Journal of Autonomous Intelligence (2024) Volume 7 Issue 5
doi: 10.32629/jai.v7i5.1466

ORIGINAL RESEARCH ARTICLE

Comparative analysis of machine learning classification algorithms for predicting olive anthracnose disease

Klimentia Kottaridi[1,*], Anna Milionis[1], Vasilis Demopoulos[1], Vasileios Nikolaidis[2], Polina C. Tsalgatidou[1], Athanasios Tsafouros[1], Anastasios Kotsiras[1], Alexandros Vithoulkas[1]

[1] *Department of Agriculture, University of the Peloponnese, 24150 Kalamata, Greece*

[2] *Accounting and Finance Department, University of the Peloponnese, 24150 Kalamata, Greece*

*** Corresponding author:** Klimentia Kottaridi, k.kottaridi@go.uop.gr

ABSTRACT

Olive anthracnose (OA) is the most damaging fungal disease of the olive tree worldwide. In the context of integrated pest management, the development of predictive models could be used for early diagnosis and control. In the current study, a dataset consisting of 58 cases, coming from 5 locations and 12 olive cultivars, was used to study the relationship between OA incidence (OAI) and 35 heterogeneous variables. These variables include orchard characteristics, olive fruit parameters, foliar and soil nutrients, soil parameters and soil texture classes. The Random Forest-Recursive Feature Elimination with Cross Validation (RF-RFECV) feature selection method identified Location, water content, P, Ca, Mg, exchangeable Mg, trace Zn, trace Cu as possible new indicators associated with OAI. The objective of this study was to investigate whether these variables have a predictive value for OAI. Six different machine learning classification algorithms, namely decision tree (DT), gradient boosting (GB), logistic regression (LR), random forest (RF), k-nearest neighbors (KNN) and support vector machine (SVM), were developed for predicting conditions leading to OAI > 0% and 10%. Grid search hyperparameter optimization was employed to optimize model parameters. The final models were evaluated in terms of several standard metrics, such as accuracy, sensitivity, specificity and ROC AUC score. Findings suggested that GB performance was superior compared to the other models for the prediction of the occurrence of OA disease (OAI > 0%) with an accuracy of 86.7%, a sensitivity of 100%, a specificity of 75% and a ROC-AUC score of 93%, while for the prediction of the spread of the disease (OAI > 10%), DT stood out with an accuracy of 86.7%, a sensitivity of 81.8%, a specificity of 100% and a ROC-AUC score of 91%.

Keywords: olive anthracnose; machine learning; forecast models; classification algorithms; soil nutrients

ARTICLE INFO

Received: 4 December 2023
Accepted: 2 January 2024
Available online: 8 March 2024

1. Introduction

Olive anthracnose (OA) caused by *Colletotrichum* species is a major fungal disease in olive oil producing countries, including Greece[1]. The first OA incidence in Greece was reported in 1920 in Corfu. In recent years, OA has become a grave concern for Greek olive oil production[2]. The most affected regions are the Peloponnese and Crete. In fact, in the harvest year 2022–2023, the incidence was so high in the Messinia region of the Peloponnese that olive mills often refused to process the damaged olive fruit, many shutting down their operations a month earlier than usual[3]. It is estimated that OA inflicts an annual loss of 300 million euros upon the olive oil sector in Greece[2]. Olive Anthracnose has a detrimental effect on the quality of olive oil as it affects its physicochemical and sensory properties[4]. A strong positive linear relationship has been found between OA incidence and acidity[5]. Furthermore, sensory

defects are present even at a low level of disease incidence[6]. The disease cycle begins with the infection of inflorescences and developing fruit through water-splashed conidia during the spring and summer seasons[7]. The infections in the developing fruit remain dormant until the fruit reaches maturity stage in autumn and winter. Severe infections can lead to rapid rotting and mummification of fruits. If mummified fruits over winter on the tree, they serve as a source of infection for new outbreaks. Olive anthracnose incidence depends on factors including cultivar susceptibility[8], environmental conditions and the virulence of the pathogen[9]. A significant increase in OA is anticipated when warm and moist conditions coincide with ripened fruit of susceptible olive cultivars[10]. Multiple *Colletotrichum* species also make disease control more challenging, as one or more species may be present in infected orchards[11]. Finally, the ability to persist and multiply without exhibiting noticeable symptoms may explain why anthracnose fungi often result in unforeseen losses in olive crops[12]. The complexity of anthracnose epidemiology highlights the necessity of ongoing research into disease management.

Many researchers[13–16] have emphasized the importance of cultural management practices for the control of OA. These practices include irrigation, sanitation, pruning and balanced nutrition. It is well known that plants suffering from nutrient stress are more susceptible to pests and diseases[16]. However, while balanced nutrition is recommended as a cultural practice, there is limited research on the role of soil and foliar amendments between fruit set and harvest as a control strategy for OA[15]. The objective of this study is to investigate whether soil and foliar parameters have an effect on the incidence of OA and thus, can be used for the early detection of the disease. To this end, we develop and compare predictive models based on high-dimensional and heterogeneous data.

Our dataset includes soil and foliar nutrients in combination with soil characteristics, the location of the orchard, olive cultivar, fruit maturity index and water content of fruit. Providing a data-driven tool can enable olive oil growers and agronomists to make better informed decisions regarding control strategies and preventative measures against olive anthracnose. In this study, the Random Forest-Recursive Feature Elimination with Cross Validation (RF-RFECV) method is used to select the important features from the original dataset. Six different classification algorithms, including decision tree (DT), gradient boosting (GB), logistic regression (LR), random forest (RF), k-nearest neighbors (KNN) and support vector machine (SVM), are developed for predicting conditions leading to OAI > 0% and 10%. Grid search hyperparameter optimization was employed to optimize model parameters. The final models are evaluated in terms of several standard metrics, such as accuracy, sensitivity, specificity and ROC AUC score.

1.1. Related work

The typical methods for diagnosing a plant disease are often untimely and may not always produce reliable results. The adoption of advanced technologies like machine learning (ML) and deep learning (DL) can address these challenges and facilitate early detection[17]. Over the last decade, there has been an increase in the number of studies relevant to this field. These studies can be divided into three categories of forecast models based on[18]: 1) image processing, 2) weather data and, 3) distinct types of data coming from heterogeneous sources. Most research is either image or weather based, with very few studies belonging to the third category.

Studies in the first category include DL forecast models based on image analysis of symptomatic fruit. Alruwaili et al.[19] proposed an enhanced convolutional neural network for the early detection and classification of 14 olive diseases (including anthracnose), based on 120 RGB images of plant leaves selected from the PlantVillage dataset, achieving an overall accuracy of 99.11%. Their model was successful in predicting anthracnose disease scoring 100% at all performance metrics (accuracy, precision, recall, F1-measure, sensitivity and specificity). Fazari et al.[20] used hyperspectral images and advanced modelling techniques of deep learning and convolutional neural networks to detect OA disease in early stages. They

achieved a very high sensitivity ratio, ranging from 85% to 100% depending on the disease progress stage after inoculation.

The research of Alves et al.[21] is among the studies that have based their models on weather data. They demonstrated that random forest had the best overall accuracy of 99.8%, compared to other classification algorithms (IBK, Naïve Bayes and SMO), when applied to predict olive anthracnose. The dataset included 2800 instances, of which 330 instances had the disease and 2470 did not. In another research, Romero et al.[13] incorporated 12-year weather data with the levels of susceptibility of nine cultivars. They developed three binary logistic models for predicting conditions leading to OAI > 0, 1 and 5% with overall accuracy of 81%, 86%, and 85%, respectively.

One study that has used heterogenous data for disease prediction is by Olivares et al.[22]. They applied the supervised machine learning methods of Orthogonal Least Squares Discriminant Analysis and Random Forest to identify the soil properties potentially associated with Banana Wilt disease incidence in Venezuela. Their dataset included 16 soil variables of 78 soil samples, 29 with low incidence and 49 with high incidence of the disease. The analysis of the receiver operating characteristics curves by random forest revealed that the combination of Zn, Fe, Ca, K, Mn and clay was able to accurately differentiate 84.1% of the banana lots with 89.80% sensitivity and specificity of 72.40%.

To our knowledge, no research has investigated whether soil nutrients are potentially associated and can be included in a predictive model for OA incidence.

2 Materials and methods

2.1. Field design

This dataset is part of a larger EU-funded research project on olive anthracnose in the region of the Peloponnese. The administrative region of the Peloponnese includes five prefectures, namely Messinia, Laconia, Argolida, Corinthia, Arcadia. The number of olive groves participating in EU-funded project was determined according to the volume of the olive oil production per prefecture. Twenty-six olive orchards were from Messinia, 19 from Laconia, 6 from Argolida, 5 from Corinthia and 2 from Arcadia. All orchards were mono-cultivars and privately-owned by single producers. The olive varieties included ten Greek ones (Koroneiki, Megaritiki, Kalamon, Manaki, Mavrolia, Asprolia, Myrtolia, Koutsourelia, Athenolia and Nemoutiana) and two Spanish (Arbequina and Picual). The collection of olive fruit was carried out during the harvest period from October 2021 to January 2022 and the maturity index of each sample was calculated immediately at the time of receipt[23].

2.2. Disease assessment

Detection of latent anthracnose disease infection was conducted on asymptomatic and macroscopically healthy detached olive drupes. Olive fruits were washed under running tap water, surface sterilized by immersion in a 5% solution of sodium hypochlorite (bleach) for 20 min, rinsed five times with sterile water and air-dried for 1 h in a laminar cabinet before wounded with a sterile needle. An aliquot of 10 μL of sterilized distilled water was inoculated on the surface of each artificial wound. Olives from each sample were then transferred into plastic containers to maintain high relative humidity and stored in a well-ventilated cabinet at 25 °C for 6 days. A completely randomized design with three replicates per treatment and 20 fruit per replicate was used.

After six days of incubation, the number of infected olive fruit was recorded, and disease incidence (OAI) was calculated according to Equation (1)[24].

$$OAI(\%) = \frac{Number\ of\ infected\ olive\ fruits}{Total\ number\ of\ olive\ fruits} \times 100 \quad (1)$$

Fruits were considered affected by *Colletotrichum* spp. when typical symptoms of anthracnose disease appeared like round and ocher or brown lesions, with profuse production of orange masses of conidia or fruit rot. The average of the three replicates was used to calculate the OAI (%) per orchard.

2.3. Soil sampling and analysis

Soil samples were collected from fifty eight olive orchards from the prefectures of Messinia, Laconia, Argolida, Corinthia and Arcadia of the Peloponnese region. Each soil sample was taken from the zone of maximum root activity, about 25 to 40 cm deep. The density of subsampling was 1–2 points per approximately 1000 m^2. The sampling points were random, and samples were taken by pressing a hand auger combination type (Eijkelkamp, the Netherlands) into the soil. Dry leaves, stems and other vegetal residuals on the soil surface were removed prior to sampling. Every sampling area contained similar soil types with trees of roughly uniform size and vigor. Subsamples of each orchard were thoroughly mixed in a plastic bucket in order to form a composite sample, which was then placed into a labeled bag until determination in the laboratory.

Soil samples were dried using forced air at ambient temperatures < 36 °C to constant weight and then passed through a 2 mm sieve (fine earth). Samples were saturated with deionized water and saturation percentage was determined[25]. Values of pH were measured in the soil/water slurry[26] using a Consort C835 multichannel analyzer. The exchangeable cations (Ca, Mg, K) were extracted with a 1 M NH4OAc solution at pH 7.00[27] and their concentration was determined by a Shimadzu AA6200 atomic absorption spectrophotometer in an air-acetylene flame. Calcium and magnesium were measured by adding La_2O_3 to both the standards and sample extraction to reach a concentration of 4500 mg L^{-1} La[28]. Phosphorus was determined colorimetrically using a Shimadzu UV-1700 UV-visible spectrophotometer according to the Olsen method[29]. Organic matter concentration was measured according to the Walkley-Black method[30]. The particle size analysis (sand, silt and clay) was performed by the hydrometer method[31].

2.4. Leaf sampling and analysis

A sample of approximately 300 leaves per orchard were collected in July 2022. Each sample was comprised of mature healthy leaves from the middle portion of nonbearing current season shoots, selected randomly and peripherally from the tree. The leaves were placed in paper bags, stored in a portable ice cooler, and transported to the laboratory.

Once in the laboratory, the leaves were pulverized in a grinder and 1 g of each sample was heated in a dry oven at 550 °C for 4 h in porcelain stoneware. The inorganic elements were extracted using 15 mL of 10% HCl solution and distilled water was added to up to 100 mL. The foliar nutrients Ca, Mg, K, Fe, Mn, Zn and Cu were determined using an atomic absorbance spectrophotometer Shimadzu AA6200. Total P[32] and B[33] were determined colorimetrically.

2.5. Dataset

The dataset had a total of 58 cases and 35 features, including numeric and categorical. The predictive targets were a) the occurrence of OA disease (OAI = 0%, OAI > 0%), where 0 indicated no occurrence (OAI = 0%) and 1 indicated occurrence (OAI > 0%) and b) the incidence of OA (OAI < 10%, OAI > 10%), where 0 indicated disease incidence lower than 10% and 1 greater than 10%.

The predictor variables (**Table 1**) included olive orchard characteristics (olive cultivar, location of the orchard), olive fruit parameters (olive fruit maturity index, water content of olive fruit), foliar nutrients (total N, P, Ca, Mg, K, Fe, Mn, Zn, Cu and B), soil parameters (saturation percentage, pH, electrical conductivity, organic matter, Olsen P), soil macronutrients, divided into exchangeable cations (Ca, Mg, K, Na) and water-soluble cations (water soluble Mg, water soluble K), soil micronutrients or trace elements (B, Fe, Mn, Zn, Cu) and soil texture indicators (sand, silt, slay and soil textural class).

Table 1. Predictor variables used in this study.

Variable	Type
Olive orchard characteristics	
Olive cultivar	Categorical
Location	Categorical
Olive fruit parameters	
Maturity index (%)	Numerical
Water content (%)	Numerical
Foliar nutrients	
Total N (%)	Numerical
P (%)	Numerical
Ca (%)	Numerical
Mg (%)	Numerical
K (%)	Numerical
Fe (ppm)	Numerical
Mn (ppm)	Numerical
Zn (ppm)	Numerical
Cu (ppm)	Numerical
B (ppm)	Numerical
Soil parameters	
SP (%)	Numerical
pH (0–14)	Numerical
EC (mS/cm)	Numerical
OM (%)	Numerical
P	Numerical
Soil macronutrients	
Exchangeable cations (ppm) mg/kg	
Ca	Numerical
Mg	Numerical
K	Numerical
Na	Numerical
Water soluble elements (ppm) mg/L	
Mg	Numerical
K	Numerical
Soil micronutrients (trace elements) (ppm) mg/kg	
B	Numerical
Fe	Numerical
Mn	Numerical
Zn	Numerical
Cu	Numerical
Soil texture indicators (%)	
Sand	Numerical
Clay	Numerical
Silt	Numerical
Soil textural class	Categorical

2.6. Data preprocessing and feature selection

The dataset did not contain any missing values or user entry errors, so no imputation or data cleaning techniques were needed. Data scaling was applied to the numerical input features by rescaling the distribution of the values so that the mean of observed values was 0 and the standard deviation was 1. One-hot encoding was used to convert categorical variables into a format that could be readily used by the machine learning algorithms[34].

To overcome the problems associated with the high dimensionality and the multicollinearity between variables, we reduced the number of features of the original dataset by employing the Random Forest-Recursive Feature Elimination with Cross Validation (RF-RFECV) method[35].

Random forest is an ensemble method that combines multiple decision trees to create a robust model. RF typically performs well with high dimensional and heterogeneous data and can identify significant predictors without making assumptions about an underlying model. According to Reif et al.[36], the recursive partitioning process of random forests allows them to capture complex interactions between features. The trees in the ensemble consider multiple attributes simultaneously and identify interactions that may not be apparent in isolated feature evaluations.

However, the presence of correlated predictors, which is a common problem of high-dimensional data sets, impacts RF's ability to identify the strongest predictors by decreasing the estimated importance scores of correlated variables. A suggested solution is the RF-RFECV algorithm which was first developed for the gene selection process using the SVM classifier[37]. RF-RFECV utilizes the inherent feature importance scores from Random Forest to rank features based on their contribution to the model's accuracy.

RFE is a wrapper-type feature selection method[38] which follows a greedy optimization approach to find a subset of features by first looking at all features in the training dataset and then successfully removing features one at a time until the appropriate number of features is left. This is accomplished by first fitting the core model's machine learning algorithm, then ranking the features according to relevance, eliminating the least important features, and finally re-fitting the model. This process is repeated until a specified number of features remains.

To find the optimal number of features 5-fold cross-validation was used with RFE to score different feature subsets and select the best scoring collection of features. To avoid overfitting and biased performance estimations due to data leakage, feature selection was only performed on the training data and not the complete dataset[39].

2.7. Proposed methodology

The aim of the current study was to optimize six classification machine learning algorithms, namely decision tree (DT), gradient boosting (GB), logistic regression (LR), random forest (RF), k-nearest neighbors (KNN) and support vector machine (SVM), to develop prediction models for the occurrence of OA disease (OAI > 0%) and its incidence level (OAI > 10%).

To evaluate a model's performance, some data (input) with known ground truth (labels) are required. In our case these labels were the values of the two binary target variables, (a) OA occurrence (0: OAI = 0%, 1: OAI > 0%), and (b) OA incidence (0: OAI < 10%, 1: OAI > 10%). The idea was to train the models on the data, for which the labels were known, and evaluate their performance on data, for which labels were unknown (unseen data to the model)[40]. For this purpose, the new dataset, after data preprocessing and feature selection, was split into 75% training data (known labels) to train the models and 25% testing data (unknown labels) to evaluate the models.

The approach involved training the models on data with known labels and evaluating them on unseen data, where labels were unknown. Following data preprocessing and feature selection, the new dataset was

divided into 75% training data (with known labels) for model training and 25% testing data (with unknown labels) for model evaluation.

Grid search with 5-fold cross validation was applied on the 75% training data for hyperparameter tuning and model selection among the six candidate models. The values of the hyperparameters, which are the parameters that control the model's learning process, have a significant impact on the predictive performance of the machine learning model[41]. The hyperparameters of the models (e.g., number of features to consider when looking for the best split for gradient boosting, number of trees in the forest for random forest, etc.) were optimized through an internal 5-fold cross-validation by grid search over a range of user-specified values and the parameters that generated the best accuracy score were selected. GridSearchCV function that comes in the Scikit-learn's model_selection package was used.

Standard 5-fold cross-validation was employed to address the overfitting issue, deal with the small sample size and increase the precision of the estimates, while still maintaining a small bias[42]. The 5-fold cross-validation was performed in the following steps: (a) the training dataset was split into 5 equal parts (folds). (b) 4 parts were used to train the model and the remaining one part to validate the model, (c) step (b) was repeated until each part was used for both the training and validation set, and (d) the performance of the model was finally computed as the average performance of the 5 estimations.

The hyperparameter setting achieving the highest 5-fold cross-validation score for each of the six learning algorithms (DT, GB, LR, RF, KNN, SVM) was used to develop the final prediction models. The final models were fitted to the entire training dataset (75% of the original data) and then tested on the held-out 25% of the data.

This methodology is summarized into 4 steps (**Figure 1**)[43]: (Step1): the original dataset was divided into a training set and an independent test set, with the test set being saved for the final model evaluation step. (Step 2): grid search was used in the second step to experiment with different hyperparameter settings. 5-fold cross-validation was employed on the training set to generate several models and performance estimates for each hyperparameter configuration. (Step 3): the entire training set was used for model fitting by selecting the hyperparameter values that corresponded to the best-performing model. (Step 4): the independent test set that was withheld in step 1 was used to evaluate the model that was obtained from step 3.

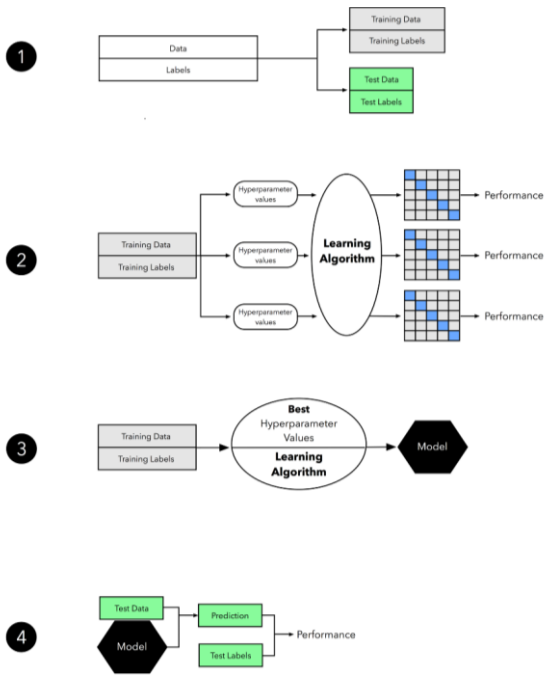

Figure 1. This image depicts model selection using grid search hyperparameter optimization with 5-fold cross-validation[44].

2.8. Machine learning models

Following the data collection and preprocessing steps, six machine learning algorithms (DT, GB, LR, RF, KNN, SVM) were developed. All machine learning algorithms were run by the open-source Jupyter Notebook App in python 3.9.12.

Decision tree (DT)[45] is a non-parametric supervised learning method used both for classification and regression. Classification trees are generally applied to output variables which are categorical and mostly binary in nature. The objective is to learn straightforward decision rules derived from the data features in order to build a model that predicts the value of the target variable. Three different node types—a root node, a child node, and a leaf node—make up the tree. The procedure begins by selecting a root node from the relationships between each input and output variable that are the strongest. The selection of a child node is then made by computing Information Gain (IG), which is given by Equation (2).

$$IG(parent, child) = Entropy(parent) - [p(c_1) \times Entropy(c_1) + p(c_2) \times Entropy(c_2) \dots] \quad (2)$$

where $\text{Entropy}(c_i) = -p(c_i) \times \log p(c_i)$ and $p(c_i)$ is a probability of child node *i*. The parent for the following generation will then be the node with the highest IG. The process will continue until all children nodes are pure, or until the IG is 0.

Gradient boosting (GB)[46] is an ensemble algorithm—a combination of weak individual models that together create a more powerful new model—based on decision trees, that is used in both regression and classification tasks. It is one of the most powerful algorithms in the field of machine learning because of its high prediction speed and accuracy. The weak learners are the individual decision trees which are connected in series and each tree tries to minimize the error of the previous tree. Boosting focuses on building up these weak learners successively and removing the observations that a learner correctly understands at each level. In essence, the emphasis is on creating new weak learners to manage the difficult observations at each step. The objective of the GB algorithm is to minimize the loss function i.e., the difference between the actual class and the predicted class, by using a gradient descent procedure. Classification algorithms frequently use logarithmic loss function whereas regression algorithms use squared errors.

Random forest (RF)[47] is an ensemble supervised machine learning algorithm that is used widely in both classification and regression problems. The random forest classifier creates a set of decision trees from a randomly selected subset of the training set and then it collects the votes from different decision trees to decide the final prediction. Basically, each model is trained independently, and the final output is based on majority voting after combining the results of all models. By developing several decision-tree models, RF takes use of the decision tree algorithm's great speed and accuracy while dealing with classification problems. There is no link between the multiple decision trees, and errors are mutually reduced, leading to more precise and reliable classification findings.

K-nearest neighbors (KNN)[48] is a non-parametric lazy learning algorithm that works well with nonlinear data since it makes no assumptions about the input. It is a simple and easy to implement supervised machine learning algorithm that can be used to address both classification and regression tasks. KNN Classifier tries to predict the correct class for the test data by calculating the distance between the test data and all the training points. It finds the nearest neighbors ranking points by increasing distance and finally votes on the predicted class labels based on the classes of the k nearest neighbors. The distance function and the value of k are the only two parameters necessary to implement KNN. The most common distance function that is used to measure similarity is the Euclidian distance and it is defined by Equation (3).

$$d_{Euclidean}(x, y) = \sqrt{\sum_i (x_i - y_i)^2} \quad (3)$$

Logistic regression (LR)[49] is a supervised learning classification algorithm used to predict the

probability of a target variable. The nature of the target variable is dichotomous, which means there can be only two possible classes (i.e., 0: uninfected, 1: infected). The function used by logistic regression to map predictions to probabilities is the sigmoid function (Equations (4) and (5)).

$$\sigma(y) = \frac{1}{(1 + e^{-y})} \tag{4}$$

$$y = b_0 x_0 + b_1 x_1 \ldots + b_n x_n \tag{5}$$

where $(x_0, x_1, \ldots, x_n)$ is an instance of the dataset and bi are the coefficients values, which are estimated and updated by stochastic gradient descent. The sigmoid function returns a probability value between 0 and 1. In order to map this probability value to a discrete class (0/1, true/false), a threshold value, called 'decision boundary' is selected. The probability values above this threshold level are mapped into class 1 and below are mapped into class 0. Generally, the decision boundary is set to 0.5.

Support vector machine (SVM)[50] is a supervised machine learning algorithm which is used in both regression and classification tasks. However, it is mostly employed to solve classification problems. The SVM algorithm's objective is to establish the decision boundary (hyperplane) that can divide a n-dimensional space into classes, allowing new data points to be easily and correctly classified. SVMs are effective in high dimensional spaces as well as in cases where the number of dimensions is greater than the number of samples. SVM algorithms use a set of mathematical functions (kernels) to transform data input into the required form. Gaussian radial basis function, linear, sigmoid and polynomial are several common kernel functions. Besides the kernel function, another important hyperparameter of SVM is the penalty parameter C which adds a penalty for each misclassified data and trades off correct classification of training examples against maximization of the decision function's margin.

2.9. Performance evaluation

The considered classification models were evaluated by calculating the metrics: the number of correctly recognized class examples (TP: true positives), the number of correctly recognized examples that do not belong to the class (TN: true negatives), and examples that either were incorrectly assigned to the class (FP: false positives) or were not recognized as class examples (FN: false negatives). Based on these metrics, the following performance measures were used to evaluate the classification models[51]:

Accuracy is the ratio of the number of correct predictions to the total number of input samples and reflects the overall effectiveness of a classifier (Equation (6)).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{6}$$

Specificity is the ratio of the correctly classified negative samples to the total number of negative samples and describes the effectiveness of a classifier to identify negative labels (Equation (7)).

$$Specificity = \frac{TN}{TN + FP} \tag{7}$$

Sensitivity is the ratio of the correctly classified positive samples to the total number of positive samples and indicates the effectiveness of a classifier to identify positive labels (Equation (8)).

$$Sensitivity = \frac{TP}{TP + FN} \tag{8}$$

While sensitivity and specificity are both important metrics in evaluating the performance of machine learning models, they represent different aspects of the model's accuracy. As sensitivity increases, specificity decreases, and vice versa. This implies that both measurements cannot be optimized at the same time. However, to choose the optimum machine learning model, it is critical to consider both sensitivity and specificity. Depending on the task at hand, one measure could be more crucial than another. In disease diagnosis, which is our case as well, it may be more important to have high sensitivity to avoid missing any

true positive cases, even if it means a higher rate of false positives[52].

In order to evaluate the performance of our machine learning models, we utilized additional common metrics known as receiver operating characteristic (ROC) curve[53] and area under the ROC curve (AUC)[54]. These metrics are widely used in the field of machine learning to assess the effectiveness of classifiers. The ROC curve illustrates the balance between sensitivity (the ability to correctly identify positive cases) and specificity (the ability to correctly identify negative cases). AUC, on the other hand, condenses the ROC curve into a single value, by measuring the entire two-dimensional area underneath the ROC curve. AUC values indicate the overall ability of the model to discriminate between classes, and range from 0 to 1, where 1 is a perfect score and 0.5 means the model is as good as random.

To deal with the small amount of data available in the current study, we employed permutation tests, a statistical method assessing classifier competence beyond accuracy[55]. These tests determine the likelihood of the observed statistic (e.g., accuracy) occurring by chance. Permutation tests assume independence between features and labels, reshuffling labels to create a null distribution. Its p-value represents the fraction of random datasets where the classifier performed as well as or better than in the original data[56]. Our study employed permutation tests on the training data to ascertain whether models with optimized hyperparameters truly captured underlying class structures, establishing a genuine connection between the data and class labels.

Following hyperparameter optimization, the final models with the best hyperparameters were fitted to the entire training dataset and evaluated on the hold-out set by measuring accuracy, specificity, sensitivity and AUC scores.

3. Results

3.1. Statistical analysis on the initial dataset

A quick overview of the dispersion and central tendency of the OA incidence (OAI) raw data is provided by the frequency distribution histogram in **Figure 2**.

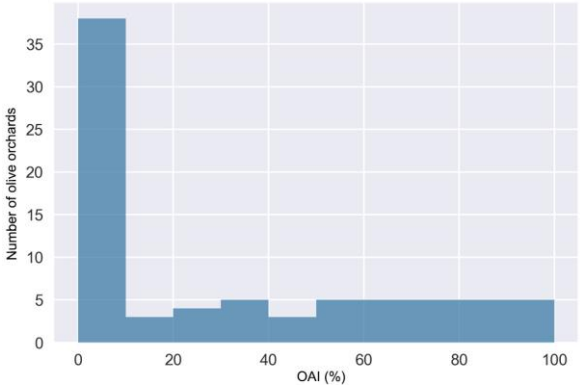

Figure 2. The frequency distribution histogram of OAI.

A relatively balanced class distribution was observed in the data, with around 47% of the total cases having OA disease (OAI > 0%) and roughly 34% having an OAI of more than 10%.

A correlation heatmap **(Figure 3)** was plotted to visualize the strength of the relationships between numerical variables. From the color-coding of the cells, it is obvious that variables such as exchangeable Ca & pH, organic matter & SP, exchangeable Na & EC, water soluble Mg & EC had strong positive correlation while variables such as trace Fe & pH and silt & sand had strong negative correlations. Relative strong or medium correlations also existed between other variables in the dataset.

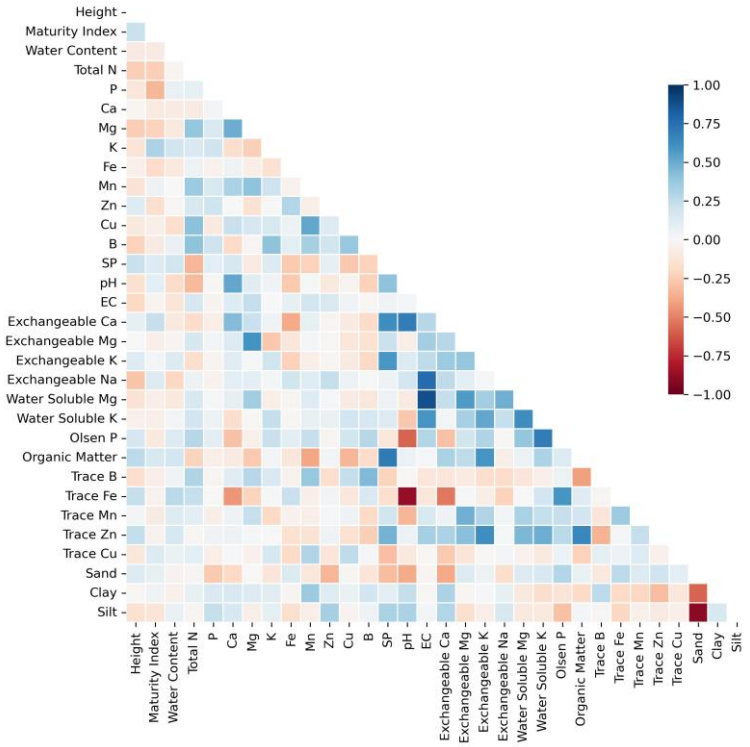

Figure 3. Heatmap correlation matrix of numeric features.

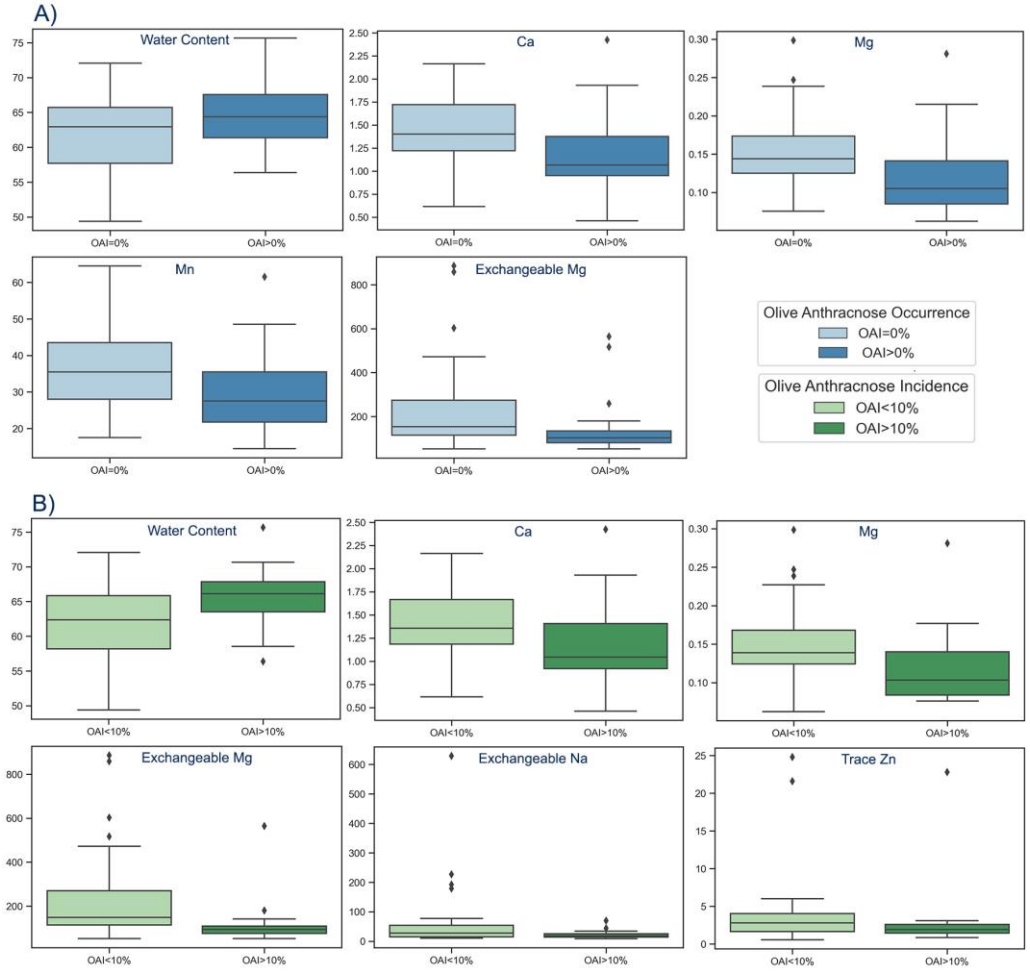

Figure 4. Boxplots visualizing statistically significant differences in **(A)** water content, Ca, Mg, Mn and exchangeable Mg between infected (OAI > 0%) and non-infected (OAI = 0%) orchards; and **(B)** water content, Ca, Mg, exchangeable Mg, exchangeable Na and trace Zn between orchards with OAI < 10% and orchards with OAI > 10%.

The univariate non-parametric Mann-Whitney U test was used to explore if there are statistically significant differences in the numerical predictor variables between infected (OAI > 0%) and not infected (OAI = 0%) cases, as well as between cases with OAI lower than 10% and those with OAI greater than 10%.

According to our findings, water content was statistically considerably higher in infected orchards (OAI > 0%) compared to non-infected (OAI = 0%), whereas Ca, Mg, Mn, and exchangeable Mg were statistically significantly lower in infected orchards compared to not infected (**Figure 4A**, **Table 2**). Additionally, Mann-Whitney results indicated statistically significantly higher values of water content in orchards with OAI > 10% compared to those with OAI < 10%, as well as significantly lower values of Ca, Mg, Mn and exchangeable Mg in orchards with OAI > 10% compared to those with OAI < 10% (**Figure 4B**, **Table 2**).

Table 2. Five number summary statistics and Mann-Whitney U tests results of statistically significant features for a) infected (OAI > 0%) and non-infected (OAI = 0%) orchards and b) orchards with OAI < 10% and OAI > 10%.

Features	Classes	Min*	Lower quartile (Q1)	Median (Q2)	Upper quartile (Q3)	Max*	*p***
A.							
Water content	OAI = 0%	49.39	57.71	62.93	65.70	72.07	0.045
	OAI > 0%	56.38	61.38	64.36	67.56	75.67	
Ca	OAI = 0%	0.62	1.22	1.40	1.72	2.16	0.005
	OAI > 0%	0.46	0.95	1.06	1.37	1.93	
Mg	OAI = 0%	0.08	0.12	0.15	0.17	0.24	0.007
	OAI > 0%	0.06	0.08	0.12	0.14	0.22	
Mn	OAI = 0%	17.50	28.00	35.50	43.50	64.50	0.043
	OAI > 0%	14.50	21.75	27.50	35.50	48.50	
Exch. Mg	OAI = 0%	53.00	115.50	154.00	274.00	472.00	0.001
	OAI > 0%	53.00	81.00	103.00	134.00	180.00	
B.							
Water content	OAI < 10%	49.39	58.17	62.38	65.84	72.07	0.012
	OAI > 10%	58.53	63.50	66.13	67.84	70.66	
Ca	OAI < 10%	0.62	1.19	1.36	1.67	2.16	0.019
	OAI > 10%	0.46	0.92	1.04	1.41	1.93	
Mg	OAI < 10%	0.06	0.12	0.14	0.17	0.23	0.007
	OAI > 10%	0.08	0.08	0.10	0.14	0.18	
Exch. Mg	OAI < 10%	53.00	114.75	148.50	270.50	472.00	0.8×10^{-4}
	OAI > 10%	53.00	75.75	94.00	108.75	142.00	
Exch. Na	OAI < 10%	10.00	15.00	27.50	53.75	77.00	0.037
	OAI > 10%	9.00	13.75	17.50	25.50	34.00	
Trace Zn	OAI < 10%	0.58	1.65	2.80	4.05	6.02	0.048
	OAI > 10%	0.86	1.43	1.92	2.60	3.10	

*Outliers were excluded, **Statistically significant at the 0.05 level.

Furthermore, we conducted a Chi-square test to investigate the association between the location of olive orchards and the occurrence of OA disease (OAI = 0%, OAI > 0%). To meet the assumptions of the Chi-square test and address small sample size concerns, the categories Argolida, Corinthia, and Arcadia were merged into a single category named “other locations”. The Chi-square test revealed a significant association between the location of olive orchards and the occurrence of olive anthracnose disease ($\chi^2 = 14.921$, $df = 2$, $p = 0.001$) (**Table 3**). Examining the adjusted residuals provided additional insights into the individual cells

within the contingency table (**Table 4**) that played a significant role in the observed associations. Adjusted residuals highlighted that the categories "Messinia" and "other locations" showed higher prevalence of infected orchards, with adjusted residuals of 3.7 and 0.7, respectively. Conversely, "Laconia" had less infected orchards than expected, with an adjusted residual of −3.3. The Chi-square test results, and the adjusted residuals support the conclusion that the distribution of olive anthracnose disease significantly varies across different locations, with Messinia showing a notably higher prevalence of the disease (**Figure 5A**).

Table 3. Chi-square test results for the association between a) the location and OA occurrence and b) the location and OA incidence.

Categorical variables	**Association**	**Test statistic**	**Degrees of freedom (*df*)**	***P*-value**
A) Location & OA occurrence	Significant	14.921	2	0.001
B) Location & OA incidence	Significant	8.216	2	0.016

Table 4. Contingency tables of a) location *OA occurrence and b) location *OA incidence.

A. OA occurrence			**OAI = 0%**	**OAI > 0%**
Location	Messinia	Count	7	19
		Expected count	13.9	12.1
		Adjusted residual	−3.7	3.7
	Laconia	Count	16	3
		Expected count	10.2	8.8
		Adjusted residual	3.3	−3.3
	Other locations	Count	8	5
		Expected count	6.9	6.1
		Adjusted residual	0.7	−0.7
B. OA incidence			**OAI < 10%**	**OAI > 10%**
Location	Messinia	Count	12	14
		Expected count	17	9
		Adjusted residual	−2.8	2.8
	Laconia	Count	16	3
		Expected count	12.4	6.6
		Adjusted residual	2.1	−2.1
	Other locations	Count	10	3
		Expected count	8.5	4.5
		Adjusted residual	1.0	−1.0

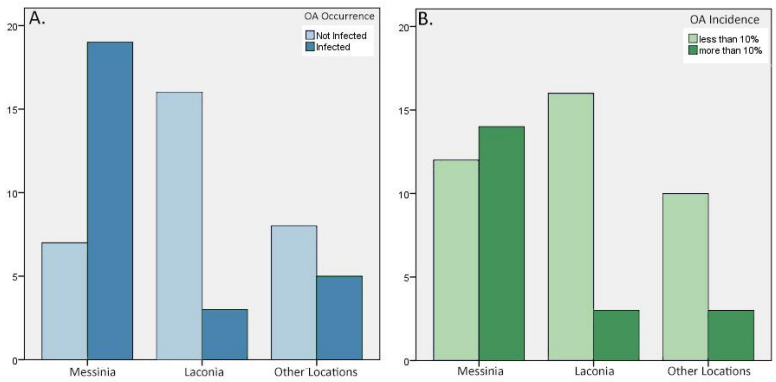

Figure 5. Distribution of **(A)** olive anthracnose occurrence across different locations; and **(B)** olive anthracnose incidence across different locations.

Similarly, we explored the association between the location of olive orchards and the incidence of olive anthracnose disease (OAI < 10%, OAI > 10%). The Chi-square test was conducted to examine the relationship between the variables, revealing a statistically significant association (χ^2 = 8.002, *df* = 2, *p* = 0.018) (**Table 3**). "Messinia" exhibited a negative adjusted residual of −2.8 for "OAI < 10%" indicating fewer orchards than expected in this category. Conversely, "Laconia" showed a positive adjusted residual of 2.1 for "OAI < 10%", suggesting a higher prevalence. The "other locations" category displayed a positive adjusted residual of 1.0 for "OAI < 10%", indicating that the number of orchards in this category was slightly higher than the expected (**Table 4**). The observed deviations, along with the statistical significance of the Chi-square test, emphasize that the incidence of olive anthracnose disease varies significantly across different locations. Specifically, "Messinia" demonstrated a relatively higher prevalence of orchards with disease incidence greater than 10%, while "Laconia" showed fewer orchards in the same category (**Figure 5B**).

Finally, Chi-square tests were employed to explore the potential associations between olive cultivar and OA occurrence and incidence, as well as between soil textural class and OA occurrence and incidence. The results were found to be not statistically significant, suggesting that there is no strong evidence of a direct relationship between the examined categorical variables and the presence or incidence of OA in the olive orchards.

3.2. Identification of important features

Six of the original thirty-three features—exchangeable Mg, Ca, Mg, location, water content, trace Cu—were selected as the final predictor variables to accurately differentiate between infected (OAI > 0%) and non-infected (OAI = 0%) orchards, based on the results of the RF-RFECV approach (**Figure 6A**). Similarly, seven features, including exchangeable Mg, water content, P, trace Cu, trace Zn, Ca and Mg were chosen as potential predictors for the distinction between the orchards with OAI < 10% and OAI > 10% (**Figure 6B**).

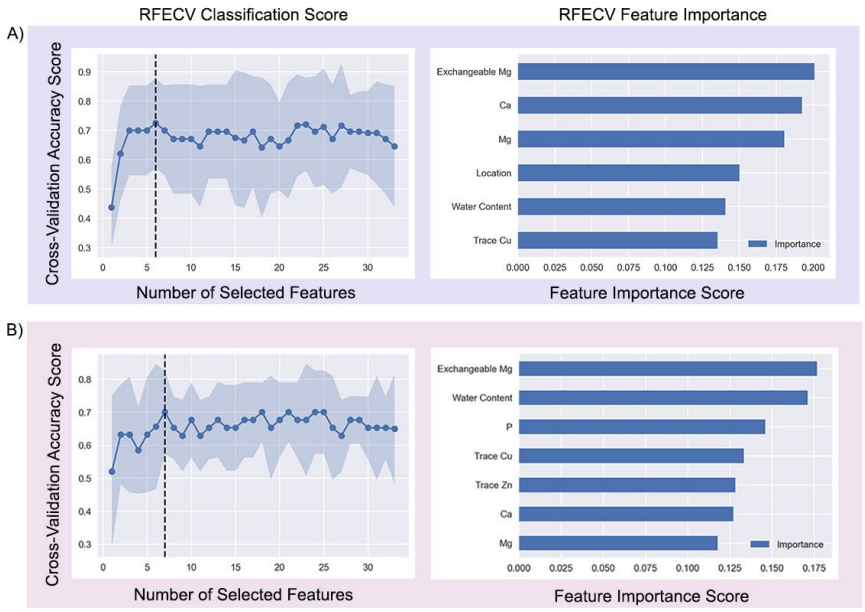

Figure 6. Recursive Feature Elimination with Cross-Validation (RFECV) to find optimal features for random forest classification of **(A)** OA occurrence (0: OAI = 0%, 1: OAI > 0%); and **(B)** OA incidence (0: OAI < 10%, 1: OAI > 10%).

Despite P and trace Cu being recognized as critical factors for the prediction of OAI, they were not found to be statistically significant, and hence not incorporated in the boxplots depicted in **Figure 4**. Supplementary boxplots (**Figure 7**) and five number summary statistics (**Table 5**) were generated to illustrate the intraclass dispersion of trace Cu among infected and non-infected samples and that of P and trace Cu among samples with OAI greater and less than 10%.

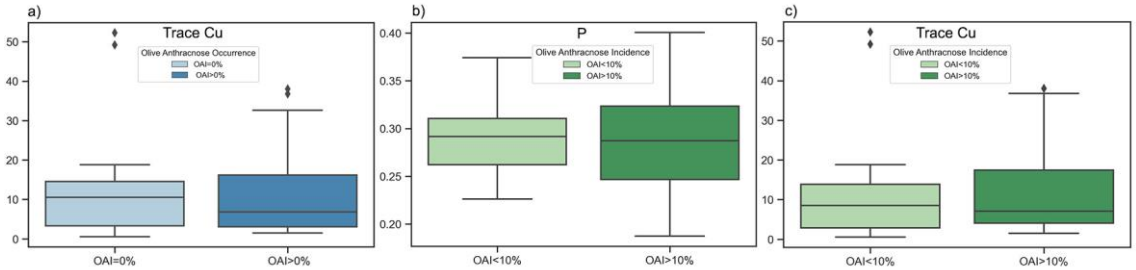

Figure 7. Boxplots visualizing differences in dispersion of **(a)** trace Cu data between infected (OAI > 0%) and non-infected (OAI = 0%) orchards; and **(b)** P data between orchards with OAI < 10% and orchards with OAI > 10%; and **(c)** trace Cu data between orchards with OAI < 10% and orchards with OAI > 10%.

Table 5. Five number summary statistics for P and trace Cu by OAI classes.

Features		Min	Lower quartile (Q1)	Median (Q2)	Upper quartile (Q3)	Max
Trace Cu	OAI = 0%	0.58	3.33	10.56	14.54	18.80
	OAI > 0%	1.48	3.09	6.80	16.17	32.60
P	OAI < 10%	0.23	0.26	0.29	0.31	0.37
	OAI > 10%	0.19	0.25	0.29	0.32	0.40
Trace Cu	OAI < 10%	0.58	2.86	8.55	13.86	18.80
	OAI > 10%	1.48	4.06	7.07	17.47	36.80

Table 6. The machine learning models hyperparameter configuration space.

Model	Hyperparameter	Search space	Best parameters	
			OAI > 0%	OAI > 10%
DT	criterion	[‘gini’, ‘entropy’]	‘entropy’	‘gini’
	max_depth	[None, 1, 2, 3, 4, 5]	None	None
	ccp_alpha	[0, 0.01, 0.1, 0.3, 1, 2]	0.01	0.1
	max_features	[‘auto’, ‘sqrt’, ‘log2’, None]	‘sqrt’	‘auto’
	min_samples_leaf	[1, 2, 3, 4]	3	1
	min_samples_split	[2, 3, 4]	2	2
GB	n_estimators	[15, 20, 22, 25]	25	20
	learning_rate	[0.1, 0.5, 0.8, 1]	0.5	0.1
	max_features	[‘auto’, ‘sqrt’, ‘log2’, None]	sqrt	sqrt
	max_depth	[None, 4, 5]	None	None
	min_samples_leaf	[1,2,3,4]	2	2
	min_samples_split	[0.5, 6, 7]	6	0.5
	subsample	[0.8, 1]	0.8	0.8
LR	penalty	[‘l1’, ‘l2’]	‘l2’	‘l1’
	C	[0.01, 0.1, 0.5, 1, 2]	0.1	0.5
	solver	[‘lbfgs’, ‘liblinear’]	‘liblinear’	‘liblinear’
	max_iter	[25, 30, 50, 100]	30	25
RF	n_estimators	[10, 30,100]	100	100
	criterion	[‘gini’, ‘entropy’]	‘gini’	‘gini’
	max_depth	[None, 1, 2, 3]	None	2
	min_samples_split	[2, 3, 4, 5, 10]	10	2
	min_samples_leaf	[1, 2, 3, 4]	1	2
	max_features	[‘auto’, ‘sqrt’, ‘log2’, None]	‘sqrt’	‘auto’
KNN	n_neighbors	[3, 4, 5, 6, 7, 8, 9]	4	7
	weights	[‘uniform’, ‘distance’]	‘distance’	‘uniform’
	metric	[‘euklidean’, ‘manhattan’]	‘manhattan’	‘manhattan’
SVM	C	[0.1, 0.5, 1, 2, 3]	0.1	0.1
	gamma	[‘scale’, 0.1, 1, 10, 100]	1	10
	kernel	[‘linear’, ‘rbf’, ‘poly’, ‘sigmoid’]	‘linear’	‘poly’

3.3. Performance of classifiers

The grid search optimization method was used to fine-tune the parameters for each model in order to optimize the accuracy score. **Table 6** shows the basics of the configuration space for the machine learning models developed for the prediction of OA occurrence (OAI > 0%) and OA incidence (OAI > 10%).

Table 7 presents the 5-fold cross-validation accuracy scores of the optimized machine learning models, used for predicting the occurrence (OAI > 0%) and incidence of OA (OAI > 10%). The results show that SVM (0.76) and GB (0.74) had the highest accuracy scores for predicting the occurrence of OA, followed by RF (0.73), DT (0.73), KNN (0.72), and LR (0.72). For predicting the incidence of OA, RF (0.81) and GB (0.77) were the models with the highest cross-validated scores, followed by KNN (0.74), DT (0.72), SVM (0.72), and LR (0.72).

Table 7. The 5-fold cross-validation accuracy of the optimized models for the prediction of OA occurrence (class 0: OAI = 0%, class 1: OAI > 0%) and incidence (class 0: OAI < 10%, class 1: OAI > 10%).

	Accuracy					
	DT	**GB**	**LR**	**RF**	**KNN**	**SVM**
OAI > 0%	0.73	0.74	0.72	0.73	0.72	0.76
OAI > 10%	0.72	0.77	0.72	0.81	0.74	0.72

After the grid search hyperparameter optimization, the models with the best hyperparameters were retrained on the complete training set (75% of the original data) and evaluated on the hold-out set (25% of the original data), using standard performance metrics (accuracy, specificity, sensitivity, AUC) (**Table 8**).

Table 8. Classification report of predictive models for OA occurrence (class 0: OAI = 0%, class 1: OAI > 0%) and OA incidence. (class 0: OAI < 10%, class 1: OAI > 10%).

Model	**Accuracy**	**Specificity**	**Sensitivity**	**AUC**
OAI > 0%				
DT	0.80	0.86	0.75	0.65
GB	0.87	1.00	0.75	0.93
LR	0.73	0.86	0.62	0.86
RF	0.80	0.86	0.75	0.93
KNN	0.73	1.00	0.50	0.90
SVM	0.80	0.86	0.75	0.86
OAI > 10%				
DT	0.87	1.00	0.81	0.91
GB	0.80	0.75	0.81	0.77
LR	0.73	0.50	0.81	0.66
RF	0.87	0.75	0.91	0.84
KNN	0.87	0.75	0.91	0.77
SVM	0.73	0.75	0.73	0.68

As shown in **Table 8**, GB classifier performed the best among all the models examined for the occurrence of OA (OAI > 0%), with an accuracy of 87%, specificity of 100%, sensitivity of 75% and AUC score of 93%. RF also had an AUC score of 93%, indicating the classifier's excellent ability to distinguish between infected and non-infected orchards, with an accuracy of 90%, a specificity of 86% and a sensitivity of 75%. Following GB and RF, SVM displayed an overall good performance, with an accuracy of 80%, a specificity of 86%, a sensitivity of 75% and an AUC score of 86%. KNN showed a remarkable specificity of 100%, indicating the classifier's excellent ability to correctly classify non-infected samples, however due to

its poor sensitivity of 50%, it was not considered effective for the identification of the disease. The low sensitivity of LR (62%) also indicated that the classifier was not useful in picking up the disease. DT classifier's AUC score (0.65) indicated a poor discrimination capacity to distinguish between infected and non-infected samples.

Among the classifiers examined for the prediction of the OA incidence (OAI > 10%), DT had the highest specificity (100%) and AUC score (91%), the highest accuracy (87%) together with RF and KNN and a sensitivity of 81%. Both RF and KNN demonstrated the highest sensitivity (91%) and the same accuracy (87%) and specificity (75%), although RF had a higher AUC score (84%) than KNN (77%). These classifiers managed to identify 91% of the orchards with OA larger than 10%. GB had a fairly good performance with an accuracy of 80%, a specificity of 75%, a sensitivity of 81% and an AUC score of 77%. LR and SVM were the least effective classifiers with poor discrimination ability indicated by their low AUC scores (66% and 68% respectively) and a low specificity (50%) of LR, which made it inappropriate for the classification of the non-infected samples.

Figure 8 shows the receiver operating characteristic (ROC) curves for the outputs of the classification models about OA occurrence (OAI > 0%) and prediction of OA incidence (OAI > 10%).

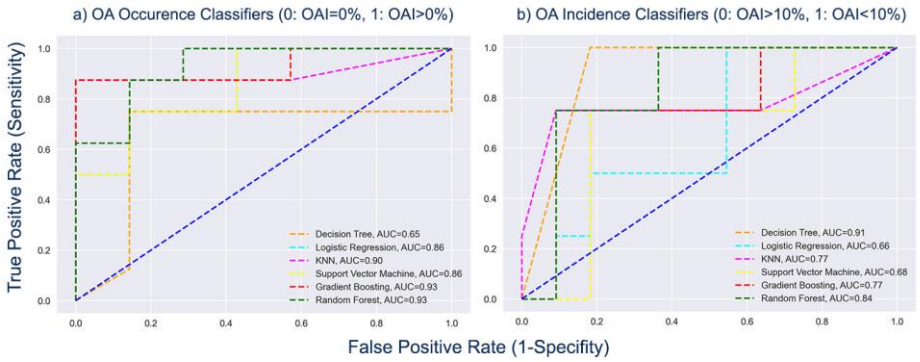

Figure 8. The comparison of the considered **(a)** OA occurrence; **(b)** OA incidence models using AUC-ROC curves.

As shown in **Figure 8a**, among the classifiers for the occurrence of OA (OAI > 0%), GB and RF achieved the highest AUC scores, equal to 0.93, demonstrating the models' excellent ability to discriminate between the infected and non-infected samples. The classifiers DT and RF achieved the greatest AUC scores, equivalent to 0.91 and 0.84 respectively, when comparing the AUC scores of the classifiers developed for the prediction of OA incidence (OAI > 10%) (**Figure 8b**), demonstrating their superior capacity to distinguish between olive orchards with OAI10% and those with OAI > 10%.

Summarizing, GB performance was superior compared to the other models for the prediction of the occurrence of OA disease (OAI > 0%) with an accuracy of 86.7%, a sensitivity of 100%, a specificity of 75% and a ROC-AUC score of 93%, while for the prediction of the spread of the disease (OAI > 10%), DT stood out with an accuracy of 86.7%, a sensitivity of 81.8%, a specificity of 100% and a ROC-AUC score of 91%. The RF classifier performed very well in both cases, with an accuracy of 80%, a sensitivity of 85.7%, a specificity of 75% and a ROC-AUC score of 93% for the prediction of the occurrence of the disease (OAI > 0%), and an accuracy of 86.7%, a sensitivity of 90.9%, a specificity of 75% and a ROC-AUC score of 84% for the prediction of the spread of the disease (OAI > 10%).

In order to verify that the best classifiers had in fact learned a significant predictive pattern in the data and that they were appropriate for the particular classification tasks, we conducted permutation tests. Specifically, we produced 1000 random permutations of the class labels for the training data sets used in the models' training. We then carried out the same 5-fold cross-validation procedure to obtain a classification accuracy score for each randomized dataset and generated a non-parametric null-distribution of accuracy

values (**Figure 9**).

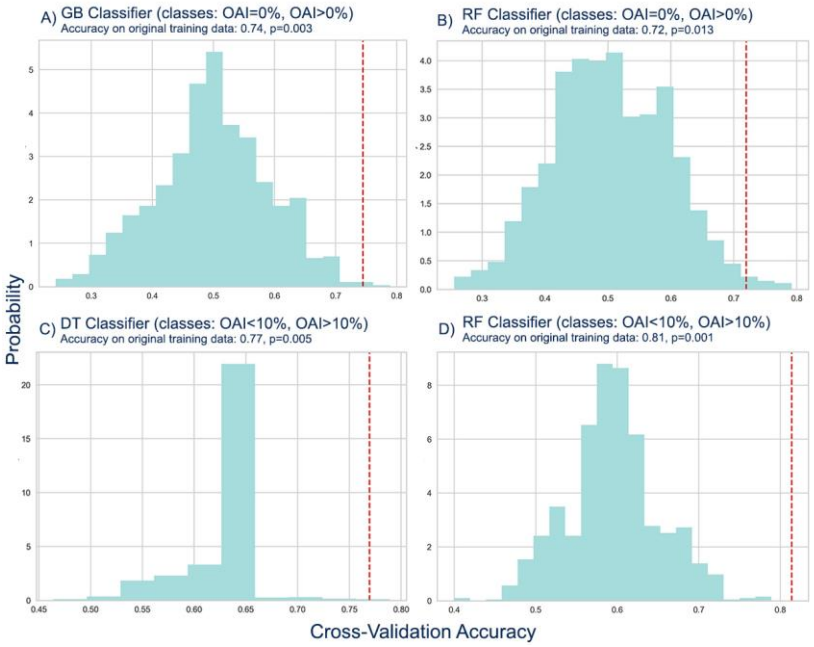

Figure 9. Generated null distributions of accuracy values from permutation tests where the class labels are randomly shuffled 1000 times and an accuracy value for each permutation is plotted. The vertical lines represent the observed accuracy values from the true class labels, **(A, B)** generated null-distributions for the GB and RF classifications of infected (OAI > 0%) against non-infected (OAI = 0%) olive orchards; **(C, D)** generated null-distributions for the DT and RF classification of olive orchards with OAI < 10% against orchards with OAI > 10%.

Based on the results of the permutation tests for the classification of the infected against non-infected orchards, we concluded that the accuracy scores of the GB (0.74) and RF (0.72) classifiers were statistically significant. With *p*-values of 0.003 and 0.013 respectively, we rejected the null hypothesis that the accuracy scores were due to chance. This means that the models' accuracy scores on the original data are likely to generalize to new unseen data. The permutation tests also provided evidence that the models were not overfitting to the training data.

The results of the permutation tests for the classification of the olive orchards with OAI < 10% against those with OAI > 10% demonstrated that the accuracy scores of the DT and RF classifiers were statistically significant with scores of 0.77 and 0.81, respectively. The null hypothesis was rejected with p-values of 0.005 and 0.001, suggesting that the accuracy scores were not due to chance. Thus, the accuracy scores for the original data are likely to be applicable to new unseen data. The results of the permutation tests indicated that the models were not overfitting to the training data.

4. Discussion

The results of this study suggested that the GB and RF showed a better performance in classifying olive orchards into infected and non-infected with OA. Moreover, when distinguishing between olive orchards with OAI values below 10% and those above 10%, the DT and RF classifiers displayed the greatest effectiveness. Previous studies in the field of agriculture have already established the efficacy of the GB[57], RF[58] and DT[59] classification algorithms.

Every classification algorithm has its own characteristics that can influence its performance on different datasets. In our study, we compared six popular classification algorithms, namely decision tree (DT), gradient boosting (GB), logistic regression (LR), random forest (RF), k-nearest neighbors (KNN) and support vector machine (SVM). The analysis revealed notable differences in the performance metrics of accuracy, sensitivity, specificity, and AUC score.

Gradient boosting and random forest algorithms demonstrated superior performance across metrics. This is because they are ensemble learning methods which combine multiple base models to create a stronger predictive model. By aggregating the predictions of multiple models, they can capture complex relationships and reduce overfitting. The decision trees employed in RF and the boosting process in GB allow them to model intricate feature interactions and non-linearities, which can be particularly beneficial in cases where the relationships between features are complex, and the dataset is of limited size.

Decision trees also performed well in our study. Despite not being an ensemble method, they recursively partition the feature space based on decision boundaries, creating tree-like structures. This provides DTs with the flexibility to capture complex relationships and non-linearities in the data.

On the other hand, LR, KNN, and SVM exhibited comparatively lower performance in our study than the ones previously discussed. Logistic regression assumes a linear relationship between the features and the target variable. As a result, this algorithm is not suitable for capturing the non-linearities of our dataset. K-nearest neighbors, which relies on the nearest neighbors for classification, is sensitive to dataset size and suffers from the curse of dimensionality. In small datasets as was the case with our study, KNN demonstrated difficulties in finding representative neighbors, leading to inferior performance. Furthermore, AUC score is a metric particularly sensitive to KNN's inability to handle imbalanced data. In our dataset, 34% of the instances had an OAI of more than 10% while 66% had an OAI less than 10%. This imbalance forced KNN to prioritize the majority class and did not allow it to distinguish accurately the minority class which in turn lowered the AUC score. Finally, support vector machines, while effective in many scenarios, can be sensitive to the choice of hyperparameters and the dataset size.

Consequently, the ensemble-based algorithms, such as GB and RF, demonstrated superior performance in terms of accuracy, sensitivity, specificity, and AUC score, showcasing their ability to handle complex relationships and mitigate overfitting.

Regarding the feature selection method, RF combined with the RFECV technique gave insights as to the importance of soil and foliar nutrient variables, considering their interactions and interdependencies. The high performance of the final models and the results obtained by the permutation tests revealed that the selected features (location, water content, Ca, Mg, exchangeable Mg, trace Cu) were effective at predicting the target class (OAI = 0%, OAI > 0%) for OA occurrence, while the chosen features (water content, P, Ca, Mg, exchangeable Mg, trace Zn, trace Cu) were able to accurately predict the target class for OA incidence (OAI < 10%, OAI > 10%).

Many of the features which were selected as variables in this study have been already linked to OAI incidence. Firstly, our findings align with previous research[15,60] which has associated severe epidemic outbreaks of OA disease with high relative humidity and frequent rainfall during the flowering and fruit development stages. Specifically, we discovered that olive fruit from infected olive orchards had significantly higher water content (Median(IQR) = 64.36 (61.38%–67.56%)) compared to non-infected (Median(IQR) = 62.93 (57.71%–65.70%)). Similarly, olive fruit from orchards with OAI greater than 10% had significantly higher water content (Median(IQR) = 66.13 (63.50%–67.84%)) compared to those with OAI < 10% (Median(IQR) = 62.38 (58.17%–65.84%)) (**Table 3**, **Figure 4**).

The location of the orchard emerged as another crucial factor for predicting OA disease, which is reasonable considering that different locations are associated with diverse microclimates and agronomical practices[7,9,10].

Furthermore, the results of this study agree with previous research, which suggests that resistance to OA is closely related to the health of plants and soil[11]. In another study, Talhinhas et al.[9] demonstrated that in certain regions of Portugal and southwest Spain, there seems to be a potential connection between increased OA occurrence and low soil pH, which in turn correlates with insufficient Ca levels. The statistical analysis

of this dataset also revealed that Ca values were significantly lower in the infected samples (Median(IQR) = 1.06 (0.95%–1.37%)) and in the samples with OAI > 10% (Median(IQR) = 1.04(0.92%–1.41%)), compared to the non-infected (Median(IQR) = 1.40 (1.22%–1.72%)) and those with OAI < 10% (Median(IQR) = 1.36 (1.19%–1.67%)), respectively (**Table 3**, **Figure 4**).

Regarding the micronutrient Cu, our findings showed that almost all the samples from olive orchards with copper levels above ≈ 19 ppm were infected with OAI greater than 10% (**Table 6**, **Figure 7**). Previous research has shown that while the application of copper-based fungicides is the recommended measure for controlling anthracnose in olive groves, overuse can cause a build-up of copper in the soil and obstruct the uptake of other nutrients[61,62].

Finally, our results showed that P levels between 0.23 to 0.37% were present in samples with OAI values both below and above 10%. It is worth mentioning that P levels that fell below 0.23% or exceeded 0.37% were only found in samples with OAI values greater than 10% (**Table 6**, **Figure 7**). Our research findings also showed that samples collected from orchards with OAI greater than 10% displayed significantly lower trace Zn values (Median(IQR) = 1.92 (1.43–2.60 mg kg^{-1})) compared to those collected from orchards with OAI < 10% (Median(IQR) = 2.80 (1.65–4.05 mg kg^{-1})) (**Table 3**, **Figure 4B**).

Magnesium and exchangeable Mg were identified as two critical factors that could potentially be used to predict the onset of OA disease. Magnesium exhibited a significant decrease in samples obtained from infected orchards (Median(IQR) = 0.12 (0.08%–0.14%)), as opposed to non-infected ones (Median(IQR) = 0.15 (0.12%–0.17%)). Similarly, samples from orchards with OAI > 10% displayed a lower Mg content (Median(IQR) = 0.10 (0.08%–0.14%)) than samples from orchards with OAI < 10% (Median(IQR) = 0.14 (0.12%–0.17%)) (**Table 3**, **Figure 4**). With regard to exchangeable Mg, its concentrations demonstrated significant reduction in samples obtained from infected orchards (Median(IQR) = 103 (81–134 mg kg^{-1})), as opposed to non-infected (Median(IQR) = 154 (115.5–274 mg kg^{-1})). Likewise, samples from orchards with OAI > 10% (Median(IQR) = 94 (75.75–108.75 mg kg^{-1})) displayed lower exchangeable Mg contents relative to orchards with OAI < 10% (Median(IQR) = 148.5 (114.75–270.5 mg kg^{-1})) (**Table 3**, **Figure 4**).

To our knowledge, there is no previous research exploring the relationship between P, Zn, Mg and OA disease. However, interactions among mineral nutrients occur frequently in the soil and at the plant level, leading to interdependencies. Consequently, a deficiency or excess of one nutrient can impact the absorption or utilization of another. The concentration of nutrients also affects plant tolerance or resistance to pathogens[63,64].

In this study, the classification algorithms selected features that confirmed that balanced nutrition is critical for the control and management of OA disease. These features could potentially have a predictive value for the determination of OA incidence.

5. Conclusion

Using machine learning methods, this study highlights the importance of balanced nutrition as a control strategy for olive anthracnose. Some of the selected features have already been shown to have an association with the disease, albeit individually. Our study proposed a model that incorporates a number of soil and foliar nutrients along with orchard and olive fruit parameters, implying their interdependency for the prediction of the disease. We acknowledge that the small dataset is a limitation of our study. However, to reduce the impact of this problem and the risk of overfitting, we applied cross-validation on the training set and evaluated the models' generalization performance on a held-out test set. Additionally, permutation tests were conducted to assess the statistical significance of the best models. In addition to working with a larger dataset, the performance of the proposed models would benefit from the inclusion of more predictive variables, such as weather data and cultivar susceptibility level.

Author contributions

Conceptualization, KK and AM; methodology, KK; software, KK and VN; validation, KK; formal analysis, KK and VN; investigation, KK, AM, PCT, AT, AK and AV; resources, KK, PCT, AT, AK and AV; data curation, KK; writing—original draft preparation, KK and AM; writing—review and editing, KK, AM, VN, PCT, AT and AV; visualization, KK; supervision, VD and VN; project administration, KK and AM; funding acquisition, VD and AM. All authors have read and agreed to the published version of the manuscript.

Conflict of interest

The authors declare no conflict of interest.

References

1. Romero J, Santa-Bárbara AE, Moral J, et al. Effect of latent and symptomatic infections by *Colletotrichum* godetiae on oil quality. European Journal of Plant Pathology 2022;163(2): 545-556. doi: 10.1007/s10658-022-02494-x
2. Kolainis S, Koletti A, Lykogianni M, et al. An integrated approach to improve plant protection against olive anthracnose caused by the *Colletotrichum* acutatum species complex. PLoS One 2020; 15(5): e0233916. doi: 10.1371/journal.pone.0233916
3. Petrogiannis A. Anthracnose has depleted 30% of this year's production in Messinia. Available online: https://www.tharrosnews.gr/2023/02/to-gloiosporio-efage-fetos-to-30-paragogis-sti-messinia/ (accessed on 12 March 2023).
4. Peres F, Talhinhas P, Afonso H, et al. Olive Oils from Fruits Infected with Different Anthracnose Pathogens Show Sensory Defects Earlier Than Chemical Degradation. Agronomy 2021; 11(6): 1041. doi: 10.3390/agronomy11061041
5. Carvalho MT, Simoes-Lopes P, Silva MJM. Influence of different olive infection rates of *Colletotrichum* acutatum on some important olive oil chemical parameters. Acta Horticulturae 2008; 791: 555-559. doi: 10.17660/ActaHortic.2008.791.85
6. Moral J, Xaviér C, Roca LF, et al. Olive Anthracnose and its effect on oil quality. Grasas Aceites 2014; 65(2): e028. doi: 10.3989/gya.110913
7. Moral J, Oliveira R, Trapero-Casas A. Elucidation of the Disease Cycle of Olive Anthracnose Caused by *Colletotrichum* acutatum. Phytopathology 2009; 99: 548-556. doi: 10.1094/PHYTO-99-5-0548
8. Moral J, Xaviér CJ, Viruega JR, et al. Variability in susceptibility to anthracnose in the World Collection of Olive Cultivars of Cordoba (Spain). Frontiers in Plant Science 2017; 8: 1892. doi: 10.3389/fpls.2017.01892
9. Talhinhas P, Loureiro A, Oliveira H. Olive anthracnose: A yield- and oil quality-degrading disease caused by several species of *Colletotrichum* that differ in virulence, host preference and geographical distribution. Molecular Plant Pathology 2018; 19: 1797-1807. doi: 10.1111/mpp.12676
10. Cacciola SO, Faedda R, Sinatra F, et al. Olive anthracnose. Journal of Plant Pathology 2012; 94(1): 29-44.
11. Sergeeva V. The role of epidemiology data in developing integrated management of anthracnose in olives—A review. Acta Horticulturae 2014; 1057: 163-168. doi: 10.17660/ActaHortic.2014.1057.19
12. Moral J, Agustí-Brisach C, Raya MC, et al. Diversity of *Colletotrichum* Species Associated with Olive Anthracnose Worldwide. Journal of Fungi 2021; 7: 741. doi: 10.3390/jof7090741
13. Romero J, Moral J, González-Domínguez E, et al. Logistic models to predict olive anthracnose under field conditions. Crop Protection 2021; 148: 105714. doi: 10.1016/j.cropro.2021.105714
14. Sergeeva V. Anthracnose in olives: symptoms, disease cycle, and management. In: Proceedings of the 4th International Conference Olivebioteq; 2011.
15. Sergeeva V. Integrated pest management of diseases in olives. Australian and New Zealand Olive Grower and Processor 2011; 80: 16-21.
16. Sergeeva V. Anthracnose management factors influencing yield and quality of olives. In: Proceedings of the Australian National Conference; 17th-19th September 2014.
17. Shoaib M, Shah B, El-Sappagh S, et al. An advanced deep learning models-based plant disease detection: A review of recent research. Frontiers in Plant Science 2023; 14: 1158933. doi: 10.3389/fpls.2023.1158933
18. Fenu G, Malloci F. Forecasting Plant and Crop Disease: An Explorative Study on Current Algorithms. Big Data and Cognitive Computing 2021; 5(2). doi: 10.3390/bdcc5010002
19. Alruwaili M, Alanazi S, Abd ElGhany S, Shehab A. An Efficient Deep Learning Model for Olive Diseases Detection. International Journal of Advanced Computer Science and Applications 2019; 10. doi: 10.14569/IJACSA.2019.0100863
20. Fazari A, Pellicer-Valero O, Gómez-Sanchís J, et al. Application of deep convolutional neural networks for the

detection of anthracnose in olives using VIS/NIR hyperspectral images. Computers and Electronics in Agriculture 2021; 187: 106252. doi: 10.1016/j.compag.2021.106252
21. Alves L, Silva R, Bernardino J. Using Data Mining to Predict Diseases in Vineyards and Olive Groves. In: Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management; 2017; pp. 282-287.
22. Olivares B, Lobo Luján D, Rey JC, et al. Identification of Soil Properties Associated with the Incidence of Banana Wilt Using Supervised Methods. Plants 2022; 11(15): 2070. doi: 10.3390/plants11152070
23. Uceda M, Frias L. Harvest dates: Evolution of the fruit oil content, oil composition and oil quality. In: Proceedings of the II Seminario Oleicola Internacional; 1975; Cordoba, Spain. pp. 125-130.
24. Tsalgatidou PC, Thomloudi EE, Baira E, et al. Integrated Genomic and Metabolomic Analysis Illuminates Key Secreted Metabolites Produced by the Novel Endophyte Bacillus halotolerans Cal.l.30 Involved in Diverse Biological Control Activities. Microorganisms 2022; 10(2): 399. doi: 10.3390/microorganisms10020399
25. Klages MG. Reproducibility of saturation percentage of soils. In: Proceedings of the Montana Academy of Sciences; 1984; 44: 67-69.
26. Kalra YP. Determination of pH of soils by different methods: collaborative study. Journal of AOAC International 1995; 78: 310-321. doi: 10.1007/BF02348343
27. Van Reeuwijk LP. Procedures for soil analysis, 6th ed. Technical Paper International Soil Reference and Information Centre; FAO/ISRIC; Wageningen the Netherlands. 2002.
28. Warncke D, Brown JR. Potassium and other basic cations. In: Recommended Chemical Soil Test Procedures for the North Central Region; Missouri Agricultural Experimental Station SB1001; Columbia, MO USA. 1982. pp. 31-33.
29. Olsen SR, Cole CV, Watanabe FS, Dean LA. Estimation of Available Phosphorus in Soils by Extraction with Sodium Bicarbonate. USDA Circular 1954; 939: 18.
30. Walkley A, Black IA. An examination of the Degtjareff method for determining soil organic matter and a proposed modification of the chromic acid titration method. Soil Science 1934; 37(1): 29–38. doi: 10.1097/00010694-193401000-00003
31. Miller RO, Kotuby-Amacher J, Rodriguez JB. Western States Laboratory Proficiency Testing Program Soil and Plant Analytical Methods; 1998.
32. Murphy J, Riley JP. A Modified Single Solution Method for the Determination of Phosphate in Natural Waters. Analytica Chimica Acta 1962; 27: 31-36. doi: 10.1016/S0003-2670(00)88444-5
33. Greweling T. Chemical analysis of plant tissue. Search 1976; 6(8): 1-35.
34. Fan C, Chen M, Wang X, et al. A Review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. Frontiers in Energy Research 2021; 9: 652801. doi: 10.3389/fenrg.2021.652801
35. Darst BF, Malecki KC, Engelman CD. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. BMC Genetics 2018; 19(Suppl 1): 65. doi: 10.1186/s12863-018-0633-8
36. Reif DM, Motsinger AA, McKinney BA, et al. Feature Selection using a Random Forests Classifier for the Integrated Analysis of Multiple Data Types. In: Proceedings of the 2006 IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology; pp. 1-8.
37. Pal M, Foody G. Feature Selection for Classification of Hyperspectral Data by SVM. IEEE Transactions on Geoscience and Remote Sensing 2010; 48: 2297-2307. doi: 10.1109/TGRS.2009.2039484
38. Akkaya B. The Effect of Recursive Feature Elimination with Cross-Validation Method on Classification Performance with Different Sizes of Datasets. In: Proceedings of the 4th International Conference on Data Science & Applications; 2021; Istanbul, Turkey.
39. Singhi S, Liu H. Feature subset selection bias for classification learning. In: Proceedings of the 23rd International Conference on Machine Learning—ICML. pp. 849-856.
40. Sarker IH. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science 2021; 2: 160. doi: 10.1007/s42979-021-00592-x
41. Li Y, Shami A. On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing 2020; 415: 295-316. doi: 10.1016/j.neucom.2020.07.061
42. Montesinos López OA, Montesinos López A, Crossa J. Overfitting, Model Tuning, and Evaluation of Prediction Performance. In: Multivariate Statistical Machine Learning Methods for Genomic Prediction. Springer; 2022. pp. 109-139.
43. Raschka S. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. arXiv 2018.
44. Sebastian Raschka. Available online: https://sebastianraschka.com/blog/2016/model-evaluation-selection-part1.html (accessed on 3 November 2023).
45. Charbuty B, Abdulazeez A. Classification Based on Decision Tree Algorithm for Machine Learning. Journal of Applied Science and Technology Trends 2021; 2(1): 20-28. doi: 10.38094/jastt20165
46. Schapire RE. The Boosting Approach to Machine Learning: An Overview. In: Denison DD, Hansen MH, Holmes CC, et al. (editors). Nonlinear Estimation and Classification. Lecture Notes in Statistics; Springer; 2003. pp. 37-64.
47. Zhu N, Zhu C, Zhou L, et al. Optimization of the Random Forest Hyperparameters for Power Industrial Control

Systems Intrusion Occurrence Using an Improved Grid Search Algorithm. Applied Sciences 2022; 12: 10456. doi: 10.3390/app122010456
48. Manish S, Parul G. A Review on Analysis of K-Nearest Neighbor Classification Machine Learning Algorithms based on Supervised Learning. International Journal of Engineering Trends and Technology 2022; 70(7): 43-48. doi: 10.14445/22315381/IJETT-V70I7P205
49. Peng J, Lee K, Ingersoll G. An Introduction to Logistic Regression Analysis and Reporting. Journal of Educational Research 2002; 96(1): 3-14. doi: 10.1080/00220670209598786
50. Nayak J, Naik B, Behera H. A Comprehensive Survey on Support Vector Machine in Data Mining Tasks: Applications & Challenges. International Journal of Database Theory and Applications 2015; 8: 169-186. doi: 10.14257/ijdta.2015.8.1.18
51. Cichosz P. Assessing the Quality of Classification Models: Performance Measures and Evaluation Procedures. Open Engineering 2011; 1: 132-158. doi: 10.2478/s13531-011-0022-9
52. Gogtay NJ, Thatte UM. Statistical Evaluation of Diagnostic Tests (Part 1): Sensitivity, Specificity, Positive and Negative Predictive Values. Journal of the Association of Physicians of India 2017; 65(6): 80-84.
53. Fawcett T. An Introduction to ROC Analysis. Pattern Recognition Letters 2006; 27: 861-874. doi: 10.1016/j.patrec.2005.10.010
54. Nahm FS. Receiver operating characteristic curve: overview and practical use for clinicians. Korean Journal of Anesthesiology 2022; 75(1): 25–36. doi: 10.4097/kja.21209
55. Braga-Neto UM, Dougherty ER. Is cross-validation valid for small-sample microarray classification? Bioinformatics 2004; 20(3): 374-380. doi: 10.1093/bioinformatics/btg419
56. Ojala M, Garriga GC. Permutation Tests for Studying Classifier Performance. Journal of Machine Learning Research 2010; 11:1833-1863.
57. Almeida RND, Greenberg M, Bundalovic-Torma C, et al. Predictive modeling of Pseudomonas syringae virulence on bean using gradient boosted decision trees. PLOS Pathogens 2022; 18(7): e1010716. doi: 10.1371/journal.ppat.1010716
58. Olivares Campos BO. Evaluation of the Incidence of Banana Wilt and its Relationship with Soil Properties. In: Banana Production in Venezuela. The Latin American Studies Book Series; Springer; 2023.
59. Ahmed K, Shahidi TR, Irfanul Alam SM, Momen AS. Rice Leaf Disease Detection Using Machine Learning Techniques. In: Proceedings of the 2019 International Conference on Sustainable Technologies for Industry 4.0 (STI); 24-25 December 2019; Dhaka, Bangladesh. pp. 1-5.
60. Moral J, Trapero A. Assessing the Susceptibility of Olive Cultivars to Anthracnose Caused by *Colletotrichum* acutatum. Plant Disease 2009; 93(10): 1028-1036. doi: 10.1094/PDIS-93-10-1028
61. Sergeeva V. Using copper sprays to control olive diseases. Australian & New Zealand Olivegrower & Processor 2010; 72: 41-42.
62. Roca L, Moral JR, Viruega A, et al. Copper fungicides in the control of olive diseases. Olea 2007; 26: 48-50.
63. Fernández-Escobar R. Olive Nutritional Status and Tolerance to Biotic and Abiotic Stresses. Frontiers in Plant Science 2019; 10: 1151. doi: 10.3389/fpls.2019.01151
64. Sergeeva V. Balanced plant nutrition may help reduce anthracnose. The Olive Press: Pests and Diseases 2011; pp. 23-24.