Classifying PDO Kalamata Olive Oil from Geographic Origins of the Messenia Region based on Statistical Machine Learning

THEODOROS ANAGNOSTOPOULOS[1], IOAKEIM SPILIOPOULOS[2]
[1]Department of Business Administration,
University of West Attica,
12241 Athens,
GREECE

[2]Department of Food Science and Technology,
University of Peloponnese,
24100 Kalamata,
GREECE

Abstract: - Kalamata is a smart city located in southeastern Greece in the Mediterranean basin and it is the capital of the Messenia regional unit. It is known for the famous Protected Designation of Origin (PDO) Kalamata olive oil produced mainly from the Koroneiki olive variety. The PDO Kalamata olive oil, established by Council regulation (EC) No 510/2006, owes its quality and special characteristics to the geographical environment, olive tree variety, and human factor. The PDO Kalamata olive oil is produced exclusively in the regional unit of Messenia, being the main profit of local farmers. However, soil chemical composition, microclimates, and agronomic factors are changed within the Messenia spatial area leading to differentiation of PDO Kalamata olive oil characteristic. In this paper, we use statistical machine learning algorithms to determine the geographical origin of Kalamata olive oil at PDO level based on synchronous excitation−emission fluorescence spectroscopy of olive oils. Evaluations of the statistical models are promising for differentiating the origin of PDO Kalamata olive oil with high values of prediction accuracy thus enabling companies that process and bottle kalamata olive oil to choose olive oil from a specific region of Messenia that fulfills certain characteristics. Concretely, the current research effort focuses on a specific olive oil variety within a limited geographic region. Intuitively, future research should also focus on validation of the proposed methodology to other olive oil varieties and production areas.

Key-Words: - PDO Kalamata olive oil, synchronous emission-excitation, fluorescence spectroscopy, statistical machine learning, data fusion, data visualization, multiclass classification, model evaluation.

Received: June 9, 2023. Revised: February 24, 2024. Accepted: April 3, 2024. Published: May 13, 2024.

1 Introduction

Smart agriculture is the dimension of the smart city concept aiming to define methods of efficient geographic cultivation in rural areas, [1]. The cropping of plants useful for cities' citizens is the main area of interest for smart farming, [2]. Specifically, in the area of olive oil farmers in the Messenia region of Greece produce the protected designation of origin (PDO) extra virgin olive oil with the name Kalamata olive oil in the rural areas of the smart city Kalamata the Koroneiki olive variety is almost exclusively cultivated which produces the extra virgin olive oil with organoleptic properties, [3], [4]. Such areas provide farmers the capability to gain more income since specific olive oil microclimates affect the quality of the selected variety, [5].

To protect olive oil quality and prevent its adulteration, global governmental agencies like the European Commission, International Olive Council, Codex Alimentarius, etc have developed standards to regulate olive oil by establishing a set of physical, chemical, and organoleptic characteristics, [6]. The traditional chemical methods to ensure olive oil quality are focused on the identification and quantification of pre-defined compounds or classes of compounds of olive oil according to the regulations of the above-mentioned global governmental agencies. These methods are time-consuming and demand expensive apparatus. The same for the detection of olive oil adulteration

although these methods fail to detect the adulteration from certain adulterants.

In recent years the non-targeted analysis has attracted much attention. This approach focuses on screening the olive oil without any prior knowledge of chemical composition. In this approach, we used analytical techniques that produce a signal which is affected by all the compounds (i.e., metabolites) present in olive oil. These methods shorten the analysis process but a vast number of data sources are required to perform data analytics based on statistical machine learning algorithms, [7].

To assess the quality of gathered olive oil there is a need to incorporate specific Internet of Things (IoT) devices, [8]. A device that is commonly used for such a process is fluorescence spectroscopy, which is calibrated accordingly to perform differences of excitation and emission radiation to the olive oil sample, [9]. Concretely, fluorescence spectrometry has been used extensively in the past years due to its efficient precision in recognizing chemical components of olive oil samples thus exploiting its overall quality, [10]. Specifically, adopted technology can access input data from olive oil sample sources to measure optimally the chemical ingredients of a given olive oil sample as well as to be able to discriminate the olive oil quality categories as well as its origin, [11]. Intuitively, fluorescence spectra technology can detect in high effectiveness adulteration of olive oil with other lower-quality oils, such as sunflower oil or soybean oil, [12]. Collecting samples from different geographical origins enables the generation of different data sources, [13]. Exploited data can be visualized and analyzed by statistical machine learning algorithms. Intuitively, the application of statistical classifiers enables the classification of olive oil samples into certain categories able to differentiate the quality of each sample, [14].

In this paper, we input synchronous emission-excitation fluorescence spectra of PDO Kalamata olive oils of different local geographic origins from Messenia to observe the resulting data sources. DPO Kalamata olive oils were from the areas of (1) Aris, (2) Thouria, (3) Verga, (4) Arfara, and (5) Meligalas. Subsequently, we input such data sets to certain statistical machine learning algorithms to assess which of them has optimal results to recognize the different local cultivations. Adopted statistical learning algorithms are evaluated with certain evaluation methods and metrics to observe an optimal classification of input olive oil samples. The outcome of the research effort is to be able to characterize the specific origin of each PDO Kalamata olive oil (within the Messenia region) thus companies that process and bottlebottle Kalamata olive oil can choose olive oil from specific regions of Messenia that fulfils superior characteristics.

The rest of the paper is organized as follows. In Section 2 it is presented the prior work in the research effort area. Section 3 defines the adopted data model. In Section 4 evaluation parameters are defined. In Section 5 experiments are performed and results are observed. Section 6 discusses the strengths and the weaknesses of the proposed research effort, while Section 7 concludes the paper and proposes future work.

2 Prior Work

Extra virgin and virgin olive oil have recently attracted consumer interest because of their quality, and its potential health benefits derived from their consumption. The high price of extra virgin olive oil and its reputation makes olive oil a target for fraudsters. Significant research has been performed in the literature in the area of olive oils' analysis, classification, authentication, origin, and adulteration. Spectroscopic techniques such as ultraviolet-visible (i.e., UV–Vis) absorption [15], [16], fluorescence spectroscopy [17], Raman spectroscopy [18], mass spectrometry [19], nuclear magnetic resonance [20] and FT-NIR [21] have been proposed to classify and detect adulteration and origin of olive oil. Classification based on statistical machine learning is used to compare virgin olive oil quality in [22]. Fluorescence spectroscopy is used along with principal component technology and factorial discriminant analysis for monitoring and classifying certain virgin olive oil varieties. Raman spectroscopy is incorporated in [23], to identify olive oil quality using classification techniques. Intuitively, the adopted method used a one-dimensional convolutional deep-learning neural network to observe optimal classification results. Portable Raman spectroscopy is used in [24], to provide quality assessment and control of several olive oil varieties. Subsequently, the proposed method adequately covers the cases of adulterated compound low-quality oils within the virgin olive oil.

Classification and authentication techniques are incorporated in [25], to distinguish the origins of virgin olive oil. Specifically, it is proposed an authentication process is proposed to analyze volatile olive oil compounds and chemometrics to assess the quality of certain olive oil varieties within a local geographic area. Statistical machine learning algorithms are incorporated in [26], to classify

specific olive oil varieties. Concretely, the adopted method uses discrimination techniques to input machine learning algorithms with spectroscopic data thus achieving effective prediction accuracy of olive oil behavior by exploiting fusion emission and absorption. Fluorescence spectroscopy is incorporated in, [27], to classify the high quality of olive oil. Intuitively, the proposed method assesses a certain thermal oxidation technique, which exploits the potentiality of an Ultra Violet (UV) fluorescence spectroscopy system to perform specific imaging classification of extra virgin olive oil varieties.

A time series classification algorithm is incorporated in [28], which can distinguish several virgin olive oil varieties. Subsequently, a statistical transformation of the generated input data sources is performed on each virgin olive oil variety to assess the ensemble classification schema thus observing optimal values of the prediction accuracy evaluation metric. A multivariate classification analysis is incorporated in [29], which can distinguish extra virgin olive oils. Concretely, the adopted method is based on Fourier Transform Infrared Spectroscopy (FTIR) along with multivariate analysis to classify virgin olive oils' geographic origins, which come from several producing countries. Adulterated olive oil, in [30], can be discriminated with the incorporation of Attenuated Total Reflection (ATR) and FTIR spectroscopy technologies. Intuitively, the proposed methods are capable of distinguishing pure samples of virgin olive oil from different oil blends by exploiting the potentiality of partial least squares discriminant analysis (PLS-DA) applied to given olive oil compounds.

Methods and applications for distinguishing several extra virgin olive oils' local geographic origins are proposed in the literature, [31]. Specifically, the classification of olive oil geographic origins is based on certain chemometric data sources. Such chemometric data are generated from several olive oils compounds, which input the fluorescence spectroscopy decision-making models to achieve optimal prediction accuracy. Synchronous scanning of chemometric data sources produced by significantly detailed fluorescence spectroscopy measurements is also supported in certain research efforts, [32]. Such knowledge is then exploited by specific statistical classification learning models, which can distinguish several varieties of edible extra virgin olive oils. Edible olive oils' premium quality is assessed in the literature, [33]. Concretely, such ability is achieved by the incorporation of synchronous fluorescence spectroscopy, which can differentiate the quantification of tocopherols from the input olive oil compounds.

Geographic origins of olive oil varieties, [34], are feasible due to the incorporation of chemometric analysis. Specifically, such advanced analytical methodology, which is applied to data sources can predict olive oil's registered designation with optimal precision taking into consideration synchronous excitation and emission of fluorescence spectra values. Rapid spectroscopic methods (Vis–NIR and FT-MIR) along with PLS analysis were applied to study thermal stress of virgin olive oils, [35]. Concretely, due to the manipulation of generated data sources to certain statistical learning models, which can evaluate optimally spectroscopic and chemometric technologies. Pattern recognition is also incorporated in extra virgin olive oil varieties classification, [36]. Intuitively, near-infrared spectrometry provides the technical methodology to assess the strengths of screening methods, which are then used to authenticate extra virgin olive oils from near local geographic origins. Shelf-Life olive oil varieties are monitored and then classified to certain geographic origins, [37]. Subsequently, IoT sensors and actuators technology is exploited to enhance fluorescence spectroscopy characteristics thus being able to correctly assess the multiclass classification process, which is based on certain statistical learning models.

There are many research approaches that deal with the origins of olive oil based on statistical machine learning. Promising efforts incorporate generated data from several chemometric technologies. However, data manipulation requires improvement to distinguish data interconnections, which can provide efficient results. In this research effort, fluorescence spectroscopy is exploited by applying enhanced data preprocessing. Such optimized data sources are then used by a statistical machine learning algorithm to perform multiclass classification to distinguish between certain local cultivations' origins of the Koroneiki olive oil variety in the smart city of Kalamata, which is located in the Messenia region, Greece.

3 Data Model

Data provided to perform analytics are synchronous emission-excitation fluorescence spectra. These spectra were recorded on a Perkin Elmer LS55 spectrofluorometer using solution 1% w/v olive oil in n-hexane, where $\Delta\lambda$ (i.e., the difference between excitation and emission wavelength) was adjusted to 30 nm, [38]. The excitation and emission slit were

tuned to 4 nm. The scan rate was 50nm/min. Each olive oil sample was measured triplicate using the new freshly prepared solution. Each measurement of an olive oil sample was statistically handled as a different sample of the same origin.

Such data have a certain structure. Specifically, observed data sources are collected from PDO Kalamata olive oil produced in a variety of local areas in the rural areas of the smart city of Kalamata in the Messenia region, Greece. Concretely, there are collected data from 29 olive oil samples from the local cultivation areas of (1) Aris, (2) Thouria, (3) Verga, (4) Arfara, and (5) Meligalas. Intuitively, local cultivation areas are classified into the following 5 classes, namely: (1) Aris: Class 1, (2) Thouria; Class 2, (3) Verga: Class 3, (4) Arfara: Class 4, and (5) Meligalas: Class 5. Subsequently, the distribution of collected data samples per class are as follows: (1) 2 samples from Aris, (2) 2 samples from Thouria, (3) 7 samples from Verga, (4) 15 samples from Arfara, and (5) 3 samples from Meligalas.

3.1 Data Structure

Synchronous emission-excitation fluorescence spectra are composed of two-dimensional coordinates, i.e., (x, y_i), where $i \in [1,5]$ is the identifier of each olive oil class, where: $i = 1$ refers to Aris, $i = 2$ refers to Thouria , $i = 3$ refers to Verga, $i = 4$ refers to Arfara, and $i = 5$ refers to Meligalas local geographic origins. Concretely, x dimension depicts the emission wavelength measured in nanometers (i.e., nm) for all sample classes while y_i dimension depicts photoluminescence intensity, which is an arbitrary net number based on internal calibration of the spectrofluorometer device for each of the i data sample classes.

3.1.1 Visualizing Initial Data Samples

Intuitively, according to the initial data sample measurements, assigned specific values for x dimension in the initial interval such as: $x \in [250, 700]$ for all data sample classes (i.e., the 5 classes of local cultivation origins). Subsequently, initial data samples are averaged according to $y_i, \forall i \in [1,5]$ values based on each olive oil class. Such average is performed to be able to provide a simple and easily understandable visualization based on initial data samples for each of the 5 classes. Concretely, averaged values in y_i dimension is varying according to the examined olive oil data sample classes, as follows: (1) in case of Aris y_1 has initial values in the interval, $y_1 \in [0.24, 752.28]$, (2) in case of Thouria y_2 has initial values in the interval, $y_2 \in [0.08, 964.05]$, (3) in case of Verga y_3 has initial values in the interval, $y_3 \in [0.06, 554.82]$, (4) in case of Arfara y_4 has initial values in the interval, $y_4 \in [0.54, 440.79]$, and (5) in case of Meligalas y_5 has initial values in the interval, $y_5 \in [0.07, 154.92]$. Figure 1, visualizes the initial data samples (i.e., synchronous photoluminescence spectra) per certain class of olive oil geographic origin. It can be observed that classes are not easily distinguished from each other based on initial data measurements. This should be treated accordingly with the data fusion process to have a clearer view of how classes could be more easily distinguished.

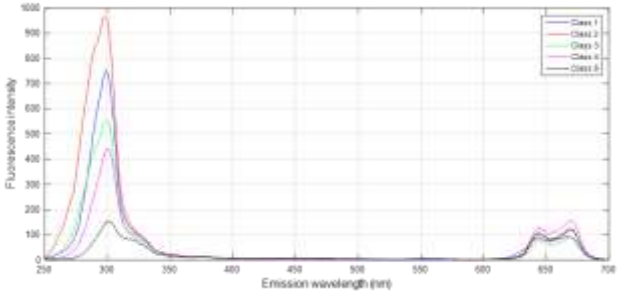

Fig. 1: Initial data (i.e., synchronous emission-excitation fluorescence spectra) are assigned to a specific class of certain local origin of PDO Kalamata olive oil. Observed spectra were recorded at $\Delta\lambda = 30$

3.2 Data Fusion

Data fusion is a widely adopted method used in machine learning literature in case there is a need to understand in depth the inherent complexity of initial data sources. Concretely, the data fusion process is applied to experimental input data sources to visualize the statistical qualitative trends of the data provided, thus being able to incorporate efficient machine learning algorithms to observe optimal results. Intuitively, initial data samples are transformed according to a specific data fusion process to remove outliers and missing values that occurred during the initial measurement process performed by the fluorescence spectroscopy device. Specifically, such a data fusion process can provide easily distinguished classes between each other in contrast to the initial data due to the adopted transformation. Intuitively, x dimension values of each data sample are transformed into s_k interval values. Such intervals form the predictive attributes, which will input the statistical learning classifier to predict the correct class of olive oil origin. Concretely, $k \in [1,5]$ is the assigned identifier of each transformed wavelength interval (i.e.,

predictive attribute) of the olive oil components. Subsequently, for $s_k, \forall k \in [1,5]$ it holds that in case of: $k = 1$ refers to $s_1 \in [250, 350]$ that is assigned to the predictive attribute of 'tocopherols', $k = 2$ refers to $s_2 \in [351, 425]$ that is assigned to the predictive attribute of 'phenolic compounds', $k = 3$ refers to $s_3 \in [426, 525]$ that is assigned to the predictive attribute of 'oxidation products of triglycerides', $k = 4$ refers to $s_4 \in [526, 600]$ that is assigned to the predictive attribute of 'oxidation products of tocopherols', and $k = 5$ refers to $s_5 \in [601, 700]$ that is assigned to the defined predictive attribute of 'chlorophylls', components.

Subsequently, measured values in y_i dimension have an arbitrary initial distribution according to the examined olive oil data sample class as produced by the synchronous photoluminescence spectra. Such values are transformed into t_k^i aggregated values for each olive oil class origin, i, and each assigned identifier, k, to each transformed wavelength interval, s_k, (i.e., a certain predictive attribute) according to specific olive oil measured compounds of local geographic origin. Concretely, it holds that t_k^i is a transformed average value that is assigned to each fused predictive attribute (i.e., s_k) of a certain data sample class. There are specific t_k^i fused values given certain data instances of the initial data observed by the 29 olive oil samples (i.e., class values) from different local reas for each of the s_k interval values, (i.e., predictive attributes).

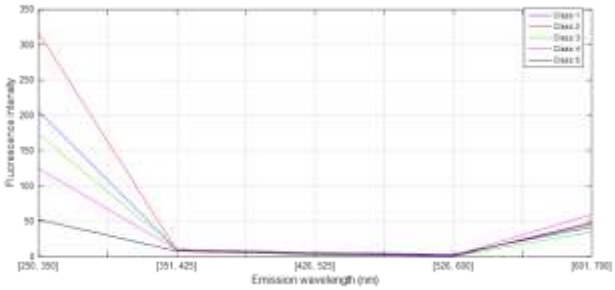

Fig. 2: Fused data (i.e., based on initial synchronous emission-excitation fluorescence spectra) are assigned to a specific class of certain local origin of PDO Kalamata olive oil. Observed spectra were recorded at $\Delta\lambda = 30$

3.2.1 Visualizing Fused Data Samples

Intuitively, according to the fused data sample measurements, for certain $\Delta\lambda = 30$, is assigned specific values for the s_k interval values (i.e., predictive attributes) according to $k \in [1,5]$ for the fused data sample classes (i.e., the 5 classes of origins). Subsequently, fused data samples are averaged according to $t_k^i, \forall i, k \in [1,5]$ values for each olive oil class. Such average is performed to be able to provide a simple and easily understandable visualization based on fused data samples for each of the 5 classes. Concretely, averaged fused values in t_k^i (i.e., predictive attributes' value range) is varying according to the examined olive oil data sample classes, as follows: (1) in case of Aris $t_k^i, i = 1, \forall k \in [1,5]$ has observed fused data values of: $t_k^1 = [206.15, 10.25, 5.61, 3.23, 41.99]$, (2) in case of Thouria, $t_k^i, i = 2, \forall k \in [1,5]$ has certain fused values, $t_k^2 = [317.04, 9.19, 4.25, 1.12, 45.77]$, (3) in case of Verga $t_k^i, i = 3, \forall k \in [1,5]$ has fused values, $t_k^3 = [173.89, 10.31, 3.78, 0.68, 34.88]$, (4) in case of Arfara $t_k^i, i = 4, \forall k \in [1,5]$ has fused values, $t_k^4 = [125.07, 9.78, 5.17, 1.86, 59.63]$, and (5) in case of Meligalas $t_k^i, i = 5, \forall k \in [1,5]$ has fused values, $t_k^5 = [52.41, 7.93, 3.14, 0.79, 48.06]$, Figure 2, visualizes the fused data samples per certain class of olive oil origin. It can be observed that classes are now more easily distinguished from each other based on fused data measurements. This is the reason for treating initial data samples with a data fusion process to have a clearer view of how classes could be more easily distinguished.

4 Evaluation Parameters

Assessing the performance of the adopted statistical machine learning algorithm, certain valuation methods and evaluation metrics should be incorporated to perform specific experiments and observe derived results.

4.1 Evaluation Method

To evaluate a statistical machine learning algorithm there are used certain evaluation methods. Authors adopt one of the widely used evaluation methods, due to its simplicity and optimum results, which is 10-fold cross-validation, [39]. Specifically, such an evaluation method divides the input dataset into 10 equal sized parts and then in a certain loop incorporates the first 9 parts to train the statistical learning classification algorithm and the remaining 1 to test the classifier. This process is repeated until all the parts are used for training and testing. The proposed evaluation method is adopted in the machine learning methodology since it provides effective results based on certain input data able to explain the observed data source's predictive analytics behavior.

Table 1. Confusion matrix

Class 1	Class 2	Class 3	Class 4	Class 5	← Classified as
A	B	C	D	E	Class 1
F	**G**	H	I	J	Class 2
K	L	**M**	N	O	Class 3
P	Q	R	**S**	T	Class 4
U	V	W	X	**Y**	Class 5

4.2 Evaluation Metrics

Given the evaluation method, which is proposed to support the experimental setup there is a need to adopt specific evaluation metrics. Such metrics are: (1) prediction accuracy, (2) correctly classified instances, and (3) confusion matrix that can assess the efficiency of a statistical classification algorithm.

4.2.1 Prediction Accuracy

The effectiveness of the adopted statistical learning algorithm is assessed by incorporating prediction accuracy evaluation metric, $a \in [0,1]$, which is defined in the following mathematical equation, (1):

$$a = \frac{tr_{pos}+tr_{neg}}{tr_{pos}+fl_{pos}+tr_{neg}+fl_{neg}} \quad (1)$$

Where, tr_{pos}, are the instances, which are classified correct as positives, and, tr_{neg}, are the instances, which are classified correct as negatives. In addition, fl_{pos}, are the instances, which are classified false are positives, and, fl_{neg}, are the instances, that are classified false as negatives. A low value of a means a weak classifier while a high value of a indicates an efficient statistical learning classifier. Concretely, experimental assessment based on the defined statistical quantities of: (1) tr_{pos}, (2) tr_{neg}, (3) fl_{pos}, and (4) fl_{neg}, which compose the prediction accuracy evaluation metric's experimental value, achieve to express the data sources' dynamics and explain the observed optimal results.

4.2.2 Correctly Classified Instances

In statistical machine learning, it is common to express prediction accuracy as a percentage thus observed results being more easily interpreted and presented. Concretely, it is used the term correctly classified instances, $c \in [0\%, 100\%]$, which is defined according to the following mathematical equation, (2):

$$c = \alpha\% \quad (2)$$

Where, a value close to 0% means that the classification algorithm is not efficient, while a value close to 100% indicates that the statistical algorithm is able to classify instances optimally.

4.2.3 Confusion Matrix

We also evaluated the adopted statistical classification algorithm with the confusion matrix evaluation metric. Confusion matrix is a special form of matrix, which in the case of a multiclass classification of 5 classes, (i.e., Class 1: Aris, Class 2: Thouria, Class 3: Verga, Class 4: Arfara, and Class 5: Meligalas) has the following encoded form, as described in Table 1.

Where, "A" quantity depicts the number of Class 1 instances, which are classified correctly as instances of Class 1. "B" quantity depicts the number of Class 1 instances, which are falsely classified as instances of Class 2. "C" quantity depicts the number of Class 1 instances, which are falsely classified as instances of Class 3. "D" quantity depicts the number of Class 1 instances, which are falsely classified as instances of Class 4. "E" quantity depicts the number of Class 1 instances, which are falsely classified as instances of Class 5. The same holds for the rest elements of the confusion matrix. A given classification model is considered efficient if it maximizes the elements of the main diagonal of the confusion matrix (i.e., "A", "G", "M", "S", and "Y") and minimizes the other elements. A confusion matrix is incorporated in machine learning evaluation methodology to support efficiently and explain in deep detail the statistical nature of output experimental results observed by the prediction accuracy evaluation metric.

5 Experiments and Results

The data model, which is based on fused data values is used to perform certain experiments and observe derived results. An experimental setup is necessary to formulate the experimental phase with certain evaluation methods and metrics and observe the results of the current research effort.

5.1 Experimental Setup

Specific parameters are incorporated to set up the experimental process. Concretely, it is defined as the number of classes, which is assigned to each data sample instance. Intuitively, predictive attributes used to describe a certain class are defined accordingly. Subsequently, a certain statistical machine learning algorithm should be adopted to perform the experiments and observe the results.

5.1.1 Multiclass Classification

Since the number of classes is 5 this classification process is characterized as a multiclass classification

problem. Specifically, 5 classes are defined as follows: (1) Class 1: Aris, (2) Class 2: Thouria, (3) Class 3: Verga, (4) Class 4: Arfara, and (5) Class 5: Meligalas. Concretely the number of predictive attributes is also 5, which are characterized as follows: (1) Predictive Attribute 1: 'tocopherols', (2) Predictive Attribute 2: 'phenolic compounds', (3) Predictive Attribute 3: 'oxidation products of triglycerides', (4) Predictive Attribute 4: 'oxidation products of tocopherols', and (5) Predictive Attribute 5: 'chlorophylls'. The number of data sample instances is 29, which have the following distribution per class: (1) 2 samples from Class 1 i.e., Aris, (2) 2 samples from Class 2, i.e., Thouria, (3) 7 samples from Class 3, i.e., Verga, (4) 15 samples from Class 4, i.e., Arfara, and (5) 3 samples from Class 5, i.e., Meligalas.

5.1.2 Logistic Statistical Learning Algorithm

To select the optimum statistical learning algorithm that is effective in this multiclass classification problem we experimented with several statistical learning classifiers available in the Weka machine learning software, [40]. Intuitively, the machine learning algorithm, which has optimal predictive behavior emerged to be a Logistic statistical learning algorithm (i.e., the implementation of the Logistic Regression algorithm in Weka machine learning software) thus it is adopted for further experimentation to observe the derived results of the current research study.

5.2 Derived Results

To evaluate the experimental phase there is a need to define a specific evaluation method (i.e., 10-fold cross-validation) and metrics used to assess the efficiency of the adopted statistical learning algorithm, which in this case is the logistic statistical machine learning algorithm. Concretely, based on certain evaluation parameters specific derived results are observed, which define the effectiveness of the incorporated experimental setup adopted in the current research effort. Intuitively, to understand the observed results and be able to explain the research effort's findings it is significant to use the incorporated evaluation method and evaluation metrics. Such knowledge would reveal the inherent complexity that exists in the provided data sources aiming to observe optimal results for the adopted machine learning algorithm.

5.2.1 Observed Prediction Accuracy

The evaluation method incorporated to evaluate the adopted machine learning multiclass classification algorithm is 10-fold cross-validation. According to this evaluation method observed prediction accuracy is: $a = 0.9655$, which is a high value for prediction accuracy thus proving that the adopted statistical learning algorithm is suitable for the examined multiclass classification problem. Concretely, the high value observed for the prediction accuracy enables the adopted machine learning algorithm to be incorporated for similar use in new unseen olive oil instances in a further future research that might extend the potentiality of the current research effort to the geographical region of interest.

5.2.2 Observed Correctly Classified Instances

According to the evaluation method of 10-fold cross-validation correctly classified instances it occurred to be: $c = 96.55\%$, which indicated that the selected statistical machine learning algorithm is an optimal choice for the examined classification problem.

5.2.3 Observed Confusion Matrix

Confusion matrix results as derived based on a 10-fold validation evaluation method for the examined multiclass classification problem. Derived results are presented in Table 2.

Table 2. Confusion matrix observed results

Class 1	Class 2	Class 3	Class 4	Class 5	← Classified as
2	0	0	0	0	Class 1
0	**2**	0	0	0	Class 2
1	0	**6**	0	0	Class 3
0	0	0	**15**	0	Class 4
0	0	0	0	**3**	Class 5

It can be observed that most of the classified instances are located in the main diagonal of Table 2. Specifically, the quantity of elements in the main diagonal depicts the significant number of certain instances, which are correctly classified. Concretely, such an optimal prediction behavior indicates a robust classification algorithm for the examined multiclass classification problem. Such a detailed confusion matrix enables the observation of experimental results in deep detail thus being able to assess the efficiency of the adopted machine learning algorithm for predicting PDO Kalamata olive oil in other provided experimental instances.

6 Discussion

Problem definition indicates a multiclass classification problem of 5 discrete classes, with 5 separate predictive attributes and a total of 29 sample instances based on the local geographic origins of the Koroneiki olive oil variety, which is

cultivated in the smart city of Kalamata in the Messenia region, Greece. Subsequently, the current research effort has achieved high values of the observed results based on certain evaluation metrics, which indicate the robustness of the examined evaluation parameters. Intuitively, the current research study has significant strengths as well as certain weaknesses, which should be presented with regard to a complete methodological research frame.

6.1 Weaknesses of the Study

Initial data as measured by the synchronous photoluminescence spectra IoT device are characterized as primitive raw data values, which should be further processed to enter a statistical machine learning algorithm to be evaluated properly. Concretely, visualizing initial data sources results in a complex plot, where there is vagueness in distinguishing the adopted 5 classes of the initial data source. Intuitively, such inefficiency results in limited evaluation capability based on the available initial data. Subsequently, classes get tangled up with each other thus making an inference assumption difficult to be applied. Exploitation of visualized initial data is not suitable for further experimentation in the current form. A fusion process is required in the initial data to remove outliers and missing values before further processing.

6.2 Strengths of the Study

The data fusion process adopted in the current research study eliminates the vagueness of the adopted 5 data classes. Concretely, fused data enabled the emergence of 5 discrete predictive attributes, which aim to face the vagueness of the initial data. Specifically, by visualizing fused data it is proved that the classes and the predictive attributes are distinguished easily, thus being able to proceed with further experimentation. Intuitively, the adopted evaluation method and metrics have proved to be effective in defining optimal derived results. Subsequently, the selection of a Logistic statistical machine learning classifier emerged to be an efficient solution to the multiclass classification problem. Concretely, the adopted classification algorithm was able to predict different classes based on the fused data sources. Such effectiveness enabled the capability of distinguishing the origin of PDO Kalamata olive oil produced in specific local areas in the rural areas of the Kalamata smart city.

7 Conclusions and Future Work

PDO Kalamata olive oil is an extra virgin olive oil produced in the province of Messenia in southeastern Greece (the name stands for capital city Kalamata). Because of different soil composition, microclimates, and agronomic factors olive oil from different areas of Messenia has diverse characteristics, although within the limits described by council regulation (EC) No 510/2006. Adopted synchronous photoluminescence spectra of olive oils IoT device can specify the different origins of PDO Kalamata olive oil. In this research effort, we use statistical machine learning algorithms to classify several geographic origins. Evaluation of the statistical models are based on certain methods and metrics, which have proved to be promising for differentiating origins thus enabling olive oil companies to choose PDO Kalamata olive oil from a specific area of Messenia with superior characteristics.

According to our research outcomes, future work should mainly focus on the incorporation of more detailed input measurement data sources based on improvements in synchronous photoluminescence spectra IoT-enabled technology, thus providing a more robust input to the selected statistical classification algorithm. Concretely, data fusion techniques should be reapplied on the more detailed initial data sources to input several statistical learning algorithms, which might result in more effective results. Intuitively, current research could be further used in more detail to verify authentication and to detect adulteration of olive oils with protected designation of origin (PDO) thus facing the fraud problem occurring in the olive oil trade. Intuitively, the current research effort focuses on a specific olive oil variety within a limited geographic region, while future research should also focus on the validation of the proposed methodology to other olive oil varieties and production areas within the Messenia region and/or in other geographic regions of Greece that are popular for the quality of their olive oil production.

References:

[1] S. Paiho, P. Tuominen, J. Rokman, M. Ylikerala, J. Pajula, and H. Siikavirta, “Opportunities of collected city data for smart cities”, *IET Smart Cities*, Volume 4, Issue 4, 2022, pages 275 – 291.

[2] J. L. D. Boer, and B. Erickson, “Setting the Record Straight on Precision, Agriculture Adoption”, *Agronomy Journal*, Volume 111, Issue 4, 2019, pages 1535 – 2139.

[3] X. Miao, J. Ma, X. Miu, H. Zhang, Y. Geng, W. Hu, Y. Deng, and N. Li, "Integrated transcriptome and proteome analysis the molecular mechanisms of nutritional quality in 'Chenggu-32' and 'Koroneiki' olives fruits (Olea europaea L.)", *Journal of Plant Physiology*, Volume 288, Issue 154072, 2023, pages 1 – 12.

[4] L. Trabelsi, B. Ncube, A. B. Hassena, M. Zouairi, F. B. Amar, and K. Gargouri, "Comparative study of productive performance of two olive oil cultivars Chemlali *Sfax and Koroneiki under arid conditions", South African Journal of Botany*, Volume 154, Issue 1, 2023, pages 356 – 364.

[5] A. Issa, M. E. Riachy, C. B. Mitri, J. Doumit, W. Skaff, and L. Karam, "Influence of geographical origin, harvesting time and processing system on the characteristics of olive-mill wastewater: A step toward reducing the environmental impact of the olive oil sector", *Environmental Technology & Innovation*, Volume 32, Issue 103365, 2023, pages 1 – 12.

[6] R. Aparicio, M. T. Morales, R. A. Ruiz, N. Tena, and D. L. G. González, "Authenticity of olive oil: Mapping and comparing official methods and promising alternatives", *Food Research International*, Volume 54, Issue 2, 2013, pages 2025 – 2038.

[7] D. I. Ellis, H. Muhamadali, S. A. Haughey, C. T. Elliott, and R. Goodacre, "Point–and–shoot: Rapid quantitative detection methods for on–site food fraud analysis–moving out of the laboratory and into the food supply chain", *Analytical Methods*, Volume 7, Issue 22, 2015, pages 9375 – 9716.

[8] P. Rajak, A. Ganguly, S, Adhikary, and S. Bhattacharya, "Internet of Things and smart sensors in agriculture: Scopes and challenges", *Journal of Agriculture and Food Research*, Volume 14, Issue 100776, 2023, pages 1 – 13.

[9] J. Krause, H. Gruger, L. Gebauer, X. Zheng, J. Knobbe, T. Pgner, A. Kicherer, R. Gruna, T. Langle, and J. Beyerer, "Smart Spectrometer–Embedded Optical Spectroscopy for Applications in Agriculture and Industry", *Sensors*, Volume 21, Issue 13, 2021, pages 1 – 18.

[10] R. Karoui, and C. Blecker, "Fluorescence Spectroscopy Measurement for Quality Assessment of Food Systems – a Review", *Food Bioprocess Technology*, Volume 4, Issue 1, 2011, pages 364 – 386.

[11] S. Khani, J. B. Ghasemi, and Z. P. Vanak, "Development of computer vision system for classification of olive oil samples with different harvesting years and estimation of chlorophyll and carotenoid contents: A comparison of the proposed method's efficiency with UV-Vis spectroscopy", *Journal of Food Composition and Analysis*, Volume 129, Issue 106078, 2024, pages 1 – 42.

[12] S. K. Drakopoulou, A. S. Kritikou, C. Baessmann, and N. Thomaidis, "Untargeted 4D-metabolomics using Trapped Ion Mobility combined with LC-HRMS in extra virgin olive oil adulteration study with lower-quality olive oils", *Food Chemistry*, Volume 434, Issue 137410, 2024, pages 1 – 9.

[13] M. E. Schiano, F. Sodano, C. Cassiano, E. Magli, S. Seccia, M. G, Rimoli, and S. Albrizio, "Monitoring of seven pesticide residues by LC-MS/MS in extra virgin olive oil samples and risk assessment for consumers", *Food Chemistry*, Volume 442, Issue 138498, 2024, pages 1 – 8.

[14] R. Reda, T. Saffaj, I. Bouzida, O. Saidi, M. Belgrir, B. Lakssir, and E. M. E. Hadrami, "Optimized variable selection and machine learning models for olive oil quality assessment using portable near infrared spectroscopy", *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, Volume 303, Issue 123213, 2023, pages. 1 – 11.

[15] K. D. T. M. Milanez, T. C.A. Nóbrega, D. S. Nascimento, M. Insausti, B. S. F. Band, and M. J. C. Pontes, "Multivariate modeling for detecting adulteration of extra virgin olive oil with soybean oil using fluorescence and UV–Vis spectroscopies: A preliminary approach", *LWT – Food Science and Technology*, Volume 85, Issue 1, 2017, pages 9 – 15.

[16] R. A. Santos, J. C. Cancilla, A. P. Pérez, A. Moral, and J. S. Torrecilla, "Quantifying binary and ternary mixtures of monovarietal extra virgin olive oils with UV–vis absorption and chemometrics", *Sensors and Actuators B: Chemical*, Volume 234, Issue 1, 2016, pages 115 – 121.

[17] I. D. Merás, J. D. Manzano, D. A. Rodríguez, and A. M. Peña, "Detection and quantification of extra virgin olive oil adulteration by means of autofluorescence excitation–emission profiles combined with multi–way classification", *Talanta*, Volume 178, Issue 1, 2018, pages 751 – 762.

[18] Y. Li, T. Fang, S. Zhu, F. Huang, Z. Chen, and Y. Wang, "Detection of olive oil adulteration with waste cooking oil via Raman spectroscopy combined with iPLS and SiPLS", *Spectrochimica Acta Part A: Molecular Biomolecular Spectroscopy*, Volume 189, Issue 1, 2018, pages 37 – 43.

[19] F. D. Girolamo, A. Masotti, I. Lante, M. Scapaticci, C. D. Calvano, C. Zambonin, M. Muraca, and L. A. Putignani, "Simple and effective mass spectrometric approach to identify the adulteration of the mediterranean diet component extra–virgin olive oil with corn oil", *International Journal of Molecular Sciences*, Volume 16, Issue 9, 2015, pages 20896 – 20912.

[20] A. Rotondo, L. Mannina, and A. Salvo, "Multiple Assignment Recovered Analysis (MARA) NMR for a Direct Food Labeling: The Case Study of Olive Oils", *Food Analytical Methods*, Volume 12, Issue 1, 2019, pages 1238 – 1245, DOI: 10.1007/s12161-019-01460-4.

[21] M. M. Mossoba, H. Azizian, A. R. F. Kia, S. R. Karunathilaka, and J. K. G. Kramer, "First Application of Newly Developed FT–NIR Spectroscopic Methodology to Predict Authenticity of Extra Virgin Olive Oil Retail Products in the USA", *Lipids*, Volume 52, Issue 5, 2017, pages 443 – 455, DOI: 10.1007/s11745-017-4250-5

[22] H. Zaroual, C. Chene, E. M. E. Hadrami, and R. Karoui, "Comparison of four classification statistical methods for characterizing virgin olive oil quality storage up to 18 months", *Food Chemistry*, Volume 370, Issue 131009, 2022, pages 1 – 16.

[23] X. Wu, S. Gao, Y. Niu, Z. Zhao, B. Xu, R. Ma, H. Liu, and Y. Zhang, "Identification of olive oil in vegetable blend oil by one-dimensional convolutional neural network combined with Raman spectroscopy", *Journal of Food Composition and Analysis*, Volume 108, Issue 104396, 2022, pages 1 – 7.

[24] I. H. A. S. Barros, L. S. Paixao, M. H. C. Nascimento, V. J. Lacerda, P. R. Figueiras, and W. Romao, "Use of portable Raman spectroscopy in the quality control of exrtra virgin olive oil and adulterated compound oils", *Vibrational Spectroscopy*, Volume 116, Issue 103299, 2021, pages 1 – 10.

[25] L.Cecchi, M. Migliorini, E. Giambanelli, A. Rosseti, A, Cane, N. Mulinacci, and F. Melani, "Authentication of the geographical origin of virgin olive oils from the main worldwide producing countries: A new combination of HS-SPME-GC-MS analysis of volatile compounds and chemometrics applied to 1217 samples", *Food Control*, Volume 112, Issue 107156, 2020, pages 1 – 10.

[26] D. Stefas, N. Gyftokostas, P. Kourelias, E. Nanou, V. Kokkinos, C. Bouras and S. Couris, "Discrimination of olive oils based on the olive cultivar origin by machine learning employing the fusion of emission and absorption spectroscopic data", *Food Control*, Volume 130, Issue 108318, 2021, pages 1 – 8.

[27] V. Rotich, D. F. A. Riza, F. Giametta, T. Suzuki, Y. Ogawa, and N. Kondo, "Thermal oxidation assessment of Italian extra virgin olive oil using an UltraViolet (UV) induced fluorescence imaging system", *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, Volume 237, Issue 118373, 2020, pages 1 – 8.

[28] A. Bagnall, L. Davis, J. Hills, and J. Lines, "Transformation Based Ensembles for Time Series Classification", *Proceedings of the 2012 SIAM International Conference on Data Mining (SDM)*, Anaheim, California, USA, April 26 – 28, 2012, pages 307 – 318.

[29] H. S. Tapp, M. Defernez, and E. K. Kemsley, "FTIR Spectroscopy and Multivariate Analysis Can Distinguish the Geographic Origins of Extra Virgin Olive Oils", *Journal of Agricultural and Food Chemistry*, Volume 51, Issue 21, 2003, pages 6110 – 6115, DOI: 10.1021/jf030232s.

[30] P. D. L. Mata, A. D. Vidal, J. M. B. Sendra, A. R. Medina, L. C. Rodriguez, and M. J. A. Canada, "Olive oil assessment in edible oil blends by means of ATR-FTIR and chemometrics", *Food Control*, Volume 23, Issue 2, 2012, pages 449 – 455.

[31] E. Sikorska, I. Khmelinskii, and M. Sikorski, "Analysis of Olive Oils by Fluorescence Spectroscopy: Methods and Applications", *InTech*, Volume 1, Issue 1, 2012, pages 63 – 88, DOI: 10.5772/30676.

[32] E. Sikorska, T. Gorecki, I. V. Khmelinskii, M. Sikorski, and J. Koziol, "Classification of edible oils using synchronous scanning fluorescence spectroscopy", *Food Chemistry*, Volume 89, Issue 2, 2005, pages 217 – 225.

[33] E. Sikorska, A. G. Swiglo, I. Khmelinskii, and M. Sirorski, "Synchronous Fluorescence of Edible Vegetable Oils. Quantification of Tocopherols", *Journal of Agriculture and Food Chemistry*, Volume 53, issue 18, 2005, pages 6988 – 6994, DOI: 10.1021/jf0507285.

[34] N. Dupuy, Y. L. Dreau, D. Ollivier, J. Artaud, C. Pinatel, and J. Kister, "Origin of French Virgin Olive Oil Registered Designation of Origins Predicted by Chemometric Analysis of Synchronous Excitation-Emission Fluorescence Spectra", *Journal of Agricultural and Food Chemistry*, Volume 53, Issue 24, 2005, pages 9361 – 9368.

[35] R. M. Maggio, E. Valli, A. Bendini, A. M. G. Caravaca, T. G. Toschi, and L. Cerretani, "A spectroscopic and chemometric study of virgin olive oils subjected to thermal stress", *Food Chemistry*, Volume 127, Issue 1, 2011, pages 216 – 221.

[36] E. Bertran, M. Blance, J. Coello, H. Iturriaga, S. Maspoch, and I. Montoliu, "Near infrared spectrometry and pattern recognition as screening methods for the authentication of virgin olive oils of very close geographical origins", *Journal of Near Infrared Spectroscopy*, Volume 8, Issue 1, 2000, pages 45 – 52.

[37] A. L. Prieto, N. Tena, R. A. Ruiz, D. L. G. Gonzalez, and E. Sikorska, "Monitoring Virgin Olive Oil Shelf-Life by Fluorescence Spectroscopy and Sensory Characteristics: A Multidimensional Study Carried Out under Simulated Market Conditions", *Foods*, Volume 9, Issue 12, 2020, pages 1 – 20, DOI: 10.3390/foods9121846

[38] The LS-55 and LS-45 Fluorescence Spectrofluorometers, Perkin Elmer, [Online]. https://resources.perkinelmer.com/lab-solutions/resources/docs/BRO_LS-55andLS-45FluorescenceSpectrophotometer.pdf (Accessed Date: February 26, 2024).

[39] E. Frank, M. A. Hall, and I. H. Witten, *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*, Morgan Kaufmann, Fourth Edition, 2016.

[40] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update", *SIGKDD Explorations*, Volume 11, Issue 1, 2009, pages 10 – 18.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed to the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare.