Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7, 93-110
2025
Publisher: Learning Gate
DOI: 10.55214/25768484.v9i7.8539

A machine learning-based classification of monocultivar olive oils—specifically Kalinjot, Ulli i bardhë Tirana, and Mixan—comparing their chemical composition

Ardiana Topi[1], Erdet Këlliçi[1], Daniel Hudhra[1], Dritan Topi[2*]
[1]European University of Tirana, Faculty of Engineering, Informatics and Architecture, Department of Informatics and Technology, Street Xhanfize Keko, Kompleksi Xhura, Tirana, 1000, Albania.
[2]University of Tirana, Department of Chemistry, Faculty of Natural Sciences, Boulevard Zogu 1, 1000, Tirana, Albania; dritan.topi@unitir.edu.al (D.T.).

Abstract: Valued for its nutritional and economic value, olive oil (OO) has been subject to adulteration practices. This study aimed to develop a classification model based on chemical composition to identify OOs from three main cultivars: Kalinjot, Ulli i bardhë Tirana, and Mixan, which comprise the majority of plantations in Albania. Eighty-five OO samples spanning different crop years and locations were studied using three different machine-learning algorithm models. The performance metrics of the k-Nearest Neighbors (kNN), Support Vector Machine (SVM), and Random Forest are discussed in terms of their classification performance. The comparison of accuracy revealed that the Random Forest model outperformed the others, achieving an accuracy of approximately 93%, compared to 81% for kNN and 78% for SVM. This significant finding, along with the clear confusion matrix of Random Forest, selects it as the preferred model for distinguishing OO based on cultivar and geographic origin. This project will help oil and oil extraction companies verify the authenticity of their products and detect adulteration practices in the Albanian oil sector.

Keywords: *Albania, kNN, Olive oil, Machine learning, Random forest, SVM.*

1. Introduction

The olive tree (*Olea europaea* L.) is an evergreen plant of significant importance to the economies and environments of countries in the Mediterranean region and other select regions worldwide. Distinguished for its fruit and oil, the olive plant has spread to other regions, such as Australia and North and South America, due to its valuable products: olive oil and table olives [1]. The western regions of Albania exhibit typical Mediterranean characteristics, where the olive tree has been cultivated (Figure 1) . Genetic studies have revealed the existence of twenty-two native olive cultivars, which are strictly present in six regions: Vlora, Berat, Elbasan, Kruja, Tirana, and Lezha. Among native cultivars, the most distinguished are Kalinjot, Mixan, Kokërrmadh Berati, Krips, and Ulliri i Bardhë Tirana, also known as Bianco di Tirana [2-4].

Olive oil (OO) is the most distinguished among vegetable oils due to its unique fatty acid profile and high content of phenolic compounds [5]. Studies suggest that OO may offer health benefits, including reducing coronary heart disease risk factors, preventing certain types of cancer, and modulating immune and inflammatory responses. It is widely recognized as a primary contributor to the Mediterranean Diet, attributed to its association with longevity and a low incidence of cardiovascular diseases among inhabitants of this region [6]. These benefits originate from the phenolic compounds, which aid in preventing several chronic diseases like atherosclerosis, cancer, chronic inflammation, strokes, and other degenerative diseases [7].

History: Received: 30 April 2025; Revised: 10 June 2025; Accepted: 13 June 2025; Published: 2 July 2025
* Correspondence: dritan.topi@unitir.edu.al

Being more expensive than other vegetable oils, the adulteration of OOs with cheap vegetable oils provides a significant economic interest. The most common OO adulteration practices involve the addition of vegetable oils, such as sunflower and soybean, as well as the addition of harmful or unnecessary substances that can alter their nature and quality [8, 9]. Other counterfeiting practices include mixing olive oil (OO) from different harvesting years, blending olive oil products with geographical designation with OO of unidentified origin, and mixing Virgin Olive oils with olive oil of low quality, such as pomace extraction [10].

Virgin olive oil (VOO) is extracted from the olive fruit solely through mechanical processes under conditions that do not alter the oil, and it undergoes no further treatment other than washing, decantation, centrifugation, and filtration [11]. It comprises triacylglycerols (~99%) and minor compounds, including phenolic compounds and sterols, aliphatic alcohols, tocopherols, and pigments [3]. Triglycerides are constituted from oleic (C18:1) and linoleic acid (C18:2), while palmitic (C16:0), palmitoleic (C16:1), stearic (C18:0), and linolenic (C18:3) fatty acids are identified in minor amounts. Other fatty acids, such as myristic acid (C14:0) and eicosanoic acid, are found in minor amounts [12]. Phenolic compounds are present in various parts of *Olea europaea* L. phenolic compounds belong to six major families of compounds: phenolic acids, flavonoids, isochromans, lignans, secoiridoids, and derivatives [5]. Tyrosol and hydroxytyrosol, and derivatives of 4-hydroxybenzoic, 4-hydroxyphenylacetic, and 4-hydroxycinnamic acids, constitute the most distinguished among them [4].

The OO fatty acid (FA) composition may differ according to cultivars, production zones, latitude, climate, variety, and fruit maturity stages. Greek, Italian, and Spanish OOs are low in linoleic and palmitic acid, while a high percentage of oleic acid. Olive oils from North African countries, e.g., Tunisia, are high in linoleic and palmitic acids and lower in oleic acid [1].

Despite the diversity of olive cultivars in Albania, the most important native cultivar is Kalinjot, approximately 50% of plantations nationally, with over 70% in southern Albania, regions of Vlora and Mallakastra (Table 3), by contributing significantly to domestic olive oil production [13, 14]. The phenolic compounds of Albanian olive oil (OO) from the distinguished native olive cultivars Kalinjot, Bardhi Tirana, Ulli i Zi, Krips Kruja, and Bardhi Kruja have been characterized [4]. The most abundant phenolic class, regardless of the cultivar, was secoiridoids, followed by phenolic alcohols. Hydroxytyrosol and tyrosol were the dominant compounds in this group, especially in Kalinjot OO [4].

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 93-110, 2025
DOI: 10.55214/25768484.v9i7.8539

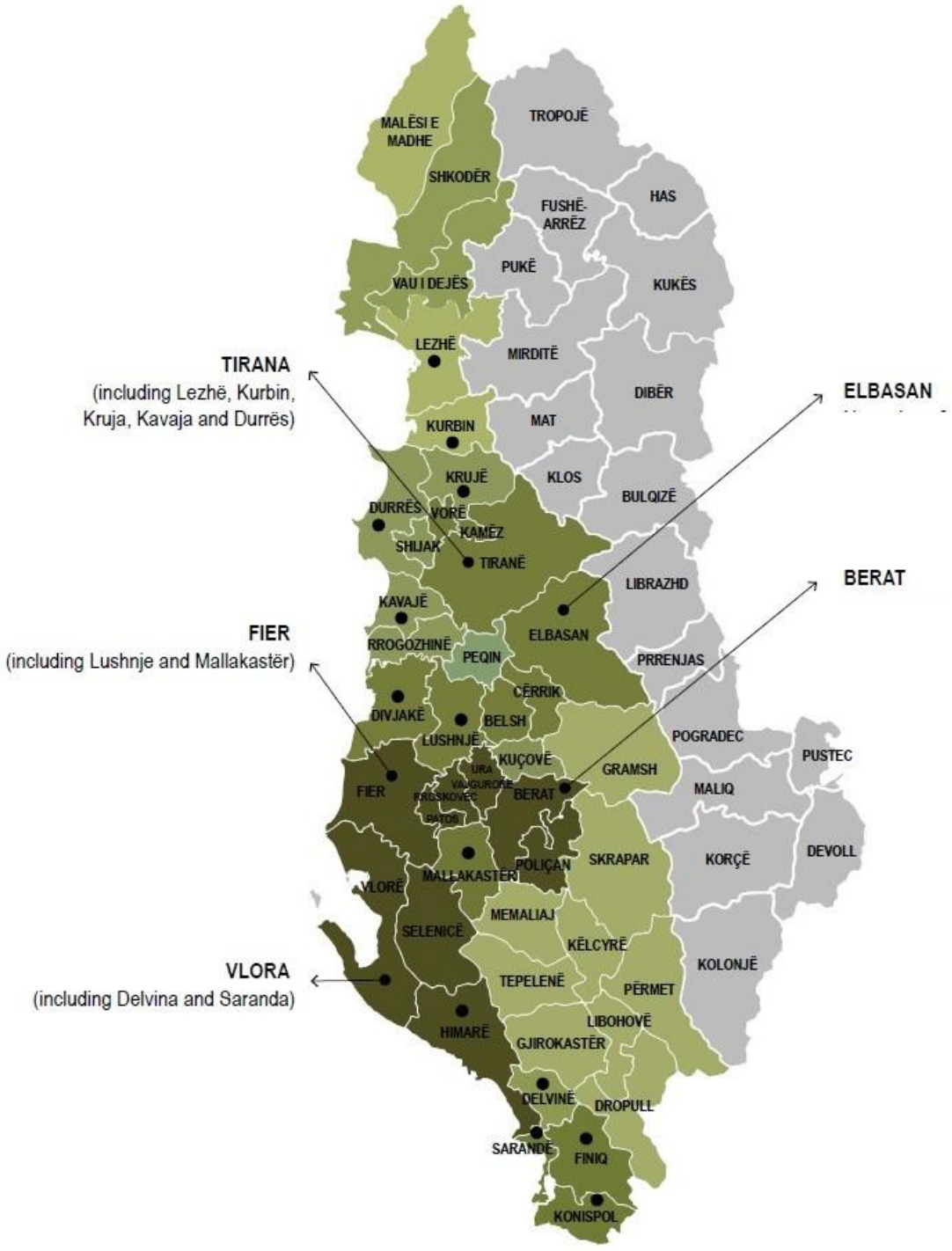

Figure 1.
Olive plantation regions of Albania.
Source: Martakos, et al. [10].

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 93-110, 2025
DOI: 10.55214/25768484.v9i7.8539

According to the national statistics, the main olive-producing areas are Berat, Elbasan, Fieri, and Vlora. Fieri is the central region for the first three indicators (total number of plants, number of plants in production, and total output). The economic potential of the main olive cultivars is remarkably high, contributing to rural economies and regions with low developing potential, such as the country's southern and inland hilly regions [15].

Establishing the OO of designated origin has raised concerns about the traceability of geographical origin. Mislabelling geographical origin or using other olive varieties can impact consumer confidence and the reputation of olive-producing regions. Unfortunately, an increasing number of low-quality OOs often end up on our tables and are not easily identifiable. Consequently, to ensure the product's authenticity, considering its geographical origin, cultivar, and agricultural practices is essential for maintaining and improving its supply chain by implementing traceability systems.

While traditional empirical methods have been relied upon to detect olive oil fraud and evaluate quality, this study pioneers a modern approach using machine learning algorithms to differentiate authentic products from counterfeits. Random Forest (RF) and k-Nearest Neighbors (kNN) algorithms, among others [16]. Evaluation of VOO from specific cultivars and regions, employing and testing different AI-based classification techniques to determine the highest accuracy [17].

2. Material and Methods

An artificial intelligence-driven system was developed, which is easily accessible, to predict the quality and origin of the VOO product based on physical and chemical properties. We identified two approaches to data. Based on seven input characteristics, including fatty acids such as palmitoleic, stearic, oleic, and linoleic acid, the algorithm can predict the VOO geographical region, represented by the output variable (in the cited study, there are three regions, e.g., Vlora, Fieri, and Berati). Therefore, our target (output) is also categorical, as noted by Aiello and Tosi [18]. The dataset #3, with 572 data objects, was closer to our study objective.

Analogously, we can use similar input characteristics in the dataset being processed for our study. Each data object (olive oil) needs to be populated with accurate data regarding its fatty acid percentage. The prediction accuracy extracted from Machine Learning algorithms increases with the data. From a confusion matrix perspective, a correct prediction can usually be interpreted as a True Positive (TP) value. For example, if the actual data is "Kalinjot" OO and the predicted data is also "Kalinjot" OO, then the algorithm has performed the prediction correctly.

2.1. Support Vector Machine

Support Vector Machine (SVM) is one of the most popular supervised machine learning algorithms. It enables the maximization of the discriminant boundary. The method divides the samples through an optimal hyperplane, maximizing the distance among the classes of data points within a multidimensional space [19, 20].

In our case study, being a "multi-class" classification, we can use one of the methods:

1-vs-rest. So, we use classifiers; for example, one of them is ["Kalinjot"] *vs* ["Berati, "Fieri," "Saranda"].

1-vs-1. In this case, to generate a classifier, we use the formula $N * \frac{N-1}{2}$, where

N- indicates the number of classes we are considering. So, for N = 4, we will have six classifiers, e.g., "Kalinjot" vs. "Berati" and "Kalinjot" *vs.* "Fieri."

2.2. The k-Nearest Neighbour

The k-nearest Neighbour (kNN) algorithm performs the classification task by predicting a new given object based on the Euclidean distance between the training point and the test observation.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 93-110, 2025
DOI: 10.55214/25768484.v9i7.8539

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \quad (1)$$

In Equation (1), n represents the number of dimensions, whereas x_i and y_i denote the data points [20].

2.3. The Dataset Algorithm Adjustment

For a data object like olive oil, the user specifies attributes such as 'Oleic acid (OA)' and 'Linoleic acid (LA),' and the algorithm determines whether the olive oil belongs to one of the following categories: Kalinjot VOO, Berati, Fieri, or Saranda region.

Step 1. Calculate the distance based on the input characteristics (e.g., OA, LA).

Step 2. Determining the rank. The first rank corresponds to the smallest distance identified in the initial step.

Step 3. The "nearest neighbor" is identified using a specific value of k. For instance, when k = 1, the top-ranked place will be chosen (Kalinjot). Conversely, with k = 4, we examine the top four ranks. If three of these four ranks correspond to Kalinjot, it indicates that these three data points are close in distance. Consequently, this leads to the prediction that the new data object will be classified as a Kalinjot object of type OO.

2.4. The Olive Oil Cultivar Predictor

The thirty-four Kalinjot VOO dataset was collected from various regions of the country. The aim is to predict the cultivar's origin, a "classification task" or "multi-class classification."

Key features include specific fatty acids and phenolic compounds that allow us to distinguish the cultivar origin, considering that soil and climate also impact the VOO chemical composition. Secondly, the ML model enables the classifier to learn from these values, uncovering hidden patterns and predicting the origin of the samples based on their geographical location.

To balance the dataset, we utilized the Synthetic Minority Oversampling Technique (SMOTE) to replicate instances of the minority class. SMOTE is an over-sampling technique that increases the occurrence of less frequent classes in the training set, generating new samples to alleviate the class imbalance [20].

Table 1.
Features, description, and value type.

Feature	Feature information	Value type
Special Fatty Acids	Uncommon or less abundant fatty acids, such as 17:0 and 17:1 (n-7), indicate additional profile variability.	Continuous
Saturated Fatty Acids (SFA)	Includes fatty acids like 14:0 (myristic), 16:0 (palmitic), 18:0 (stearic), 20:0 (arachidic), 21:0 (henicosanoic), 22:0 (behenic).	Continuous
Monounsaturated Fatty Acids (MUFA)	Fatty acids with one double bond, such as 16:1(n-9), 16:1(n-7), 18:1(n-9)cis, 18:1(n-9)trans, 18:1(n-7), 20:1(n-9).	Continuous
Polyunsaturated Fatty Acids (PUFA)	Fatty acids with multiple double bonds, including 18:2(n-6)cis, 18:2(n-6)trans, 18:3(n-3), 20:2(n-6), 20:3(n-3).	Continuous
Fatty Acid Ratios	Ratios such as (n-6)/(n-3) and 18:1/18:2 reflect the balance between omega-6 and omega-3 fatty acids or specific fatty acids.	Continuous

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 93-110, 2025
DOI: 10.55214/25768484.v9i7.8539

2.5. Training and Testing Datasets

The *scikit-learn* library in Python was used. The dataset was split into 20% for testing and the remaining 80% for training the model (Figure 2). The model was trained on *X_train* and *Y_train* in the training dataset, while the prediction model used *X_test* and compared the predicted results with *Y_test*.

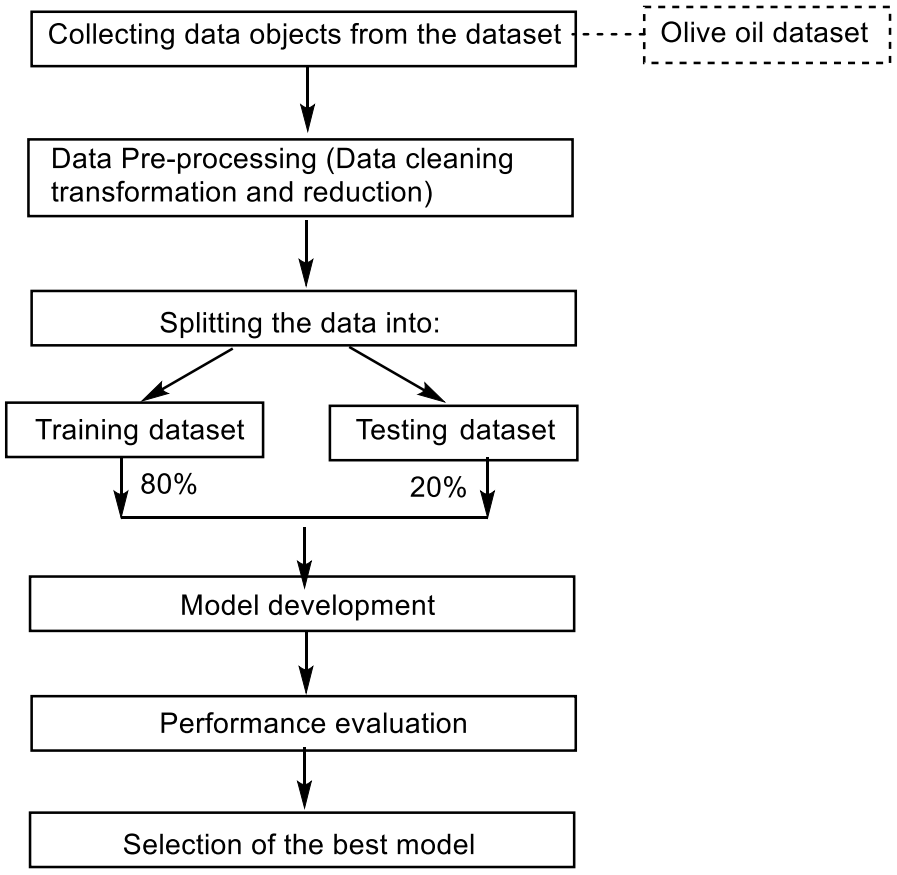

Figure 2.
The research methodology used.

2.6. Feature Selection and Extraction

In the Olive Oil dataset, feature scaling is crucial because our chosen machine learning algorithms must calculate distances between data points. The *scikit-learn* Python module integrates the newest machine-learning algorithms in supervised and unsupervised cases [20]. The standardized data means $\mu = 0$ and $\sigma = 1$. The following formula for the calculation of the standard score of sample x :

$$z_i = \frac{x_i - \mu}{\sigma}$$

Where x_i is each value, μ is the mean value of the training samples, and σ is the standard deviation of the training samples. Conversely, Principal Component Analysis (PCA) is a dimensionality reduction method that reduces the number of input features while preserving as much information as possible from the original dataset. PCA converts the twenty-four original features into smaller principal components, each representing a sizeable portion of the dataset's variance [20].

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 93-110, 2025
DOI: 10.55214/25768484.v9i7.8539

Table 2.
Post-standardization statistical data (mean and StDev) for the twenty-four features.

	1	2	3	4	5	6
Count	28.0	28.0	28.0	28.0	28.0	28.0
Mean	0.0	0.0	0.0	0.0	0.0	0.0
StDev	0.0	1.0	1.0	1.0	1.0	1.0
Min	0.0	-1.7	-2.6	-1.2	-1.5	-1.6
25%	0.0	-0.6	-0.5	-0.7	-1.1	-0.9
50%	0.0	-0.3	-0.0	-0.3	0.6	0.4
75%	0.0	0.8	0.7	0.2	0.8	0.7
Max	0.0	2.2	1.4	2.4	1.2	1.3
8 rows x 24 columns						

2.7. Evaluation Metrics

Several metrics help assess a model's ability to correctly classify the OO into their respective origins [21].

Accuracy: indicating the number of correctly classified instances over the total number.

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

After performing SMOTE, our target class achieved a well-balanced score, indicating accuracy —a good metric for measuring how often the classifier predicts correctly.

Precision: expressing the proportion of positive instances predicted, predicted as positive.

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

Recall the proportion of current positive instances precisely predicted as positive.

$$Recall = \frac{TP}{TP + FN}$$

F1 score: calculated as the balanced mean of recall and precision.

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

3. Results and Discussions

Application of ML models for olive oil classification based on chemical composition has been successfully applied recently [22].

3.1. Feature importance using Random Forest

The importance of features for predicting olive oil cultivar origin using the Random Forest classification model is presented in Figure 3. The dataset comprised OO samples from the Kalinjot cultivar collected from various regions in Albania, including Vlora, Berat, and Fier (resulting in three target classes). The analysis aimed to classify the Kalinjot VOO, utilizing chemical characteristics such as fatty acid (FA) profiles, polyunsaturated fatty acids (PUFA), monounsaturated fatty acids (MUFA), and the 18:1/18:2 ratio as input features.

Among these features, the ratio of 18:1 to 18:2 was identified as the most crucial in distinguishing the Kalinjot VOO. The MUFA group significantly influenced the origin of our target variable, indicating that the ratio among fatty acid (FA) groups is specific to the region. The same conclusion was drawn for the PUFA and SFA groups, reflecting the broader FA composition within the same cultivar. The feature f14:0 and specific trans-FA have the lowest importance scores, suggesting they contribute minimally to differentiating the samples based on geography.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 93-110, 2025
DOI: 10.55214/25768484.v9i7.8539

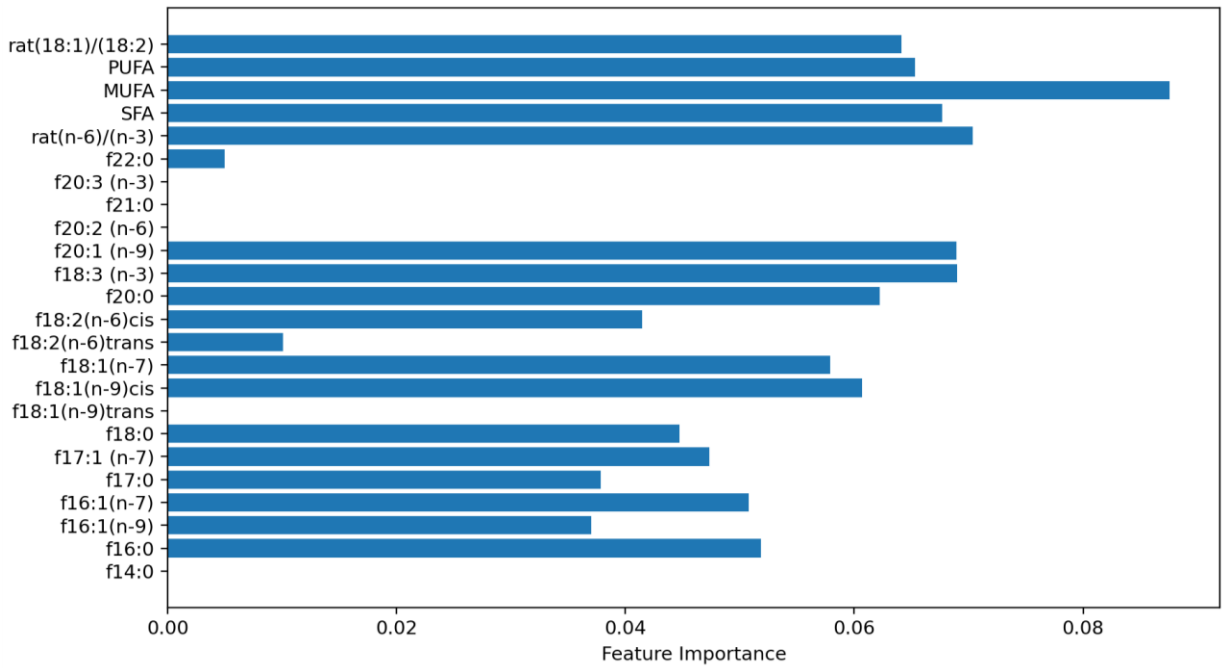

Figure 3.
Random Forest model and feature importance on the Olive Oil dataset.

The Gini impurity metric assessed the role of every feature in reducing uncertainty during classification. Features with higher importance scores appear more frequently in decision-making nodes within the model, which is pivotal in achieving accurate predictions of the olive oil's origin. This highlights the importance of utilizing well-chosen chemical markers to achieve robust classification outcomes. In conclusion, the feature importance analysis highlights the key chemical indicators that facilitate a reliable classification of Kalinjot OO based on their geographical origin.

MUFA — holds the highest importance, indicating it plays a key role in determining the origin of olive oil. SFA (Saturated Fatty Acids) and PUFA – also demonstrate strong importance. Other major contributors include ratios like rat(18:1)/(18:2) and fatty acids such as f20:2(n-6) and f18:2(n-6)cis.

3.2. Evaluation Metrics Analysis

The evaluation metrics provide critical insights into the strengths and weaknesses of machine learning models in predicting the origin of Albanian mono-cultivar OO. Logistic Regression and SVC demonstrated moderate overall performance, achieving an accuracy of 72.4%. Despite these models show balanced precision and recall for specific classes, they struggle to correctly identify samples from Vlora, as evidenced by low precision and recall values for this class. While conceptually simple, the k-Nearest Neighbours algorithm underperformed with an accuracy of 50%. However, after hyperparameter tuning, kNN performed better (Table 3-7).

Similarly, the Decision Tree model exhibited overfitting tendencies, resulting in an overall accuracy of 50% and inconsistent predictions across regions. Random Forest, a more robust ensemble method, showed slight improvements for specific classes but failed to generalize effectively, yielding 50% accuracy. Gradient Boosting provided a more nuanced balance in classifying oils. However, its overall accuracy remained constrained by challenges in capturing the underlying data distribution, particularly for samples from Berati and Vlora. These findings underline the limitations of standard classifiers in

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 93-110, 2025
DOI: 10.55214/25768484.v9i7.8539

tackling regionally diverse and imbalanced datasets. Issaad, et al. [23] successfully demonstrated that ML models can distinguish the OO produced in different regions and different cultivars from Algeria.

Table 3.
Classification report on the SUPPORT VECTOR CLASSIFIER model accuracy.

Accuracy: 72.50 %				
	Precision	**Recall**	**F1-score**	**Support**
	0.6	0.724	0.571	
Berati	0.00	0.00	0.00	2
Fieri	1.00	0.75	0.86	4
Vlora	0.40	1.00	0.57	2

Table 4.
Classification report on the kNN model accuracy.

Accuracy: 50.00 %				
	Precision	**Recall**	**F1-score**	**Support**
	0.5	0.5	0.467	
Berati	0.00	0.00	0.00	2
Fieri	0.50	0.75	0.60	4
Vlora	1.00	0.50	0.67	2

Table 5.
Classification report on the RANDOM FOREST model accuracy.

Accuracy: 72.40 %				
	Precision	**Recall**	**F1-score**	**Support**
	0.625	0.625	0.625	
Berati	0.00	0.00	0.00	2
Fieri	0.75	0.75	0.75	4
Vlora	0.33	0.50	0.40	2

Note: The last column ("Support") provides the number of samples in each class.

The Support Vector Classifier (SVC) model correctly classifies OO samples from Fieri (3) and Vlora (2) but fails to classify Berat samples with an accuracy of 62.5%. The k-Nearest Neighbors (KNN) algorithm correctly classifies OO from the Fier region and Vlora but fails to classify OO from Berat with an accuracy of 57.1%. The Random Forest model was the best performer, correctly classifying OO from the Fier region and Vlora, but failed to classify OO originating from Berat with an 66.7% accuracy.

3.3. Confusion Matrix Analysis

The confusion matrices revealed key patterns in the misclassification tendencies of the models, shedding light on the regional separability of the OO samples. Across most models, samples originating from Fieri were consistently classified with higher accuracy, suggesting more distinct feature patterns for this region. Conversely, OO samples from Berati and Vlora were frequently misclassified, indicating overlapping characteristics. For example, Logistic Regression and SVC exhibited significant confusion in differentiating OO samples from the Vlora, often misclassifying them as belonging to Fieri. Gradient Boosting showed marginal improvement in class distribution but struggled with borderline cases between Berati and Fieri. Despite its ensemble nature, Random Forest demonstrated similar biases to the individual Decision Tree model, particularly in under classifying Vlora samples.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 93-110, 2025
DOI: 10.55214/25768484.v9i7.8539

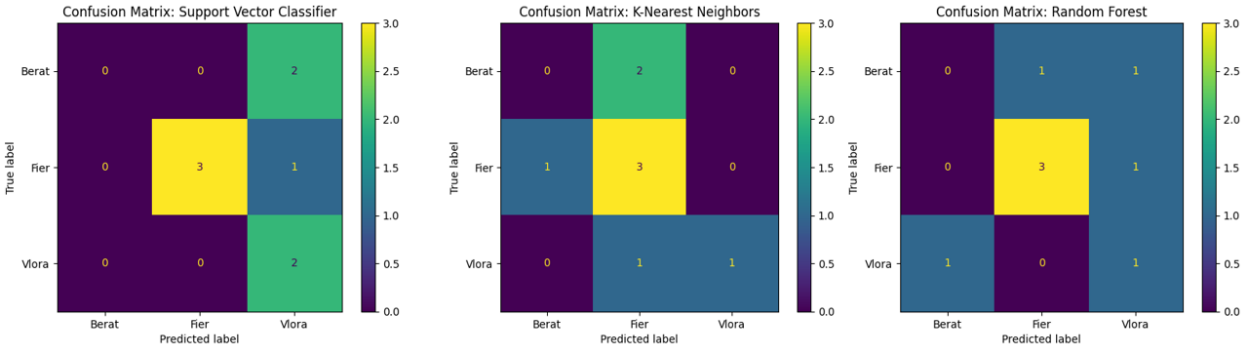

Figure 4.
Visualization of confusion matrices created using the Confusion Matrix Display class from the sklearn—metrics library.

Since our dataset is small, consisting of only thirty-four instances, the observed trends reflect the dataset's complexity and the limitations of training on such as the small sample size. This highlights the importance of adopting advanced pre-processing techniques, in addition to feature scaling, such as dimensionality reduction or to improve class separability and enhance model performance. To achieve more excellent predictive reliability, future work should explore more sophisticated algorithms and consider collecting a larger, more representative dataset to capture the full variability of the features. However, the Confusion Matrices and Model Comparisons concluded that each confusion matrix provides a visual breakdown of how each classification model performed on three olive oil origin labels: Berat, Fier, and Vlora.

3.4. Hyperparameter Tuning for the K-Nearest Neighbor Classifier

The hyperparameters of different models were tuned to optimize their performance. Precisely, for Logistic Regression, the regularization strength of the Support Vector Machines (SVM), the penalty parameter C; in the Decision Tree Model, the maximum depth, and minimum samples per split; while in the Random Forest, the number of estimators and maximum features.

In the post-tuning evaluation of ML models, it was observed that the kNN algorithm, with its optimized hyperparameters, performed better than the other models (www.geeksforgeeks.org/k-nearest-neighbours). The Euclidean distance between two points in the hyperplane was used to estimate the k-NN for a given data point. This metric, along with Manhattan distance, is a particular case of Minkowski distance:

$$d(x,y) = \left(\sum_{i=1}^{n}(x_i - y_i)^p\right)^{\frac{1}{p}}$$

Setting the Minkowski metric (p) to 2 produced the Euclidean distance. According to *scikit-learn*, the value of the other main kNN hyperparameter (the number of k neighbors) was set to 5 by default (Table 8). Experiments with k values ranging from 1 to 28 determined the optimal k value to be 4, with an accuracy of 72.4% (Figures 5 and 6).

Table 6.
The k-Nearest Neighbor training and testing accuracy.

Value of k	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Accuracy	0.5	0.375	0.5	0.724	0.5	0.5	0.5	0.625	0.375	0.375	0.375	0.125	0.125	0.125	0.25

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 93-110, 2025
DOI: 10.55214/25768484.v9i7.8539

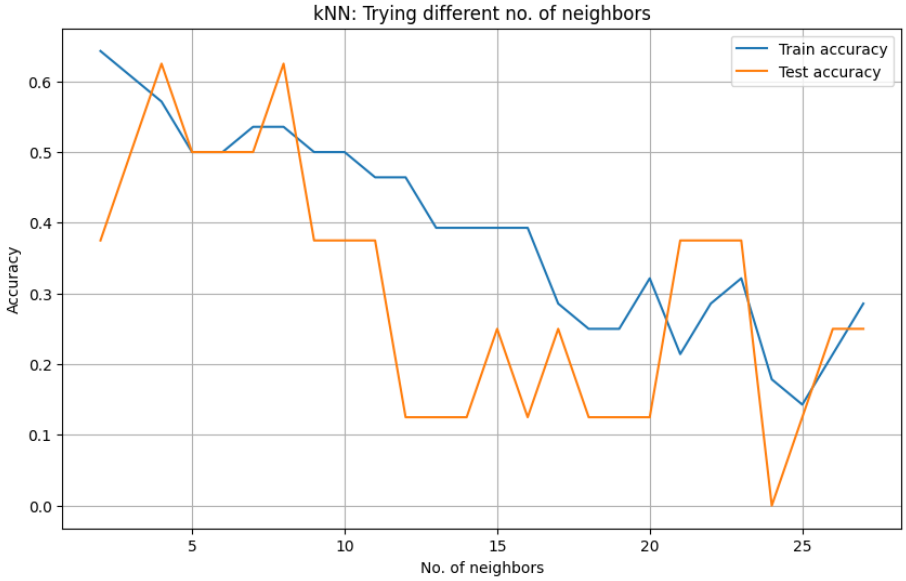

Figure 5.
The k-nearest Neighbor value and its corresponding accuracy.

A second method for estimating the optimal value of k is to use the error rate chart for k values within the specified range (Figure 6). The value k = 4 represents the lowest point (error), indicating the optimal value.

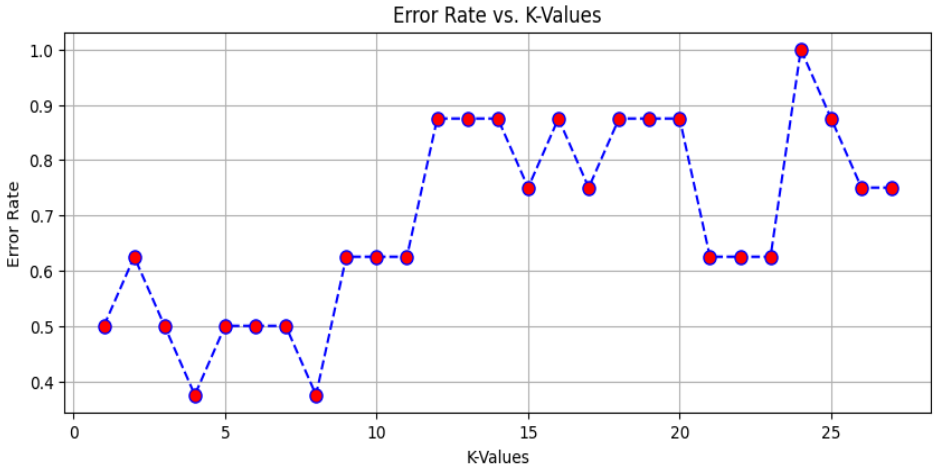

Figure 6.
The k-nearest Neighbor value and its corresponding Error Rate.

The Hyperparameter Tuning plots in KNN (k-value) demonstrate how the choice of a number of neighbors (k) affects KNN model performance:

The Accuracy Plot indicates that training accuracy steadily decreases as k increases (as expected) while testing accuracy fluctuates significantly. Peaks at k = 3 and 6, suggesting small k values work better and drop severely for larger values of k. The Error Rate Plot indicates that the lowest error rates

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 93-110, 2025
DOI: 10.55214/25768484.v9i7.8539

occur at k = 3 and k = 6, reinforcing earlier accuracy findings. Beyond k = 10, the error remains high, suggesting poor generalization.

3.5. Hyperparameter Searching Procedure

The Grid Search process involves the creation of a hyperparameters grid by searching for the combination that yields the highest validation score. The kNN classifier was applied to optimize the parameters of the neighbors, allowing for control over the influence of neighbors on the classification decision. Using 5-fold cross-validation (*KFold(n_splits=5)*), the data splitted into training and testing sets to evaluate each candidate value. Finally, the model was assessed using unverified data during each iteration, reducing the overfitting risk (usually characteristic of low values of k).

After evaluating all combinations, GridSearchCV identified the optimal number of neighbors, which was subsequently used to train the final kNN model on the entire training dataset. This contributes to a significant improvement in accuracy and performance.

The Grid Search, which favors Randomized Search, was connected with a limited number of hyperparameters, although it consumes considerable computational resources. We might train different models in the future, increasing the number of hyperparameters and their values.

3.6. Graphical User Interface (GUI)

The graphical user interface (GUI) is designed to classify the origin of Kalinjot OO samples based on their chemical composition. It provides input fields for fatty acid concentrations. The user enters the values, and by clicking the "Predict Origin" button, the application employs a pre-trained machine learning model and a scaler to prepare the inputs and determine the origin of the olive oil sample. The predicted region appears below the button in real time (Figure 7).

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 93-110, 2025
DOI: 10.55214/25768484.v9i7.8539

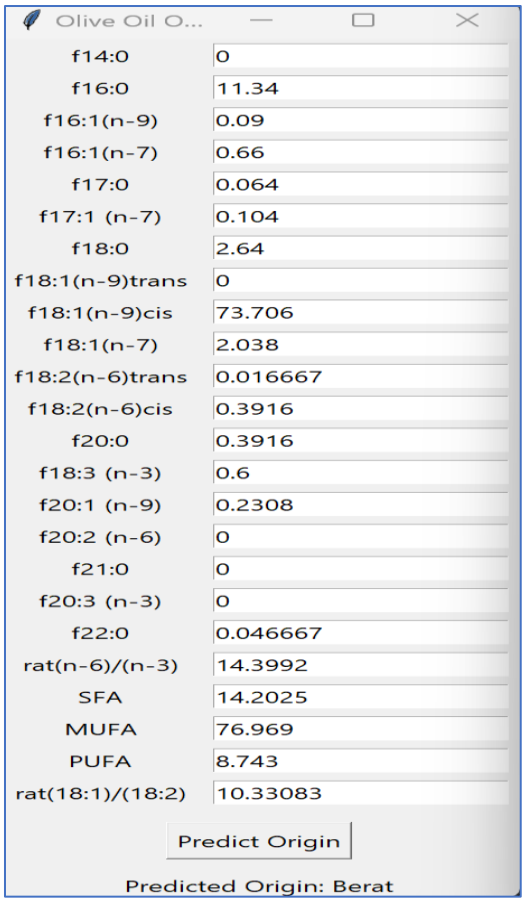

Figure 7.
Implementing the GUI.

The dataset used in this study was applied to the Kalinjot VOO, comprising samples from the Berat, Fier, and Vlora regions. However, we anticipate a significant increase in the dataset size through consecutive harvesting seasons, potentially covering additional regions and incorporating seasonal variations. The RF algorithm is particularly suited for this type of expansion as it can scale effectively and manage large datasets without sacrificing performance. This interface suggests a working end-user tool for real-time prediction of OO origin.

3.7. Model Comparison accuracy of Random Forest vs SVM and kNN for Three Olive Cultivars

A model performance comparison for the Classification Task revealed that Random Forest achieved an accuracy of approximately 93%, making it the highest among the evaluated models. This model produced a clean confusion matrix, indicating that it makes relatively few classification errors and handles both true positives and true negatives effectively.

Random Forest aggregates predictions from multiple decision trees by reducing the risk of overfitting and improving generalization.

Regarding robustness to Noise and Overfitting, averaging multiple trees tends to be less sensitive to outliers and noise in the data. Concerning 'Feature Importance,' the model can naturally rank the importance of input features, aiding in model interpretability and further tuning. Employing the Support Vector Machine (SVM), it was concluded that the accuracy is slightly lower than Random

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 93-110, 2025
DOI: 10.55214/25768484.v9i7.8539

Forest, while the model's strength lies in its effectiveness in high-dimensional spaces. The k-nearest Neighbors (KNN) model's accuracy was slightly lower compared to the Random Forest model (Figure 8). The model's strength lies in its simplicity and intuition, making no assumptions about the underlying data distribution. It is also adaptable to small datasets with well-separated classes. Random Forest's high accuracy and clean confusion matrix make it the preferred model for this task.

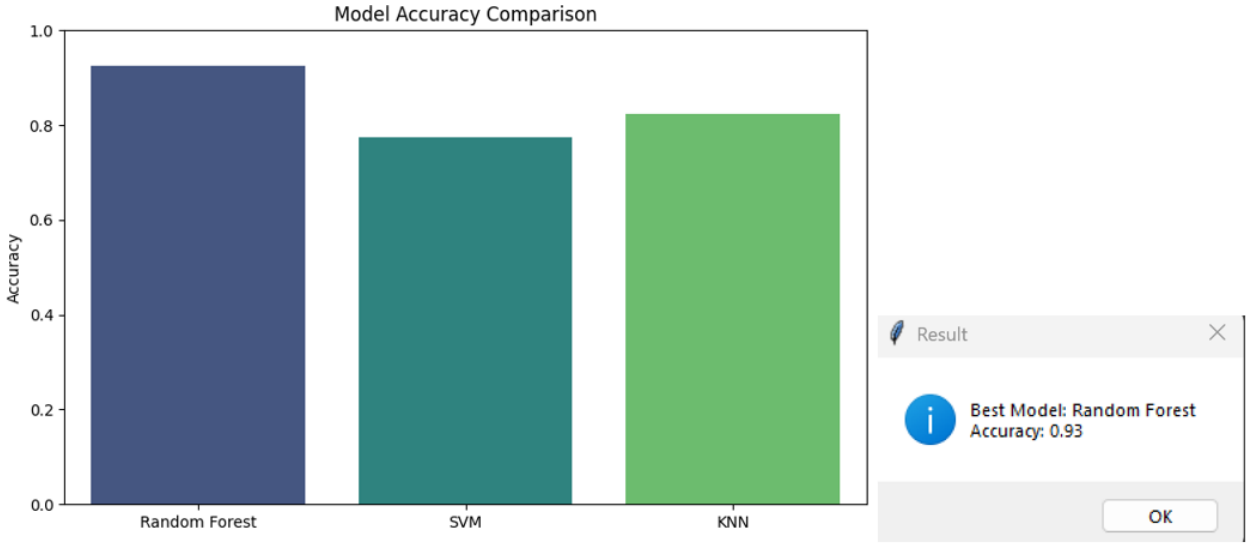

Figure 8.
The machine learning-based classification of three mono-cultivar OOs, Kalinjot, Ulli i bardhë, and Mixan.

This PCA plot reduces the many chemical features into two principal components (PC1 and PC2) (Figure 9). The plot shows Kalinjot OO (green) is well-separated from Ulli i Bardhë OO (orange) and Mixan OO (blue). In contrast, Mixan OO and Ulli i Bardhë OO are closer in feature space, which could explain why fewer samples are available and why they may be harder to distinguish than Kalinjot. Another explanation is that the regions associated with the above-mentioned Olive cultivars are geographically close and share similar climate characteristics.

Principal Component Analysis (PCA) is a statistical technique for dimensionality reduction that converts a large set of variables into a smaller set while retaining most of the original information. In our study, the data set employed to describe different types of olive oils includes twelve chemical features (e.g., fatty acid levels) for each sample. The PCA transforms these high-dimensional data into fewer dimensions (here, 2D) that retain the most critical information. These new dimensions are referred to as Principal Components (PC1 and PC2). PC1 captures the direction of the most significant variation in the data, while PC2 is the direction of the second most significant variation, orthogonal to PC1. Kalinjot OO (green) is well-separated from Ulli i Bardhë OO (orange) and Mixan OO (blue), supporting the efficacy of the 2D PCA plot in describing the differences between olive oils. Kalinjot OO forms a distinct cluster, far from the other two mono cultivar OO, indicating Kalinjot's distinct chemical composition from the others, at least in terms of the main features captured by PC1 and PC2. Machine learning models can facilitate easier classification or differentiation in the quality control and authentication process. Mixan OO (blue) and Ulli i Bardhë OO (orange) are plotted close together, meaning that their chemical profiles are more similar. It may require explanation, referring to a limited number of samples analyzed, which are more difficult to distinguish from Kalinjot OO due to their similar chemical composition. The PCA plot shows that Kalinjot OOs are chemically distinct and easily

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 93-110, 2025
DOI: 10.55214/25768484.v9i7.8539

recognizable compared to the other two monocular OOs, Mixan and Ulli I Bardhë. This type of analysis is beneficial for understanding product variation, quality control, and classification tasks in fields such as food science or chemistry.

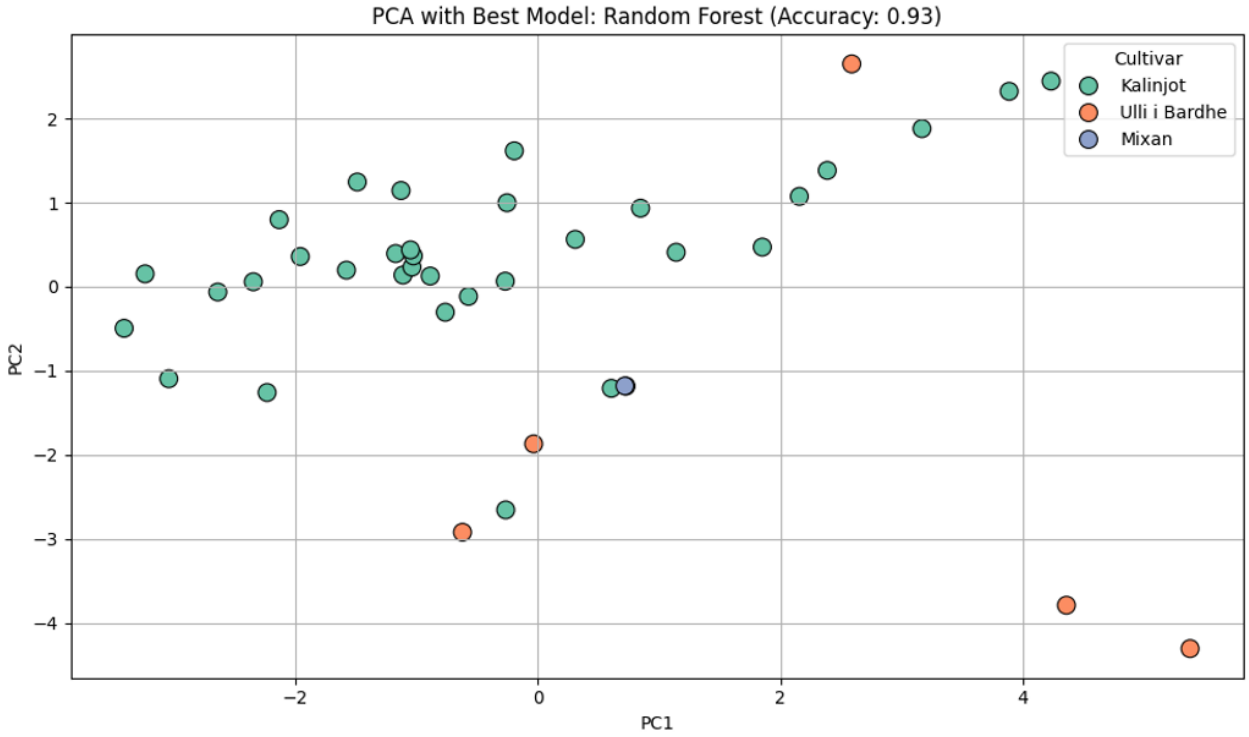

Figure 9.
PCA plot refers to the Random Forest model.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 93-110, 2025
DOI: 10.55214/25768484.v9i7.8539

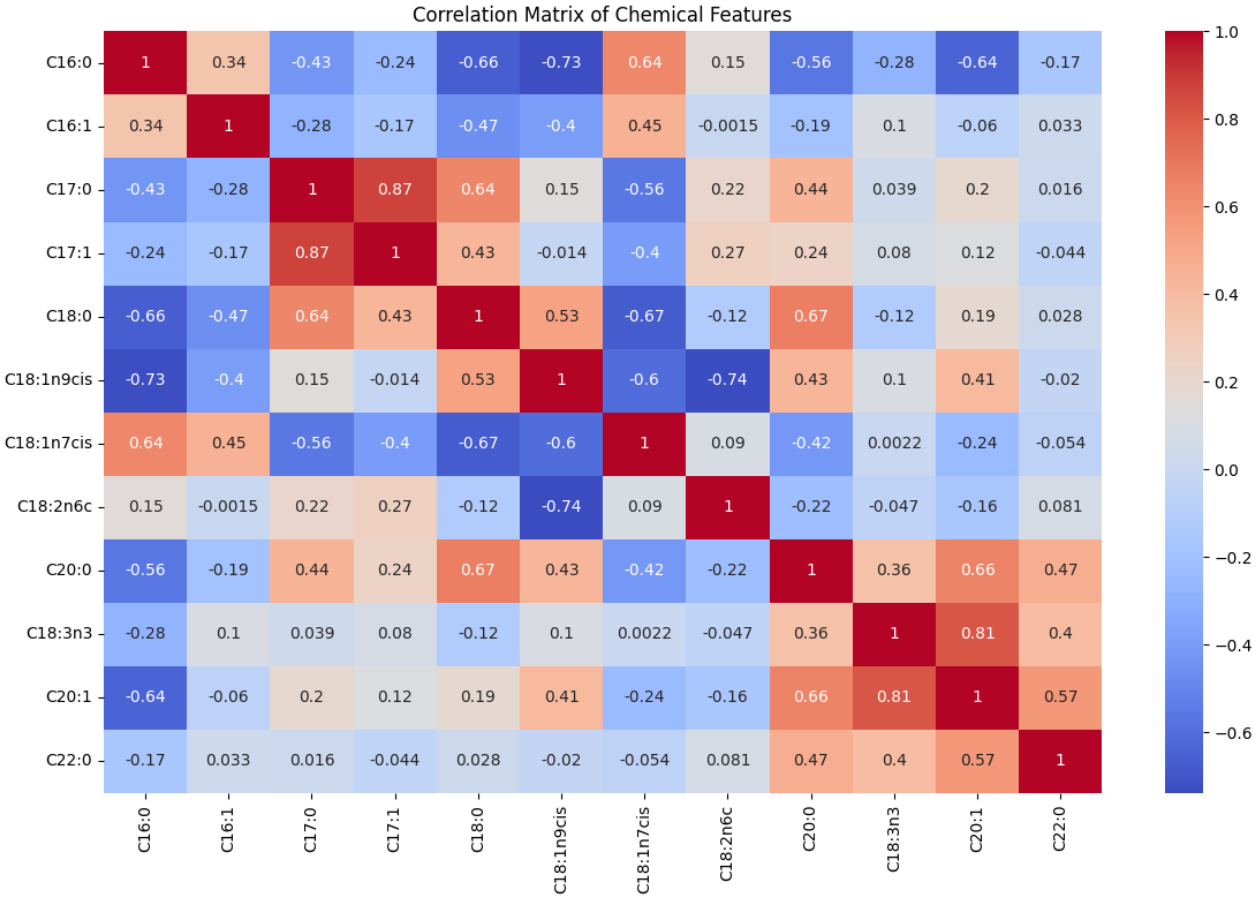

Figure 10.
The chemical features correlation heatmap.

The heatmap (Figure 10) on chemical features reveals a strong negative correlation between C18:1(n-9)-cis and C16:0, suggesting that higher levels of oleic acid are associated with lower levels of palmitic acid. Meanwhile, C18:2(n-6) cis is moderately positively correlated with C18:3(n-3), suggesting that they co-occur more frequently in some cultivars. These relationships help the model differentiate cultivars more effectively.

This heatmap visually displays the pairwise correlations between different chemical features—specifically, fatty acids—in the samples. Each cell represents the correlation coefficient between two compounds, with the color intensity reflecting the strength and direction of the relationship. A positive correlation (typically indicated by warmer colors) indicates that the concentrations of the two compounds tend to increase or decrease in tandem. In comparison, a negative correlation (usually depicted with cooler colors) indicates that as one variable increases, the other tends to decrease in value.

The oleic acid (C18:1(n-9) cis) exhibits a strong negative correlation with palmitic acid (C16:0). This suggests that in the analyzed samples, when oleic acid levels are high, palmitic acid levels tend to be low, and vice versa. This inverse relationship may reflect metabolic trade-offs or preferences in biosynthetic pathways within the plant or cultivar. Another example is linoleic acid, C18:2(n-6)cis, which has a moderate positive correlation with alpha-linolenic acid (C18:3(n-3)). This suggests that these two polyunsaturated fatty acids often co-occur because they share similar biosynthetic pathways

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 93-110, 2025
DOI: 10.55214/25768484.v9i7.8539

or are influenced by similar environmental or genetic factors. Their co-occurrence in certain cultivars can point to specific metabolic profiles associated with those varieties.

Understanding these interrelationships between chemical components is essential for the model. By capturing how these compounds vary together, the model gains a better understanding of the underlying chemistry that distinguishes one cultivar from another. This enables more accurate classification or prediction tasks in areas such as plant breeding, nutritional analysis, or quality control. The Random Forest model performs perfectly on this dataset (based on this matrix), suggesting distinct chemical signatures for each cultivar.

4. Conclusions

Identifying olive oil (OO) by cultivar and origin enhances the success of products both locally and globally. Kalinjot olive oil parameters can be adjusted based on geographic origin to verify quality and prevent adulteration. The Random Forest model achieves approximately 93% accuracy in classifying origin, with strong confusion matrix results and good generalization, requiring minimal tuning. SVM and KNN perform adequately but are less suitable due to lower accuracy and sensitivity to parameters. The research shows that Logistic Regression, KNN, and SVM effectively identify Kalinjot VOO sources, improving quality control and traceability of Albanian mono-cultivar olive oils. This helps companies verify product authenticity. Machine learning algorithms can extract complex data, aid decision-making, and boost quality and competitiveness. As the dataset grows, prediction accuracy improves, with the model achieving 100% accuracy in the current dataset. The Random Forest model distinguishes the chemical signatures of three olive varietals effectively, due to their unique profiles. Results for Mixan and Ulli i Bardhë cultivars may be affected by the limited number of samples, through overfitting and untested generalization. Overall, Random Forest is the most effective model, correctly classifying all three types according to the confusion matrix.

Transparency:

The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.

Copyright:

References

[1] D. Boskou, *Olive and olive oil bioactive constituents*. Amsterdam, Netherlands: Elsevier, 2015, pp. 1-30.

[2] D. Topi, F. Thomaj, and E. Halimi, "Virgin olive oil production from the major olive varieties in Albania," Poljoprivreda i Sumarstvo, vol. 58, no. 2, p. 87, 2012.

[3] D. Topi, A. Amanpour, H. Kelebek, and S. Selli, "Screening of aroma profiles in Albanian cvs. Kalinjot and Bardhi Tirana olive oils using purge and trap extraction technique," Rivista Italiana delle Sostanze Grasse, vol. 96, no. 2, 2019.

[4] D. Topi, G. Gamze, K. Hasim, and S. and Selli, "Comparative elucidation of phenolic compounds in Albanian olive oils using LC-DAD-ESI-MS/MS," Journal of Liquid Chromatography & Related Technologies, vol. 43, no. 5-6, pp. 203-212, 2020/04/02 2020. https://doi.org/10.1080/10826076.2019.1711117

[5] R. Malheiro, N. Rodrigues, and J. A. Pereira, "Olive oil phenolic composition as affected by geographic origin, olive cultivar, and cultivation systems," in Olive and Olive Oil Bioactive Constituents: Elsevier, 2015, pp. 93-121.

[6] M.-I. Covas, M. Fitó, and R. de la Torre, "Minor bioactive olive oil components and health: Key data for their role in providing health benefits in humans," Olive and Olive Oil Bioactive Constituents, pp. 31-52, 2015.

[7] J. Llupa, U. Gašić, I. Brčeski, P. Demertzis, V. Tešević, and D. Topi, "LC-MS/MScharacterization of phenolic compounds in the quince (cydonia oblonga mill.) and sweet cherry (prunus avium l.) fruit juices," Agriculture and Forestry, vol. 68, no. 2, pp. 193-205, 2022.

[8] O. Abbas and V. Baeten, *Advances in identifying adulterated vegetable oils. advances in food authenticity testing, woodhead publishing series in food science, technology, and nutrition, edited by gerard downey*. Amsterdam, Netherlands: Elsevier, 2016, pp. 519-542.

[9] E. Nanou, N. Pliatsika, and S. Couris, "Rapid authentication and detection of olive oil adulteration using laser-induced breakdown spectroscopy," Molecules, vol. 28, no. 24, p. 7960, 2023. https://doi.org/10.3390/molecules28247960

[10] I. C. Martakos, M. G. Kostakis, M. E. Dasenaki, and N. S. Thomaidis, "Authenticity and chemometrics of olive oil," in Chemometrics and Authenticity of Foods of Plant Origin: CRC Press, 2022, pp. 36-55.

[11] C. Sanmartin and F. Mencarelli, "Prediction of fruity aroma intensity and defect presence in virgin olive oil using an electronic nose," Sensors, vol. 2, no. 7, p. 2298, 2021.

[12] H. Topi *et al.*, " IS 2010: Curriculum guidelines for undergraduate degree programs in information systems," ACM & AIS, 2012.

[13] S. Velo and D. Topi, "Study of Kalinjoti extra virgin olive oils, fatty acids profiles and trans-isomers," Journal of Hygienic Engineering and Design, vol. 12, pp. 129-133, 2015.

[14] S. Velo and D. Topi, "Characterization of Kalinjoti and Nisioti monocultivar Virgin olive oils produced in Albania," Asian Journal of Chemistry, vol. 29, no. 6, pp. 1347-1350, 04/09 2017. https://doi.org/10.14233/ajchem.2017.20498

[15] INSTAT, "Olives, area and production," Ministry of Agriculture and Rural Development; processing: INSTAT, 2025.

[16] V. Sheth, U. Tripathi, and A. Sharma, "A comparative analysis of machine learning algorithms for classification purpose," Procedia Computer Science, vol. 215, pp. 422-431, 2022/01/01/ 2022. https://doi.org/10.1016/j.procs.2022.12.044

[17] F. Lonetti, F. Martelli, and G. Resta, "Artificial neural networks applied to olive oil production and characterization: A systematic review," Intelligent Systems with Applications, vol. 26, p. 200525, 2025/06/01/ 2025. https://doi.org/10.1016/j.iswa.2025.200525

[18] G. Aiello and D. Tosi, "An artificial intelligence-based tool to predict "unhealthy" wine and olive oil," Journal of Agriculture and Food Research, vol. 16, p. 101179, 2024/06/01/ 2024. https://doi.org/10.1016/j.jafr.2024.101179

[19] B. Vega-Márquez, I. Nepomuceno-Chamorro, N. Jurado-Campos, and C. Rubio-Escudero, "Deep learning techniques to improve the performance of olive oil classification," Frontiers in chemistry, vol. 7, p. 929, 2020.

[20] Y. Yakar and K. Karadağ, "Identifying olive oil fraud and adulteration using machine learning algorithms," Química Nova, vol. 45, no. 10, pp. 1245-1250, 2022.

[21] J. Niyogisubizo, J. de Dieu Ninteretse, E. Nziyumva, M. Nshimiyimana, E. Murwanashyaka, and E. Habiyakare, "Towardspredicting the quality of red wine using novel machine learning methods for classification, data visualization, and analysis," in *Artificial Intelligence and Applications*, 2025, vol. 3, no. 1, pp. 31-42.

[22] V. Skiada, P. Katsaris, M. E. Kambouris, V. Gkisakis, and Y. Manoussopoulos, "Classification of olive cultivars by machine learning based on olive oil chemical composition," Food Chemistry, vol. 429, p. 136793, 2023/12/15/ 2023. https://doi.org/10.1016/j.foodchem.2023.136793

[23] F. Z. Issaad *et al.*, "Classification of algerian olive oils: Physicochemical properties, polyphenols and fatty acid composition combined with machine learning models," Journal of Food Composition and Analysis, vol. 125, p. 105812, 2024/01/01/ 2024. https://doi.org/10.1016/j.jfca.2023.105812

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 93-110, 2025
DOI: 10.55214/25768484.v9i7.8539