EEET ECOLOGICAL ENGINEERING
& ENVIRONMENTAL TECHNOLOGY

Ecological Engineering & Environmental Technology, 2025, 26(7), 299–315
https://doi.org/10.12912/27197050/205194
ISSN 2719–7050, License CC-BY 4.0

Received: 2025.05.01
Accepted: 2025.05.30
Published: 2025.06.15

Sustainable date palm agriculture in Moroccan oases: AI and machine learning for prediction of Bayoud disease

Youssef El Hilali Alaoui[1,2*], Said Bouda[1], Zakaria Khoudi[3],
Abdellatif Boutagayout[4], Hajar Mhammedi[1,2],
Houria Ben Amar[2], Eimad Dine Tariq Bouhlali[2], Adil Essarioui[2]

[1] Laboratory of Agro-Industrial and Medical Biotechnology, Faculty of Sciences and Technics, Sultan Moulay Slimane University, B.P. 523, Beni Mellal, Morocco
[2] Oasis System Research Unit, Regional Center of Agricultural Research of Errachidia, National Institute of Agricultural Research, PO. Box 415, Rabat 10090, Morocco
[3] Equipe des Mathematiques et Interactions, Faculty of Sciences and Technics, Sultan Moulay Slimane University, Beni Mellal, Morocco
[4] Health Environment and agroecosystem sustainability, Moulay Ismail University, Faculty of Science, P.B. 11201, Zitoune, Meknes
* Corresponding author's e-mail: saidbouda@yahoo.fr

ABSTRACT

Bayoud disease, caused by *Fusarium oxysporum f.* sp. *albedinis* (Foa), threatens date palm cultivation, especially in North Africa, where date production is crucial for food security and economic stability. Current management strategies, including resistant cultivars and chemical treatments, have limited success because of the genetic adaptability and survival of the pathogen in soil. Early detection remains challenging because current methods rely on visible symptoms that appear after significant damage. This study explored the potential of machine learning (ML) to predict the soil suppressiveness of Bayoud disease by analyzing microbial metabolic activity using Biolog SF-P2 assays. Sixty *Fusarium* isolates from suppressive and conducive soils in Morocco were assessed using 95 different carbon sources. Six ML models, logistic regression, naive bayes, k-nearest neighbors (KNN), XGBoost, Gradient Boosting, and support vector machine (SVM) were applied to the data. The results showed that the XGBoost and Gradient Boosting models achieved the highest predictive accuracy, with AUC values exceeding 90%, indicating strong classification ability. SHAP analysis identified key metabolic markers linked to disease suppression, thereby highlighting the role of microbial communities in pathogen resistance. This study established a data-driven framework for predicting soil suppressiveness and facilitating proactive disease management.

Keywords: Bayoud disease, oasis ecosystems, Biolog SF-P2 assays, date palm, early detection, machine learning.

INTRODUCTION

In North Africa, and especially Morocco and Algeria, date palm (*Phoenix dactylifera*) is one of the mainstays of food security and an important component of economic development. Unfortunately, this key crop is in danger from Bayoud disease, which is caused by the fungus *Fusarium oxysporum* f. sp. albedinis (Foa) causing vascular wilt of date palm The disease has already wiped-out millions of date palms in the region and threatens agricultural production and the local economy to a large extent [El Hilalai Alaoui et al., 2024; Essarioui et al., 2018] The pathogen persists in the soil causing complications in its management as it also gets transmitted through water and contaminated tools [Sedra, 2018].

Current management strategies for Bayoud disease of date palm include resistant cultivars, and on the other side, chemical treatments. However, these methods have been proven to have limited success. The genetic plasticity of the pathogen makes it difficult for resistant cultivars to adapt, and pesticide use gives rise to serious environmental and sustainability issues [El modafar, 2010; Benzohar et al., 2015]. Also, another significant bottleneck in the management of this disease is that all the current diagnostic techniques are clinically symptomatic and only identify the infection after substantial plant damage, leaving little time for their management.

Soil microbiome and pathogen suppression represent an attractive alternative and promising approach for the control of Bayoud disease. For example, "suppressive" soils can prevent *Fusarium oxysporum* growth, while "conducive" soils can facilitate this pathogen [Ou et al., 2019; Todorović et al., 2023]. This property of a few soils to suppress pathogen development capacity is related to particular plant–microbial communities able to regulate the rhizosphere environment and control pathogens. Yet, the molecular and metabolic mechanisms that mediate these interactions are largely unknown [Philippot et al, 2024. Trivedi et al, 2020].

Identifying the profiles of community metabolism using assays such as Biolog SF-P2 has recently been able to distinguish between suppressive and conducive soils [El Hilali Alaoui et al., 2024; Pereyra, 2021]. Nevertheless, the need for rigorous diagnostic tools to assess soil suppressiveness remains an important, but largely unmet, need.

The rise of machine learning (ML) technologies has substantially advanced the analysis of large-scale data generated from microbial activity. ML models can learn to reveal relationships in more complex datasets, and subsequently, in a supervised fashion, predict the behavior of microbial communities from their associated characteristics, employing algorithms such as logistic regression, random forests, XGBoost, and support vector machines (SVM) [Wanget al., 2024. Mohseni et al., 2024]. These models could be applied in the context of Bayoud disease to predict the soil suppressive potential of *Fusarium oxysporum* using microbial metabolic profiles.

By leveraging ML models, accurate soil suppressiveness can be predicted and knowledge on the microbial traits associated with this resistance will be obtained. For instance, we could use feature importance analysis via SHAP (SHapley Additive exPlanations) to determine specific metabolic markers connected to control of disease [Chen et al., 2024; Li et al., 2024].

The main aim of this study was the exploration of ML models in predicting soil suppressiveness towards Bayoud disease using the microbial metabolic activity [El Hilali Alaoui et al., 2025] as determined by Biolog SF-P2 assays and consequently creating a data driven the framework which should allow the classification of soils into either potential suppressors or activators of pathogen proliferation. To this end, we hypothesized that ML models are able to identify autosomal metabolic indicators that distinguish suppressive versus conducive soils and allow for predictive rather than retrospective approaches to disease management. This framework may be significant in the prevention of the disease, as well as in providing targeted control strategies

MATERIALS AND METHODS

Study sites and soil sampling

This study was conducted in two distinct regions of southern Morocco (Figure 1): Marrakech, characterized by soils naturally suppressive to Bayoud disease, and Zagora, where soils are conducive to the pathogen [Sedra, 1989]. These regions were selected due to their contrasting edaphic and climatic conditions, which influence microbial community structure and dynamics.

In each region, six soil samples were collected, totaling 12 samples. Each sample was composed of two soil cores, taken at a depth of 0–20 cm and spaced 1 meter apart around the base of adult date palm trees. The samples were dried, sieved, and stored at -20 °C before further analysis. This sampling approach aimed to assess the variability in soil microbial communities and their potential role in Bayoud disease suppression or conduciveness.

Isolation and identification of fusarium isolates

To characterize the microbial composition of the soils, 60 Fusarium isolates were obtained, with 30 isolates from Marrakech and 30 from Zagora. Soil suspensions were prepared by mixing 5 g of soil from each sample in 50 mL of sterile deionized water, followed by agitation at 175 rpm for 60 minutes at 4 °C (Figure 1). Serial dilutions up to 10^{-3} were plated onto peptone pentachloronitrobenzene (PCNB) Agar Medium (PPA), a

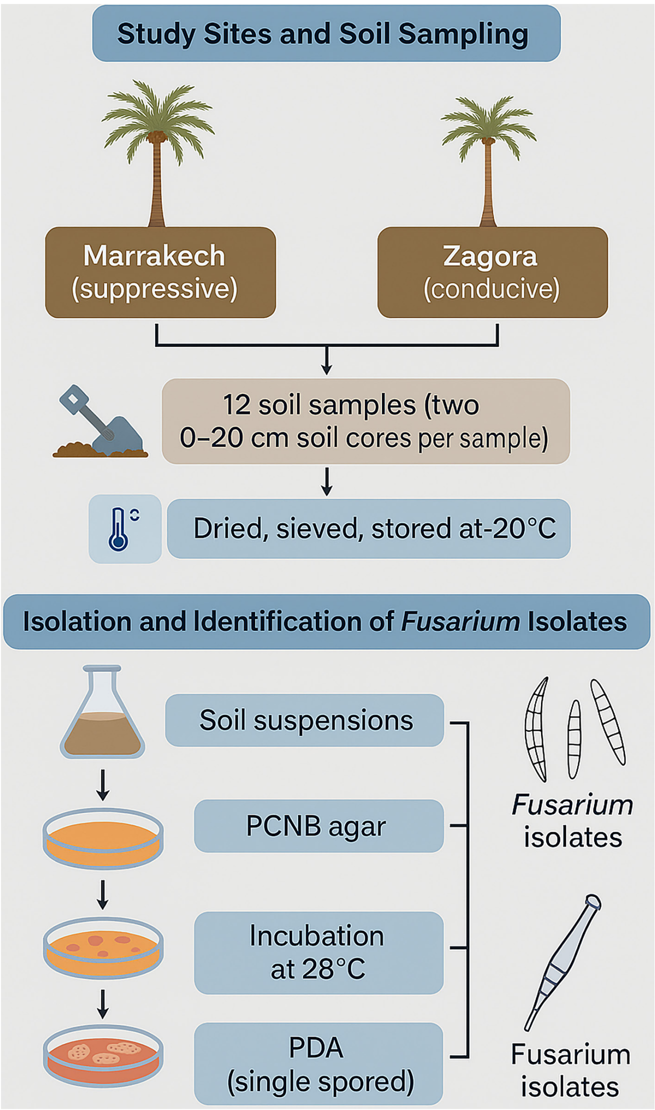


Figure 1. Detailed workflow of soil sampling and *Fusarium oxysporum* isolation from suppressive and conducive soils in Moroccan oasis agroecosystems

selective medium for *Fusarium* species [Leslie and Summerell, 2006; Summerell et al., 2003].

Plates were incubated at 28 °C for 3–4 days, and fungal colonies presenting morphological characteristics of *Fusarium* were transferred to potato dextrose agar (PDA) for further identification (Figure 2). Identification was performed based on colony morphology, pigmentation, and macroconidial shape, following the criteria described by Leslie and Summerell [2006]. All isolates were single-spored before being used for further microbiological analysis (Figure 2).

Carbon source utilization analysis using Biolog SFP2 plates

Biolog SF-P2 plates (Biolog, Inc. Hayward, CA) were used to determine the carbon sources utilized by each *Fusarium isolate* (n = 60). Biolog SF-P2 microplates directly assess the growth of an isolate on 95 isolates on individual carbon sources by comparing turbidity in each well to water control. In the SF-P2 panel, 95 carbon substrates were tested which cover 11 carbon groups (the number of substrates for each group in parentheses): alcohol (3), amide (3), amine (1), amino acid (9), aromatic compound (4), carbohydrate (41), carboxylic acid (15), ester (3), phosphorylated compound (8), and polymer (8). Fungal suspensions were prepared by swabbing spores/mycelium of 10 days old pure culture of *Fusarium* (grown on PDA) into 1.5 ml of 0.2% carrageenan. Sprays were adjusted at an OD of 590 nm of 0.20–0.24 and diluted in 0.2% carrageenan (13.5 mL). 100 µl of the resulting microbial suspension was injected into each well of a Biolog plate. Plates were incubated at 28versusC for five days [18]. The Biolog SF-P2 plate generally consists of 95 different carbon compounds, which were used on the Biolog SF-P2 plate and reading of these data was performed over 24 h 36, h, 48 h, 60 h, 72 h, 84 h, and 120 h post-inoculation. For standardisation of absorbance values, the absorbance of each well (only contained water) was removed, then negative values were transformed to zero (Figure 3) [El Hilali Alaoui et al., 2025].

Growth parameters calculated

Several growth indicators were determined to evaluate the metabolic activity of *Fusarium* isolates:

- Niche width (NW): The total number of carbon substrates utilized by each isolate.
- Average growth (AG): The mean absorbance value across all utilized substrates.
- Total growth (TG): The sum of absorbance values for all utilized substrates.
- Growth rate: The rate of increase in total absorbance per hour.

These parameters were used to compare the metabolic capacities of *Fusarium* isolates from suppressive and conducive soils and to investigate their potential role in Bayoud disease suppression or conduciveness [Ahmad et al., 2023].

Modeling and data analysis

The *Fusarium* soil dataset consists of 60 isolates from two regions (Marrakech and Zagora), with 30 isolates per region recorded across eight time points. Each *Fusarium* soil is described by 105 features, including 95 carbon source utilization values obtained from Biolog SF-P2 tests, 11 carbon group classifications, and three key microbiological parameters: NW, TG, and AG. Thus, considering these elements, the *Fusarium* soil dataset contains 50,400 numerical values. [El Hilali Alaoui et al., 2025].

Before model training, all numerical features were normalized using Min-Max scaling to standardize the input range and reduce bias from highly variable carbon utilization rates. This step was essential to ensure comparability between different carbon sources and to enhance the convergence of machine learning algorithms.

To predict the presence of *Fusarium oxysporum f.* sp. *albedinis*, six supervised machine learning models were tested: logistic regression

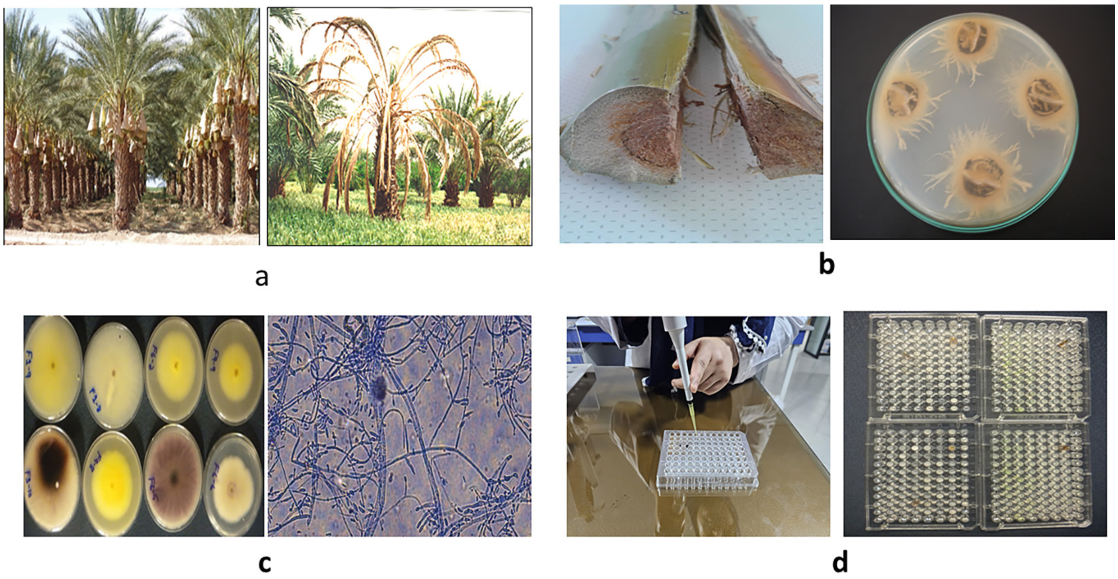


Figure 2. Experimental process for isolating, identifying, and studying the nutritional behavior of *Fusarium oxysporum f.* sp. albedinis from date palm plants and *Fusarium* isolates from the studied soils. a: Cultivated date palm plots, showing one healthy plot (left) and another affected by Bayoud disease (right), with a suppressive soil for Bayoud (right) and a conducive soil (left). b: Image of the rachis of a date palm infected by Bayoud disease, clearly showing the Fusarium infection. c: Photos showing the colonies of *Fusarium* from the studied soils, cultivated on selective medium in Petri dishes with microscopic characteristics of *Fusarium* (mycelium, conidiophores, microconidia, and macroconidia) d: Use of Biolog SF-P2 plates to test the metabolic behavior of *Fusarium* isolates

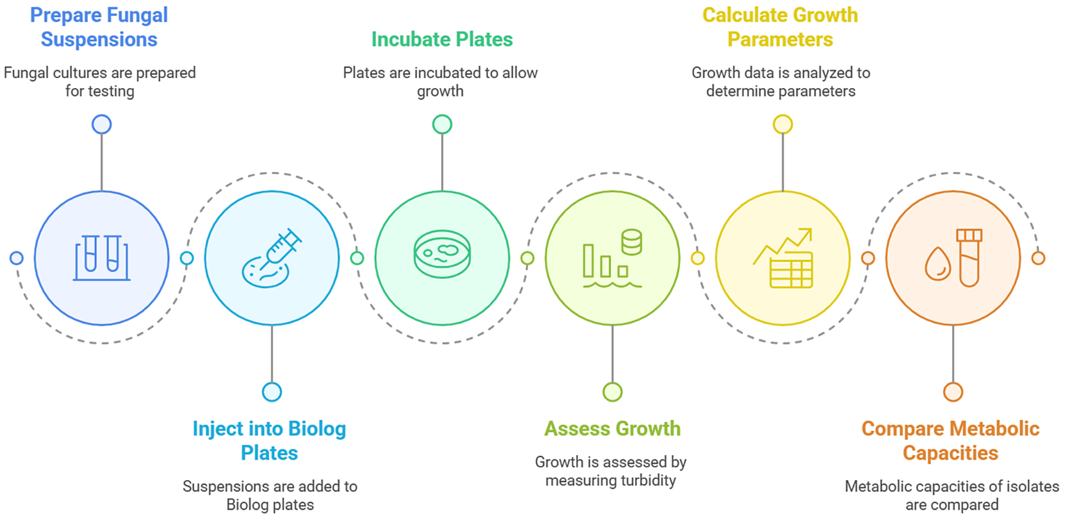


Figure 3. Detailed workflow for fungal culture preparation, growth assessment, and metabolic capacity comparison for *Fusarium oxysporum* isolates using Biolog SF-P2 plates

(LR), naive bayes (NB), k-nearest neighbors (KNN), XGBoost (XGB), gradient boosting (GB), and support vector machine (SVM). These models were selected to capture both linear and non-linear relationships in the dataset, leveraging different classification paradigms to assess the robustness of predictive performance (Figure 4).

The dataset was divided into 80% training and 20% testing, ensuring a balanced representation of suppressive and conducive soils [Ahmad et al, 2023]. Model performance was optimized through 5-fold cross-validation, which mitigates overfitting and enhances generalizability. Hyperparameter tuning was performed using Grid Search CV, optimizing key parameters such as the number of estimators for boosting models, kernel types for SVM, and the number of neighbors for KNN.

The predictive performance of each model was evaluated using accuracy, F1-score, and AUC-ROC, which provide complementary insights into classification effectiveness. To further interpret model predictions, Feature importance scores were computed, identifying the most influential carbon sources and microbiological parameters in distinguishing between suppressive and conducive soil environments.

Machine learning models applied

This section summarizes the machine learning models used to predict soil suppressiveness against *Fusarium oxysporum f.* sp. *albedinis (Foa),* explaining their mechanisms, mathematical foundations, and relevance to soil microbial profile analysis [Nemer et al, 2024].

Logistic regression (LR)

The Logistic Regression model is a statistical classification technique used to predict binary outcomes. It models the probability that a given sample belongs to a particular class using the sigmoid function, which transforms linear combinations of input features into probabilities between 0 and 1.

The mathematical formulation of logistic regression is given by:

$$P(Y=1|X)=\frac{1}{1+e^{-(\beta_0+\sum\beta_i X_i)}} \tag{1}$$

where: $P(Y = 1 \mid X)$ is the probability of the positive class, β_0 is the intercept, β_i are the model coefficients, X_i are the input features, e is Euler's number.

This model was chosen because it provides interpretable coefficients and is efficient for datasets with well-separated classes, such as the prediction of soil suppressiveness based on microbiological parameters.

Naive bayes (NB)

The Naive Bayes classifier is a probabilistic model based on Bayes' theorem and the assumption that features are conditionally independent given the class label. It computes the posterior probability of a sample belonging to a class using:

$$P(Y \mid X)=\frac{P(X \mid Y)P(Y)}{P(X)} \tag{2}$$

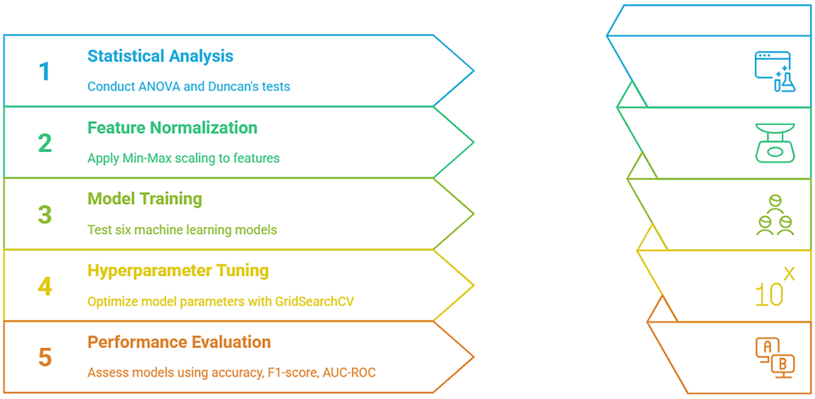


Figure 4. Steps in data analysis and machine learning model optimization for predicting *Fusarium oxysporum f.* sp. *albedinis* (Foa) infection

where: $P(Y \mid X)$ is the posterior probability of class YY given the feature vector XX, $P(Y|X)$ is the likelihood, $P(Y)$ is the prior probability of the class, $P(X)$ is the normalizing constant.

Naive bayes is computationally efficient and works well with high-dimensional data, making it suitable for classifying soil microbial profiles based on carbon utilization patterns.

K-nearest neighbors (KNN)

The K-Nearest Neighbors algorithm is a non-parametric classification method that assigns a class label to a new sample based on the majority vote of its K nearest neighbors in feature space. The distance between points is typically measured using Euclidean distance:

$$d(X, X') = \sqrt{\sum_{i=1}^{n} \left(X_i - X_i'\right)^2} \tag{3}$$

where: X and X' are two data points in an n-dimensional space.

This model was used due to its ability to capture non-linear decision boundaries in high-dimensional spaces, such as microbial metabolic profiles.

XGBoost (XGB) and gradient boosting (GB)

Both XGBoost and gradient boosting are ensemble learning techniques based on decision trees, where multiple weak learners are trained sequentially to minimize prediction errors. The update rule for boosting is given by:

$$F_m(X) = F_{m-1}(X) + \gamma h_m(X) \tag{4}$$

where: $F_m(X)$ is the updated model at iteration mm, $F_{m-1}(X)$ is the previous iteration, γ is the learning rate, $h_m(X)$ is the new weak learner (decision tree).

XGBoost introduces regularization and parallel computation, making it more efficient than traditional boosting models. These models were selected for their high predictive power and ability to handle complex relationships in the dataset.

Support vector machine (SVM)

The Support Vector Machine classifier finds an optimal hyperplane that maximizes the margin between two classes. The decision function is defined as:

$$f(X) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i Y_i K(X_i, X) + b\right) \tag{5}$$

where: α_i are the Lagrange multipliers, Y_i are class labels, $K(X_i)$ is the kernel function, b is the bias term.

For non-linearly separable data, SVM uses kernel functions such as the radial basis function (RBF):

$$K(X, X') = e^{-\gamma \left\| X - X' \right\|^2} \tag{6}$$

This model was chosen due to its robustness in high-dimensional spaces and its ability to handle non-linear relationships in microbial metabolic data.

This section provides a structured explanation of each machine learning model used in the study, along with their mathematical foundations and relevance to soil microbial analysis (Table 1).

Statistical test explanation

To compare the performance of Gradient Boosting with other models (XGBoost, SVM, Naïve Bayes, KNN, Logistic Regression), a paired t-test was performed. This test allows for the comparison of accuracy differences between gradient boosting and each other model across the folds of 10-fold cross-validation. The paired t-test was chosen because it effectively compares the differences between paired samples, ensuring a robust statistical evaluation of performance differences. The assumptions of normality for the differences in accuracy across folds were validated using the Central Limit Theorem, which holds given 10 folds.

Distribution of the target variable

One of the first aspects analyzed in this study is the distribution of the target variable, which differentiates between suppressive and conducive soils in relation to *Fusarium oxysporum f.* sp. *albedinis (Foa)* infection. The dataset presents a balanced distribution of both classes, with an equal number of suppressive and conducive soil samples. This balance is particularly advantageous for machine learning applications, as it helps prevent model bias toward a dominant class and enhances the generalizability of predictions.

EVALUATION METRICS FOR BINARY CLASSIFICATION IN SOIL ANALYSIS

When distinguishing between suppressive and conductive soils, evaluating the model's predictive performance using appropriate metrics is essential. These metrics offer a detailed assessment of how well the model differentiates between the two soil types, helping to ensure accurate classification and reliable decision-making [Steurer et al., 2021; Cabot et al., 2023].

A well-performing model should not only correctly classify soils but also minimize misclassification errors, particularly false positives (incorrectly identifying conducive soil as suppressive) and false negatives (failing to detect suppressive soil). Since misclassifications can lead to incorrect soil management practices such as unnecessary treatments or overlooking harmful soil conditions it is critical to use multiple evaluation

Table 1. Justification, advantages, and limitations of the models used in the study

Model	Justification	Advantages in the context of the study	Limitations in the context of the study
Logistic regression (LR)	Used to model the relationship between metabolic characteristics (carbon source utilization) and the presence of *Fusarium oxysporum* in suppressive or conducive soils.	Easy to interpret; Effective for binary classification problems; Handles linear relationships well.	Assumes a linear relationship between the explanatory variables and the target probability; Less effective with complex non-linear relationships.
Naive bayes (NB)	Based on Bayes' theorem; classifies isolates based on the conditional probability of metabolic characteristics.	Fast to train, even with a large number of variables; Suitable for high-dimensional datasets; Performs well with small datasets.	Assumption of independence between features is rarely true in microbiology; Sensitive to correlations between variables.
K-Nearest neighbors (KNN)	Classifies isolates based on similarity to the nearest 'neighbors' in the feature space.	Non-parametric model, suitable for complex relationships; Easy to understand and implement; Adapts well to small datasets.	Slow for large datasets (requires distance calculation); Sensitive to noise and the choice of 'k' value.
XGBoost (XGB)	An ensemble model based on boosting, combining multiple weak decision trees to improve accuracy.	Highly effective with complex non-linear relationships; Handles missing values and complex variable interactions well; Optimized for fast training.	Risk of overfitting if the model is too complex; Requires careful tuning of hyperparameters.
Gradient boosting (GB)	Boosting-based model with a customizable loss function.	High performance for complex classification problems; Models non-linear relationships well; Works well with small and medium-sized datasets.	Slower to train than XGBoost; Prone to overfitting if too many estimators are used.
Support vector machine (SVM)	Used to separate classes by maximizing the margin between them in a high-dimensional space.	Effective for binary classification with small datasets; Works well in high-dimensional spaces; Handles non-linear problems with kernel functions.	Computationally expensive for large datasets; Hard to interpret; Sensitive to kernel and regularization parameter choices.

metrics to get a comprehensive understanding of model effectiveness.

By analyzing metrics such as accuracy, precision, recall, F1-score, and AUC-ROC, researchers and agronomists can determine whether the model provides a balanced and reliable classification. These measures help in identifying weaknesses in the model and making improvements for better soil health monitoring and sustainable agricultural practices.

Accuracy

Accuracy represents the overall proportion of correctly classified soil samples relative to the total number of samples. It is a useful metric when the dataset is balanced (our case), but it may be misleading in cases of class imbalance.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

where: *TP* (true positives): Suppressive soils correctly classified as suppressive, *TN* (true negatives): Conducive soils correctly classified as conducive, *FP* (false positives): Conducive soils incorrectly classified as suppressive, *FN* (false negatives): Suppressive soils incorrectly classified as conducive

Precision

Precision quantifies how many of the samples predicted as suppressive soils are suppressive. It helps assess the reliability of positive predictions.

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

Recall

Recall measures the proportion of actual suppressive soils that the model correctly identifies. It is crucial when missing suppressive soils could have serious consequences.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{9}$$

F1-score

The F1-score is the harmonic mean of precision and recall, balancing both aspects in a single measure. It is particularly useful when the dataset is imbalanced.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{10}$$

Receiver operating characteristic (ROC) curve & Area under the curve (AUC-ROC)

The ROC curve plots the true positive rate (Recall) against the false positive rate (FPR) at various classification thresholds, showing the trade-off between correctly detecting suppressive soils and misclassifying conducive soils.

$$False\ Positive\ Rate = \frac{FP}{FP + TN} \tag{11}$$

The Area Under the Curve (AUC-ROC) quantifies the model's ability to distinguish between suppressive and conducive soils:

- AUC close to 1: The model has high discriminative power.
- AUC around 0.5: The model performs no better than random chance.
- AUC close to 0: The model is making incorrect predictions systematically.

RESULTS

Machine learning models for classification of suppressive and conducive soils

The performance of several machine learning models was evaluated for the classification of suppressive (Class 0) and conducive (Class 1) soils, using precision, recall, F1-score, and overall accuracy as evaluation metrics. As shown in Table 2, the results highlight clear differences in model performance.

Gradient Boosting and XGBoost achieved the highest scores across all metrics, each recording an F1-score of 0.99 for both classes. The overall accuracy reached 0.99 for Gradient Boosting and 0.98 for XGBoost, indicating strong and consistent classification results for both suppressive and conducive soils.

Logistic regression also showed high performance, with an overall accuracy of 0.98. Precision and recall were equally distributed across both classes, confirming the model's consistent behavior in identifying each soil type.

SVM followed with an accuracy of 0.96, maintaining stable precision and recall values. The KNN model achieved a slightly lower accuracy of 0.94. Although its recall for conducive soils remained high, recall for suppressive soils dropped to 0.89, reflecting a higher rate of misclassification in that category.

Naïve Bayes reported the lowest performance among the evaluated models. Its overall accuracy was 0.79, with a recall of 0.60 for suppressive soils and 0.98 for conducive soils, indicating an imbalance in its classification capacity. All numerical results supporting these observations are detailed in Table 2.

Paired t-test results for comparisons of gradient boosting with other models

The results of the paired t-test showed that Gradient Boosting significantly outperformed all the other models ($p < 0.05$; Table 3). The largest difference was observed against Naïve Bayes, with a mean difference of 0.204, t-statistic of 125.15, and p-value less than 0.001, reflecting a substantial accuracy gap (0.99 vs. 0.79). The smallest differences were found between XGBoost and logistic regression, both with a mean difference of 0.01. Although these differences are small, they remain statistically significant, with t-statistics of 4.74 and p-values less than 0.01.

Moderate differences were observed between KNN and SVM, with mean differences of 0.055 and 0.035, respectively. Both comparisons showed highly significant results ($p < 0.001$). These results indicate that Gradient Boosting is consistently superior across all the models tested, confirming its strong classification performance.

Comparative analysis of machine learning models for soil classification based on ROC curves

Figure 5 presents ROC curves for six machine learning models applied to soil classification: SVM, NB, KNN, LR, XGBoost, and Gradient Boosting. Each model's curve displays the trade-off between the true positive rate (TPR) and false positive rate (FPR), with the area under the curve (AUC) serving as a quantitative measure of classification performance.

SVM shows a high AUC of 0.96, indicating a strong separation between soil classes. Its kernel-based transformations allow it to capture complex patterns effectively. Naïve Bayes, with an AUC of 0.79, underperforms due to its assumption of feature independence, which limits its ability to capture feature interactions in soil profiles. KNN achieves a high AUC of 0.94, benefiting from local pattern recognition but is sensitive to hyperparameters like k and distance metrics, and struggles with scalability.

Logistic Regression delivers an AUC of 0.963, reflecting solid performance, though its assumption of linear separability of features may limit its effectiveness compared to more flexible models. XGBoost and Gradient Boosting both reach near-perfect AUCs of 0.99. These ensemble methods excel at handling non-linear relationships and feature interactions, with XGBoost's robustness to overfitting stemming from regularization and tree-pruning techniques, although training cost increases with model complexity. Gradient Boosting's iterative refinement ensures high accuracy but risks overfitting if hyperparameters aren't carefully tuned.

Overall, ensemble methods (XGBoost, gradient boosting) and kernel-based models (SVM) show the best performance, particularly for complex soil classification tasks. KNN provides strong results through local pattern recognition, while Logistic regression offers interpretability. Naïve Bayes, despite its lower performance, serves as a useful benchmark model due to its simplicity. These findings highlight that non-linear and ensemble methods are well-suited for soil classification, with XGBoost and Gradient Boosting offering the highest accuracy and robustness.

Table 2. Comparative results of machine learning models for suppressive and conducive soil classification

Model	Precision Classe 0	Rappel Classe 0	F1-Score Classe 0	Precision Classe 1	Rappel Classe 1	F1-Score Classe 1	Accuracy
SVM	0.94	0.98	0.96	0.98	0.94	0.96	0.96
XGBoost	1.00	0.98	0.99	0.98	1.00	0.99	0.98
Naïve Bayes	0.97	0.60	0.74	0.72	0.98	0.83	0.79
KNN	0.98	0.89	0.93	0.91	0.98	0.94	0.94
Logistic reg.	0.98	0.98	0.98	0.98	0.98	0.98	0.98
Gradient boosting	0.98	1.00	0.99	1.00	0.98	0.99	0.99

Table 3. Paired t-test results for comparisons of gradient boosting with other models (XGBoost, SVM, KNN, Logistic regression, Naïve Bayes)

Comparison	Mean difference ($\bar{d}$)	Std Dev (s_d)	t-statistic	p-value	Significant (α = 0.05)?
GB vs. XGBoost	0.01	0.00667	4.74	lt 0.01	Yes
GB vs. SVM	0.035	0.00957	11.55	lt 0.001	Yes
GB vs. KNN	0.055	0.00972	17.92	lt 0.001	Yes
GB vs. Logistic regression	0.01	0.00667	4.74	lt 0.01	Yes
GB vs. Naïve Bayes	0.204	0.00516	125.15	lt 0.001	Yes

Confusion matrix for XGBoost model

The confusion matrix for the XGBoost model is presented in Figure 6. Out of the total samples, the model correctly classified 69 suppressive soils as suppressive and 67 conducive soils as conducive. Misclassifications included 5 suppressive soils labeled as conducive and 3 conducive soils labeled as suppressive. These values align with the performance metrics reported in Table 2 and support the model's high classification accuracy.

Feature importance analysis from the gradient boosting model

The feature importance scores generated by the Gradient Boosting model are presented in Figure 7.

These scores indicate the relative contribution of each input variable to the model's predictions in the classification of suppressive and conducive soils. The figure displays horizontal bars representing the importance values, with longer bars corresponding to higher impact features. The top three features show markedly higher importance scores compared to the remaining variables, indicating a concentration of predictive influence among a limited subset of inputs. Feature importance in the Gradient Boosting model is calculated using Gini importance, which reflects the total reduction in node impurity contributed by each feature across all trees in the ensemble. Lower-ranked features show progressively smaller importance values [Otchere et al., 2022].

Feature importance analysis from SHAP (SHapley additive explanations)

SHAP values were calculated to assess the contribution of each input feature to the predictions made by the soil classification models. The SHAP summary plot (Figure 8) displays the

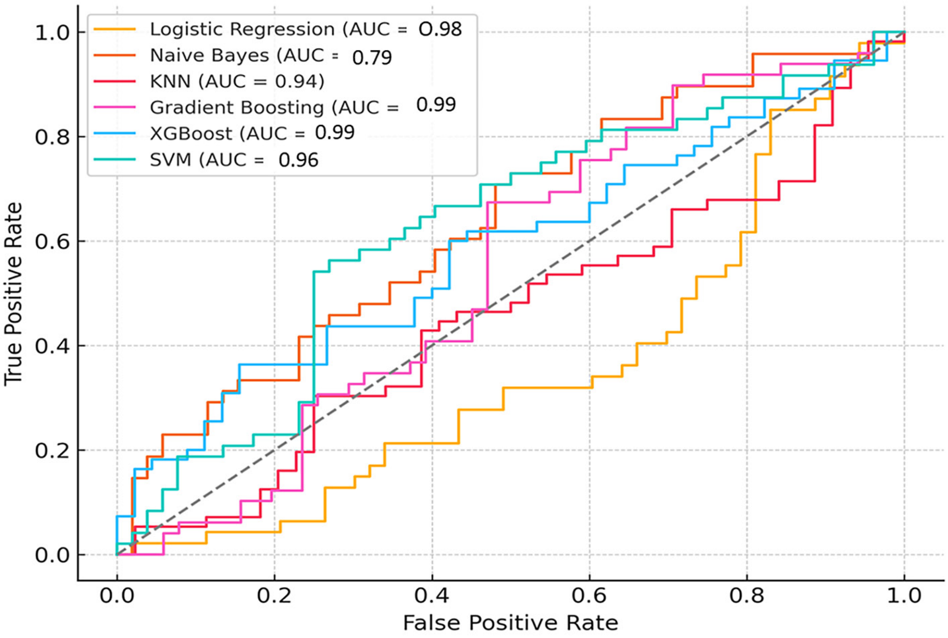


Figure 5. Comparative performance of machine learning models based on ROC curves for soil classification

magnitude and direction of each feature's impact on the model's output. Features with the highest SHAP values (it's worthy to mention these features) were consistent with those ranked most important in the Gradient Boosting feature importance analysis. These top-ranked features show the strongest influence, either increasing or decreasing the predicted probability of a given class. Several features exhibit a broad range of SHAP values across the dataset, indicating variable effects on the model's output depending on the input value. The SHAP analysis also reveals that a small number of features (mention it please) accounted for the majority of the predictive contribution.

Correlation matrix analysis for soil classification models

The correlation matrix for the input features used in the soil classification models is shown in Figure 9. The matrix displays pairwise Pearson correlation coefficients, ranging from -1 to 1, representing the linear relationships between variables.

The matrix displays pairwise Pearson correlation coefficients, ranging from -1 to 1, representing the linear relationships between variables. Several feature pairs exhibit strong positive correlations, with coefficients above 0.8, such as the relationship between B_7 and C_2 (r = 0.79), indicating a significant linear relationship between Fusarium isolates and carbon source utilization. A smaller number of feature pairs show moderate to high negative correlations, like between C_1 and A_6 (r = -0.24), suggesting a weak negative association. Most remaining feature combinations demonstrate weak or negligible correlations, with values close to zero. These results indicate varying degrees of linear association among the Fusarium soil features.

DISCUSSION

Bayoud disease, which is caused by *Fusarium oxysporum f.* sp., *Cactoblastis cactorum* (Berg), and *Cactoblastis albedinis* (Foa) continues to be an important pest of date palm, especially in North Africa. These results suggest that functional profiles of stimulated microbes reflect the metabolic activity patterns within the soil and can be effective predictors of soil suppressiveness, which can be used as a foundation for better modeling for disease management. This is an important stride towards the multidisciplinary merger of microbial ecology and machine learning in agriculture-related soil-borne pathogens.

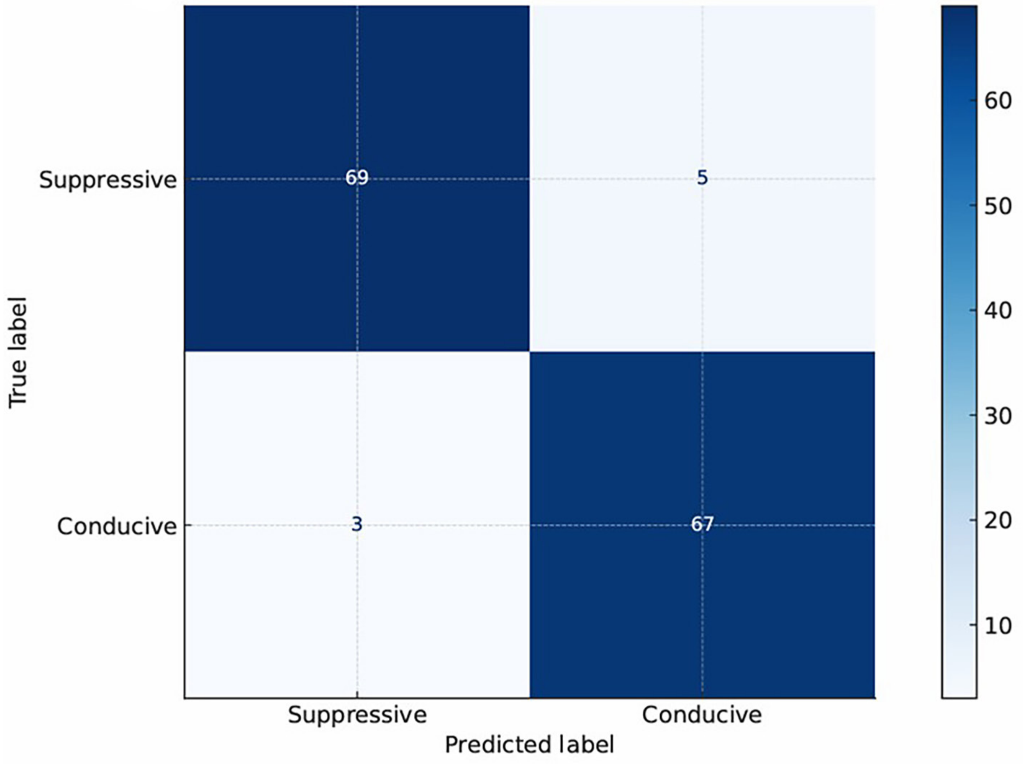


Figure 6. Confusion matrix for XGBoost model – classification of suppressive and conducive soils

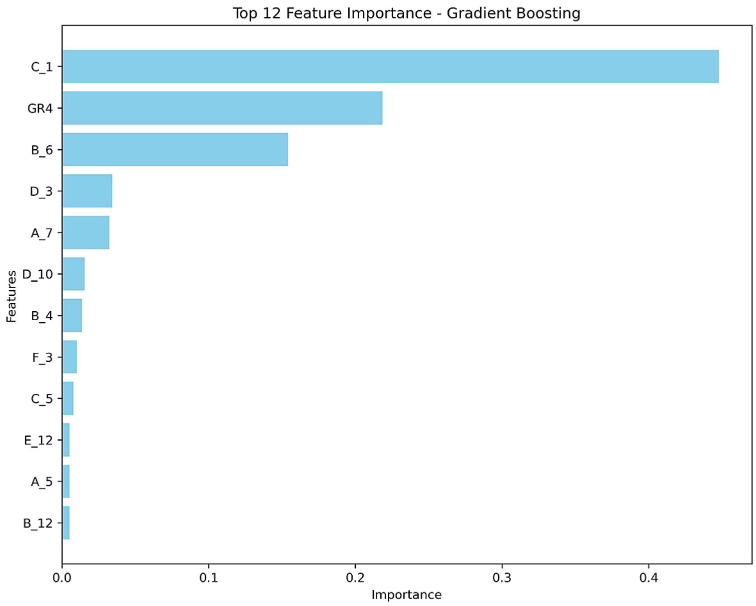


Figure 7. Feature importance analysis for soil classification models using Gradient Boosting – key variables influencing disease suppression and conduciveness; give the significance of features: meaning of each, the codes on the x-axis of the graph (A, C, D, GR4, etc.) represent the observations corresponding to the sources of carbon and carbon group classifications from the Biolog SF-P2 plates. Each of these codes is associated with specific *Fusarium* soil isolates

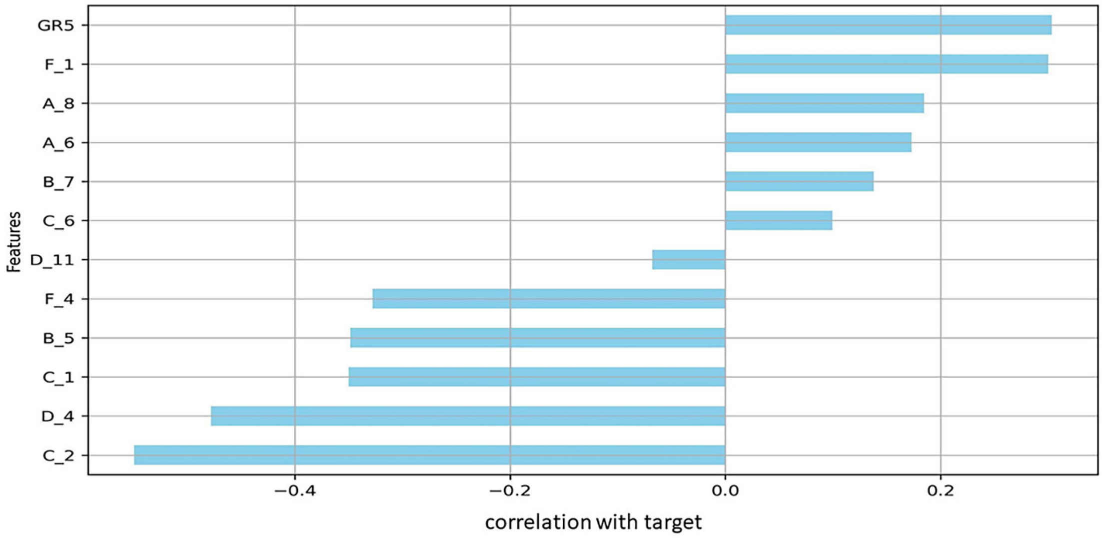
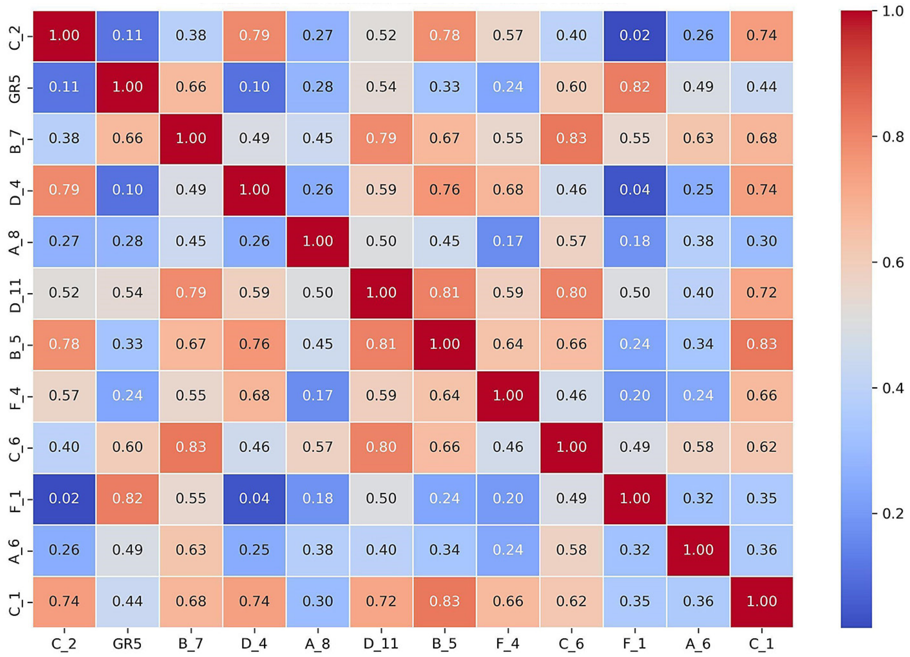


Figure 8. SHAP value analysis for soil classification models – understanding feature contributions and non-linear interactions. Abbreviations names as shown in Figure 7

The strong predictive performance of XG-Boost and Gradient Boosting emphasizes the benefit of ensemble-based approaches for complex microbial metabolic data. The AUC values of both models were greater than 0.99, confirming that our models could capture non-linear relationships and complex feature interactions. Our findings confirm other studies that found that ensemble methods outperform more classical statistical approaches for problems in the fields of plant pathology and soil health assessment [Wang and Zou, 2024]. That consistent prediction performance of SVM and KNN strongly evidence supports microbial metabolic diversity of predictive index may be an essential driver of soil suppressiveness. Although the performance of Naive Bayes is still relatively low, this reinforces the concept that microbial metabolic



	C_2	GR5	B_7	D_4	A_8	D_11	B_5	F_4	C_6	F_1	A_6	C_1
C_2	1.00	0.11	0.38	0.79	0.27	0.52	0.78	0.57	0.40	0.02	0.26	0.74
GR5	0.11	1.00	0.66	0.10	0.28	0.54	0.33	0.24	0.60	0.82	0.49	0.44
B_7	0.38	0.66	1.00	0.49	0.45	0.79	0.67	0.55	0.83	0.55	0.63	0.68
D_4	0.79	0.10	0.49	1.00	0.26	0.59	0.76	0.68	0.46	0.04	0.25	0.74
A_8	0.27	0.28	0.45	0.26	1.00	0.50	0.45	0.17	0.57	0.18	0.38	0.30
D_11	0.52	0.54	0.79	0.59	0.50	1.00	0.81	0.59	0.80	0.50	0.40	0.72
B_5	0.78	0.33	0.67	0.76	0.45	0.81	1.00	0.64	0.66	0.24	0.34	0.83
F_4	0.57	0.24	0.55	0.68	0.17	0.59	0.64	1.00	0.46	0.20	0.24	0.66
C_6	0.40	0.60	0.83	0.46	0.57	0.80	0.66	0.46	1.00	0.49	0.58	0.62
F_1	0.02	0.82	0.55	0.04	0.18	0.50	0.24	0.20	0.49	1.00	0.32	0.35
A_6	0.26	0.49	0.63	0.25	0.38	0.40	0.34	0.24	0.58	0.32	1.00	0.36
C_1	0.74	0.44	0.68	0.74	0.30	0.72	0.83	0.66	0.62	0.35	0.36	1.00

Figure 9. Correlation matrix of input features in soil classification models – insights into redundancy, multicollinearity, and model performance; abbreviations names as shown in Figure 7

interactions are not independent, making probabilistic models less appropriate for this form of data [Li et al., 2023].

A comparison of these two approaches showed that the most important features influencing soil suppressiveness were carbohydrate utilization and amino acid metabolism (SHAP analysis). This observation agrees with the studies that demonstrated that microbial competition for carbon resources increases soil resistance to pathogens [Malik et al., 2020; Xu et al., 2024). This carbohydrate metabolic dominance also suggests the presence of a competitive microbial environment for readily available carbon sources that may inhibit the colonization of *Foa* by creating an antagonistic soil environment. In a similar vein, greater amino acid metabolism may also reflect elevated microbial activity and metabolic flexibility, which in turn may further aid in pathogen suppression.

Classification models used molecular ecological networks of metabolic profiles to separate suppressive and conducive soil types. Suppressive soils exhibited greater microbial diversity, wider niche width, and increased utilization of carbohydrates. These patterns are consistent with the concept of "soil immunity" that suggests diverse microbial communities form ecological barriers that inhibit pathogen establishment [Li et al., 2024; Ding et al., 2022]. The additional evidence of functional redundancy and microbial competition proposed as the mechanisms of pathogen inhibition was further proved by the presence of higher growth rates and total metabolic activity in the suppressive soils [Kinkel et al., 2011].

The correlation analysis showcased additional information on the basic biological mechanisms of soil suppressiveness. Our results showed both a wide breadth of niche width among groups as well as strong positive correlations between carbohydrate utilization patterns, niche width and total growth, further indicating that metabolic versatility facilitates soil resilience. On the other hand, low associations between overall amino acid utilization patterns and pathogen suppression suggest that amino acid metabolism may not be a primary player in disease suppression. These findings conform to previously observed work, where soil suppressiveness was driven by metabolic diversity, rather than the diversity of individual metabolites [Bi et al., 2021].

The present study has proven the feasibility to use machine learning modeling to predict soil suppressiveness directly from information on microbial metabolic activity, thus constituting a useful tool for early disease risk assessment. Ensemble models can capture complex interactions between microbes, which is a key advantage over traditional diagnostic approaches that often depend on identification based on visual symptoms or pathogen isolation [Leslie and Summerell, 2006; Summerell et al., 2003]. The identification of key metabolic pathways linked with pathogen suppression additionally provides practical applications for defining targeted soil health practices [Jagadesh et al., 2024]. For instance, enhancing the soil resilience to Foa and lowering the chemical control requirement are could be accomplished by encouraging beneficial microbe's consortia that benefit carbohydrate and amino acids metabolism [Vishwakarma et al., 2020]. This is in agreement with previous studies [Mendes et al., 2011; Raaijmakers and Mazzola, 2016] that have shown the effects of soil microbial diversity and metabolic activity in soil disease suppression (Table 1). Pseudomonas spp. have been recognized as important indicators of soil suppressiveness based on their ability to produce antibiotics and/or effectively colonize roots [Mendes et al., 2018]. Berendsen et al. [2012] stated that Increased microbial diversity improves the resilience of soil through ecological competition, thereby restricting pathogen establishment. This study build on these results by revealing the metabolic pathways including especially carbohydrate and amino acid metabolism that are critical to soil suppressiveness.

This excellent predictive performance of XGBoost and Gradient Boosting is consistent with previous plant pathology studies showing that ensemble models offer superior predictive performance compared to tree models [Pereyra, 2021]. As previously noted, ensemble methods have a more reliable performance than linear models when predicting plant disease outbreaks because they consideration of complex interactions of features [Wang and Li, 2025]. Here we extend these findings to soil-borne diseases providing evidence that microbial functional diversity measured through redox enzyme activity can be modeled at ensemble scales.

This is a tiny step in the right direction, but does have several issues. However, the data is only for two areas (Marrakech and Zagora), potentially affecting the generalizability of the models. To confirm the robustness of the predictive models, a sampling of soils over multiple regions with diverse soil should be conducted for future studies. Furthermore, the temporal fluctuations of microbial populations, which could also lead to changes in soil suppressiveness happened to be not considered in the existing model. Longitudinal data could enhance predictive performance by capturing the dynamics of microbial interactions that may change according to various environmental conditions and seasons.

There are also limitations in terms of the focus on patterns of metabolic activity, Metabolic profiles are functional, but cannot give the complete picture of microbial community structure and genetic diversity. Metagenomic and metabolomic data are often complementary, and their combined use has the potential to provide a broader view of the mechanisms of soil suppressiveness [Yang et al., 2023]. In addition, machine learning can be computationally expensive, and balancing model complexity with soil health assessment needs to be carefully weighed. XGBoost and Gradient Boosting are ensemble models that consume a lot of resources in decoding the parameters which may restrict their adoptability for large-scale agricultural use [Araújo et al., 2023].

Thus, the incorporation of machine learning into microbial metabolic profiling could have powerful applications to sustainable agriculture. It provides early warning to detect a risk of disease; so farmers could take preventative action before they see any signs of symptoms. Second, it encourages less expensive chemical-based treatments that can disturb beneficial soil microbial communities. Finally, this same set of metabolic markers most likely associated with soil suppressiveness can also guide the formulation of biological applications and soil management options to induce microbial resilience. In the end, ML-based soil diagnostics may help promote scientific technologies for sustainable agricultural practices that enhance soil quality and mitigate crop losses due to soil-borne pathogens [Sharma et al., 2024; Prasad et al., 2024].

CONCLUSIONS

This study indicates the strong predictive skills of machine learning algorithms, especially XGBoost and Gradient Boosting, in classifying soils dependent on their suppressiveness to *Fusarium oxysporum f. sp. albedinis (Foa)*. Remarkably high AUC-ROC values (surpassing 0.99) and balanced precision-recall metrics validate their trustworthiness in distinguishing between soil types. SHAP analysis further recognized key metabolic markers linked to disease suppression, underscoring the importance of microbial diversity, carbohydrate utilization, and practical redundancy in suppressive soils.

The ability to categorize soils based on their metabolic profiles represents a significant stride ahead in sustainable disease administration, offering early identification of Bayoud disease risks while reducing the necessity for chemical treatments. Yet the scope of sampling requires broadening and integration of metagenomic information would strengthen the model's resilience. Accessible AI-driven diagnostic tools for farmers will enrich soil health management and reinforce agricultural adaptability to climate change. Comprehending soil metabolic profiles at many locations and times utilizing metagenomic data could substantially boost the model's skill in forecasting Bayoud disease risk, supporting proactive and targeted administration decisions by farmers to improve yields and bolster ecosystem services.

Acknowledgments

The authors express their sincere gratitude to the National Institute of Agronomic Research (INRA) for providing material, logistical, and human resources throughout this study. Special thanks are extended to the field and laboratory teams for their essential contributions, particularly during field visits to southern Morocco for sample collection and laboratory procedures. This study could not have been accomplished without the support and expertise of all those involved.

REFERENCES

1. Ahmad, F., Ghani Khan, M. U., Tahir, A., Tipu, M. Y., Rabbani, M., Shabbir, M. Z. (2023). Two phase feature-ranking for new soil dataset for *Coxiella burnetii* persistence and classification using machine learning models. *Scientific Reports*, *13*(1), 29. https://doi.org/10.1038/s41598-022-26956-8
2. Alaoui, Y. E. H., Bouda, S., Amar, H. B., Bouhlali, E. D. T., Khoulassa, S., Boutagayout, A., Essarioui, A. (2024). Enhancing date palm (*Phoenix dactylifera* L.) resilience: Strategies against Bayoud spread in Moroccan oases. In *IOP Conference Series: Earth and Environmental Science 1398*(1), 012001. IOP Publishing. https://doi.org/10.1088/1755-1315/1398/1/012001
3. Alaoui, Y. E. H., Bouda, S., Chabaa, S., Alami, M. E., Khoudi, Z., Essarioui, A. (2024). Digital modeling of soil-borne fusarium's nutritional: Investigating microbiological and microecological dynamics in Moroccan agroecosystems. In *ITM Web of Conferences 69*, 01008. EDP Sciences. https://doi.org/10.1051/itmconf/20246901008
4. Araújo, S. O., Peres, R. S., Ramalho, J. C., Lidon, F., Barata, J. (2023). Machine learning applications in agriculture: current trends, challenges, and future perspectives. *Agronomy*, *13*(12), 2976. https://doi.org/10.3390/agronomy13122976
5. Benzohra, I. E., Megateli, M., Berdja, R. (2015). Bayoud disease of date palm in Algeria: History, epidemiology and integrated disease management. *African Journal of Biotechnology*, *14*(7), 542–550.
6. Berendsen, R. L., Pieterse, C. M., Bakker, P. A. (2012). The rhizosphere microbiome and plant health. *Trends in plant science*, *17*(8), 478–486. https://doi.org/10.1016/j.tplants.2012.04.001
7. Bi, B., Wang, K., Zhang, H., Wang, Y., Fei, H., Pan, R., Han, F. (2021). Plants use rhizosphere metabolites to regulate soil microbial diversity. *Land Degradation & Development*, *32*(18), 5267–5280. https://doi.org/10.1002/ldr.4107
8. Cabot, J. H., Ross, E. G. (2023). Evaluating prediction model performance. *Surgery*, *174*(3), 723–726. https://doi.org/10.1080/09599916.2020.1858937
9. Chen, C., Quan, J., Chen, X., Yang, T., Yu, C., Ye, S.,..., Weng, Y. (2024). Explore key genes of Crohn's disease based on glycerophospholipid metabolism: A comprehensive analysis Utilizing Mendelian Randomization, Multi-Omics integration, Machine Learning, and SHAP methodology. *International Immunopharmacology*, *141*, 112905. https://doi.org/10.1016/j.intimp.2024.112905
10. Ding, M., Shang, N., He, Y., Hu, B., Cai, H., Dai, B.,..., Huang, Y. (2022). Long-term continuous-cropping system altered the microbial communities and nutrient cycles in karst agroecosystems. https://doi.org/10.21203/rs.3.rs-2174000/v1
11. El Hilali Alaoui, Y., Bouda, S., Boutagayout, A., Haddioui, A., Ben Amar, H., Bouhlali, E. D. T.,..., Essarioui, A. (2025). Nutrient profiles among soil borne Fusarium species reveal variations in carbon source use across oasis ecosystems in Morocco. *International Journal of Environmental Studies*, 1–18. https://doi.org/10.1080/00207233.2025.2456461
12. El Modafar, C. (2010). Mechanisms of date palm resistance to Bayoud disease: current state of knowledge and research prospects. *Physiological and Molecular Plant Pathology*, *74*(5–6), 287–294. https://doi.org/10.1016/j.pmpp.2010.06.008
13. Essarioui, A., Houria, B.-A., Khoulassa, S., *et al.* (2018). Gestion du Bayoud du palmier dattier dans les oasis marocaines. *Revue Marocaine des Sciences Agronomiques et Vétérinaires*, *6*(4), 537–543.

14. Hassan, S. M. (2018). Management of diseases of date palm. *Date Palm Pests and Diseases*, *139*. https://doi.org/10.1007/978-94-017-9694-1_8
15. Jagadesh, M., Dash, M., Kumari, A., Singh, S. K., Verma, K. K., Kumar, P.,..., Sharma, S. K. (2024). Revealing the hidden world of soil microbes: metagenomic insights into plant, bacteria, and fungi interactions for sustainable agriculture and ecosystem restoration. *Microbiological Research*, 127764. https://doi.org/10.1016/j.micres.2024.127764
16. Kinkel, L. L., Bakker, M. G., Schlatter, D. C. (2011). A coevolutionary framework for managing disease-suppressive soils. *Annual review of phytopathology*, *49*(1), 47–67. https://doi.org/10.1146/annurev-phyto-072910-095232
17. Leslie, J. F., Summerell, B. A. (2006). Fusarium laboratory workshops—A recent history. *Mycotoxin Research*, *22*(2), 73–74. https://doi.org/10.1007/bf02956766
18. Li, J., Zhang, Y., He, S., Tang, Y. (2024). Interpretable mortality prediction model for ICU patients with pneumonia: using shapley additive explanation method. *BMC Pulmonary Medicine*, *24*(1), 447. https://doi.org/10.1186/s12890-024-03252-x
19. Li, X., Zheng, X., Yadav, N., Saha, S., Salama, E. S., Li, X.,..., Jeon, B. H. (2024). Rational management of the plant microbiome for the Second Green Revolution. *Plant Communications*, *5*(4). https://doi.org/10.1016/j.xplc.2024.100812
20. Li, Z., Selim, A., Kuehn, S. (2023). Statistical prediction of microbial metabolic traits from genomes. *PLoS Computational Biology*, *19*(12), e1011705. https://doi.org/10.1371/journal.pcbi.1011705
21. Malik, A. A., Martiny, J. B., Brodie, E. L., Martiny, A. C., Treseder, K. K., Allison, S. D. (2020). Defining trait-based microbial strategies with consequences for soil carbon cycling under climate change. *The ISME journal*, *14*(1), 1–9. https://doi.org/10.1038/s41396-019-0510-0
22. Mendes, L.W., Raaijmakers, J.M., de Hollander, M., Mendes, R., Tsai, S.M. (2018). Influence of resistance breeding in common bean on rhizosphere microbiome composition and function. *The ISME journal*, *12*(1), 212–224. https://doi.org/10.1038/ismej.2017.158
23. Mohseni, P., Ghorbani, A. (2024). Exploring the synergy of artificial intelligence in microbiology: Advancements, challenges, and future prospects. *Computational and Structural Biotechnology Reports*, 100005. https://doi.org/10.1016/j.csbr.2024.100005
24. Nemer, Z.N., Raheem, S.F., Alabbas, M. (2024). Comparison of classification of different machine learning algorithms in the diagnosis and detect of diabetes. *International Journal of Computing and Digital Systems*, *16*(1), 1–10.
25. Otchere, D.A., Ganat, T.O.A., Ojero, J.O., Tackie-Otoo, B.N., Taki, M.Y. (2022). Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions. *Journal of Petroleum Science and Engineering*, *208*, 109244. https://doi.org/10.1016/j.petrol.2021.109244
26. Ou, Y., Penton, C.R., Geisen, S., Shen, Z., Sun, Y., Lv, N.,..., Shen, Q. (2019). Deciphering underlying drivers of disease suppressiveness against pathogenic *Fusarium oxysporum*. *Frontiers in microbiology*, *10*, 2535. https://doi.org/10.3389/fmicb.2019.02535
27. Pereyra, M. (2021). *Evaluating the Effects of Antagonistic Interactions on Pathogen Inhibition by Streptomyces Isolates from a Disease Suppressive Soil* (Master's thesis, University of Minnesota).
28. Philippot, L., Chenu, C., Kappler, A., Rillig, M. C., Fierer, N. (2024). The interplay between microbial communities and soil properties. *Nature Reviews Microbiology*, *22*(4), 226–239. https://doi.org/10.1038/s41579-023-00980-5
29. Prasad, S. R., Thyagaraju, G. S. (2024). Leaf analysis based early plant disease detection using Internet of Things, Machine Learning and Deep Learning: A comprehensive review. *Journal of Integrated Science and Technology*, *12*(2), 734–734. https://doi.org/10.62110/sciencein.jist.2025.v13.1015
30. Sedra, M. H. (1989). Resistance des sols aux maladies : mise en evidence de la resistance d'un sol de la palmeraie de Marrakech aux fusarioses vasculaires. *Awamia*.
31. Sharma, P., Sharma, P., Thakur, N. (2024). Sustainable farming practices and soil health: A pathway to achieving SDGs and future prospects. *Discover Sustainability*, *5*(1), 250. https://doi.org/10.1007/s43621-024-00447-4
32. Steurer, M., Hill, R. J., Pfeifer, N. (2021). Metrics for evaluating the performance of machine learning based automated valuation models. *Journal of Property Research*, *38*(2), 99–129.
33. Summerell, B.A., Salleh, B., Leslie, J.F. (2003). A utilitarian approach to Fusarium identification. *Plant disease*, *87*(2), 117–128. https://doi.org/10.1094/pdis.2003.87.2.117
34. Todorović, I., Moënne-Loccoz, Y., Raičević, V., Jovičić-Petrović, J., Muller, D. (2023). Microbial diversity in soils suppressive to Fusarium diseases. *Frontiers in Plant Science*, *14*, 1228749. https://doi.org/10.3389/fpls.2023.1228749
35. Trivedi, P., Leach, J. E., Tringe, S. G., Sa, T., Singh, B. K. (2020). Plant–microbiome interactions: from community assembly to plant health. *Nature Reviews Microbiology*, *18*(11), 607–621. https://doi.org/10.1038/s41579-020-0412-1

36. Vishwakarma, K., Kumar, N., Shandilya, C., Mohapatra, S., Bhayana, S., Varma, A. (2020). Revisiting plant–microbe interactions and microbial consortia application for enhancing sustainable agriculture: a review. *Frontiers in Microbiology*, *11*, 560406. https://doi.org/10.3389/fmicb.2020.560406
37. Wang, L., Yin, H., Li, Y., Yang, Z., Wang, Y., Liu, X. (2024). Prediction of microbial activity and abundance using interpretable machine learning models in the hyporheic zone of effluent-dominated receiving rivers. *Journal of Environmental Management*, *357*, 120627. https://doi.org/10.2139/ssrn.4676862
38. Wang, M., Li, T. (2025). Pest and disease prediction and management for sugarcane using a hybrid autoregressive integrated moving average—a long short-term memory model. *Agriculture*, *15*(5), 500. https://doi.org/10.3390/agriculture15050500
39. Wang, Y., Zou, Q. (2024). Deep learning meta-analysis for predicting plant soil-borne fungal disease occurrence from soil microbiome data. *Applied Soil Ecology*, *202*, 105532. https://doi.org/10.1016/j.apsoil.2024.105532
40. Xu, Q., Zhang, H., Vandenkoornhuyse, P., Guo, S., Kuzyakov, Y., Shen, Q., Ling, N. (2024). Carbon starvation raises capacities in bacterial antibiotic resistance and viral auxiliary carbon metabolism in soils. *Proceedings of the National Academy of Sciences*, *121*(16), e2318160121. https://doi.org/10.1073/pnas.2318160121
41. Yang, J., He, J., Jia, L., Gu, H. (2023). Integrating metagenomics and metabolomics to study the response of microbiota in black soil degradation. *Science of the Total Environment*, *899*, 165486. https://doi.org/10.1016/j.scitotenv.2023.165486