Machine learning to identify olive-tree cultivars

João Mendes[1,3,6][0000−0003−0979−8314], José Lima[1,2,6][0000−0001−7902−1207], Lino Costa[3][0000−0003−4772−4404], Nuno Rodrigues[4,6][0000−0002−9305−0976], Diego Brandão[5][0000−0003−3874−784X], Paulo Leitão[1,6][0000−0002−2151−7944], and Ana I. Pereira[1,3,6][0000−0003−3803−2043]

[1] Research Centre in Digitalization and Intelligent Robotics (CeDRI), Instituto Politécnico de Bragança, 5300-253 Bragança, Portugal {joao.cmendes, jllima, pleitao, apereira}@ipb.pt

[2] INESC TEC - INESC Technology and Science Porto, Portugal

[3] ALGORITMI Center, University of Minho, 4710-057 Braga, Portugal lac@dps.uminho.pt

[4] Centro de Investigação de Montanha (CIMO), Instituto Politécnico de Bragança, 5300-253 Bragança, Portugal. nunorodrigues@ipb.pt

[5] Celso Suckow da Fonseca Federal Center of Technological Education, Rio de Janeiro, Brazil diego.brandao@cefet-rj.br

[6] Laboratório para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal

Abstract. The identification of olive-tree cultivars is a lengthy and expensive process, therefore, the proposed work presents a new strategy for identifying different cultivars of olive trees using their leaf and machine learning algorithms. In this initial case, four autochthonous cultivars of the Trás-os-Montes region in Portugal are identified (Cobrançosa, Madural, Negrinha e Verdeal). With the use of this type of algorithm, it is expected to replace the previous techniques, saving time and resources for farmers. Three different machine learning algorithms (Decision Tree, SVM, Random Forest) were also compared and the results show an overall accuracy rate of the best algorithm (Random Forest) of approximately 93%.

Keywords: Machine Learning · Identification · Leaf · Cultivars · Varieties.

1 Introduction

Some researchers estimate that people have consumed olive oil and olives for 6000 years, making the olive tree one of the first fruit trees to be domesticated [14]. However, its origin is not subject to consensus. It is considered acceptable to say that the olive tree is native to the entire Mediterranean basin with its origin in Asia Minor, where it is highly abundant and grows in dense forests [1].

Since its domestication, olive cultivation has expanded worldwide, being cultivated in dozens of countries on all continents except for Antarctica. Its production is mainly used (90%) for the production of olive oil, with the rest being processed as table olives. All of the countries where they are cultivated, it is possible to highlight 23 countries [14] that are responsible for about 85% of all world olive production. According to the International Olive Council (IOC), of these most representative countries, it is estimated that there are about 139 domesticated cultivars. This number grows significantly if we also consider wild species, making a total of more than 2,000 different cultivars [13].

Olive production is still a developing process despite thousands of years of domestication, breaking production records practically every year. In addition to the production records, the economic valuation also increased, reaching 1.23 billion dollars in world trade transactions in 2020. Portugal is an active part of this process, being responsible for about 6.9% of this value, highlighting as third-ranked among the countries that export the Most [2].

Olive growing is the main activity in Portugal, contributing economically, socially, and environmentally to the country. Looking at data from the National Statistics Institute (INE) for 2019 [22], there are 361,483 hectares of olive groves in the country, on around 118,450 farms, with a particular focus on inland regions, with Alentejo, Trás-os-Montes, Beira Interior, Ribatejo and Oeste being the central producing regions. According to provisional data from the INE [23], the year 2021 was another year of records, with a production of 2.25 million hectoliters of olive oil being expected. Within these dimensions and similarly to the other fruit trees, the olive tree also has different cultivars, each one being more adapted to specific climatic and geographical conditions. The use of different cultivars for the production of olive oil makes it possible to ensure a more harmonious composition of the oil from an organoleptic point of view since each one of them has unique chemical and physical characteristics [19, 31]. In addition to the taste, different cultivars also allow to improve the season or date of harvest by selecting types with different maturation periods.

Formerly it was enough to crush the olives and make olive oil. Consumers are showing more and more interest in the composition of the products they buy, olive oil being no exception. There is a need for the farmer's knowledge because there is a great variety of olives. He/She must be able to distinguish the lots and know how to specify the type of cultivars present. This identification makes it possible to add value to the product compared to others. This process can be facilitated in recent olive groves that the producer has already planted. However, it may cause some difficulties in the case of olive groves with some age and where there is no certainty of the cultivars present. In such cases, it is necessary to use identification techniques.

The most of the techniques applied for the identification of cultivars use genetic analysis [5–7, 15]. These techniques have a high index of reliability. However, they involve time-consuming and expensive processes requiring laboratory analysis, preventing identification on the farm itself. According to research carried out in recent years, other identification techniques have been observed, using

artificial vision. These have encouraging hit rates, as suggested by the authors [8, 30, 39]. On the other hand, these techniques always use the tree's fruit (olives and seeds) to perform the classification. Thus, they are restricted to a specific time of the year when producers are busy harvesting the fruit. In most cases, it is not possible to have the time necessary for them to assist in the identification process.

Analyzing the problems that arise from the techniques presented above, the focus of this work is to ensure an innovative form of identification, on-site, with minimal impact on the tree and that can be carried out at any season of the year. In this way and analyzing what is being done in other crops, the presented solution involves using machine learning algorithms to identify the leaves of the tree. This solution solves practically all the problems shown above and guarantees an instant classification without having to resort to specialized technicians. It is possible to do it any time of the year since the olive tree is a permanent leaf tree. We also guarantee that the impact on the tree is null as it is not necessary to take samples to ensure identification.

This implementation becomes easier considering the advances in the area of artificial intelligence algorithms. Following the hardware developments, this type of algorithm has proven to be highly qualified for solving this type of task, being applied to various species such as pistachios [20], grapes[28], apples [27], among others and other examples [18, 26].

In this way, it is the purpose of this work to analyze various types of machine learning algorithms, comparing them to each other, in order to understand which one presents better results for the intended classification. To this end, a survey of related works are carried out, composed not only of the most used methods for the identification of cultivars but also which are the most used machine learning algorithms for image classification, which will be presented in section 2. Then, in the section 3, the proposed methodology will be discussed. In the section 4, the main results and their analysis will be presented, and finally, the section 5 will address the conclusion and main future works.

2 Related Works

A literature review was conducted on the various identification techniques applied to identify olive variety. For this purpose, two databases were used (Scopus and Web of Science), where two terms, "Olive identification cultivars" or "Olive variety identification", were searched. The search resulted in 921 and 536 documents in the WOS and Scopus databases. After their collection, the R software was used, combined with the Bibliometrix tool [4], where trends and main groupings of keywords were analyzed.

This first research resulted in several keywords indicating that most of the published articles use genetic identification techniques, microsatellites [37], simple-sequence repeats (SSRs) markers [10], and mainly random amplified polymorphic DNA (RAPDS) [16] and single nucleotide polymorphisms (SNPs) [35]. This type of technique achieves high rates of effectiveness, however, as this is not the focus

of the article, a second research was carried out in which the term "learning" was added to the terms previously used. This second iteration resulted in an article that fits the theme, in the article [36], the authors present a method of identifying olive tree cultivars using deep learning algorithms and ISSR markers. They use a dataset with 800 training and 200 test images to train a convolutional neural network. As a result, they have an overall accuracy of 89.57% across four different leaf cultivars.

With no more published articles identified in this area, it was necessary to carry out a new search, leaving aside the term "olive" and adding "Leaf". By searching the words "identification cultivars learning leaf" or "variety identification learning leaf" within this group, it was possible to verify the existence of some articles that use tree leaves and artificial intelligence algorithms to identify cultivars, as is the case of Liu, et all , who present in their paper [27] a new method for classifying apple cultivars. In this study, a TensorFlow model was developed capable of identifying fourteen distinct apple cultivars with an overall accuracy of 97.11%. The authors of the article used a self-built dataset with 12,435 images.

Similarly, the authors of the article [38] developed a system based on convolutional neural network to identify twelve distinct bean cultivars belonging to three species. The results obtained were divided into three different classes, evaluated by classifying species (level I) with an accuracy of 95.86%, cultivars from the same species (level II) with 91.37%, and cultivars from different species (level III) with 86.87%.

Another example is the grapevine leaves, presented in the article [32] using the six most representative cultivars of that region with 240 images for training and 60 for testing (using data-augmentation techniques). The authors present a model of Deep Convolutional Neural Network, based on the transfer learning technique with a VGG16 structure, in which the structure was modified, adding the global average pooling layer, dense layers, a batch normalization layer, and a dropout layer. The results obtained by the authors are pretty encouraging, achieving accuracy in recognizing and grouping different cultivars of grapevine with an average classification of over 99%

As can be seen, the application of artificial intelligence algorithms in agriculture and, more specifically in identifying both cultivars and species is still a recent process and development. Thus, the number of published articles is still restricted to some species and specific algorithms. There is, therefore, the possibility of exploring the behavior in other species and other types of algorithms to improve identification efficiencies, thus making this type of process increasingly reliable.

3 Method and Materials

As the main objective of this work is to develop an artificial intelligence system capable of identifying the different cultivars of olive trees present in the region, there is a set of steps to be taken and implemented, thus ensuring the correct

functioning of the system. Therefore, this section was divided into subsections: Dataset, Classification Models, Evaluation Metrics, and Methodology.

3.1 Dataset

Machine learning (ML) systems work similarly to humans, learning from experience. However, in the case of ML, this experience is controlled, with the programmer responsible for all of it. The experience we talk about here is nothing more than the dataset made available to the algorithms for their training and testing, on which all the success or failure of the application will depend. It is a consensus in the literature that the performance presented by the machine learning algorithms is entirely dependent on the dataset that is used [24]. In this way, and since it is an innovative approach to the problem in question, it was necessary to create a sufficiently large dataset of images to ensure the operation of supervised classification algorithms.

The process of creating the dataset consists of several stages. The initial phase is collecting leaves, prepared in partnership with certified technical personnel in the area, in monovarietal olive groves, to be sure of the cultivars collected.

This collection process followed all seasons of the year, analyzing all times of growth and recession of leaves and other factors that impact their development, such as water or nutrient stress, thus ensuring correct identification, regardless of the phase and conditions of the tree. In this way, samples were collected from several olive groves in the district of Bragança, in the most random way possible, leaves from the inside and outside of the canopy, leaves of the year and older leaves, with an average of 20-25 leaves per tree being collected.

After collecting the leaves, it is necessary to digitize them. This process was carried out with the help of a digital camera with a resolution of 2610×4640 pixels. All samples were photographed in RGB type and JPG format. The background image was white to facilitate their treatment and application in the algorithms.

After digitizing the samples, their pre-processing was realized. Depending on the sample size, some processes were carried out. The first step was the cutting of the images. This was done autonomously, using threshold and find contour techniques to identify the shape of the leaf and then cut it out. For this purpose, the image was changed to gray format, and the adaptive gaussian-weighted [11] threshold method was used. After this transformation, the OpenCV library was used to find the image contour using the findContours function. It allows identifying the sheet's shape that will define the area to be cut. After obtaining the area size to be cropped, it was applied to the original image (RGB). The last step of the pre-processing was to resize the images to a resolution of 299×299 pixels to facilitate the algorithms' processing.

Once the pre-processing was completed, the result obtained was a dataset with approximately 1500 images of the region's four most representative categories (Fig.1). A balance was also made between the categories, ensuring that they all had the same number of images.

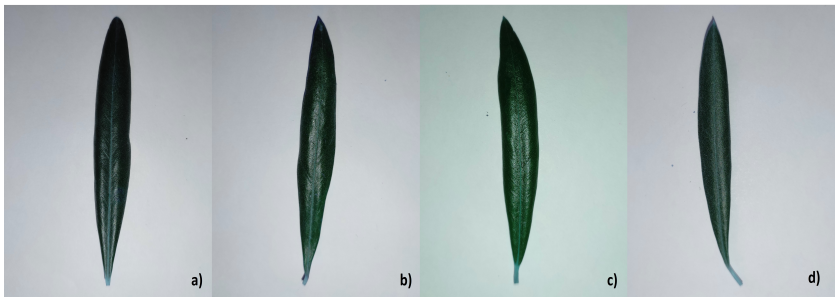

Fig. 1. Dataset cultivars: a) Cobrançosa; b) Madural; c) Negrinha; d) Verdeal

3.2 Classification Models

The identification of olive tree cultivars is a typical classification problem. In this type of problem, the main objective is to predict the category of an unobserved data based on the algorithm's input data. Three supervised classification algorithms [25] will be compared for the intended effect in this case.

The choice of these algorithms was based on a bibliographic research process, the articles presented in the Scopus and WOS databases were gathered, in the time interval from 2018 to 2022 that had the terms "Machine learning classification comparison" or "Machine learning classification leaf". This resulted in a universe of 6096 articles, which were later analyzed with the bibliometrix software. After the analysis, the most used author keywords were used to choose the algorithms and it was concluded that the most used algorithms are the Support vector machine (SVM), and Random Forest (RF). Aside from these, the Decision Tree algorithm (CART) was also selected due to its straightforward interpretation, easily identifying which variables have the most significant influence on the classification of images. In addition to this advantage, it is a method that does not require a long training process and, therefore, can save a lot of modeling time [40].

Decision Tree Decision trees are one of several supervised machine learning methods, introduced in 1986 by JR Quinlan [34]. It is obtained by recursively partitioning the data space and fitting a simple prediction model within each partition. As a result, the partitioning can be graphically represented as a decision tree, which can be used for classification and regression. Classification trees are designed for dependent variables that take on a finite number of unordered values, with prediction error measured in terms of the cost of misclassification [29].

Support Vector Machine Originally proposed to build a linear classifier, the support vector machine (SVM) algorithm was developed by Vapnik in 1963 [21]. Used as a supervised machine learning algorithm, SVM can be applied to linear

and non-linear data, making it quite versatile in the range of possible uses. The way this algorithm works is based on the creation of a decision limit between two classes that allows later the prediction of labels of one or more characteristic vectors. This decision limit is called Hyperplane, and its positioning is calculated in function of the closest data points of each of the classes (support vectors). When used in nonlinear functions, this algorithm uses several kernel functions (linear, polynomial, radial and sigmoid) to maximize the margins between the Hyperplanes [3].

Random Forest The Random Forest (RF) algorithms are the most recent of those chosen for this comparison, published for the first time in 2001 in the article by Leo Breiman [12]. This type of algorithms has a field of applications similar to the previous ones, being able to be used both in regression or in classification problems [9]. Its operation follows a simple but very effective concept, the wisdom of crowds, i.e., the combination of several decision trees (not correlated with each other) operated as a committee, will always produce better results than any of the individual constituents. The constitution of RF models is composed of several decision trees, each one of which is trained on a sample of the original training data, and searches only on a randomly selected subset of the input variables to determine a split. To elaborate the final classification, each of the trees casts a unit vote of the obtained class, and, finally, the classifier's output is determined by the majority of the votes of the trees [17].

3.3 Evaluation Metrics

Since this is a classification problem, all the used metrics will be based on the confusion matrix generated for each algorithm, so it is essential to define all its constituents. Based on the predefined constitution for any type of binary problem, in this case too, the confusion matrix will be composed of True Positives (Predicted Positive and truly Positive), True Negatives (Predicted Negative and truly Negative), False Positives (Predicted Positive but truly Negative) and False Negatives (Predicted Negative but truly Positive). However, in this specific case, we will have a confusion matrix depending on the olive tree categories, which will resemble as presented in Table 1.

Table 1. Confusion matrix example.

		Predict Label			
		Cobrançosa (c)	Madural (m)	Negrinha (n)	Verdeal (v)
True Label	Cobrançosa (c)	N_{cc}	N_{cm}	N_{cn}	N_{cv}
	Madural (m)	N_{mc}	N_{mm}	N_{mn}	N_{mv}
	Negrinha (n)	N_{nc}	N_{nm}	N_{nn}	N_{nv}
	Verdeal (v)	N_{vc}	N_{vm}	N_{vn}	N_{vv}

As possible to seen in Table 1, the arrangement is a little different from the traditional True Positives/True Negatives, being necessary to add some values to arrive at some categories, such as:

- Cobrançosa True Positives (TP) = N_{cc}
- Cobrançosa False Positive (FP) = $N_{mc} + N_{nc} + N_{vc}$
- Cobrançosa True Negative (TN) = $N_{mm} + N_{mn} + N_{mv} + N_{nm} + N_{nn} + N_{nv} + N_{vm} + N_{vn} + N_{vv}$
- Cobrançosa False Negative (FN) = $N_{cm} + N_{cn} + N_{cv}$

Where N_{cc} represents the number of predicted Cobrançosa that are actually Cobrançosa and so on. Once the confusion matrix is explained, all the other metrics that will be based on that with very simplified calculations, as follows:

Accuracy The accuracy is the most used metric in this type of problem, its formula describes the number of correct predictions as a function of all predictions made and can be calculated as (Eq.1):

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

Precision The precision is used to measure positive patterns that are correctly predicted from the total predicted patterns in a positive class and is calculated using (Eq.2):

$$Precision(p) = \frac{TP}{TP + FP} \tag{2}$$

Recall The recall (Eq.3) is used to measure the fraction of True Positives that were correctly classified:

$$Recall(r) = \frac{TP}{TP + TN} \tag{3}$$

F-score The traditional F-score (or F1) is the harmonic mean of precision and recall and is defined by (Eq.4):

$$F - score = \frac{2 * Precision(p) * Recall(r)}{Precision(p) + Recall(r)} \tag{4}$$

3.4 Methodology

All the algorithms presented in this work were tested, trained, and implemented on a computer equipped with an Intel(R) Core(TM) i7-10875H processor, with a RAM DDR4 32GB memory and Python version 3.8.12 using the machine learning library scikit-learn [33]. The methodology used is shown in Figure 2.

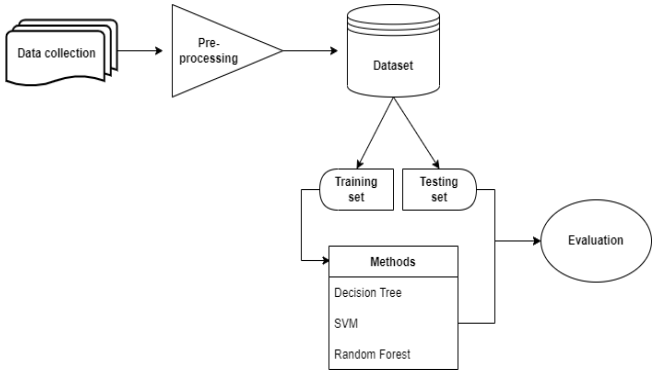

Fig. 2. Proposed system methodology

```
{'C': 0.1,
 'break_ties': False,
 'cache_size': 200,
 'class_weight': None,
 'coef0': 0.0,
 'decision_function_shape': 'ovr',
 'degree': 3,
 'gamma': 0.001,
 'kernel': 'poly',
 'max_iter': -1,
 'probability': False,
 'random_state': None,
 'shrinking': True,
 'tol': 0.001,
 'verbose': False}
```

Fig. 3. SVM best hyper-parameters.

```
{'ccp_alpha': 0.0,
 'class_weight': None,
 'criterion': 'gini',
 'max_depth': None,
 'max_features': None,
 'max_leaf_nodes': None,
 'min_impurity_decrease': 0.0,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'random_state': 0,
 'splitter': 'best'}
```

Fig. 4. Decision Tree best hyper-parameters.

```
{'bootstrap': True,
 'ccp_alpha': 0.0,
 'class_weight': None,
 'criterion': 'gini',
 'max_depth': None,
 'max_features': 'auto',
 'max_leaf_nodes': None,
 'max_samples': None,
 'min_impurity_decrease': 0.0,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'n_estimators': 100,
 'n_jobs': None,
 'oob_score': False,
 'random_state': 0,
 'verbose': 0,
 'warm_start': False}
```

Fig. 5. Random Forest best hyper-parameters.

After collecting and pre-processing the dataset, the training and test sets were divided. Here, it was stipulated *a priori* to use 90% of the dataset data for training and the remaining 10% for testing the algorithms. This division was done randomly using the split function of the library used. In this way, cross-validation of 5 times was applied to each of the algorithms either in the training stage or in the test stage.

The algorithm training process was an adapted process where some adjustments were made to improve its performance. This adjustment was made with the GridSearchCV function that allows to study the behavior of the fitting function according to the hyperparameters used, and in this way optimize their behavior. The first algorithm to be tested was the SVM, where the parameters of "C" (Regularization parameter), Gamma (kernel coefficient), and even the kernel to be used were tested, since it is a non-linear problem. After the evaluation with the GridSearchCV function and with a five times Cross-validation, the following combination of hyper-parameters was obtained in Figure 3.

Observing Figure 3, it is possible to verify that from all the used combinations, the one that produced the best results was C = 0.1, Gamma = 0.001, and kernel ="poly", with the remaining hyperparameters with default values. The exact process was carried out for the remaining DT (Figure 4) and RF (Figure 5) algorithms. In these cases, the hyperparameters that were tested were the minimum number of samples needed to split an internal node (min samples split), the minimum number of samples required to be at a leaf node (min samples leaf) and the maximum depth of the tree (max depth). Adding one more in the case of RF, the number of trees in the "forest" (n estimators), obtaining ideal values similar to those adopted by default in the respective algorithms.

4 Results and Discussion

After the optimization of the algorithms parameters was completed, their training was carried out by adopting the respective best parameters values. Once the training process was completed, they were then evaluated in the test set with the metrics chosen in subsection 3.3. To ensure that the algorithm generalizes correctly and can identify the leaves regardless of their position, color, or dimensions, the test set was randomly divided, and five-time cross-validation was used. The results obtained are presented in the following subsections:

4.1 Decision Tree

The training of this specific algorithm with the training dataset (1224 images) took approximately 17.414 seconds. Predicting the ratings in the test set (136 images) took about 1.281 seconds. The confusion matrix is presented in Table 2 was obtained through the average of the five tests performed.

Table 2. Decision Tree confusion matrix.

		Predict Label			
		Cobrançosa	Madural	Negrinha	Verdeal
True Label	Cobrançosa	22	4	2	6
	Madural	5	26	1	2
	Negrinha	4	2	27	1
	Verdeal	4	2	1	27

As seen from Table 2, the most problematic category is Cobrançosa, with only 22 hits in the set of 34 possible while the remaining categories have a similar rate of hits between them. For better understanding the behavior of the algorithm in the classification, the other metrics were applied, noting that here the results are the averages of the five iterations performed (Table 3).

Analyzing the data in Table 3, it is possible to see that the behavior of the metrics is similar for all categories, being the species Negrinha the one that

Table 3. Decision Tree metrics.

	Precision	Recall	F-score	Accuracy
Cobrançosa	0.66	0.65	0.65	0.75
Madural	0.78	0.78	0.78	
Negrinha	0.86	0.80	0.83	
Verdeal	0.74	0.78	0.76	

allows a better identification. The species Cobrançosa as it had already been perceptible in the confusion matrix is the species that cause the most identification difficulties in this algorithm. Observing the accuracy as a whole, it is noticeable that the Decision Tree still has some limitations in classifying of the four categories.

4.2 SVM

The training of the SVM algorithm took approximately 64.329 seconds, 3.5 times more than the Decision Tree when compared to the time required to make the test set predictions. The SVM also needs more time, using approximately 8.479 seconds, about 6.5 times more than the competitor. The confusion matrix resulting from the predictions of the SVM method is presented in Table 4.

Table 4. SVM confusion matrix.

		Predict Label			
		Cobrançosa	Madural	Negrinha	Verdeal
True Label	Cobrançosa	25	4	1	4
	Madural	4	28	0	2
	Negrinha	2	1	30	1
	Verdeal	5	1	1	27

In agreement with the previous results, the SVM algorithm also has some difficulties in identifying the Cobrançosa category, which is the one with the fewest hits, but already with an improvement of 3 images. Looking at the remaining categories, there is an improvement, albeit slight, in all categories, reaching the point of identifying 30 out of 34 possible examples in category Negrinha. The results of the remaining evaluation metrics are presented in Table 5.

As observed in the confusion matrix, from Table 5, it is possible to verify an increase in the performance of the SVM algorithm when compared to the Decision Tree. Variety Negrinha continues to be the most easily identifiable, achieving a harmonic average (F-score) of 0.9. When analyzing the accuracy of the SVM algorithm, it was noticed that despite the difficulty in identifying the first variety, there was a significant improvement compared to the previous method, with an increase of 0.06.

Table 5. SVM metrics.

	Precision	Recall	F-score	Accuracy
Cobrançosa	0.71	0.75	0.73	0.81
Madural	0.84	0.83	0.84	
Negrinha	0.92	0.88	0.90	
Verdeal	0.80	0.78	0.79	

4.3 Random Forest

The latest algorithm to be tested, Random Forest, took approximately 104.493 seconds to perform its training, six times more when compared to the Decision Tree. To perform the test set classifications, this algorithm took about 1.318 seconds, roughly the same time as the Decision Tree method. The results obtained by this latest algorithm are presented in the following confusion matrix Table 6.

Table 6. Random Forest confusion matrix.

		Predict Label			
		Cobrançosa	Madural	Negrinha	Verdeal
True Label	Cobrançosa	30	1	0	3
	Madural	1	33	0	0
	Negrinha	1	0	33	0
	Verdeal	3	0	0	31

Observing the values in Table 6, it is possible to verify that the Random Forest algorithm has a higher hit rate than its competitors, achieving 33 out of 34 hits in two categories. Looking deeper, Table 7 summarizes the remaining evaluation metrics presented below.

Table 7. Random Forest metrics.

	Precision	Recall	F-score	Accuracy
Cobrançosa	0.87	0.88	0.88	0.93
Madural	0.98	0.95	0.97	
Negrinha	0.97	0.98	0.97	
Verdeal	0.91	0.90	0.90	

After analyzing Table 7, it is possible to verify a significant improvement in the system's overall accuracy. Despite continuing to be the most difficult to identify, the Cobrançosa variety presented an F-score of 0.88, that represents an improvement of 0.15 compared to the SVM algorithm.

In short, after analyzing the results obtained, mainly confusion matrices, we can confirm that there are several mistakes in the identification of cultivars. This

failure is undoubtedly linked to the problem's difficulty, emphasizing once again that the visible differences in the leaves are very tenuous, making its identification very complex. In general, it is possible to classify the category Cobrançosa as the most challenging and most susceptible to errors. On the other hand, category Negrinha was the one that showed the highest hit rate in all algorithms.

Regarding the comparison of the algorithms, the method with higher observed accuracy, according to the used metrics, despite having a training time much higher than its competitors, the Random Forest algorithm guaranteed an overall accuracy of 93%, making the identification of olive tree species very reliable.

5 Conclusion and future works

The traditional methods of identifying the different cultivars of olive trees resort to genetic analysis. Such processes require time for execution and high associated costs. These factors make species identification a complicated process for farmers who need to do so. With the approach explored here, it is possible to make an identification on the spot, in real-time, and without associated costs. It is undoubtedly an added value for all producers who want to differentiate their oils. In addition to the producers, this approach is also an asset for tourism, giving tourists the ability to understand what type of olive grove they visit and what kind of oil will result from it.

As it was possible to verify throughout the article, there are currently no published articles that use only the image of the leaves to identify the olive tree cultivars. Thus, it becomes difficult to evaluate the effectiveness of the system when compared to other approaches. Observing the leaves of the different types, the difficulty in finding their differences is remarkable. However, analyzing the results obtained through the application of the three algorithms, it became clear that at least one of them can make this identification, not guaranteeing the efficiency rate of genetic methods; however, with an accuracy of 93%, which gives us the motivation to continue in this direction.

In this way, as main future works, it is intended to explore the classification algorithms in depth, approaching the aspect of deep learning. In addition to optimizing the classification model, web and smart devices interfaces will be developed giving the producers the possibility of testing in the field.

6 Acknowledgements

This work was carried out under the Project "OleaChain: Competências para a sustentabilidade e inovação da cadeia de valor do olival tradicional no Norte Interior de Portugal" (NORTE-06-3559-FSE-000188), an operation to hire highly qualified human resources, funded by NORTE 2020 through the European Social Fund (ESF). The authors are grateful to the Foundation for Science and Technology (FCT,Portugal) for financial support through national funds FCT/MCTES

(PIDDAC) to CeDRI (UIDB/05757/2020 and UIDP/05757/2020) and SusTEC (LA/P/0007/2021).

References

1. International olive oil. https://www.internationaloliveoil.org/olive-world/olive-tree/, accessed: 2022-06-07
2. The observatory of economic complexity. https://oec.world/en/profile/hs/olive-oil-fractions-refined-not-chemically-modifie, accessed: 2022-05-10
3. Ahmad, I., Basheri, M., Iqbal, M.J., Rahim, A.: Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection. IEEE access **6**, 33789–33795 (2018)
4. Aria, M., Cuccurullo, C.: bibliometrix: An r-tool for comprehensive science mapping analysis. Journal of informetrics **11**(4), 959–975 (2017)
5. Bautista, R., Crespillo, R., Cánovas, F.M., Gonzalo Claros, M.: Identification of olive-tree cultivars with scar markers. Euphytica **129**(1), 33–41 (2003)
6. Besnard, G., Baradat, P., Bervillé, A.: Olive cultivar identification using nuclear rapds and mitochondrial rflps. In: International Symposium on Molecular Markers for Characterizing Genotypes and Identifying Cultivars in Horticulture 546. pp. 317–324 (2000)
7. Besnard, G., Breton, C., Baradat, P., Khadari, B., Bervillé, A.: Cultivar identification in olive based on rapd markers. Journal of the American Society for Horticultural Science **126**(6), 668–675 (2001)
8. Beyaz, A., Özkaya, M.T., İçen, D.: Identification of some spanish olive cultivars using image processing techniques. Scientia Horticulturae **225**, 286–292 (2017)
9. Biau, G., Scornet, E.: A random forest guided tour. Test **25**(2), 197–227 (2016)
10. Bracci, T., Sebastiani, L., Busconi, M., Fogher, C., Belaj, A., Trujillo, I.: Ssr markers reveal the uniqueness of olive cultivars from the italian region of liguria. Scientia Horticulturae **122**(2), 209–215 (2009)
11. Bradski, G.: The OpenCV Library. Dr. Dobb's Journal of Software Tools (2000)
12. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
13. Breton, C., Terral, J.F., Pinatel, C., Médail, F., Bonhomme, F., Bervillé, A.: The origins of the domestication of the olive tree. Comptes rendus biologies **332**(12), 1059–1064 (2009)
14. Council, I.O.O.: World catalogue of olive varieties. International Olive Oil Council, 2000, Madrid, Spain (2000)
15. Ergulen, E., Ozkaya, M., Ulger, S., Ozilbey, N.: Identification of some turkish olive cultivars by using rapd-pcr technique. In: IV International Symposium on Olive Growing 586. pp. 91–95 (2000)
16. Fabbri, A., Hormaza, J., Polito, V.: Random amplified polymorphic dna analysis of olive (olea europaea l.) cultivars. Journal of the American Society for Horticultural Science **120**(3), 538–542 (1995)
17. Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R.: Random forests for land cover classification. Pattern recognition letters **27**(4), 294–300 (2006)
18. Grinblat, G.L., Uzal, L.C., Larese, M.G., Granitto, P.M.: Deep learning for plant identification using vein morphological patterns. Computers and Electronics in Agriculture **127**, 418–424 (2016)
19. Guinda, A., Lanzón, A., Albi, T.: Differences in hydrocarbons of virgin olive oils obtained from several olive varieties. Journal of Agricultural and Food Chemistry **44**(7), 1723–1726 (1996)

20. Heidary-Sharifabad, A., Zarchi, M.S., Emadi, S., Zarei, G.: An efficient deep learning model for cultivar identification of a pistachio tree. British Food Journal (2021)
21. Huang, S., Cai, N., Pacheco, P.P., Narrandes, S., Wang, Y., Xu, W.: Applications of support vector machine (svm) learning in cancer genomics. Cancer genomics & proteomics **15**(1), 41–51 (2018)
22. INE: Instituto nacional de estatística, estatísticas agrícolas de base. https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_base_dados, accessed: 2022-05-11
23. INE: Instituto nacional de estatística, previsões agrícolas. https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_destaques&DESTAQUESdest_boui=526211517&DESTAQUESmodo=2, accessed: 2022-05-11
24. Jain, A., Patel, H., Nagalapatti, L., Gupta, N., Mehta, S., Guttula, S., Mujumdar, S., Afzal, S., Sharma Mittal, R., Munigala, V.: Overview and importance of data quality for machine learning tasks. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 3561–3562 (2020)
25. Kotsiantis, S.B., Zaharakis, I., Pintelas, P., et al.: Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering **160**(1), 3–24 (2007)
26. Larese, M.G., Granitto, P.M.: Hybrid consensus learning for legume species and cultivars classification. In: European Conference on Computer Vision. pp. 201–214. Springer (2014)
27. Liu, C., Han, J., Chen, B., Mao, J., Xue, Z., Li, S.: A novel identification method for apple (malus domestica borkh.) cultivars based on a deep convolutional neural network with leaf image input. Symmetry **12**(2), 217 (2020)
28. Liu, Y., Su, J., Shen, L., Lu, N., Fang, Y., Liu, F., Song, Y., Su, B.: Development of a mobile application for identification of grapevine (vitis vinifera l.) cultivars via deep learning. International Journal of Agricultural and Biological Engineering **14**(5), 172–179 (2021)
29. Loh, W.Y.: Classification and regression trees. Wiley interdisciplinary reviews: data mining and knowledge discovery **1**(1), 14–23 (2011)
30. Martínez, S.S., Gila, D.M., Beyaz, A., Ortega, J.G., García, J.G.: A computer vision approach based on endocarp features for the identification of olive cultivars. Computers and Electronics in Agriculture **154**, 341–346 (2018)
31. Montaño, A., Sánchez, A., Casado, F., De Castro, A., Rejano, L.: Chemical profile of industrially fermented green olives of different varieties. Food Chemistry **82**(2), 297–302 (2003)
32. Nasiri, A., Taheri-Garavand, A., Fanourakis, D., Zhang, Y.D., Nikoloudakis, N.: Automated grapevine cultivar identification via leaf imaging and deep convolutional neural networks: a proof-of-concept study employing primary iranian varieties. Plants **10**(8), 1628 (2021)
33. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
34. Quinlan, J.R.: Induction of decision trees. Machine learning **1**(1), 81–106 (1986)
35. Reale, S., Doveri, S., Díaz, A., Angiolillo, A., Lucentini, L., Pilla, F., Martín, A., Donini, P., Lee, D.: Snp-based markers for discriminating olive (olea europaea l.) cultivars. Genome **49**(9), 1193–1205 (2006)
36. Sesli, M., Yegenoğlu, E., Altıntaa, V.: Determination of olive cultivars by deep learning and issr markers. Journal of Environmental Biology **41**(2), 426–431 (2020)

37. Shahriari, M., Omrani, A., Falahati-Anbaran, M., Ghareyazei, B., Nankali, A.: Identification of iranian olive cultivars by using rapd and microsatellite markers. In: V International Symposium on Olive Growing 791. pp. 109–115 (2004)
38. Tavakoli, H., Alirezazadeh, P., Hedayatipour, A., Nasib, A.B., Landwehr, N.: Leaf image-based classification of some common bean cultivars using discriminative convolutional neural networks. Computers and electronics in agriculture **181**, 105935 (2021)
39. Vanloot, P., Bertrand, D., Pinatel, C., Artaud, J., Dupuy, N.: Artificial vision and chemometrics analyses of olive stones for varietal identification of five french cultivars. Computers and Electronics in Agriculture **102**, 98–105 (2014)
40. Zhao, Y., Zhang, Y.: Comparison of decision tree methods for finding active objects. Advances in Space Research **41**(12), 1955–1959 (2008)