020; Silva et al. 2019). Also, ML algorithms are used to enhance crop yield predictions by analyzing environmental conditions, soil properties, and genetic factors (Folorunso et al. 2023; Yan and Wang 2023; Yang et al. 2021). For instance, Galatali et al. (2021) demonstrated the production of disease-free olive seedlings using artificial intelligence and biotechnological methods, emphasizing the potential of ML in germplasm and agriculture. Similarly, ML algorithms were applied to predict soybean seed yield using hyperspectral reflectance, highlighting the robustness of ML techniques in plant breeding (Yoosefzadeh-Najafabadi et al. 2021).

In plant studies, the optimization process is crucial, and the utilization of new computational methods, such as machine learning algorithms with a reduced number of inputs, has gained prominence in recent years (Xie et al. 2022). When conducting research involving plant seedlings and employing fewer inputs with various machine learning techniques, it is essential to consider the selection of an appropriate model to enhance the precision of result predictions (Dinga et al. 2020). This aligns with the trend observed in plant science, where machine learning methods like support vector regression, random forest regression, and neural networks have been successfully applied to predict traits such as flood resistance in plants (Xie et al. 2022) and survival rates of seedlings under extreme conditions (Vásquez-Coronel et al. 2022).

In this study, we aimed to contribute to the conservation of *C. monogyna* L. as a genetic resource using synthetic seed technology and determine the effect of different storage times, storage temperatures, and hormone compositions on synthetic seed regeneration percentages. Machine learning models such as Decision Tree, Gaussian Process, Multilayer Perceptron, Random Forest, and XGBoost were used to predict the regeneration results of synthetic seeds. The research aims to identify the most effective techniques by comparing

model performance with accuracy, confusion matrices, F1 scores, cross-validation averages, and feature importance. Surface and contour plot analyses were also used to better understand the regeneration process.

Material and methods

Plant material

In this study, micropropagated cultures of *Crataegus monogyna* subsp. *monogyna* previously in in vitro conditions were used as explant source. Plant material was originated from plants with orange-colored fruit that grew naturally in Van, Türkiye.

Encapsulation of propagules, storage, and regeneration of synthetic seeds

The propagules were coated in two different encapsulation protocols. The encapsulation matrix consisted of MS (Murashige and Skoog, 1962) basal salt mixture and MS vitamins as nutrient basal solution (artificial endosperm), 30 g/L sucrose and 3% sodium alginate as solidifier in both protocols. The first protocol included 2 mg/L IAA (MS+IAA) supplementation, whereas the second protocol featured a hormone-free (HF-MS) composition.

Nodal segments of approximately 2 mm were cut as propagules and mixed with an encapsulation matrix. The propagules with encapsulation matrix were individually dropped into a 100 mM CaCl₂ solution using a micropipette with sterile enlarged tips to ensure the proper size of the synthetic seed. CaCl₂ solution was shaken gently until the gel matrix hardened to form well-shaped beads. The beads were rinsed with sterile distilled water in a sterile metal strainer to remove residual debris from the CaCl₂ solution. Then, the synthetic seeds were dried on sterile blotters and ready for storage or culture (Fig. 1A, B). The encapsulated propagules (synthetic seeds) were stored in sterile centrifuge tubes with a volume of 50 ml at −20 °C, 4 °C, or 24 °C for one, two, or three months in dark conditions (Table 1).

After the storage period, the synthetic seeds were transferred to glass jars containing 50 ml of the culture media for regeneration. In both experiments, the media used for the encapsulation matrix were also used for the culture of synthetic seeds. The regeneration media were solidified with 3.7 g/L gelrite, and the pH was adjusted to 5.7–5.8 before autoclaving. The synthetic seeds were cultured at 24 ± 1 °C and 16 h light/8 h dark period.

The experiments were designed based on completely randomized plots with 10 replicates and 5 synthetic seeds per replicate. The effects of hormone (HF-MS or MS+IAA),

storage period (1, 2, or 3 months), and storage temperature (−20 °C, 4 °C, or 24 °C) on synthetic seed regeneration were examined, and mean regeneration percentages depending on these factors were recorded. Regeneration means of the synthetic seeds were compared using analysis of variance (ANOVA). Significant differences for regeneration of the synthetic seeds between hormone, storage temperature, storage period, and interaction of them were examined using the least significant difference (LSD) test. The percentage data were arcsine-transformed before analyses and JMP® program (SAS Institute, Cary, NC) ver. 8.00 was used for all the analyses.

Machine learning methods

In predicting the regeneration of synthetic seeds and assessing the importance of features affecting this process, several machine learning models were employed. All coding and model implementation were conducted in Python, utilizing the Scikit-learn (Sklearn) library. The regeneration data, originally recorded as percentage ratios (e.g., 0%, 20%, 40%), were transformed into binary values: a regeneration percentage of 0% was coded as 0, while all other percentages were coded as 1. This transformation enabled the use of classification models to distinguish between non-germinated and germinated seeds. To mitigate the impact of class imbalance, Stratified K-Fold Cross-Validation was applied, ensuring that class distributions were maintained across training and testing sets. Hyperparameter tuning was performed using the grid search method to optimize model performance.

The following machine learning models were employed in this study: Decision Tree (DT), Gaussian Process (GP), Multi-Layer Perceptron (MLP), Random Forest (RF), and Extreme Gradient Boosting (XGBoost). These models were selected for their diverse methodologies, allowing for a comprehensive evaluation of predictive accuracy and efficiency.

Each model was evaluated against a set of key performance metrics, including accuracy, F1 score, precision, and recall (Alkhadar et al. 2021). These metrics provide deeper insights into the model's predictive capability beyond overall accuracy and ensure a balance between precision and recall. The metrics were derived from the confusion matrices generated for each model (Okawa et al. 2022).

Accuracy: This metric represents the proportion of true results, both true positives (TP) and true negatives (TN), in relation to the total number of cases tested. While accuracy offers an initial assessment of model performance, it is not always sufficient, particularly in cases of imbalanced data. The metrics of TP, TN, false positives (FP), and false negatives (FN) are essential for a thorough evaluation.

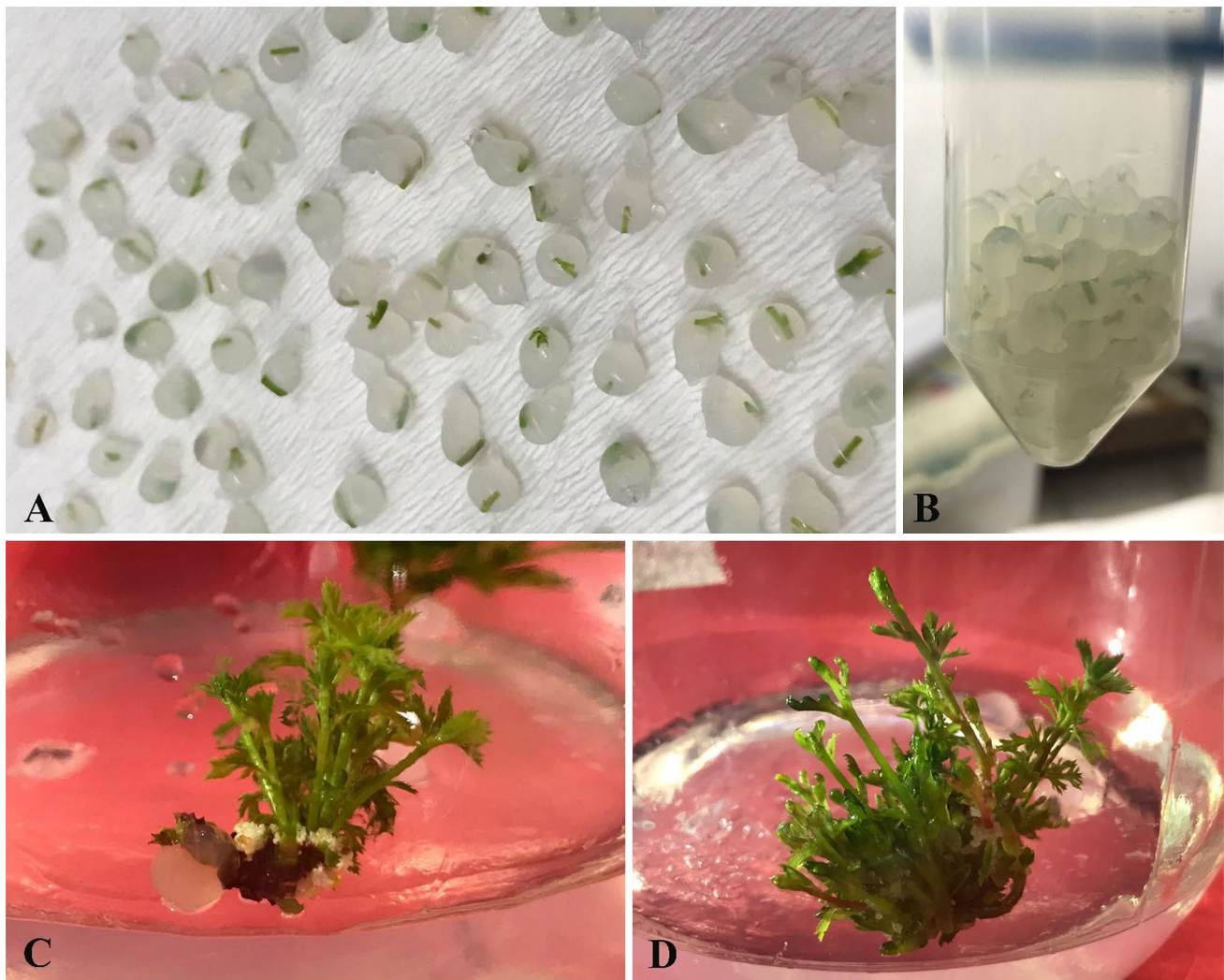


Fig. 1 **A** Encapsulated propagules, **B** Storage of synthetic seeds in a sterile centrifuge tube, **C** The shoot development of synthetic seeds stored at 4 °C for 1 month and regenerated in HF-MS medium, **D** The

shoot development of synthetic seeds stored at 24 °C for 2 months and regenerated in HF-MS medium

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

F1 Score: The F1 score is a statistical measure that provides a balanced assessment of a model's performance, considering both precision and recall. It is particularly valuable when there is a need to weigh the importance of correctly identifying positive instances against avoiding false positives.

$$\text{F1 Score} = 2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall})$$

Precision: The ratio of correctly predicted positive observations to the total predicted positives, reflecting the model's ability to avoid false positives.

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

Recall: The ratio of correctly predicted positive observations to all observations in the actual class, indicating the model's capacity to capture true positives.

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

By employing these evaluation metrics, a more detailed and reliable assessment of model performance was achieved, ensuring that predictions were not solely dependent on overall accuracy but also considered the balance between false positives and false negatives.

Decision tree (DT)

DTs are supervised learning methods applied to both classification and regression problems where a tree structure is modeled: the nodes represent an attribute, the branches

Table 1 Medium of encapsulation matrix and growth culture, storage temperature, and storage period

Medium (Encapsulation matrix and growth culture)	Storage temperature (°C)	Storage period
HF-MS	-20	1 month
		2 months
		3 months
	4	1 month
		2 months
		3 months
	24	1 month
		2 months
		3 months
MS+IAA	-20	1 month
		2 months
		3 months
	4	1 month
		2 months
		3 months
	24	1 month
		2 months
		3 months

HF-MS: Hormon-free Murashige and Skoog medium, MS+IAA: Murashige and Skoog medium supplemented with 2 mg/L IAA

represent the value, and the leaves represent a class (Amir et al. 2018; Dima et al. 2018). The DT algorithm was implemented using the Classification and Regression Tree (CART) approach (Breiman et al. 1984), which recursively partitions the data by minimizing the Gini impurity:

$$G = \sum p(i) \times (1 - p(i)) \quad (1)$$

where $p(i)$ represents the proportion of instances belonging to class i at a given node. Lower Gini values indicate purer nodes, improving classification accuracy (Biau and Scornet 2016).

Gaussian process (GP)

GPs are non-parametric, probabilistic models used for regression and classification tasks. They work by placing a distribution over functions and returning the mean and covariance of the underlying distribution as a prediction (Rasmussen 2003). Predictions are drawn from a Gaussian distribution:

$$f(x) \sim GP(m(x), k(x, x')) \quad (2)$$

where $m(x)$ is the mean function and $k(x, x')$ is the kernel function, which determines the similarity between data points. The Radial Basis Function (RBF) and Matérn kernels were tested for optimal performance (De and Rasmussen 2008; Rasmussen 2003).

Multi-layer perceptron (MLP)

Multi-Layer Perceptron (MLP) is an artificial neural network model that consists of multiple layers: an input layer, one or more hidden layers, and an output layer. The network learns by adjusting its internal parameters to minimize prediction errors, a process known as backpropagation (Rumelhart et al. 1986). Each layer consists of neurons interconnected through weighted edges, where the output of each neuron is computed as:

$$y = \sigma(Wx + b) \quad (3)$$

where W represents the weights, x the input, b the bias, and $\sigma(\cdot)$ the activation function.

Random forest (RF)

Random Forest (RF) is an ensemble learning method that aggregates multiple decision trees to improve predictive accuracy and reduce overfitting (Breiman 2001). Each tree is trained on a different subset of the data, and predictions are made independently. The final output is determined by majority voting in classification tasks or averaging in regression tasks:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^B T_b(x) \quad (4)$$

where B is the number of trees and $T_b(x)$ represents the prediction of each tree (Breiman 2001; Desalvo and Mohri 2016).

Extreme gradient boosting (XGBoost)

XGBoost is an advanced ensemble learning method that tries to improve accuracy and efficiency further. It builds a sequence of decision trees, where each tree attempts to correct the errors made by the previous ones, following the principles of gradient boosting (Shetty et al. 2023). The objective function of XGBoost incorporates both loss minimization and regularization, ensuring a balance between predictive accuracy and model complexity:

$$L = \sum l(y_i, \hat{y}_i) + \lambda \|w\|^2 \quad (5)$$

where $l(y_i, \hat{y}_i)$ represents the loss function, and $\lambda \|w\|^2$ is a regularization term that controls model complexity to reduce overfitting (Patil et al. 2023; Ser and Bati 2023).

Feature importances

Feature importance analysis is crucial for understanding the contribution of various factors in predicting regeneration outcomes. In this study, feature importance for models such as Decision Tree, Random Forest, and XGBoost was determined using their built-in mechanisms. These models inherently provide interpretative measures, including Gini importance and gain-based importance, which offer insights into the relative significance of each feature.

For models without built-in interpretative mechanisms, specifically Gaussian Process (GP) and Multi-Layer Perceptron (MLP), SHAP (SHapley Additive exPlanations) analysis was employed. SHAP assigns importance values to features based on their contributions to model predictions, providing a consistent and detailed understanding of feature impact. This method, as proposed by Lundberg et al. (2017), enabled the evaluation of feature importance for GP and MLP, offering a robust framework to interpret these models.

Surface and contour plot analysis

Surface and contour plot analysis was conducted to visualize the relationship between storage conditions (temperature, period) and regeneration percentage, in addition to machine learning analysis. The analysis was conducted using the statsmodels library to fit a regression model with the ordinary least squares (OLS) method. The predicted regeneration percentage were then used to construct surface and contour plots with matplotlib, allowing for the identification of trends and interactions among the studied factors. The optimal conditions for synthetic seed regeneration were better explained through surface plots. These plots were developed using a combination of machine learning predictions and experimental data to identify trends and interactions among different factors (Budelsky and Galatowitsch 1999; Harimoto et al. 2022; Imanishi and Imanishi 2014).

Results

Synthetic seeds regenerated within 2 to 3 weeks following the storage the process. Figure 1A, B illustrates the regeneration of synthetic seeds. The synthetic seeds were cultured after storage in two different medium HF-MS and MS+IAA. Although the regeneration in the HF-MS medium was observed with a percentage of 20.44%, no regeneration occurred in the MS+IAA. The mean value of synthetic seeds according to storage temperature was significantly different. The synthetic seeds cultured after storage at 4 °C showed the best result with 20.33%. Storage of synthetic seeds at 24 °C decreased the regeneration performance

Table 2 Synthetic seed regeneration percentages of hormone × storage temperature interaction

	HF-MS	MS+IAA	Mean
−20 °C	0 (0)c	0 (0)c	0 (0)C
4 °C	40.66 (36.4)a	0 (0)c	20.33 (18.2)A
24 °C	20.66 (21.2)b	0 (0)c	10.33 (10.6)B
Mean	20.44 (19.2)A	0 (0)B	

LSD_{hormone × storage temperature} = 5.7, LSD_{hormone} = 3.3, LSD_{storage temperature} = 4, ($p < 0.05$). All values indicated in parentheses were arcsine transformed of percentage values. HF-MS: Hormon-free Murashige and Skoog medium, MS+IAA: Murashige and Skoog medium supplemented with 2 mg/L IAA. Different lowercase letters indicate significant differences between hormone × storage temperature interactions, different uppercase letters within a column indicate significant differences between storage temperatures, and different uppercase letters within a row indicate significant differences between hormone compositions ($p < 0.05$)

Table 3 Synthetic seed regeneration percentages of hormone × storage period interaction

	HF-MS	MS+IAA	Mean
1 month	26.66 (24.4)a	0 (0)c	13.33 (12.2)A
2 months	17.33 (16.9)b	0 (0)c	8.66 (8.4)A
3 months	17.33 (16.4)b	0 (0)c	8.66 (8.2)A
Mean	20.44 (19.2)A	0 (0)B	

LSD_{hormone × storage period} = 5.7, LSD_{hormone} = 3.3, LSD_{storage period} = 4, ($p < 0.05$). All values indicated in parentheses were arcsine transformed of percentage values. HF-MS: Hormon-free Murashige and Skoog medium, MS+IAA: Murashige and Skoog medium supplemented with 2 mg/L IAA. Different lowercase letters indicate significant differences between hormone × storage period interactions, different uppercase letters within a column indicate significant differences between storage periods, and different uppercase letters within a row indicate significant differences between hormone compositions ($p < 0.05$)

(10.33%). Storage at −20 °C regardless of medium ingredient, inhibited the regeneration of synthetic seeds (Table 2). Interactions of hormone × storage temperature exhibited statistically important differences. The synthetic seeds stored at 4 °C and cultured HF-MS medium showed the best regeneration performance (40.66%). The culture in HF-MS after storage at 24 °C gave the second-best regeneration performance with the percentage of 20.66%. The regeneration was not observed in any of the other hormone × storage temperature interactions (Table 2).

The synthetic seeds were stored at different temperatures for 1, 2, and 3 months to test short-term conservation ability. It was detected that there were no significant differences between storage periods for the regeneration of synthetic seeds (Table 3). The best regeneration percentage (26.66%) for hormone-storage period interaction was obtained from the synthetic seeds stored for 1 month and cultured in HF-MS medium. The synthetic seeds were stored for 2 and 3 months and cultured in HF-MS medium, followed by a percentage of 17.33% (Table 3).

The mean regeneration percentage of storage temperature-storage period interaction were significantly different. Storage at 4 °C for 1 month exhibited the best results with 25% regeneration percentage. None of the synthetic seeds stored at -20 °C regenerated regardless of storage time. Storage at 24 °C for 3 months showed the lowest regeneration performance (4%) (Table 4).

The mean regeneration percentages of triple factors interactions were compared for examination of the simultaneous effect of hormone, storage temperature, and storage period. The differences between the mean regeneration values of the synthetic seeds were statistically significant. The synthetic seeds cultured in MS+IAA or stored at -20 °C, regardless of other factors, failed to regenerate. Culture in HF-MS after storage at 4 °C for 1 month provided the best regeneration results with 50%. Increasing the storage time at 24 °C before cultivation in HF-MS medium decreased the regeneration performance of the synthetic seeds. Correspondingly, the synthetic seeds cultured in HF-MS after storage at 24 °C for 3 months showed the lowest regeneration percentage (Table 5).

The machine learning results

In this study, we evaluated several machine learning models to see how well they could predict whether seeds would regenerate. We focused on key performance metrics, including accuracy, F1 score, precision, recall, and feature importance, to provide a comprehensive evaluation of each model. The models we tested, Decision Tree (DT), Multi-Layer Perceptron (MLP), Random Forest (RF), and Extreme Gradient Boosting (XGBoost), all performed remarkably well, each achieving an accuracy rate of 97.3% in predicting regeneration outcomes. These models' F1 scores of 0.947 indicate that they performed exceptionally well, not only in terms of accuracy but also in terms of precision and recall balance (Table 6). This demonstrates that the models can

Table 4 Synthetic seed regeneration percentages of storage temperature × storage period interaction

	1 month	2 months	3 months	Mean
-20 °C	0 (0)e	0 (0)e	0 (0)e	0 (0)C
4 °C	25 (21.7)a	14 (13.6)bc	22 (19.3)ab	20.33 (18.2)A
24 °C	15 (14.9)abc	12 (11.7)cd	4 (5.3)de	10.33 (10.6)B
Mean	13.33 (12.2)A	8.66 (8.4)A	8.66 (8.2)A	

LSD_{storage temperature × storage period} = 7, LSD_{storage temperature} = 4, LSD_{storage period} = 4, ($p < 0.05$). All values indicated in parentheses were arcsine transformed of percentage values. Different lowercase letters indicate significant differences between storage temperature × storage period interactions, different uppercase letters within a column indicate significant differences between storage temperatures, and different uppercase letters within a row indicate significant differences between storage periods ($p < 0.05$)

effectively predict conditions under which seeds will successfully regenerate versus those under which they will not.

In contrast, the GP model exhibited lower accuracy at 75.7%, with an F1 score of 0.667, indicating weaker overall performance. The confusion matrix for the GP model shows a higher number of false positives, with 9 instances where non-regenerated was incorrectly predicted as regeneration. Despite its lower precision (0.500), the GP model achieved a perfect recall of 1.0, correctly identifying all regeneration cases. However, this model's lower accuracy and F1 score suggest it was less reliable in predicting non-regeneration outcomes compared to the other models. Interestingly, the GP and MLP models had the highest mean cross-validation scores (0.876), implying that they might generalize better across different data splits, potentially offering more robustness when applied to varied datasets (Table 6).

When comparing feature importance across models, some differences emerge in the factors driving regeneration predictions with a 47.7% contribution to the model's decisions. Storage temperature was the most influential feature in the DT model, closely followed by hormone with 47.1% and storage period made up only 5.2% of the total, which was less significant. The RF model exhibited a similar pattern, with hormone (53.9%) as the most significant feature,

Table 5 Synthetic seed regeneration percentages of hormone × storage temperature × storage period interaction

H	ST	SP	Regeneration (%)	H	ST	SP	Regeneration (%)
HF-MS	-20	1 m	0 (0)e	MS+IAA	-20	1 m	0 (0)e
HF-MS	-20	2 m	0 (0)e	MS+IAA	-20	2 m	0 (0)e
HF-MS	-20	3 m	0 (0)e	MS+IAA	-20	3 m	0 (0)e
HF-MS	4	1 m	50 (43.5)a	MS+IAA	4	1 m	0 (0)e
HF-MS	4	2 m	28 (27.2)c	MS+IAA	4	2 m	0 (0)e
HF-MS	4	3 m	44 (38.5)ab	MS+IAA	4	3 m	0 (0)e
HF-MS	24	1 m	30 (29.8)bc	MS+IAA	24	1 m	0 (0)e
HF-MS	24	2 m	24 (23.4)c	MS+IAA	24	2 m	0 (0)e
HF-MS	24	3 m	8 (10.6)d	MS+IAA	24	3 m	0 (0)e

LSD_{hormone × storage temperature × storage period} = 9.9, ($p < 0.05$), H: Hormone, ST: Storage temperature, SP: Storage period. All values indicated in parentheses were arcsine transformed of percentage values. HF-MS: Hormon-free Murashige and Skoog medium, MS+IAA: Murashige and Skoog medium supplemented with 2 mg/L IAA. Different letters within a column indicate significant differences between hormone × storage temperature × storage period interactions ($p < 0.05$)

Table 6 Confusion matrices and machine learning accuracy results

Confusion matrix		Y-predicted		Regeneration (%)				
Model	Y-true	0	1	Accuracy	F-1 score	Precision	Recall	Mean CV
DT	0	27	1	0.973	0.947	0.900	1	0.862
	1	0	9					
GP	0	19	9	0.757	0.667	0.500	1	0.876
	1	0	9					
MLP	0	27	1	0.973	0.947	0.900	1	0.876
	1	0	9					
RF	0	27	1	0.973	0.947	0.900	1	0.869
	1	0	9					
XGBoost	0	27	1	0.973	0.947	0.900	1	0.869
	1	0	9					

Table 7 Feature importances

Model	Features		
	Storage temperature	Hormone	Storage period
DT	0.477	0.471	0.052
GP (SHAP)	0.154	0.241	−0.002
MLP (SHAP)	0.201	0.284	0.059
RF	0.407	0.539	0.054
XGBoost	0.517	0.356	0.128

followed by storage temperature (40.7%). Using SHAP values for analysis, the MLP model revealed that hormone was the most significant feature (28.4%), followed by storage temperature (20.1%) and storage period (5.9%). In addition, the XGBoost model ranked storage temperature (51.7%) as the most important factor, ahead of hormone (35.5%) and storage period (12.8%). Meanwhile, the GP model, also using SHAP values, emphasized hormone (24.1%) and storage temperature (15.4%), with the storage period showing a slightly negative impact (−0.27%), suggesting it may not consistently influence predictions (Table 7).

Surface and contour plot analysis

The surface and contour plots effectively illustrated the relationship between storage conditions (temperature, period) and regeneration percentage of synthetic seeds in *C. monogyna* L., across two different media: HF-MS and MS+IAA. The plots revealed that storage at 4 °C consistently yielded the highest regeneration percentage, particularly in HF-MS. In contrast, the MS+IAA showed significantly lower regeneration percentage. This observation aligns with the

regression analysis results, where the medium coefficient of −9.485 ($p < 0.001$) indicated that regeneration percentage in MS+IAA were substantially lower compared to HF-MS (Table 8).

The regression model, which explained 45.4% of the variance in regeneration percentage ($R^2 = 0.454$), also demonstrated that temperature had a positive initial effect on regeneration percentage with a coefficient of 0.616 ($p < 0.001$). However, the quadratic term (Temperature^2) had a negative effect of −0.031 ($p < 0.001$), indicating that regeneration percentages decreased at higher temperatures. This trend was observed at 24 °C in both media, where regeneration percentages declined significantly (Table 8).

Moreover, the interaction effect between hormone and storage temperature had a negative coefficient of −0.517 ($p < 0.001$), suggesting that the negative impact of higher temperatures was more pronounced in the MS+IAA compared to HF-MS (Table 8). This interaction effect further emphasizes the challenges of using MS+IAA for optimal regeneration at higher temperatures. These visual insights, supported by the regression analysis, confirm that 4 °C is the most suitable temperature for storing *C. monogyna* L. synthetic seeds, with HF-MS performing better than MS+IAA (Fig. 2).

Discussion

The regeneration of propagules and the storage of artificial seeds are essential components of synthetic seed technology. For synthetic seed technology and its commercial use,

Table 8 Regression model summary

Variable	Coefficient	Standard error	t-Statistic	p-value	95% confidence interval
Intercept	28.879	2.318	12.456	<0.001	24.304 to 33.455
Hormone	−9.485	1.155	−8.211	<0.001	−11.765 to −7.206
Temperature	0.616	0.091	6.749	<0.001	0.436 to 0.796
I (Temperature^2)	−0.031	0.005	−6.023	<0.001	−0.041 to −0.021
Hormone × Temperature	−0.517	0.127	−4.072	<0.001	−0.767 to −0.266
I (Hormone^2)	−9.485	1.155	−8.211	<0.001	−11.765 to −7.206

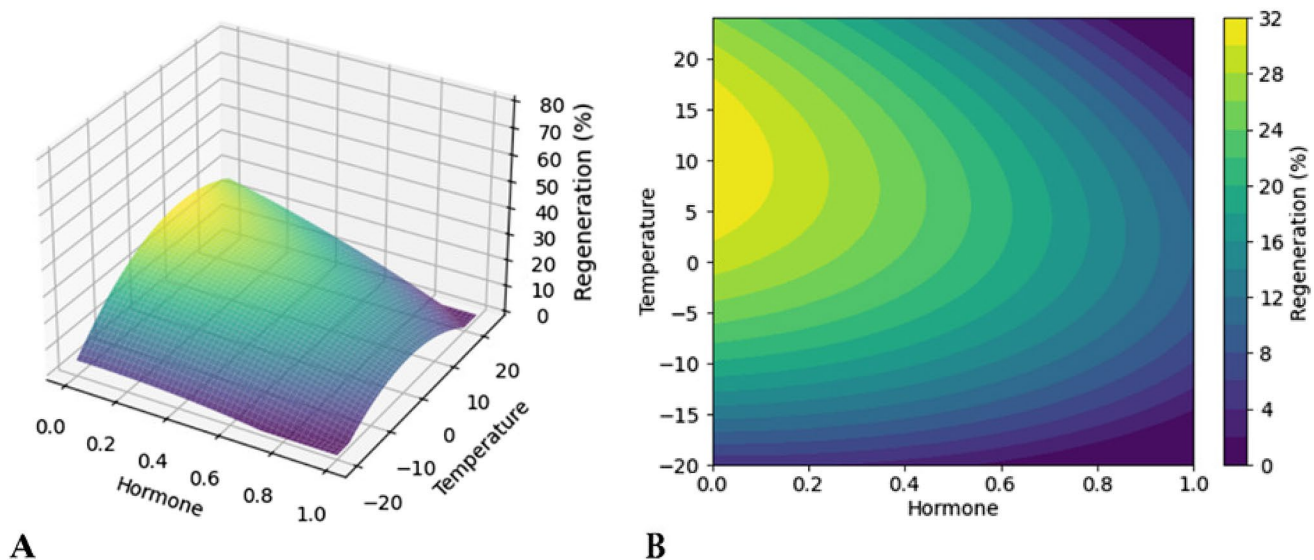


Fig. 2 A Surface and B Contour plot analysis

appropriate storage conditions, such as temperature and time of storage, that allow for high regeneration percentage, are crucial (Regni et al. 2024). Therefore, we have stored the synthetic seeds at -20°C , 4°C , or 24°C for one, two, or three months. To the best of our knowledge, the present study is the first report on assessing the effects of the storage period, storage temperature, and hormone for regeneration of encapsulated nodal segments of *C. monogyna*.

The storage of synthetic seeds at 4°C gave the best regeneration results with the percentage of 20.33% in our study. Our results confirm that synthetic seeds of *C. monogyna* can be effectively stored at 4°C for short-term periods. Several papers have reported similar outcomes. Regni et al. (2024) indicated that encapsulated clump's bases of blackberry stored at 4°C for 30, 60, or 120 days were better than those stored at 25°C . Rathore and Kheni (2017) observed 72% regeneration on the synthetic seeds from the lateral meristems of *Withania coagulans* after 60 days of storage at 4°C . Bertsouklis and Tsopele (2023) showed that the synthetic seeds of *Cerastium candidissimum* Corr., an endangered endemic Greek species, could be stored at 4°C for 2 months without losing viability. Trejo-Tapia et al. (2019) stated that at low temperatures, the metabolic activities in synthetic seeds decrease, allowing them to be stored for different durations. In contrast, Shilpha et al. (2021) stated that synthetic seeds of *Solanum trilobatum* incubated at 25°C exhibited a better regrowth percentage than compare to those stored at 8°C and 4°C . Bhattacharyya et al. (2018) stored encapsulated protocorm-like bodies of *Ansellia africana* at 4, 8, and 25°C for 15, 30, 45, 60, 75, and 90 days. They informed that, although the synthetic seeds stored at 25°C converted to plantlets in all the storage periods, the synthetic seeds stored at 8°C converted only in the 75- and

90-day periods, and no conversion to plantlets was observed at 4°C . These reports indicate that the variation in optimal temperature for the synthetic seed storage is strongly associated with the plant species.

Another factor influenced by the plant species in the short-term preservation of synthetic seeds is the storage duration. One of the requirements for synthetic seed technology and its commercial use is clear storage periods that keep seeds alive during the exchange of germplasm (Regni et al. 2024). In this study, synthetic seeds were successfully stored for up to 3 months. Different storage durations depending on different types and storage temperatures have been previously reported by Behera et al. (2021) for 30 days in *Paederia foetida*, Regni et al. (2024) for 120 days in blackberry, Ikhlaiq et al. (2010) for 60 days in olive, Javed et al. (2017) for 8 weeks in *Erythrina variegata*, and Bhattacharyya et al. (2018) for 90 days in *Ansellia africana*.

We observed that increasing of storage duration typically resulted in decreased regeneration percentages. This decrease was particularly observed in synthetic seeds stored at 24°C . This outcome was reported by Bhattacharyya et al. (2018) for *Ansellia africana*, Nugrahani et al. (2018) for *Chrysanthemum*, and Javed et al. (2017) for *Erythrina variegata*.

The application of machine learning (ML) methods to predict regeneration outcomes based on features such as storage temperature, hormone composition, and storage period represents a novel approach in this field. ML algorithms, particularly classification-based models, are becoming increasingly common in ecological contexts, especially for endangered species, where maximizing regeneration conditions is crucial for conservation efforts (Olden et al. 2008). Unlike regression models, classification approaches

allow for more granular predictions of regeneration success. The use of confusion matrices in this context provides a comprehensive framework for evaluating model performance, as it enables the calculation of key metrics such as accuracy, precision, recall, and the F1 score, which collectively offer deep insights into the model's predictive capabilities (Chicco and Jurman 2020; Pomme et al. 2022). Cross-validation serves as a crucial method for estimating the accuracy of the employed models, providing a systematic way to assess their performance on unseen data. This technique helps in selecting the best-performing models by minimizing the risk of overfitting and ensuring that the models generalize well to new data (Pomme et al. 2022).

DT, RF, MLP, and XGBoost achieved high accuracy rates, demonstrating superior prediction accuracy and F1 scores compared to the GP. These models enhance our understanding of synthetic seed regeneration processes by providing more reliable predictions, which can be especially valuable for optimizing regeneration conditions in conservation efforts. RF and XGBoost, as ensemble methods, reduce overfitting by aggregating multiple decision trees, contributing to their robustness (Breiman 2001; Chen and Guestrin 2016). The trees in RF are trained on different subsets of data, and this method's ability to consider random subsets of features at each split further strengthens its performance and generalizability (Breiman 2001; Désir et al. 2013). Tuning parameters such as the number of trees and maximum depth can optimize RF's performance (Liu et al. 2015). XGBoost, on the other hand, builds a sequence of decision trees that correct the errors of previous ones, making it particularly effective for handling complex datasets. Its support for parallel processing and regularization techniques further enhances its utility in large-scale applications (Jiang et al. 2019; Wu et al. 2022). Moreover, XGBoost's strong outlier management and prevention of overfitting have cemented its position as a leading tool in machine learning (Patil et al. 2023).

The strong performance of MLP highlights neural networks' ability to learn complex, non-linear relationships. MLPs are particularly effective for handling sparse or high-dimensional data, making them a reliable choice in various applications (Dara et al. 2022). While simpler, the DT also achieved comparable accuracy, likely due to the structured dataset (Loh 2011; Quinlan 1986). DT are widely recognized for their simplicity and interpretability, making them a popular choice in classification problems (Aaboub et al. 2023; Pratama et al. 2024). In contrast, the GP model underperformed, possibly due to limitations in the chosen kernel functions despite hyperparameter optimization. Kernel functions such as the Radial Basis Function (RBF) or Matérn kernel play a critical role in GP models by defining the similarity between data points, and their selection significantly

impacts model performance (Rasmussen 2003). However, GP models excel in quantifying uncertainty, making them particularly useful in small dataset binary classification tasks where careful uncertainty estimation is required (De and Rasmussen 2008; Rasmussen 2003).

The success of ML models in predicting regeneration outcomes is further supported by various studies. For instance, numerous studies strongly support the success of ML models in predicting germination outcomes. Okumuş et al. (2024) demonstrated the robustness of ML algorithms in analyzing forage pea germination under temperature and salt stress, revealing how these models can effectively capture stress impacts on seed viability. Similarly, Benlioğlu et al. (2024) applied ML to assess drought tolerance in tetraploid wheat, highlighting elastic net models as particularly effective in predicting germination speed and water absorption. Expanding on these capabilities, Genze et al. (2020) incorporated temporal data through convolutional LSTM networks, enhancing predictive accuracy for grain crop germination. Machine learning also holds promise for conserving endangered species, such as *Crataegus* spp. (hawthorn), where traditional propagation techniques often yield insufficient outcomes (Şimşek et al. 2024). By leveraging ML techniques, researchers can better predict and optimize germination outcomes, contributing to the preservation of such species.

The feature importance analysis revealed that temperature was the most significant factor influencing regeneration percentage, followed closely by the hormone. This finding is consistent with existing literature highlighting environmental conditions' critical role on seed germination (Javaid et al. 2022). Feature importance analysis indicated that storage temperature, particularly cold storage at 4 °C, was the most influential variable, in line with prior research showing that lower temperatures enhance seed viability (Baskin and Baskin 2014). As revealed through SHAP, the negative impact of high concentrations of IAA (2 mg/l) on germination supported existing findings on the inhibitory effects of auxins on seed germination (Shuai et al. 2017; Ye et al. 2016).

The surface and contour plot analysis visually confirmed the non-linear relationships between storage temperature, hormone, and regeneration percentages. The surface plot showed that regeneration decreases at higher temperatures, particularly in the MS+IAA, consistent with prior studies on temperature's role in seed dormancy and germination (Bewley et al. 2013; Fenner and Thompson 2005). The regression model supports this, highlighting a significant negative interaction between hormone and storage temperature (Coefficient = -5.168, $p < 0.001$), suggesting that IAA intensifies the negative temperature effect, as previously noted by Chen and Guestrin (2016). The contour plot further

illustrated these trends, validating the feature importance outcomes from the machine learning models (Lundberg et al. 2017). Additionally, the negative quadratic term for storage temperature ($I(\text{Temperature}^2)$) aligned with findings by Baskin and Baskin (2014), indicating a peak germination point at optimal lower temperatures. These visual analyses complement the machine learning interpretations, showing how contour plots can bridge complex interactions (Lundberg et al. 2017).

Conclusion

In conclusion, this study highlights the significant potential of synthetic seed technology for *C. monogyna* L. in enhancing regeneration percentage and supporting genetic conservation efforts. We have proposed a novel methodology to optimize propagation strategies by leveraging sodium alginate for encapsulation and employing machine learning algorithms to predict regeneration outcomes based on crucial factors such as storage temperature, storage duration, and hormone concentration. These findings underscore the importance of environmental conditions in synthetic seed viability and the effectiveness of ML methods in plant sciences. Ultimately, this research contributes to the broader goal of preserving genetic diversity and improving the sustainability of plant propagation practices in the face of environmental challenges.

Author contributions All authors contributed to the study conception and design. Material preparation and experimental labour were performed by Metin KOÇAK, Cansu KUZĞUN, Sibel TURAN SİRKE and Mehtap YILDIZ. Data processing and analysis were performed by Metin KOÇAK and Mustafa Can Yılmaz. The first draft of the manuscript was written by Metin KOÇAK, Mustafa Can Yılmaz, Cansu KUZĞUN and Sibel TURAN SİRKE, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK). The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability The data that support the findings of this study are available on request from the corresponding author.

Declarations

Competing interests The authors have no relevant financial or non-financial interests to disclose.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate

if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Aaboub F, Chamlal H, Ouaderhman T (2023) Statistical analysis of various splitting criteria for decision trees *. JACT 17. <https://doi.org/10.1177/17483026231198181>
- Alkhadar H, Macluskey M, White S, Ellis I, Gardner A (2021) Comparison of machine learning algorithms for the prediction of five-year survival in oral squamous cell carcinoma. J Oral Pathol 50(4):378–384. <https://doi.org/10.1111/jop.13135>
- Amir WM, Ahmad W, Um N, Allah M, Shafiq M, Ibrahim M, Mamat M, Aleng A, Husein A, Arif M, Nawi A, Yaqoob MA (2018) Using decision tree and response surface methodology to review chronic periodontitis profile in Pakistan: a preliminary case study. IJET 7 (3.28) (2018) 76–79. www.sciencepubco.com/index.php/IJET
- Bahorun T, Aumjaud E, Ramphul H, Rycha M, Luximon-Ramma A, Trotin F, Aruoma O (2003) Phenolic constituents and antioxidant capacities of *Crataegus monogyna* (hawthorn) callus extracts. Nahrung/food 47(3):191–198. <https://doi.org/10.1002/food.200390045>
- Baskin CC, Baskin JM (2014) Seeds: ecology, biogeography, and evolution of dormancy and germination. In seeds: ecology, biogeography, and evolution of dormancy and germination. <https://doi.org/10.1016/C2013-0-00597-X>
- Behera B, Behera S, Shasmita S, Mohapatra D, Barik D P and Naik S K (2021) Regeneration of plants from alginate-encapsulated axenic nodal segments of *Paederia foetida* L.-A medicinally important and vulnerable plant species. J Plant Biotechnol 48(4):255–263
- Benlioglu B, Demirel F, Türkoğlu A, Haliloğlu K, Özaktan H, Kujawa S, Piekutowska M, Wojciechowski T, Niedbala G (2024) Insights into drought tolerance of tetraploid wheat genotypes in the germination stage using machine learning algorithms. Agriculture (Switzerland) 14(2). <https://doi.org/10.3390/agriculture14020206>
- Bernatoniene J, Trumbeckaitė S, Majienė D, Baniene R, Baliutytė G, Savickas A, Toleikis A (2009) The effect of *Crataegus* fruit extract and some of its flavonoids on mitochondrial oxidative phosphorylation in the heart. Phytother Res 23:1701–1707. <https://doi.org/10.1002/ptr.2815>
- Bertsouklis K, Tsopela S (2023) In vitro Propagation of three populations of the endangered, Greek endemic *Cerastium candidissimum* and short-term storability of Alginate-Encapsulated shoot explants for exploitation and conservation. Horticulturae 9:273. <https://doi.org/10.3390/horticulturae9020273>
- Bewley JD, Bradford KJ, Hilhorst HWM, Nonogaki H (2013) Seeds: physiology of development, germination and dormancy, 3rd edition. In seeds: physiology of development, germination and dormancy, 3rd Edition (Vol. 9781461446934). <https://doi.org/10.1007/978-1-4614-4693-4>
- Bhattacharyya P, Kumar V, Van Staden J (2018) In vitro encapsulation based short term storage and assessment of genetic homogeneity in regenerated *Ansellia africana* (Leopard orchid) using gene targeted molecular markers. PCTOC 133:299–310
- Biau G, Scornet E (2016) A random forest guided tour. TEST 25(2):197–227. <https://doi.org/10.1007/s11749-016-0481-7>
- Breiman L (2001) Random forests. machine learning, Springer 45:5–32

- Breiman L, Ihaka R (1984) Nonlinear discriminant analysis via scaling and ACE. Department of Statistics, University of California, Davis One Shields Avenue Davis, CA, USA
- Budelsky RA, Galatowitsch SM (1999) Effects of moisture, temperature, and time on seed germination of five wetland Carices: Implications for restoration. *Restor Ecol* 7(1):86–97. <https://doi.org/10.1046/j.1526-100X.1999.07110.x>
- Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system. Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining, 13–17-August-2016. <http://doi.org/10.1145/2939672.2939785>
- Chicco D, Jurman G (2020) The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21(1). <https://doi.org/10.1186/s12864-019-6413-7>
- Christensen K, Zarrei M, Kuzmina M, Talent N, Lin C, Dickinson T (2014) *Crataegus ×ninae-celottiae* and *C. ×cogswellii* (Rosaceae, Maleae), two spontaneously formed intersectional nothospecies. *Phytokeys* 36:1–26. <https://doi.org/10.3897/phytokeys.36.6784>
- Dara S, Dhamercherla S, Jadav SS, Babu CM, Ahsan MJ (2022) Machine Learning in Drug Discovery: A Review. *Artif Intell Rev* 55(3):1947–1999. <https://doi.org/10.1007/s10462-021-10058-4>
- Daud N, Taha RM, Hasbullah NA (2008) Artificial seed production from encapsulated micro shoots of *Saintpaulia ionantha* Wendl. (African violet). *J Appl Sci* 8(24) 4662–4667
- De HM, Rasmussen CE (2008) Approximations for binary gaussian process classification hannes nickisch. in journal of machine learning research (Vol. 9).
- DeSalvo G, Mohri M (2016) Random composite forests. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 30, No. 1)
- Désir C, Bernard S, Petitjean C, Heutte L (2013) One class random forests. *Pattern Recognit* 46(12):3490–3506. <https://doi.org/10.1016/j.patcog.2013.05.022>
- Dima S, Wang KJ, Chen KH, Huang YK, Chang WJ, Lee SY, Teng NC (2018) Decision tree approach to the impact of parents' oral health on dental caries experience in children: a cross-sectional study. *Int J Environ Res Public Health* 15(4). <https://doi.org/10.3390/ijerph15040692>
- Dinga R, Schmaal L, Penninx BWJH, Veltman DJ, Marquand AF (2020) Controlling for effects of confounding variables on machine learning predictions. *BioRxiv*. <https://doi.org/10.1101/2020.08.17.255034>
- Dirk A, Van Dijk J, Kootstra G, Kruijer W, De Ridder D (2021) Machine learning in plant science and plant breeding. *ISCIENCE* 24:101890. <https://doi.org/10.1016/j.isci.2021.101890>
- Dönmez AA (2007) Taxonomic notes on the genus *Crataegus* (Rosaceae) in Turkey. *Bot J Linn* 155(2):231–240
- Fenner M, Thompson K (2005) The ecology of seeds. In the ecology of Seeds. Cambridge University Press. <https://doi.org/10.1017/CBO9780511614101>
- Fichtner A, Wissemann V (2021) Biological flora of the british isles: *Crataegus monogyna*. *J Ecol* 109(1):541–571. <https://doi.org/10.1111/1365-2745.13554>
- Folorunso O, Ojo O, Busari M, Adebayo M, Joshua A, Folorunso D, Ugwunna CO, Olabanjo O, Olabanjo O (2023) Exploring machine learning models for soil nutrient properties prediction: a systematic review. *Big Data Cogn Comput* 7:113. <https://doi.org/10.3390/bdcc7020113>
- Galatali S, Balci MA, Akguller O, Kaya E (2021) Production of disease-free olive seedlings with artificial intelligence and biotechnological methods. *Eur J Biol* 2(3) 79–84. <https://doi.org/10.24018/ejbio.2021.2.3.172>
- Genze N, Bharti R, Grieb M, Schultheiss SJ, Grimm DG (2020) Accurate machine learning-based germination detection, prediction and quality assessment of three grain crops. *Plant Methods* 16(1). <https://doi.org/10.1186/s13007-020-00699-x>
- Ghosh B, Haque SM (2019) Synthetic seeds: An alternative approach for clonal propagation to avoiding the heterozygosity problem of natural botanical seeds. In: Faisal M, Alatar AA (eds) *Synthetic Seeds*. Springer, Switzerland, pp 77–112
- Harimoto S, Kataoka H, Kubo M (2022) Seed germinability in the endangered *Pulsatilla cernua* (Ranunculaceae) following cold storage for up to 31 years. *Plant Species Biol* 37(5):308–314. <https://doi.org/10.1111/1442-1984.12382>
- Hesami M, Alizadeh M, Jones AMP, Torkamaneh D (2022) Machine learning: its challenges and opportunities in plant system biology. *AMBB* 106:3507–3530. <https://doi.org/10.1007/s00253-022-11963-6>
- Ikhlaq M, Hafiz I A, Micheli M, Ahmad T, Abbasi N A and Standardi A (2010) In vitro storage of synthetic seeds: effect of different storage conditions and intervals on their conversion ability. *AJB* 9(35)
- Imanishi A, Imanishi J (2014) Seed dormancy and germination traits of an endangered aquatic plant species, *Euryale ferox* Salisb. (Nymphaeaceae). *Aquat Bot* 119:80–83. <https://doi.org/10.1016/j.aquabot.2014.08.001>
- Jang BK, Cho JS, Lee CH (2020) Synthetic seed technology development and production studies for storage, transport, and industrialization of bracken spores. *Plants* 9(9):1079
- Javadi MM, Mahmood A, Alshaya DS, AlKahtani MDF, Waheed H, Wasaya A, Khan SA, Naqve M, Haider I, Shahid MA, Nadeem MA, Azmat S, Khan BA, Balal RM, Attia KA, Fiaz S (2022) Influence of environmental factors on seed germination and seedling characteristics of perennial ryegrass (*Lolium perenne* L.). *Sci Rep* 12(1). <https://doi.org/10.1038/s41598-022-13416-6>
- Javed SB, Alatar AA, Anis M, Faisal M (2017) Synthetic seeds production and germination studies, for short term storage and long distance transport of *Erythrina variegata* L.: a multipurpose tree legume. *Ind Crops Prod* 105:41–46
- Jiang Y, Tong G, Yin H, Xiong N (2019) A pedestrian detection method based on genetic algorithm for optimize XGBoost training parameters. *IEEE Access* 7:118310–118321. <https://doi.org/10.1109/ACCESS.2019.2936454>
- Kuhn T, Jancso B, Ruprecht E (2021) Hawthorn (*Crataegus* l.) taxa and their hybrids in north-western romania: a recommendation for national identification keys based on morphometric analyses. *Contribuții Botanice* 55 7–26. <https://doi.org/10.24193/contrib.bot.55.1>
- Lediuk K, Damascos M, Puntieri J, Svriz M (2014) Differences in phenology and fruit characteristic between invasive and native woody species favor exotic species invasiveness. *Plant Ecol* 215(12):1455–1467. <https://doi.org/10.1007/s11258-014-0402-3>
- Liu C, Chan Y, Alam Kazmi SH, Fu H (2015) Financial fraud detection model: based on random forest. *IJEF* 7(7). <https://doi.org/10.5539/ijef.v7n7p178>
- Loh WY (2011) Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1). <https://doi.org/10.1002/widm.8>
- Lundberg SM, Allen PG, Lee SI (2017) A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst* 30. <https://github.com/slundberg/shap>
- Martinelli F, Perrone A, Yousefi S, Papini A, Castiglione S, Guarino F, Salami S (2021) Botanical, phytochemical, anti-microbial and pharmaceutical characteristics of hawthorn (*Crataegus monogyna* jacq.), *Rosaceae*. *Mol* 26(23) 7266. <https://doi.org/10.3390/molecules26237266>
- Murashige T, Skoog F (1962) A revised medium for rapid growth and bio assays with tobacco tissue cultures. *Physiol Plant* 15:473–497
- Niazian M, Niedbala G (2020) Machine learning for plant breeding and biotechnology. *Machine learning for plant breeding and*

- biotechnology. Agriculture 10(10) 436. <https://doi.org/10.3390/agriculture10100436>
- Nugrahani P, Moeljani IR, Lydiana I (2018) Encapsulation and germination of synthetic seeds of Chrysanthemum. In international conference on science and technology (ICST 2018) (pp. 126–129). Atlantis Press
- Okawa T, Mizuno T, Hanabusa S, Ikeda T, Mizokami F, Koseki T, Takahashi K, Yuzawa Y, Tsuboi N, Yamada S, Kameya Y (2022) Prediction model of acute kidney injury induced by cisplatin in older adults using a machine learning algorithm. PLoS ONE, 17(1). <https://doi.org/10.1371/journal.pone.0262021>
- Okumuş O, Say A, Eren B, Demirel F, Uzun S, Yaman M, Aydın A (2024) Using machine learning algorithms to investigate the impact of temperature treatment and salt stress on four forage peas (*Pisum sativum* var. arvense L.). Horticulturae 10(6). <https://doi.org/10.3390/horticulturae10060656>
- Olden JD, Lawler JJ, Poff NL (2008) Machine learning methods without tears: a primer for ecologists. Q Rev Biol 83(2):171–193. <https://doi.org/10.1086/587826>
- Özcan M, Haciseferoğlu H, Marakoğlu T, Arslan D (2005) Hawthorn (*Crataegus* spp.) fruit: some physical and chemical properties. J Food Eng 69:409–413
- Patil P, Kale G, Bivalkar N, Kolhatkar A (2023) Comparative analysis of weighted ensemble and majority voting algorithms for intrusion detection in openstack cloud environments. IJACSA 14:12
- Pomme LE, Bourqui R, Giot R, Auber D (2022) Relative confusion matrix: Efficient comparison of decision models. Proceedings of the international conference on information visualisation, 2022–July. <https://doi.org/10.1109/IV56949.2022.00025>
- Pratama MR, Suryana AL, Alfiansyah G, Olivia Z, Nurmawati I, Desartianto P (2024) Diagnosis of stroke and diabetes mellitus with classification techniques using decision tree method. IJHISI 2(1) 1–8. <https://doi.org/10.47134/ijhis.v2i1.36>
- Quinlan JR (1986) Induction of decision trees" Machine Learning. Springer 1:81–106
- Rasmussen CE (2003) Gaussian Processes in Machine Learning. In LNAI (3176). Springer-Verlag. <http://www.tuebingen.mpg.de/~carl>
- Rathore MS, Khani J (2017) Alginate encapsulation and in vitro plantlet regeneration in critically endangered medicinal plant, *Withania coagulans* (Stocks) Dunal. Proc Natl Acad Sci, India Sect B Biol Sci 87(1):129–134
- Regni L, Micheli M, Pino A M D, Facchin SL, Rabica E, Camilloni L, Cesarini A and Proietti P (2024) Blackberry synthetic seeds storage: effects of temperature, time, and sowing substrate. PCTOC 158(1) 17
- Rumelhart D, Hinton G and Williams R (1986) Learning representations by back-propagating errors. Nature 323:533–536. <https://doi.org/10.1038/323533a0>
- Ser G, Bati CT (2023) Modelling overdispersed seed germination data: Xgboost's performance. JAPS 33(4)
- Shetty S, Umesh P, Shetty A (2023) The effectiveness of machine learning-based multi-model ensemble predictions of CMIP6 in Western Ghats of India. Int J Climatol 43(11):5029–5054. <https://doi.org/10.1002/joc.8131>
- Shilpha J, Pandian S, Larga MJV, Sohn SI, Ramesh M (2021) Short-term storage of *Solanum trilobatum* L. synthetic seeds and evaluation of genetic homogeneity using SCoT markers. Plant Biotechnol Rep 15:651–661
- Shuai H, Meng Y, Luo X, Chen F, Zhou W, Dai Y, Qi Y, Du J, Yang F, Liu J, Yang W, Shu K (2017) Exogenous auxin represses soybean seed germination through decreasing the gibberellin/abscisic acid (GA/ABA) ratio. Sci Rep 7(1). <https://doi.org/10.1038/s41598-017-13093-w>
- Silva JCF, Teixeira RM, Silva FF, Brommonschenkel SH, Fontes EPB (2019) Machine learning approaches and their current application in plant molecular biology: a systematic review. Plant Sci 284:37–47. <https://doi.org/10.1016/j.plantsci.2019.03.020>
- Şimşek Ö, Dalda Şekerci A, Isak MA, Bulut F, İzgü T, Tütüncü M, Dönmez D (2024) Optimizing micropropagation and rooting protocols for diverse lavender genotypes: a synergistic approach integrating machine learning techniques. Hortic 10(1). <https://doi.org/10.3390/horticulturae10010052>
- Trejo-Tapia G, León-Romero Y, Montoya-Medina EB et al (2019) Perspectives of synthetic seed technology for conservation and mass propagation of the medicinal plant *Castilleja tenuiflora* Benth. In: Faisal M, Alatar AA (eds) Synthetic seeds. Springer International Publishing Cham pp 335–345. https://doi.org/10.1007/978-3-03-0-24631-0_15
- Urban J, Matoušková M, Robb W, Jelinek B, Úradníček L (2023) Effect of drought on photosynthesis of trees and shrubs in habitat corridors. Forests 14(8):1521. <https://doi.org/10.3390/f14081521>
- Vásquez-Coronel J, Altamirano-Fernández A, Espinoza-Meza S, Rodríguez-Gallardo M (2022) Prediction of survival of *Pinus radiata* seedlings subjected to physical-water restriction extreme using learning neural networks. J Phys: Conf Ser, 2153(1). <https://doi.org/10.1088/1742-6596/2153/1/012015>
- Wu J, Chang H, Tong R, Wu X (2022) A novel method for handling missing data in health care real-world study: Optimal Intact Subset Method. <https://doi.org/10.21203/rs.3.rs-1761573/v1>
- Xie X, Zhang X, Shen J, Du K (2022) Poplar's waterlogging resistance modeling and evaluating: exploring and perfecting the feasibility of machine learning methods in plant science. Front Plant Sci 13. <https://doi.org/10.3389/fpls.2022.821365>
- Yan J, Wang X (2023) Machine learning bridges omics sciences and plant breeding. Trends Plant Sci 28(2):199–210. <https://doi.org/10.1016/j.tplants.2022.08.018>
- Yanakiyeva V (2023) Study on the antimicrobial activity of medicinal plant extracts and emulsion products with integrated herbal extracts. J Agric and For 69(4). <https://doi.org/10.17707/agricultforest.69.4.06>
- Yang C, Gunay B, Shi Z, Shen W (2021) Machine learning-based prognostics for central heating and cooling plant equipment health monitoring. T-ASE 18(1):346–355. <https://doi.org/10.1109/TASE.2020.2998586>
- Ye Y, Gong Z, Lu X, Miao D, Shi J, Lu J, Zhao Y (2016) Germostatin resistance locus 1 encodes a PHD finger protein involved in auxin-mediated seed dormancy and germination. Plant J 85(1):3–15. <https://doi.org/10.1111/tpj.13086>
- Yoosefzadeh-Najafabadi M, Earl HJ, Tulpan D, Sulik J, Eskandari M (2021) Application of machine learning algorithms in plant breeding: predicting yield from hyperspectral reflectance in soybean. Front Plant Sci 11. <https://doi.org/10.3389/fpls.2020.624273>
- Yoosefzadeh Najafabadi M, Hesami M, Eskandari M (2023) Machine learning-assisted approaches in modernized plant breeding programs. Genes 14(4):777. <https://doi.org/10.3390/genes14040777>