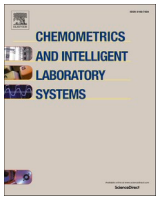
Chemometrics and Intelligent Laboratory Systems 250 (2024) 105136

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics



# A new sub-class linear discriminant for miniature spectrometer based food analysis

Omar Nibouche [a,*], Fayas Asharindavida [a], Hui Wang [b], Jordan Vincent [a], Jun Liu [a], Saskia van Ruth [c,d], Paul Maguire [a], Enayet Rahman [a,e]

[a] Faculty of Computing, Engineering and the Built Environment at Ulster University, Belfast, UK
[b] School of Electronics, Electrical Engineering and Computer Science at Queen's University Belfast, UK
[c] School of Agriculture and Food Science, University College Dublin, Dublin, Ireland
[d] Department of Agrotechnology and Food Sciences, Wageningen University & Research, Wageningen, the Netherlands
[e] Research Centre for Biomedical Engineering, City, University of London, London, UK

ARTICLE INFO

*Index Terms:*
Dimensionality reduction
Feature extraction
Linear discriminant analysis
Subclass discriminant analysis
Classification
Food analysis
Miniature spectrometers

ABSTRACT

The well-known and extensively studied Linear Discriminant Analysis (LDA) can have its performance lowered in scenarios where data is not homoscedastic or not Gaussian. That is, the classical assumptions when LDA models are built are not applicable, and consequently LDA projections would not be able to extract the needed features to explain the intrinsic structure of data and for classes to be separated. As with many real word data sets, data obtained using miniature spectrometers can suffer from such drawbacks which would limit the deployment of such technology needed for food analysis. The solution presented in the paper is to divide classes into subclasses and to use means of sub classes, classes, and data in the suggested between classes scatter metric. Further, samples belonging to the same subclass are used to build a measure of within subclass scatterness. Such a solution solves the shortcoming of the classical LDA. The obtained results when using the proposed solution on food data and on general machine learning datasets show that the work in this paper compares well to and is very competitive with similar sub-class LDA algorithms in the literature. An extension to a Hilbert space is also presented; and the kernel version of the presented solution can be fused with its linear counter parts to yield improved classification rates.

## 1. Introduction

Chemometrics which is the chemical discipline that uses Machine Learning (ML) underpinned by mathematical and statistical methods can benefit from the well-known techniques of feature detection, classification including classification via Discriminating Analysis (DA), regression including regression via partial least square (PLSR) and its discriminant analysis variant, namely, PLS-Discriminant Analysis (PLS-DA), and clustering. Such techniques and algorithms can be used to improve the performance of the analysis of chemical data [1], and they can be regarded as a set of statistical based estimation processes which aim to establish a relationship between one or more dependent response variables and one or more independent variables [2–4]. Few of such ML techniques are based on feature extraction which underpins a generic class of methods for reducing the size of information required to describe a relatively large data set whilst at the same time retaining the part required to describe such data and classify it with an appropriate level of accuracy [2–4]. Feature extraction via DA is a well-researched, studied and deployed approach in many applications of ML [3,5–8]. Algorithms belonging to such a class exhibit ease of implementation, automatic extraction of features and low dimensionality of the processed data, good separation and representation of classes [6,9–15]. From their classical deployment in face recognition, such algorithms were used with success in computer vision, biometrics, spectral data analysis [24–27], agriculture and crop analysis [28], fault diagnosis [29], the medical field [30] and food authentication [24].

Further, their ease of implementation can be seen as a consequence of their assumed linearity; although their extension to non-linear algorithms has been well discussed in the open literature [16,17]. They are also holistic in their intrinsic nature as all data is considered when models are designed and implemented. An underpinning assumption is that the data in every class can be joined in one cluster. Data can be

* Corresponding author.
*E-mail address:* o.nibouche@ulster.ac.uk (O. Nibouche).

https://doi.org/10.1016/j.chemolab.2024.105136
Received 22 June 2023; Received in revised form 22 April 2024; Accepted 3 May 2024
Available online 4 May 2024

divided into classes or groups where a datum shares more similarities with points in its neighbourhood than to other data in other groups and classes. In other words, points in a particular class can be segregated from points in other classes based on latent distributions and as such, data of the same class are likely to share the same distribution [14,18–21]. And this is a fundamental point in many ML techniques including clustering and LDA which turns out also to be their drawback as class distributions may not always be homoscedastic in practice [18,20,22]. When such an issue can be tackled in clustering using mixture models as in Gaussian Mixture Models, the well-known Expectation-Maximization algorithm, and distribution based clustering algorithms [7,12,17,18,23], LDA can benefit from relaxing such constraints by using an approach in which a class can be divided into further clusters termed subclasses and then redefine an LDA like criterion that incorporate the newly obtained information. This may be regarded as an attempt to solve problems of nonlinearity and resolve the unimodality and homoscedasticity assumptions causing the limitation of LDA [17,18,20].

### 1.1. Food quality and analysis

Food quality establishment and authentication can be based on spectral data acquisition and analysis [1,24–26,31]. Such an application can be well suited for a linear classifier deployment scenario in which classes can be further divided into subclasses. For instance, one may investigate a food item spectrum to establish its authenticity; however, further divisions can be based on season of production, geographical location, (sub-) species, method of production and processing technologies which is a clear case where subclasses rather than classes would come into play. All such analysis can be very useful in the context of the detection of and fight against food fraud. While there is a variety of different technologies and methods useful for such an analysis including spectroscopy (vibrational and fluorescence), genomics & proteomics, liquid/gas chromatography & mass spectrometry, isotopic ratios and bioinformatics, all and each of such methods has its advantages and disadvantages and is capable of making different determinations with differing accuracies. Nevertheless, portability, cost, processing time, classification rate and size are key requirements for the system to be adopted by consumers. Under such circumstances, portable and miniaturised spectroscopy is an appropriate technology as it is portable, affordable, non-destructive and exhibits short processing times. The shortening of the processing time is the obvious consequence of bringing the device to the food sample rather than dispatching the sample to a lab for analysis [25]; a relative wide *deployability* can be expected due to the availability of miniature spectrometers within the price range that render them at the reach of normal citizens [26,32]. Obviously, the last two points can signal a shift in the food fraud fight realm where the matter can swing from being tackled by professionals in labs to consumers at the forefront in the fight against food fraud and exhibiting cost-effective hand-held portable or miniature devices. Such technology can identify the unique “fingerprint” in agri-food products and one can generate spectral data of a food sample at a quick rate beyond what can be managed by labs employing bulky spectrometers, and not needing a lengthy training or education. All of that would equate to a major step forward in which food quality analysis is on citizens doorstep [25]. The technology comes with its own challenges though. One must deal with relatively large and easy to acquire amounts of data, the devices reduced immunity to ambient and background noise, their relative low accuracy rates, and limited wavelength range; all of which would call for augmenting miniature spectrometer with machine learning components.

### 1.2. Contribution of the paper

One matter which is not always covered in food analysis papers in the field of chemometrics is that data would usually be collected in one session; and is then split to training and testing subsets. In this manuscript, there is an emphasis on the separation of training data from testing data which are acquired during different sessions. That is, a point which may be observed in spectral data collected using miniature devices is that such spectrometers are not immune to distortions. Hence, it is important to take that matter into consideration; and as such, the concept of sub-class is replacing that of class in the classical LDA classifier. Further, in the ensuing development, the terms ‘session data’ can simply be replaced by ‘cluster’ which is commonly used in ML.

In this paper, a new sub-class based LDA solution is presented. Once the pre-clustering is applied, the new approach combines both the distances of sub-class means to the class means and the class means to the data mean at the same time to build a between scatter metric; and attempts to reduce sub-class scatterness by reducing the distance between the sub-classes’ samples and the means of their respective sub-classes in the LDA feature space. The method is termed Combined Sub-Class LDA (CSDA); and it can be further generalised when one builds a scatterness metric from the classical LDA between class scatter and the between sub-class scatter matrices. Such an extension can be useful and shows the flexibility and capability to generalise a sub-class LDA. Such a method is termed Balanced Sub-Class LDA (BSDA). A further extension of the proposed CSDA LDA to a Hilbert space and the use of the kernel version of the presented solution is also shown. Such a kernel-based extension is termed KSDA. And as both the linear sub-class and non-linear kernel-based version can be appreciated as attempting to solve the somewhat almost similar issues of homoscedasticity and non-linearity of data, and where the extension to a Hilbert space may be competitive, an ensemble of subclass learners can be merged using $l_0$, $l_1$ and $l_2$ regression-based representation algorithms to improve the classification rates. The remainder of this paper is organised as follows: in section 2, the LDA algorithm is revisited, and background and previously published work in the field of sub-class LDA is covered. The new solutions are shown in section 3. This includes the new linear sub-class and non-linear kernel based sub-class LDA algorithms. The new algorithms are tested on food and generic ML data in section 4. For testing and analysing the performance of the presented algorithms, food data is used for its suitability to test the underpinning idea behind the work. As the presented solution can find further applications in the generic field of ML, the extensive testing includes comparison of performance obtained using generic ML datasets of similar sub-class LDA variants available in the research literature.

## 2. LDA revisited and previous work

LDA is a well-researched and documented algorithm; mainly due its perceived easy implementation and well understood assumptions and modelling. This is a matter that stems from the algorithm’s used statistics for modelling purposes. LDA is underpinned by low order statistics encompassing mean and (co-) variance; hence avoiding stability issues that regression may encounter when using a small number of well separated training samples for estimation [2,3,14,16,17,32]. LDA statistical quantities are estimated from the data used to build the model; and the algorithm’s assumptions are straightforward; data adhere to and is of a Gaussian distribution and samples within the same class would adhere to the same and homoscedastic distribution, and as such are generated from multivariate Gaussian distributions of a common covariance matrix although exhibiting different means. LDA is a supervised learning technique which yields a projection into a lower dimensionality space than the original space in which data is represented. That is, as with the Principal Component Analysis (PCA), dimensionality reduction plays a crucial role in feature selection. This can be achieved via the use of a projection matrix $\widetilde{\Psi} \in \mathbb{R}^{F\times D}$ where $D \ll F$ and the $D$ vectors $[\widetilde{\boldsymbol{\psi}}_1 \ldots \widetilde{\boldsymbol{\psi}}_D]$ in $\widetilde{\Psi}$ are all in $\mathbb{R}^F$ and are used for mapping data from the original feature space $\mathbb{R}^F$ to a space of a lower dimensionality $\mathbb{R}^D$. The projection of data using $\widetilde{\Psi}$ can lead to relatively high classification rates. Discriminant analysis can be based on the

Fisher-Rao's criterion [3,5,21,22]which is the maximixation of $\frac{|\boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v}|}{|\boldsymbol{v}^T\boldsymbol{B}\boldsymbol{v}|}$; the matricess $\boldsymbol{A}$ and $\boldsymbol{B}$ are positive semi definite. In the remainder of the paper, subscripts $A$ and $B$ would be used to refer to matrices used in the numerator and denominator parts of $\frac{|\boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v}|}{|\boldsymbol{v}^T\boldsymbol{B}\boldsymbol{v}|}$, respectively. A linear combination of the generalised eigenvectors $\boldsymbol{v}$ in $\boldsymbol{B}\boldsymbol{v} = \lambda\boldsymbol{A}\boldsymbol{v}$ would best linearly classify the data; where $\lambda$ is the generalised eigenvalue associated with. If matrix $\boldsymbol{B}$ is invertible, the matter turns then to be a classical eigen value decomposition problem [3,5,21,22]. It is well known that to compute the projection matrix $\widetilde{\Psi}$, LDA uses a between-class scatter matrix termed $\boldsymbol{S}_b$ as a measure of class separability and uses a within-class scatter matrix termed $\boldsymbol{S}_w$ as a measure of class compactness. The goal of LDA is to find a matrix $\widetilde{\Psi}$ such that the measure of the between class scatter matrix $\boldsymbol{S}_b$ in the new space is maximised and simultaneously, the measure of the within-class scatter matrix $\boldsymbol{S}_w$ in the new space is minimised:

$$\boldsymbol{S}_w = \boldsymbol{S}_B^{LDA} = \frac{1}{N}\sum_{i=1}^{C}\sum_{j=1}^{N_i}\left(\boldsymbol{x}_{ij}-\boldsymbol{m}_i\right)\left(\boldsymbol{x}_{ij}-\boldsymbol{m}_i\right)^T \tag{1}$$

$$\boldsymbol{S}_b = \boldsymbol{S}_A^{LDA} = \frac{1}{N}\sum_{i=1}^{C}N_i(\boldsymbol{m}_i-\boldsymbol{m})(\boldsymbol{m}_i-\boldsymbol{m})^T = \sum_{i=1}^{C}P_i(\boldsymbol{m}_i-\boldsymbol{m})(\boldsymbol{m}_i-\boldsymbol{m})^T \tag{2}$$

Where the total variance is $\boldsymbol{S}_t = \boldsymbol{S}_b + \boldsymbol{S}_w = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i-\boldsymbol{m})(\boldsymbol{x}_i-\boldsymbol{m})^T$ which is the well-known covariance matrix used in PCA for reducing the dimensionality of data. Parameter $N$ is the total number of samples available for training in the set $\boldsymbol{X} = \{\boldsymbol{x}_1,..\boldsymbol{x}_i..,\boldsymbol{x}_N\}$ with class membership labels $\{L_1..,L_i..,L_N\}$. To indicate class membership of a sample $\boldsymbol{x}_i$, any of the labels $L_i$ can take one value between 1 and $C$, and where $C$ is the number of classes. In equations (1) and (2), $\boldsymbol{x}_{ij}$ is the $j$th sample of the $i$th class, $N_i$ is the number of samples in class $i$, $P_i = N_i/N$ is the $i$th class prior, $\boldsymbol{m}$ is the sample-average computed over the $N$ samples available in training, and $\boldsymbol{m}_i$ is the mean of the samples in class $i$. The original space in which samples $\boldsymbol{x}_i$ are represented is $\mathbb{R}^F$ and $\in \mathbb{R}^{F\times N}$.

For maximising a between cluster measure and minimising a within cluster measure, LDA can be expressed as an optimisation process as follows:

$$\underset{\Psi}{argmax}J^{LDA}(\Psi) = \underset{\Psi}{argmax}\frac{Trace\left(\Psi^T\boldsymbol{S}_A^{LDA}\Psi\right)}{Trace\left(\Psi^T\boldsymbol{S}_B^{LDA}\Psi\right)} \tag{3}$$

The solution $\widetilde{\Psi}$ $[\widetilde{\boldsymbol{\psi}}_1\ldots\widetilde{\boldsymbol{\psi}}_D]$ of (3) is the result of the maximization of the objective function which may be subject to further constraints imposed on the properties of $[\widetilde{\boldsymbol{\psi}}_1\ldots\widetilde{\boldsymbol{\psi}}_D]$; and t he parameter $D$ is the result of the optimisation operation. Hence, once data is projected into the new space spanned by the column vectors of $\widetilde{\Psi}$, the dimension of any vector $\widetilde{\boldsymbol{x}}_i = \widetilde{\Psi}^T\boldsymbol{x}_i \in R^D$ is reduced from $F$ as in the original space of data (as $\boldsymbol{x}_i \in R^F$) to $D$ in the new transform space. As such, LDA can be regarded as a mapping from the $F$-dimensional space in which data is represented to a $D$-dimensional space. The upper threshold of $D$ is $C$-1 which is the highest possible rank of $\boldsymbol{S}_b$. Although, such an upper limit can make computation efficient, it may lack classification power as reported in Refs. [16–18,21,22,33].

### 2.1. Subclass LDA algorithms

It is really the classical assumption that LDA can be derived from Gaussian distributions of classes of a common covariance matrix and with different means which has been blamed in research for the method's shortcomings and sub-optimal projections. Therefore, it can be assumed that the sought LDA projections cannot attain a high between class variance combined at the same time with a low within class variance as the algorithm aims for. Further to the non-*Gaussianity*, LDA may not be able to tackle and successfully preserve the data underlying complex structure if the assumption of homoscedasticity does not hold. That is, data non-homoscedasticity is a practical case encountered in many applications which can turn out to be very restrictive for LDA deployment [7,9,14,16–18,34]. A strategy for tackling such restrictiveness is to resort for probabilistic modelling when representing Gaussian distributed subpopulations within a larger population using Gaussian mixture models (GMM); and to embed such information which may be termed 'multimodality' into the LDA optimisation problem. That is, GMM is used in an unsupervised learning paradigm in which subpopulations, sub-classes (or clusters within classes) and modalities are learnt, and hence, classes can be modelled as mixtures of Gaussian subclasses [7,17,23]. In another approach, one can tackle multimodality as a nonlinearity problem which can be solved using kernel-based non-linear LDA as in the K-LDA algorithms. In such algorithms [16,17,35], a kernel is used to transform data into a Hilbert space where data can be linear and class-separable. LDA can then be employed in the new Hilbert space for classification, and the algorithm is implemented using the so-called kernel trick. The matter can be further improved when coupled with clustering which is yet another unsupervised learning approach.

For going beyond the traditional restrictions of LDA, augmenting it with a clustering step is yet another appealing extension. With a proper clustering step, data can be represented according to its inherent multiple subclasses structure. The approach departs from the assumption that at class level, classes would contain the discriminative information needed for classification purposes. Instead, information at sub-class level would be used where sub-class homoscedasticity can be expected or even attained [10,14,17,18,20–22,34,36–38]. Subclass Discriminant Analysis has been proposed as an improvement that would make LDA more suitable for real-world data where the unimodality assumption does not hold. It is also an approach which can increase the dimension of the projected data and as such, may help to solve the small sample size problem. Compared with classical LDA, the dimensionality is increased by a factor equal to the number of clusters per class. It is also an approach which comes with the advantages of relatively short computation times and classification accuracy as it mainly involves matrices manipulation.

By pre-clustering, one introduces on equations (1) and (2) the parameters $S_i$ and $N_{ij}$ as the number of subclasses in the $i$th class and number of samples in the $j$th subclass of the $i$th class, respectively, to yield:

$$\boldsymbol{S}_{ws} = \frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{S_i}\sum_{k=1}^{N_{ij}}\left(\boldsymbol{x}_{ijk}-\boldsymbol{m}_{ij}\right)\left(\boldsymbol{x}_{ijk}-\boldsymbol{m}_{ij}\right)^T \tag{4}$$

$$\boldsymbol{S}_{bs} = \frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{S_i}N_{ij}\left(\boldsymbol{m}_{ij}-\boldsymbol{m}\right)\left(\boldsymbol{m}_{ij}-\boldsymbol{m}\right)^T = \sum_{i=1}^{c}\sum_{j=1}^{S_i}P_{ij}\left(\boldsymbol{m}_{ij}-\boldsymbol{m}\right)\left(\boldsymbol{m}_{ij}-\boldsymbol{m}\right)^T \tag{5}$$

Parameter $k$ in sample $\boldsymbol{x}_{ijk}$ is used to refer to the $k$th sample in the $j$th subclass of the $i$th class and $\boldsymbol{m}_{ij}$ refers to the mean of the $j$th subclass of the $i$th class. One can show that the total variance used in PCA is $\boldsymbol{S}_t = \boldsymbol{S}_{ws} + \boldsymbol{S}_{bs}$, with $\boldsymbol{S}_{bs} = \boldsymbol{S}_b + \sum_{i=1}^{C}\sum_{j=1}^{S_i}P_{ij}\left(\boldsymbol{m}_i-\boldsymbol{m}_{ij}\right)\left(\boldsymbol{m}_i-\boldsymbol{m}_{ij}\right)^T$ and $\boldsymbol{S}_w = \boldsymbol{S}_{ws} + \sum_{i=1}^{C}\sum_{j=1}^{S_i}P_{ij}\left(\boldsymbol{m}_{ij}-\boldsymbol{m}_i\right)\left(\boldsymbol{m}_{ij}-\boldsymbol{m}_i\right)^T$. By considering sub-classes and data means, the optimisation operation in (3) can be revised to be:

$$arg\ \underset{\Psi}{max}J^{MDA}(\Psi) = \underset{\Psi}{argmax}\frac{Trace\left(\Psi^T\boldsymbol{S}_A^{MDA}\Psi\right)}{Trace\left(\Psi^T\boldsymbol{S}_B^{MDA}\Psi\right)} = \underset{\Psi}{argmax}\frac{Trace(\Psi^T\boldsymbol{S}_{bs}\Psi)}{Trace(\Psi^T\boldsymbol{S}_{ws}\Psi)} \tag{6}$$

The Mixture Discriminant Analysis (MDA) in (6) models classes as mixtures as of subclasses and its metrics take account of $\boldsymbol{S}_B^{MDA} = \boldsymbol{S}_{ws}$ and $\boldsymbol{S}_A^{MDA} = \boldsymbol{S}_{bs}$ to build measures of within-subclass scaterness and within subclass scaterness, respectively [7,18]. That is, MDA solely employs sub-classes. The authors in Refs. [21,22] identified the reason for which a generalised eigendecomposition criterion like the Fisher Rao criterion

which simultaneously maximises $\|\boldsymbol{\Psi}^T\boldsymbol{A}\boldsymbol{\Psi}\|$ and minimises $\|\boldsymbol{\Psi}^T\boldsymbol{B}\boldsymbol{\Psi}\|$ may yield a below par projection. A suggested solution was to pre-cluster the data according to the stability criterion defined in Ref. [21] as:

$$\Theta = \frac{1}{r}\sum_{i=1}^{r}\sum_{j=1}^{i}\cos\theta_{ij}^{\,2} = \frac{1}{r}\sum_{i=1}^{r}\sum_{j=1}^{i}\left(\boldsymbol{\psi}_{A_i}^T\boldsymbol{\psi}_{B_j}\right)^2 \tag{7}$$

Where $r \leq p$, $\theta_{ij}$ is the angle between eigenvectors $\boldsymbol{\psi}_{A_i}$ and $\boldsymbol{\psi}_{B_j}$ of matrices $\boldsymbol{A}$ of rank $p$ and $\boldsymbol{B}$ assumed to be of rank $q$, respectively, and with $q \geq p$. The eigenvectors in matrices $\Psi_A = \left[\boldsymbol{\psi}_{A_1}, \ldots, \boldsymbol{\psi}_{A_p}\right]$ and $\Psi_B = \left[\boldsymbol{\psi}_{B_1}, \ldots, \boldsymbol{\psi}_{B_q}\right]$ are ordered by descending order of their corresponding eigenvalues. The stability criterion can dictate the number of clusters needed for the Fisher Rao criterion to work; and if $\Theta \neq 0$, the vectors resulting from maximising $\frac{|\boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v}|}{|\boldsymbol{v}^T\boldsymbol{B}\boldsymbol{v}|}$ do not guarantee the minimisation of the Bayes error and can indicate a 'conflict' between maximising $\left\|\boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v}\right\|$ and minimising $\left\|\boldsymbol{v}^T\boldsymbol{B}\boldsymbol{v}\right\|$ at the same time. That is, the basis vectors obtained by the generalised eigenvalue decomposition process are not guaranteed to be correct when the smallest angle between the $i$th eigenvector given by the metric to be maximised and the first $i$ eigenvectors given by the metric to be minimised is close to zero. To mitigate such a conflict, Subclass Discriminant Analysis (SDA) starts by selecting the optimal number of subclasses either using the leave-one-out-test (LOOT) criterion or iteratively, using a normalised quantity of $\Theta$ shown in (7) [21]. Then SDA redefines the between scatter metric as depending on matrix:

$$\boldsymbol{S}_A^{SDA} = \boldsymbol{S}_{bsb} = \sum_{i=1}^{c-1}\sum_{j=1}^{S_i}\sum_{k=i+1}^{c}\sum_{l=1}^{S_k} P_{ij}P_{kl}\left(\boldsymbol{m}_{ij}-\boldsymbol{m}_{kl}\right)\left(\boldsymbol{m}_{ij}-\boldsymbol{m}_{kl}\right)^T \tag{8}$$

The metric based on matrix $\boldsymbol{S}_A^{SDA}$ (termed $\boldsymbol{S}_{bsb}$ in Ref. [18]) attempts to maximise the distance between the means of subclasses of different classes only as in $\frac{1}{2}\sum_{i=1}^{c}\sum_{j=1}^{S_i}\sum_{\substack{k=1\\k\neq i}}^{c}\sum_{l=1}^{S_k} P_{ij}P_{kl}\left(\boldsymbol{m}_{ij}-\boldsymbol{m}_{kl}\right)\left(\boldsymbol{m}_{ij}-\boldsymbol{m}_{kl}\right)^T$.

The within class scatterness metric is based on $\boldsymbol{S}_A^{SDA} = \boldsymbol{S}_t$ to yield SDA's optimisation as follows:

$$\underset{\Psi}{argmax} J^{SDA}(\Psi) = \underset{\Psi}{arg\,max}\frac{Trace\left(\Psi^T\boldsymbol{S}_A^{SDA}\Psi\right)}{Trace\left(\Psi^T\boldsymbol{S}_B^{SDA}\Psi\right)} = \underset{\Psi}{arg\,max}\frac{Trace(\Psi^T\boldsymbol{S}_{bsb}\Psi)}{Trace(\Psi^T\boldsymbol{S}_t\Psi)} \tag{9}$$

The Mixture Subclass Discriminant Analysis (MSDA) suggests that the optimisation should use $\boldsymbol{S}_A^{MSDA} = \boldsymbol{S}_{bsb}$ as in (8) & (9), and $\boldsymbol{S}_B^{MSDA}$ should either be $\boldsymbol{S}_{WS}$ or $\boldsymbol{S}_{WS} + \boldsymbol{S}_{bsb}$. The analysis in Ref. [18] has based its argument on the properties of $\frac{\|\boldsymbol{v}\boldsymbol{A}\boldsymbol{v}\|_2}{\|\boldsymbol{v}^T\boldsymbol{B}\boldsymbol{v}\|_2}$ and $\frac{\left\|\boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v}\right\|_2}{\left\|\boldsymbol{v}^T(\boldsymbol{A}+\boldsymbol{B})\boldsymbol{v}\right\|_2}$ which attain the same maximum point and has taken account of the fact that $\boldsymbol{S}_t = \boldsymbol{S}_{ws} + \boldsymbol{S}_{bsb} + \boldsymbol{S}_{bsw}$ to stipulate that one should aim to optimise:

$$\underset{\Psi}{argmax} J^{MSDA}(\Psi) = \underset{\Psi}{arg\,max}\frac{Trace\left(\Psi^T\boldsymbol{S}_A^{MSDA}\Psi\right)}{Trace\left(\Psi^T\boldsymbol{S}_B^{MSDA}\Psi\right)} = \underset{\Psi}{arg\,max}\frac{Trace(\Psi^T\boldsymbol{S}_{bsb}\Psi)}{Trace(\Psi^T(\boldsymbol{S}_{bsb}+\boldsymbol{S}_{ws})\Psi)} \tag{10}$$

which is equivalent to $\underset{\Psi}{arg\,max}\frac{Trace\left(\Psi^T\boldsymbol{S}_{bsb}\Psi\right)}{Trace\left(\Psi^T\boldsymbol{S}_{ws}\Psi\right)}$ and where $\boldsymbol{S}_{bsw} = \frac{1}{2}\sum_{i=1}^{c}\sum_{j=1}^{S_i}\sum_{k=1}^{S_i} P_{ij}P_{ik}\left(\boldsymbol{m}_{ij}-\boldsymbol{m}_{ik}\right)\left(\boldsymbol{m}_{ij}-\boldsymbol{m}_{ik}\right)^T$ is a matrix built using the distance between the means of the subclasses within the same class. The optimisation in (10) and its equivalent form minimise the Bayes error by treating the Gaussian homoscedastic subclasses as the main classes. MSDA criterion departs from that of SDA as one should omit considering $\boldsymbol{S}_{bsw}$ and therefore the MSDA and SDA algorithms would attain different maxima [18].

### 2.2. Separability-oriented LDA

The matter of relationships between the various definition of metrics can be even further relaxed. In the Separability-oriented Subclass Discriminant Analysis (SSDA) set of algorithms, the within-class scatterness is shown to include two quantities; that is, local scatterness at subclass level termed $\boldsymbol{S}_{sw_1}$ and global scatterness at class level termed $\boldsymbol{S}_{sw_2}$ [20]. The local scatterness measures the degree to which data instances in a subclass of a class are scattered around the subclass mean and is defined as:

$$\boldsymbol{S}_{sw_1} = \sum_{i=1}^{c}\sum_{j=1}^{S_i}\sum_{k=1}^{N_{ij}}\left(\boldsymbol{x}_{ijk}-\boldsymbol{m}_{ij}\right)\left(\boldsymbol{x}_{ijk}-\boldsymbol{m}_{ij}\right)^T \tag{11}$$

The global scatterness in the other hand measures the degree to which subclass means in a class are scattered around the class mean:

$$\boldsymbol{S}_{sw_2} = \sum_{i=1}^{C} P_i\sum_{j=1}^{S_i}\left(\boldsymbol{m}_{ij}-\boldsymbol{m}_i\right)\left(\boldsymbol{m}_{ij}-\boldsymbol{m}_i\right)^T \tag{12}$$

and

$$\boldsymbol{S}_B^{SSDA} = \boldsymbol{S}_{sw} = \boldsymbol{S}_{sw_1} + \boldsymbol{S}_{sw_2} \tag{13}$$

It is worth pointing out that although $\boldsymbol{S}_{sw}$ in (13) would define a measure of scatterness, it is different from $\boldsymbol{S}_{ws}$ in (4). Further, although $\boldsymbol{S}_{sw_2}$ presents some measure of global scatterness, it departs from the term $\sum_{i=1}^{C}\sum_{j=1}^{S_i} P_{ij}\left(\boldsymbol{m}_i-\boldsymbol{m}_{ij}\right)\left(\boldsymbol{m}_i-\boldsymbol{m}_{ij}\right)^T$ which appears in the relationship between $\boldsymbol{S}_{bs}$ and $\boldsymbol{S}_b$. As a matter of fact, it favours clustering and separability regardless of the introduced sub-classes balance and classes subclass membership. This is clear when considering the term $P_i$ instead of $P_{ij}$ in (12). Building on this approach, SSDA introduced a measure of between class scatterness as:

$$\boldsymbol{S}_A^{SSDA} = \boldsymbol{S}_{sb} = \sum_{i=1}^{c} P_i\sum_{j=1}^{S_i}\left(\boldsymbol{m}_{ij}-\boldsymbol{m}\right)\left(\boldsymbol{m}_{ij}-\boldsymbol{m}\right)^T \tag{14}$$

As in (12), the term $P_i$ in (14) favours any clustering strategy rather than balance in the number of subclusters per class. It is again a matter which can be contrasted with (5) where the term $P_{ij}$ is used in the definition of $\boldsymbol{S}_{bs}$. SSDA comes with three algorithms, namely, SSDA-1, SSDA-2 and SSDA-3.

For the SSDA-1 variant, $S_A^{SSDA-1} = S_A^{SSDA}$ and $S_B^{SSDA-1} = S_B^{LDA}$. Hence the optimisation in SSDA-1 is given as:

$$\underset{\Psi}{argmax} J^{SSDA-1}(\Psi) = \underset{\Psi}{arg\,max}\frac{Trace\left(\Psi^T\boldsymbol{S}_A^{SSDA}\Psi\right)}{Trace\left(\Psi^T\boldsymbol{S}_B^{LDA}\Psi\right)} = \underset{\Psi}{arg\,max}\frac{Trace(\Psi^T\boldsymbol{S}_{sb}\Psi)}{Trace(\Psi^T\boldsymbol{S}_w\Psi)} \tag{15}$$

In SSDA-2, $\boldsymbol{S}_B^{SSDA}$ is used as a within-class scatter matrix and $\boldsymbol{S}_A^{LDA}$ as the between-class scatter matrix; the algorithm optimisation is expressed as:

$$\underset{\Psi}{argmax} J^{SSDA-2}(\Psi) = \underset{\Psi}{arg\,max} \frac{Trace\left(\Psi^T S_A^{LDA} \Psi\right)}{Trace\left(\Psi^T S_B^{SSDA} \Psi\right)} = \underset{\Psi}{arg\,max} \frac{Trace(\Psi^T \boldsymbol{S}_b \Psi)}{Trace(\Psi^T \boldsymbol{S}_{sw} \Psi)} \tag{16}$$

And for the third variant of SSDA, the optimisation function is expressed as:

$$\underset{\Psi}{argmax} J^{SSDA-3}(\Psi) = \underset{\Psi}{arg\,max} \frac{Trace\left(\Psi^T S_A^{SSDA} \Psi\right)}{Trace\left(\Psi^T S_B^{LDA} \Psi\right)} = \underset{\Psi}{arg\,max} \frac{Trace(\Psi^T \boldsymbol{S}_{sb} \Psi)}{Trace(\Psi^T \boldsymbol{S}_w \Psi)} \tag{17}$$

For all variants of the SSDA algorithm, a hierarchical agglomerative clustering step is carried out first which is then followed by one of the optimisation operations in either (15), (16) or (17).

## 3. A new sub-class linear data analysis

In all the LDA like algorithms, some compactness measure is used which takes account of individual instances, the mean of the class, and the means of sub-classes. In SSDA, the equality $\boldsymbol{S}_{sw_1} = N\boldsymbol{S}_{ws}$ suggests that the scatterness of samples belonging to sub-classes around sub-class means plays a more important role in the term $\boldsymbol{S}_{sw}$ than the global scatterness of subclass means around the class mean as defined in $\boldsymbol{S}_{sw_2}$. Further, when classes are balanced and priors are constant, the matrix $\boldsymbol{S}_{sw_2}$ is reduced to $\sum_{i=1}^{C}\sum_{j=1}^{S_i} P_{ij}\left(\boldsymbol{m}_{ij} - \boldsymbol{m}_i\right)\left(\boldsymbol{m}_{ij} - \boldsymbol{m}_i\right)^T$ which is the difference term in $\boldsymbol{S}_w - \boldsymbol{S}_{ws}$ and $\boldsymbol{S}_{bs} - \boldsymbol{S}_b$. This shows some involvement of such term in both matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ used in the maximization of $\frac{|\boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v}|}{|\boldsymbol{v}^T\boldsymbol{B}\boldsymbol{v}|}$; hence in between-class and within-class scatterness metrics. For instance, it is part of a term that should be maximised if $\boldsymbol{S}_{bs}$ is adopted as between sub-class scatter matrix and at the same time it should be minimised if $\boldsymbol{S}_w$ is the adopted as measure of compactness.

### 3.1. New combined and Balanced Sub-class LDA algorithms

In the suggested new sub-class LDA termed Combined Sub-class LDA (CSDA), one argues for a regular use of class membership and sub-class membership priors instead of favouring any clustering of classes to sub-classes regardless of clusters membership. This would allow for a logical development with very clear links between all used matrices and the work presented in the literature. CSDA adopts a within-subclass compactness metric to bring samples of subclasses as close as possible to their respective subclass means. Hence, such measure of compactness would only use $\boldsymbol{S}_{ws}$ to minimise scatterness as it was also the choice selected in MSDA to address a relative shortcoming of the SDA algorithm [17]. That is, in MSDA, one omits to involve $\boldsymbol{S}_{bsw}$ termed intra-subclass scatter of means as it encompasses the scatter of means of subclasses within the same classes. Its minimisation may lead to the minimisation of the $\boldsymbol{S}_{bs}$ scatter. Building on the use of $\boldsymbol{S}_{ws}$ to minimise scatterness (or add compactness), and for maximising data scatterness, the proposed CSDA employs:

$$\boldsymbol{S}_A^{CSDA} = \sum_{i=1}^{c}\sum_{\substack{j=1 \\ j\neq i}}^{c}\sum_{k=1}^{S_j} P_i P_{jk}\left(\boldsymbol{m}_i - \boldsymbol{m}_{jk}\right)\left(\boldsymbol{m}_i - \boldsymbol{m}_{jk}\right)^T \tag{18}$$

which can be expressed as the scatter of means of subclasses belonging to other classes but not within the same class. In fact, (18) excludes the term $\sum_{i=1}^{c}\sum_{j=1}^{S_i} P_i P_{ij}\left(\boldsymbol{m}_i - \boldsymbol{m}_{ij}\right)\left(\boldsymbol{m}_i - \boldsymbol{m}_{ij}\right)^T$. And with $\boldsymbol{S}_B^{CSDA} = \boldsymbol{S}_{ws}$, CSDA objective function is as follows:

$$\underset{\Psi}{argmax} J^{CSDA}(\Psi) = \underset{\Psi}{arg\,max} \frac{Trace\left(\Psi^T \boldsymbol{S}_A^{CSDA} \Psi\right)}{Trace\left(\Psi^T \boldsymbol{S}_B^{CSDA} \Psi\right)} \tag{19}$$

In the transformed space, such an optimisation attempts to move means of sub-classes as far away as possible from the means of classes except the means of the classes encompassing them; and bring the samples of sub-classes as close as possible to the means of such sub-classes. Furthermore, $\boldsymbol{S}_A^{CSDA}$ can also be expressed as:

$$\sum_{i=1}^{c}\sum_{\substack{j=1 \\ j\neq i}}^{c}\sum_{k=1}^{S_j} P_i P_{jk}\left(\boldsymbol{m}_i - \boldsymbol{m}_{jk}\right)\left(\boldsymbol{m}_i - \boldsymbol{m}_{jk}\right)^T = \boldsymbol{S}_b + \boldsymbol{S}_{bsb} \tag{20}$$

That is, CSDA would include a projection in which there is a combination of class means being distant from each other; and sub-class means being far from each other as well. As in MSDA, (19) and (20) stipulate that scatter matrix $\boldsymbol{S}_{bsw}$ is not involved. An extension of (20) is to employ a weighted sum which can be exploited in the proposed Balanced Subclass LDA (BSDA). That is, one can make the contribution of scatter matrices $\boldsymbol{S}_b$ and $\boldsymbol{S}_{bsb}$ depending on weighting scalars $\alpha_1$ and $\alpha_2$ as follows:

$$\boldsymbol{S}_A^{BSDA} = \alpha_1 \boldsymbol{S}_b + \alpha_2 \boldsymbol{S}_{bsb} \tag{21}$$

Further to adding the proposed solution to the wealth of Sub-lass LDA available algorithms, one embeds more flexibility in BSDA as the matter of fixing the weight parameters can be addressed during validation. Hence, in BSDA, one can analyse data to establish whether it is better to transform data so that the class means are projected far from each other; and this can be useful if the data can be clustered into well-defined classes, or to emphasise dividing classes into sub-classes if data is not divided into well-defined classes. Furthermore, the weighted sum in (21) can also stipulate the involvement of $\boldsymbol{S}_A^{CSDA}$ as it is the sum of the two scatter matrices, and their difference $\boldsymbol{S}_{bsb} - \boldsymbol{S}_b = \sum_{i=1}^{C}\sum_{j=1}^{S_i} P_{ij}\left(\boldsymbol{m}_i - \boldsymbol{m}_{ij}\right)\left(\boldsymbol{m}_i - \boldsymbol{m}_{ij}\right)^T$. BSDA adopts $\boldsymbol{S}_B^{BSDA} = \boldsymbol{S}_{ws}$ to build a measure of compactness, and hence the objective function of the algorithm is as follows:

$$\underset{\Psi}{argmax} J^{BSDA}(\Psi) = \underset{\Psi}{arg\,max} \frac{Trace\left(\Psi^T \boldsymbol{S}_A^{BSDA} \Psi\right)}{Trace(\Psi^T \boldsymbol{S}_{ws} \Psi)} \tag{22}$$

### 3.2. A nonlinear extension: kernel CSDA

A nonlinear extension of CSDA based on kernel functions can solve the non-linearity issue of data at subclass level. Without prior knowledge the nonlinear feature mapping explicitly, the formulation is based on the so-called kernel trick and the use of dot products to map the partitioned data into a high- or even infinite-dimensional feature space where the data are expected to be linearly separable. The extension is termed Kernel CSDA (KSDA). To deal with cases where the data is still non-linear at sub-class level, KSDA employs a nonlinear feature mapping $\varnothing$ to map the samples in the original feature space $\mathbb{R}^F$ to a Hilbert space $\mathscr{H}$ defined as:

$$\begin{cases} \varnothing(\bullet) : \mathbb{R}^F \rightarrow \mathscr{H} \\ x \rightarrow \varnothing(x) \text{ and } \mathbf{X} \rightarrow \varnothing(\mathbf{X}) \end{cases} \tag{23}$$

where $\boldsymbol{x} \in \mathbb{R}^F$ is a training sample in data $\boldsymbol{X}$. The annotation $\varnothing$ is used to indicate mapping of data or samples in the Hilbert space as $\boldsymbol{x}^{\varnothing} = \varnothing(\boldsymbol{x})$ and $\boldsymbol{X}^{\varnothing} = \varnothing(\boldsymbol{X})$. However, the dot product can be used instead of using mapped samples which may not be feasible due to the high or even the infinite dimensionality of $\mathscr{H}$. The dot product in $\mathscr{H}$ is equivalent to applying the kernel function $\varkappa$ in the original space if the kerel used satisfies Mercer's condition [16,17,35]. That is, for two samples $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^F$, one can have:

$$\varkappa(\boldsymbol{x}, \boldsymbol{y}) = \varnothing(\boldsymbol{x})^T \varnothing(\boldsymbol{y}) \tag{24}$$

The data $\mathbf{X}$ is divided into sub-class blocks $\mathbf{X} = \left[\boldsymbol{X}_{1,1}, \boldsymbol{X}_{1,2}, \ldots, \boldsymbol{X}_{ij}, \ldots, \boldsymbol{X}_{C,S_C}\right] = \left[\boldsymbol{x}_{1,1,1}\ldots\boldsymbol{x}_{1,1,N_{1,1}}, \boldsymbol{x}_{1,2,1}\ldots\boldsymbol{x}_{1,2,N_{1,2}}, \ldots, \boldsymbol{x}_{i,j,1}\ldots\boldsymbol{x}_{i,j,N_{i,j}}, \ldots, \boldsymbol{x}_{C,S_C,1}\ldots \boldsymbol{x}_{C,S_C,N_{C,S_C}}\right]$ where $\boldsymbol{X}_{ij}$ refers to the block of samples in the $j$th subclass of the $i$th class. The matrices $\boldsymbol{S}_A^{CSDA}$ and $\boldsymbol{S}_B^{CSDA}$ in the Hilbert space $\mathscr{H}$ are redefined as:

$$\boldsymbol{S}_A^{KCDA} = \boldsymbol{S}_A^{CSDA\varnothing} = \sum_{i=1}^{c}\sum_{\substack{j=1\\ j\neq i}}^{c}\sum_{k=1}^{S_j} P_i P_{jk}\left(\boldsymbol{m}_i^{\varnothing} - \boldsymbol{m}_{jk}^{\varnothing}\right)\left(\boldsymbol{m}_i^{\varnothing} - \boldsymbol{m}_{jk}^{\varnothing}\right)^T \tag{25}$$

$$\boldsymbol{S}_B^{KCDA} = \boldsymbol{S}_B^{CSDA\varnothing} = \frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{S_i}\sum_{k=1}^{N_{ij}}\left(\varnothing\left(\boldsymbol{x}_{ijk}\right) - \boldsymbol{m}_{ij}^{\varnothing}\right)\left(\varnothing\left(\boldsymbol{x}_{ijk}\right) - \boldsymbol{m}_{ij}^{\varnothing}\right)^T \tag{26}$$

The class means and sub-class means in $\mathscr{H}$ are computed as the average of the class samples and subclass samples mapped in $\mathscr{H}$, respectively: $\boldsymbol{m}_i^{\varnothing} = \frac{1}{N_i}\sum_{j=1}^{N_i}\varnothing\left(\boldsymbol{x}_{ij}\right)$ and $\boldsymbol{m}_{ij}^{\varnothing} = \frac{1}{N_{ij}}\sum_{k=1}^{N_{ij}}\varnothing\left(\boldsymbol{x}_{ijk}\right)$ and where the indices $i$ and $j$ refer to the $i$th class and $j$th sub-class, respectively, and index $k$ refers to a sample's membership in a subclass. The optimisation problem leads to a general eigen decomposition solution which involves vectors in $\boldsymbol{\Psi}$ and values in $\boldsymbol{\Lambda}$ as follows:

$$\boldsymbol{S}_A^{KCDA\varnothing}\,\Psi = \boldsymbol{S}_B^{KCDA}\Psi\Lambda^{\varnothing} \tag{27}$$

A solution can be obtained when taking account of the fact that according to the Representer's theorem condition [16,17,35], the sought projection matrix can be defined as a linear combination of the training samples $\varnothing(\mathbf{X})$ in $\mathscr{H}$. There exists a coefficient matrix $\Gamma$ such that:

$$\boldsymbol{\Psi} = \varnothing(\mathbf{X})\Gamma \tag{28}$$

hence,

$$\varnothing(\mathbf{X})^T\boldsymbol{S}_A^{KCDA}\varnothing\,(\mathbf{X})\Gamma = \varnothing(\mathbf{X})^T\boldsymbol{S}_B^{KCDA}\varnothing(\mathbf{X})\boldsymbol{\Gamma\Lambda}^{\varnothing} \tag{29}$$

And by terming $\varnothing(\mathbf{X})^T\boldsymbol{S}_A^{KCDA}\varnothing(\mathbf{X})$ and $\varnothing(\mathbf{X})^T\boldsymbol{S}_B^{KCDA}\varnothing(\mathbf{X})$ matrices $\boldsymbol{S}_A^{\varkappa}$ and $\boldsymbol{S}_B^{\varkappa}$ respectively, one would have:

$$\boldsymbol{S}_A^{\varkappa} = \sum_{i=1}^{C}\sum_{\substack{j=1\\ j\neq i}}^{C}\sum_{k=1}^{S_j} p_i p_{jk}\left(\mathrm{K}_i\mathrm{P}_i - \mathrm{K}_{jk}\mathrm{P}_{jk}\right)\left(\mathrm{K}_i\mathrm{P}_i - \mathrm{K}_{jk}\mathrm{P}_{jk}\right)^T \tag{30}$$

and

$$\boldsymbol{S}_B^{\varkappa} = \frac{1}{N}\sum_{i=1}^{C}\sum_{j=1}^{S_i}\left(\left(\mathrm{K}_{ij}\right)\left(\mathrm{I}_{ij} - \mathbf{P}_{ij}\right)\mathrm{K}_{ij}^T\right) \tag{31}$$

$\mathrm{I}_{ij}$ and $\mathbf{P}_{ij}$ are the identity and all-$^1/_{N_{ij}}$ matrices, respectively, and both of a $N_{ij} \times N_{ij}$ dimensionality. Further, $\mathbf{K}_i = \varnothing^T(\boldsymbol{X})\varnothing(\boldsymbol{X}_i) \in \mathbb{R}^{N\times N_i}$ and $\mathbf{K}_{ij} = \varnothing^T(\boldsymbol{X})\varnothing\left(\boldsymbol{X}_{ij}\right) \in \mathbb{R}^{N\times N_{ij}}$. The development in (30) and (31) shows that only dot products are involved in the calculation of $\boldsymbol{S}_A^{\varkappa}$ and $\boldsymbol{S}_B^{\varkappa}$. Hence the kernel trick can be used to solve (27) within the general eigendecomposition framework.

## 4. Performance and results analysis

To appreciate the performance of the suggested solution, it is deployed in a context where it uses several types of data and is compared to similar solutions and works in the literature [17,18,20,21,25,26,39]. That is, the modified sub-class LDA presented in this paper can find applications in the general field of ML; and is also compared to other sub-class LDA algorithms. Such a comparison can be based on general ML datasets. Further, for comparison purposes, food-based datasets acquired using miniature spectrometers are used. This includes datasets collected by the authors and other sets used in similar works in the literature.

### 4.1. The data

For the empirical analysis of machine learning algorithms and to compare the performance to those of sub-class LDA variants, datasets from the publicly accessible UCI (University of California Irvine) machine learning repository are used. The UCI data has been widely used within the machine learning community for applications, involving clustering analysis and labelled data of various attributes, sizes, and dimensionalities [40,55].

Using miniature spectrometers, the authors collected a further two food-based spectral datasets, namely, the Olive Oil Dataset and the Apple Dataset. The Olive Oil dataset (OOD) has been prepared by increasingly mixing OO with Vegetable Oil (VO) with steps between 5 % and 10 % in terms of decreasing OO purity. That is, one begun data collection with a 100 % pure OO set of samples from which the purity was decreased by adding more VO. The purities of the samples used are 100 %, 90 %–95 %, 80 %–85 %, 70 %–75 %, 60 %–65 % and 50 %–55 %. To simplify the annotations used, those 6 mixtures are referred to as the 100 %, 90 %, 80 %, 70 %, 60 % and 50 % classes.

Beyond the class of pure OO (100 % purity), all samples have been prepared in the lab where ten data collection sessions have taken place. The data collection phase took three weeks; and as it is common in ML, sessions are well distinct from each other; and only once a session is completed that data collection for a new session would commence. It is worth pointing out that over the time it took for the 10 sessions of data collection to complete, ambient and background environments changed; sessions took place in mornings, afternoons, and early evenings and further distortions may include any lighting induced distortions and the window blinds were randomly kept open or close. In addition, all experiments have been conducted under ambient room temperature to mimic the use of such a device by an ordinary consumer for a relatively wide deployment scenario. Oil samples were collected using an STS UV portable spectrometer from Ocean Insight Inc covering the range from 190 nm to 650 nm. In each session, 380 scans would be carried out, hence, leading to a dataset of 3800 samples. Each scan yielded 1024 wavelength intensity pixels per sample. The adopted integration time is 1 s; and the calibration of the device has been carried out by the manufacturer and which was further updated using the manufacturer's calibration. In between the data collection sessions, dark measurements were taken to correct for the differences in background lighting. As the sensor used has no built-in light source, a 45° diffuse reflectance DR-probe with an integrated Tungsten Halogen light source has allowed for elegant reflectance measurements. The position of the probe and the oil surface was kept at 1.5 cm across the measurements. This makes the measurement results more reliable and consistent. The spectrometer used has an optical resolution of 1.5 nm Full Width Half Maximum (FWHM) with a slit size of 25 μm [41,42]. Figs. 1 and 2 show the plot of the oil classes and their PCA representations, respectively. The devices used for data collection are shown in Fig. 3.

The Apple Dataset (AD) has been prepared from four types of Apples available in a local Tesco supermarket which is part of a British multinational groceries and general merchandise retailer. The selected four types were Ariane, Braeburn, Golden Delicious and Gala. Except the Ariane apples which were cultivated in France, all remaining apples originated from Belgium. The spectral data was collected using an STS-NIR spectrometer covering the 645–1085 nm range, an integration time of 500 ms, a resolution of 1.5 nm, a slit size of 25 μm and a HL-2000 light source in the reflection mode [41,42]. Two sessions of data acquisition have been completed with a one-day interval between the two sessions. Three Apples were selected for each type and cut into four pieces. Ten spectral data readings were taken for each piece making 120 reading per Apple type. For the four Apple types, 480 spectral data readings have been completed per session which completed a dataset of 960 spectral

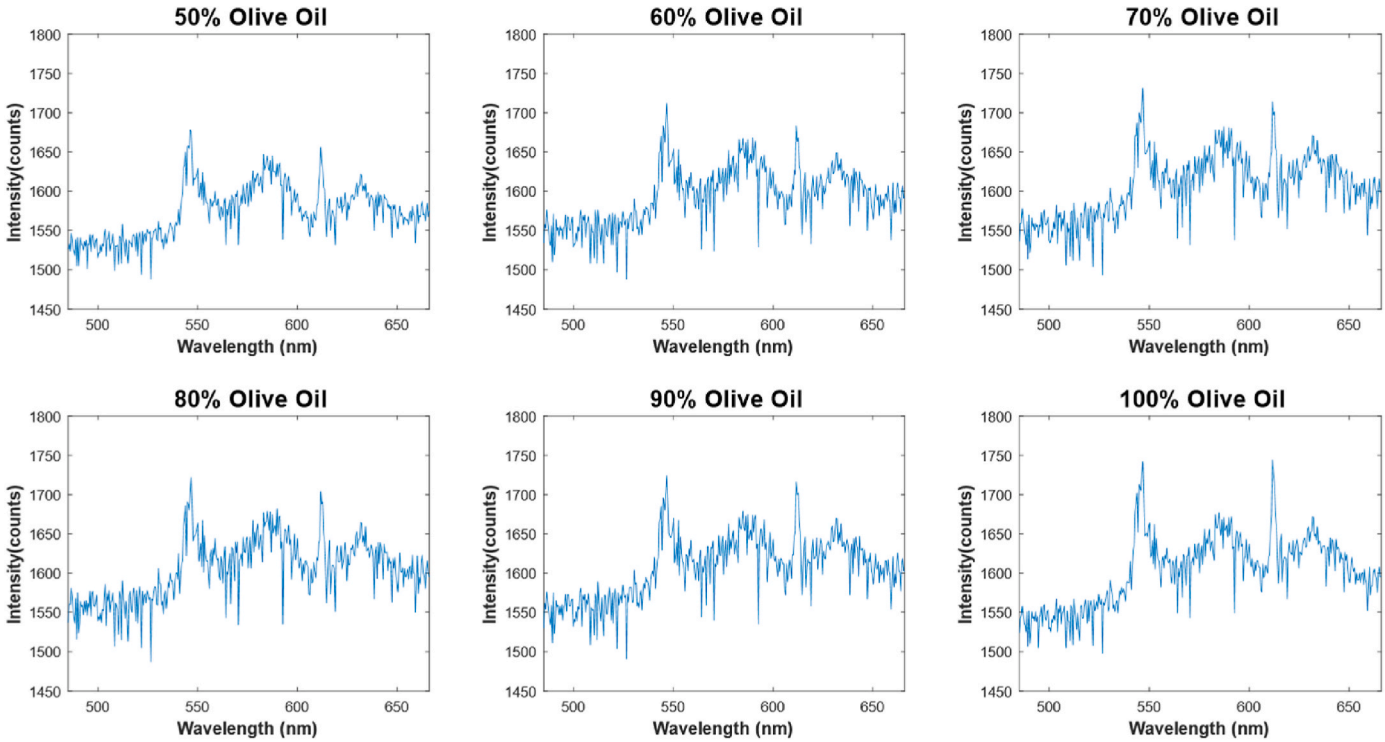



**Fig. 1.** Plot of oil classes.

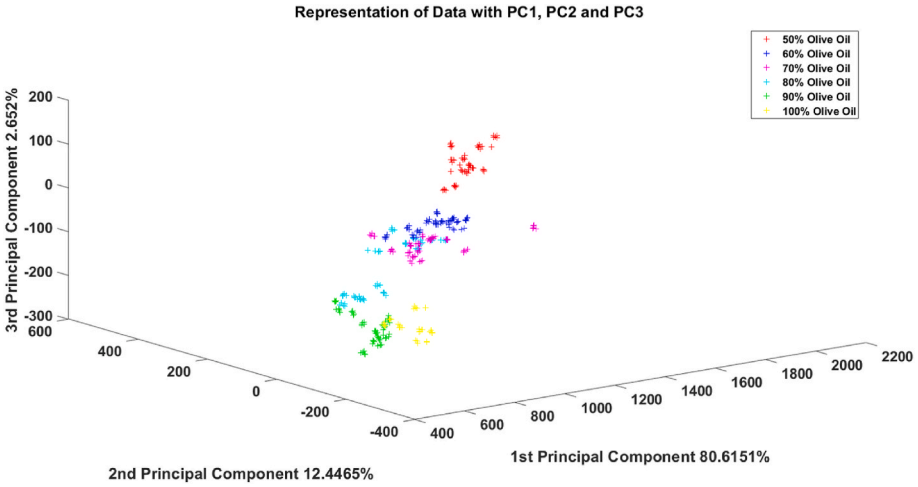



**Fig. 2.** PCA plot of oil classes.


**Fig. 3.** Miniature spectrometers used for data collection [41,42].

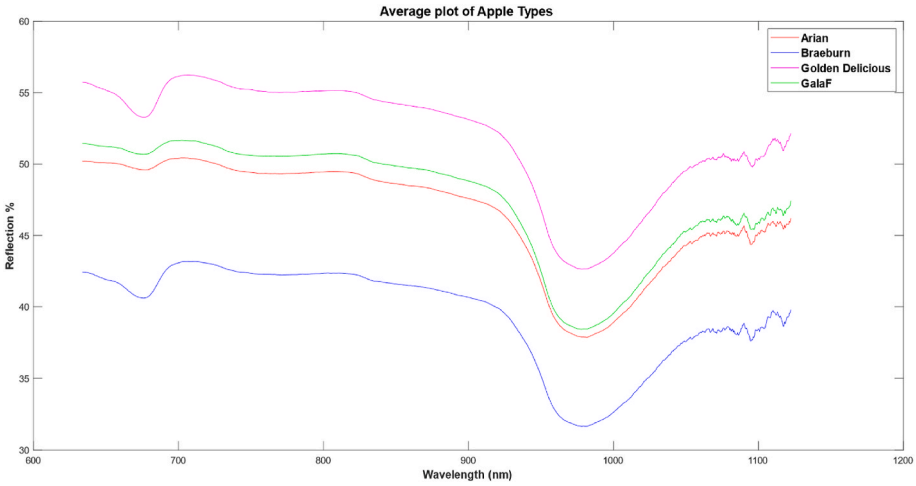



**Fig. 4.** Average plot of the four Apple types.

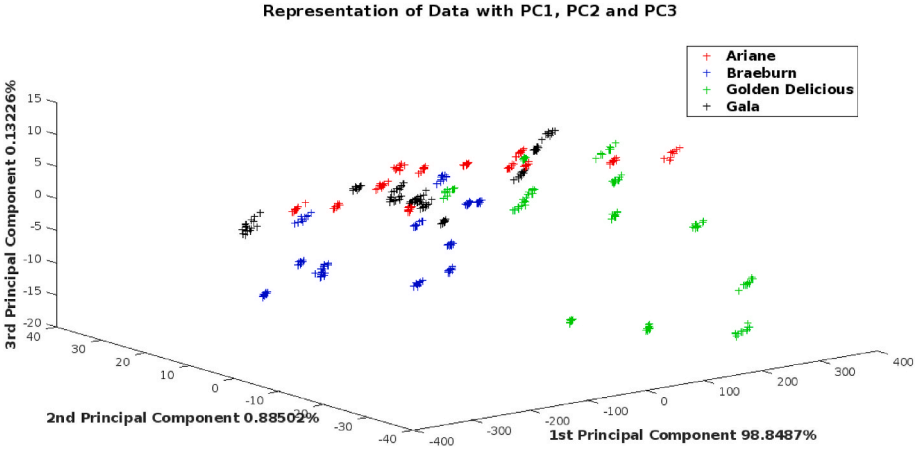



**Fig. 5.** PCA plot of apple data.

data belong to four classes. Samples' dimensionality is 1024. Figs. 4 and 5 show the plot of the apple classes and their PCA representations, respectively.

Raw data acquired from the STS-UV and NIR portable spectrometers is pre-processed before it is passed into further classification steps. The pre-processing is based on a combination of mean & deviation normalisation, Multiplicative Scatter Correction (MSC) for which the mean and standard deviation of the training data are taken into account to MSC-correct test data, Savitzky-Golay (SG) filters and Daubechies Wavelet filters [1,24,26,27,31,43,44].

In the research literature [39], the Cultivar Dataset has been used to establish the performance of miniature spectroscopy in discriminating the three important cultivars of barley, chickpea and sorghum in the Ethiopian agricultural system. The work in Ref. [39] addressed the hypothesis of the suitability of miniature NIR for dry cultivar identification; and for that purpose, a Consumer Physics SCIO spectrometer covering the range 740–1070 nm has been employed. All grain samples from a seasonal harvest were dried before processing and then scanned in a similar position. A total of 50 grains per cultivar were used, and 24 cultivars, 19 cultivars and 10 cultivars of barley, chickpea and sorghum, respectively. This yielded 1200 samples of barely, 950 samples of chickpea and 500 samples of sorghum. For comparison purposes, the authors also used the Milk Dataset in Ref. [25] which is another set of data collected using a miniature spectrometer. The milk database includes 87 full fat and pasteurized retail milk samples collected from grocery stores in The Netherlands over a period of 8 weeks; of which 37 samples of organic milks and the remaining 50 samples are conventional non-organic milk samples. All samples have been analysed using a Consumer Physics Scio device along with its accessory accessories to allow for the submergence of the device in a glass beaker of milk. The spectrometer has been used to collect transmission data in the 740–1070 nm range. Further, calibration with the device white reference has been carried out after analysis of each set of three samples.

The major advantage of the four spectral datasets in the context of the presented work is that they used a miniature spectroscopy technology. The spectrometers used are cost effective and relatively in the price range of consumers; and therefore, available to many which makes the deployment of the technology even more suitable for citizen science. The matter is to investigate the power of such technology and whether the pertinent task of food analysis can be carried out by the crowd to establish if such food products adhere to regulations. However, in the datasets the authors collected, only cross session classification is attempted. Subclasses are taken as sessions data which is a very plausible deployment scenario and is a matter not emphasized in research papers. It is also a point which coincides with the common practice in ML where data belonging to the same session is not used to build models and test them at the same time.

All data used for analysis is summarised in Table 1. Although datasets may include continuous feature values, data is solely used for classification. When needed, to emphasise the concept of sub-class classification, datasets classes can be divided into clusters where a cluster is used as a sub-class. This is also the adopted approach with the

**Table 1**
Description of data used for analysis.

| Dataset | Description | Samples in Dataset | Classes | Num. Clusters | Num.of Samples for Training & Testing |
|---|---|---|---|---|---|
| Breast Tissue | Impedance measurements for classification | 106 | 6 | 3 | 74 & 32 |
| Ecoli | Protein localisation data with both continuous and binary features for classification | 332 | 5 | 3 | 229 & 98 |
| Forest Types | Remote sensing data for classification | 523 | 4 | 3 | 366 & 157 |
| Ionosphere | Radar data of continuous values for classification | 351 | 2 | 3 | 246 & 105 |
| Parkinsons | Biomedical data with continuous features for classification | 195 | 2 | 3 | 137 & 58 |
| QSAR-biodeg | Chemometrics data for classification | 1055 | 2 | 3 | 739 & 316 |
| SPECTF heart | Tomography images for binary classification | 267 | 2 | 3 | 187 & 80 |
| WDBC | biomedical images of continuous variables for classification | 569 | 2 | 3 | 398 & 171 |
| RWQ | Physiochemical and sensory data | 1599 | 6 | 3 | 1119 & 480 |
| Milk | Spectral data for classification | 87 | 2 | 2 | 61 & 26 |
| Barley | Spectral data for classification | 1200 | 24 | 2 | 840 & 360 |
| ChickPeas | Spectral data for classification | 950 | 19 | 2 | 665 & 285 |
| Olive Oil | Discrete spectral data | 3800 | 6 | 10 | 3420 & 380 |
| Apple | Discrete spectral data | 960 | 4 | 2 | 480 & 480 |

apple data as there are only two collection sessions; hence, the classes have been further divided into 2 clusters each. Further, all the datasets in Table 1 have been divided into training and testing samples with almost 70 % of samples used for model building and about 30 % of samples dedicated to testing. However, this was carried out with the exception of OOD and apple data. For olive oil data, the emphasis is clear on the use of sessions data rather than clusters. All samples from 9 sessions would be used for training and the 10th session's data used for testing; which is then repeated 10 times in a round-robin style.

### 4.2. Results analysis

The first carried out experiment aims at analysing the performance of the presented algorithm and compares it to other and similar LDA variants in a generic ML setting for which such algorithms have been devised. That is, as part of bringing the matter of food analysis using miniature spectrometer to the generic ML field in its broadest sense, the suggested solution should also compete with generic and similar solutions devised for a rather different set of machine learning applications, purposes, and aim. Hence, the suggested algorithm is compared to open research published classifiers in Table 2 on various sets from the UCI database. A clear point one can take from such table is that subclass LDAs tend to improve the performance of the classical LDA. By creating three clusters of each class, LDA do not attain higher rates than the subclass LDA algorithms although there are cases where LDA performs as good as its subclass variants. The suggested CSDA compares well with subclass LDA algorithms in the literature. This is clear in Table 2 where CSDA can be compared to the best of all LDA variants in the table in the second column from the right (except BSDA) which shows that in the context of the presented experiment results, CSDA can yield the better rates of all similar algorithms in the table in about 50 % of all cases. Furthermore, BSDA has always the tendency of improving the performance of CSDA although none of the algorithms in Table 2 is the outright best. Nevertheless, in average, BSDA would be the top performing algorithm.

Building up on such results, OOD was used to gauge the performance of the suggested algorithm. The scenario here is to use 9 sets of data to build a model and the tenth set which has not been used in modelling as a test set. The process is repeated 10 times in a round-robin style so that all data can be used for testing, which is a plausible deployment scenario is which relatively large amounts of data can be available mainly due to the ease of use of the technology and the large community of citizens who would be the technology users. The performance of such classifiers in Table 2 and throughout the remaining sections of the manuscript is gauged in terms of accuracy. As with Table 2, the LDA variant algorithms in Table 3 have been implemented by the authors who also used Matlab implementations of SVM, KNN and PLS-DA. Only the performance of such algorithms is shown in this paper within the limits of the presented analysis. Such table shows that the suggested CSDA is quite competitive with similar algorithms in the literature. Its performance can be improved using BSDA in average, although only slightly in the context of the experiment and results exhibited in Table 3. In the authors' implementation of SVM, a Gaussian kernel was used; PLS-DA uses 16 LVs and KNN is taking 9 neighbours into consideration to make a classification decision. Such values and parameters yielded the highest performance these 3 classifiers could attain to the extent of the authors' analysis. The matter is further detailed in Table 4 which shows the result of the 3 best performing algorithms in Table 3. The performance of CSDA is compared to SSDA-1 on the 10 sets; and in average, they have attained similar rates. To underline the flexibility of using subclasses in LDA, BSDA is shown to attain a higher rate when $\mathbf{0.1}(\boldsymbol{S}_{bsb}-\boldsymbol{S}_b)$ is added to $\boldsymbol{S}_A^{CSDA}$ which is referred to as BSDA -0.1 in Table 4. A further tuning of the parameters in (12) is shown as the maximum rates attained by BSDA. However, this is only a slight improvement in average over CSDA and BSDA-0.1.

**Table 2**
LDA variants performance on various UCI subsets.

| | LDA | MDA | SDA | MSDA | SSDA-1 | SSDA-2 | SSDA-3 | CSDA | Max. LDA variants | BSDA |
|---|---|---|---|---|---|---|---|---|---|---|
| Breast Tissue | **71.43** | 57.14 | 47.62 | 66.67 | 66.67 | 66.67 | 66.67 | **71.43** | 71.43 | 80.95 |
| Ecoli | **93.94** | 87.88 | 90.91 | 90.91 | **93.94** | 90.91 | 90.91 | 90.91 | 93.94 | 93.94 |
| Forest Types | 88.46 | 92.31 | 88.46 | 86.54 | 90.38 | 90.38 | 90.38 | **94.23** | 94.23 | 96.15 |
| Ionosphere | 91.43 | 94.29 | 85.71 | 94.29 | 91.43 | 91.43 | **100** | 97.14 | 100 | 100 |
| Parkinsons | **94.74** | 89.47 | **94.74** | **94.74** | **94.74** | 84.21 | 84.21 | **94.74** | 94.74 | 94.74 |
| QSAR-biodeg | 88.57 | 86.67 | **89.52** | 87.62 | 88.57 | 88.57 | 87.62 | 87.62 | 89.52 | 88.57 |
| SPECTF heart | 76.92 | **80.77** | **80.77** | 69.23 | **80.77** | **80.77** | **80.77** | **80.77** | 84.62 | 80.77 |
| WDBC | 97.35 | 98.23 | 98.23 | **99.12** | 96.46 | **99.12** | 97.35 | **99.12** | 99.12 | 99.12 |
| RWQ | 60.19 | 59.56 | 59.56 | 59.25 | 58.93 | **62.07** | 59.87 | **62.07** | 62.07 | 63.01 |
| average | 84.78 | 82.92 | 81.72 | 83.15 | 84.65 | 83.79 | 84.2 | **86.45** | 87.74 | **88.58** |

**Table 3**
Performance of the proposed algorithm on the OOD.

| | LDA | MDA | SDA | MSDA | SSDA-1 | SSDA-2 | SSDA-3 | CSDA | BSDA | SVM | KNN | PLS-DA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Avg | 82.39 | 83.92 | 83.13 | 50.63 | **85.05** | 73.61 | 48.92 | **85.32** | **85.84** | 79.16 | 67.55 | 67.55 |

**Table 4**
Performance of SSDA-1, CSDA and BSDA.

| | SSDA-1 | CSDA | BSDA -0.1 | Max. of CSDA & BSDA -0.1 | BSDA |
|---|---|---|---|---|---|
| Set-1 | **96.32** | 93.42 | 92.37 | 93.42 | 96.05 |
| Set-2 | 67.11 | 64.47 | **68.16** | 68.16 | 68.16 |
| Set-3 | 86.84 | 84.21 | **90.53** | 90.53 | 90.53 |
| Set-4 | 71.05 | 67.89 | **72.63** | 72.63 | 72.63 |
| Set-5 | 82.63 | **91.05** | 83.95 | 91.05 | 91.05 |
| Set-6 | **90.79** | 90.26 | **90.79** | 90.79 | 92.11 |
| Set-7 | 93.95 | **95.79** | 94.47 | 95.79 | 95.79 |
| Set-8 | 84.21 | **86.58** | 85 | 86.58 | 87.37 |
| Set-9 | 91.32 | **93.95** | 91.58 | 93.95 | 95 |
| Set-10 | 86.32 | 85.53 | **88.95** | 88.95 | 88.95 |
| Average | 85.05 | 85.32 | **85.84** | **87.19** | **87.76** |

Further to the results obtained using miniature spectroscopy data, the performance is also gauged on the milk database. Two clusters per class are used to create subclasses which yielded the results in in Table 5; and CSDA exhibits better results when compared to similar algorithms in the literature. Interesting however is to compare the performance of the algorithms in Table 5 with similar work in the literature where a miniature spectroscopy technology was used. For instance, the authors in Ref. [45] reported an accuracy of 73 % when distinguishing organic from non-organic milk with external validation on NIR spectral data. With PLS-DA and cross validation, the accuracy reported in Ref. [25] is 89 %. That is, the result shown in Table 5 go some way in confirming the findings in Refs. [25,45] that a reasonable classification rate can be attained using miniature spectrometers although remaining inferior to the performance obtained using benchtop spectrometers in distinguishing organic from non-organic milk application. Table 6 further shows the improvements introduced by using subclass LDA classification, the improved performance of both CSDA and BSDA, and the results of using classifiers fusion. Albeit simple fusion rules have been used including minimum (MIN), sum (SUM) and Dempster Shafer Theorem (DST) based fusion [46], the results have outperformed a tunned BSDA. The classifiers involved in fusion at score level are SVM, LDA and KNN. It is also worth pointing out that other fusion rules failed to yield better classification rates that what is reported in Table 6 and have as such been omitted. The authors also omitted various combination of classifiers including tree based classifiers, as they did not improve the reported results in Table 6. That is, although a careful fusion can improve the results, this has not been a consistent occurrence, and a subclass LDA classification remains a very competitive alternative.

Taking this analysis further, the authors have also gauged the performance of the proposed classifier when combined with regression based representations as in the sparse, regularised and collaborative representations in Refs. [47–53]:

$$\widehat{\boldsymbol{\rho}} = arg\min_{\rho} \|\boldsymbol{\rho}\|_1 \; s.t. \; \|\boldsymbol{A\rho} - \boldsymbol{y}\|_2 \leq \varepsilon \tag{32}$$

$$\widehat{\boldsymbol{\rho}} = arg\min_{\rho} \|\boldsymbol{A\rho} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{\rho}\|_1 \tag{33}$$

$$\widehat{\boldsymbol{\rho}} = arg\min_{\rho} \|\boldsymbol{A\rho} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{\rho}\|_2^2 \tag{34}$$

Such representations may be able to solve the misalignment or pose changes distortions in image classification, and plane transformation, and can show some robustness when dealing with signal corruption [47]. In Refs. [51,54], it is shown that when the aim is to recover or approximate a query sample from a limited number of observations as a linear combination of measurements of the same class, one can attain an improved classification rate and a low representation error. Moreover, the ensuing representation can further improve classification results when combined with supervised and unsupervised classification-based projections, and even when random projections are employed, which are subsequently followed by regression-based representations in the transform domain. The results exhibited in Tables 6–8 follow such approach; and the used regression-based representations are shown in equations (32)–(34).

The sought representation $\widehat{\boldsymbol{\rho}}$ in (32) is sparse and is an optimisation attempt to reduce its $l_0$ norm which simply measures the sparseness of the sought solution. That is, the sparse representation-based classification (SRC) regression computes the size of the support of the signal which is the number of its nonzero entries. Such optimisation can be formulated as an $l_1$ minimisation subject to quadratic constraint as in (32) which is a convex minimisation problem. The annotation followed when equation (32) is used in the remainder of the paper is *classifier*+ $l_0$ to indicate that the projection takes place first then is followed by the regression-based representation. Equation (34) seeks the classical solution of a Collaborative Representation based Classification (CRC) with an $l_2$ regularisation term and an ensuing solution given as $\widehat{\boldsymbol{\rho}} = \left(\boldsymbol{XX}^{\boldsymbol{T}} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}$. The representation in (33) uses an $l_1$ regularisation term; hence when used in conjuncture with a classifier the annotation *classifier*+ $l_1$ is used. Similarly, the annotation *classifier*+ $l_2$ indicates that the CRC representation is used once data is projected in the transform domain. Further, in (32), (33) and (34), the parameter $\lambda$ is a tuning parameter, $\boldsymbol{y}$ is the query sample, $\boldsymbol{X}$ is a matrix containing the training sample and acts as a code dictionary, and $\varepsilon$ is an arbitrarily very small quantity which considers noise, hence the representation error and the sparsity of the representation $\widehat{\boldsymbol{\rho}}$. Tables 6–8 exhibit the results of using the kernel classifiers introduced in this paper, KCDA. The tables

**Table 5**
Performance of the algorithms in establishing the organic feature of milk.

| Classifier | LDA | MDA | SDA | MSDA | SSDA-1 | SSDA-2 | SSDA-3 | CSDA |
|---|---|---|---|---|---|---|---|---|
| Performance | 73.38 | 75.54 | 74.77 | 72.46 | 73.85 | 75.54 | 73.38 | **77.08** |

**Table 6**
Performance of the algorithms on the cultivar database in [39].

| | LDA | MDA | SDA | MSDA | SSDA-1 | SSDA-2 | SSDA-3 | CSDA | BSDA | [39] | SVM | KNN | PLSDA | SUM | MIN | DST |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Barley | 85 | 85 | 85.83 | 86.67 | 88.33 | 86.67 | 86.67 | 85 | 87.5 | 86.9 | 85.83 | 76.67 | 60.83 | 89.17 | **90** | 89.17 |
| Chickpeas | 94.74 | 93.68 | 94.74 | 95.79 | 92.63 | 94.74 | 90.53 | 94.74 | **97.89** | 94.7 | 90.53 | 87.37 | 77.89 | 92.63 | 91.58 | 93.68 |

**Table 7**
KCDA and CSDA results on the apple dataset with session S1 used for testing.

| CSDA | KCDA-*d2* | KCDA-*d3* | KCDA-*d4* | KCDA-*d5* |
|---|---|---|---|---|
| **85.83** | **86.04** | **80.21** | 65.63 | 68.75 |
| CSDA+ $l_0$ | KCDA-*d2*+ $l_0$ | KCDA-*d3*+ $l_0$ | KCDA-*d4*+ $l_0$ | KCDA-*d5*+ $l_0$ |
| 77.08 | **81.04** | 70.42 | 42.5 | 56.04 |
| CSDA+ $l_1$ | KCDA-*d2*+ $l_1$ | KCDA-*d3*+ $l_1$ | KCDA-*d4*+ $l_1$ | KCDA-*d5*+ $l_1$ |
| 76.67 | **85.63** | 77.29 | 50.63 | 65 |
| CSDA+ $l_2$ | KCDA-*d2*+ $l_2$ | KCDA-*d3*+ $l_2$ | KCDA-*d4*+ $l_2$ | KCDA-*d5*+ $l_2$ |
| 79.58 | **82.29** | 81.25 | 70 | 71.67 |

**Table 8**
KCDA and CSDA results on the apple dataset with session S2 used for testing.

| CSDA | KCDA-*d2* | KCDA-*d3* | KCDA-*d4* | KCDA-*d5* |
|---|---|---|---|---|
| **94.17** | 71.04 | 83.96 | 88.33 | 75.21 |
| CSDA+ $l_0$ | KCDA-*d2*+ $l_0$ | KCDA-*d3*+ $l_0$ | KCDA-*d4*+ $l_0$ | KCDA-*d5*+ $l_o$ |
| 86.46 | 86.45 | 85.83 | 90.42 | 87.08 |
| CSDA+ $l_1$ | KCDA-*d2*+ $l_1$ | KCDA-*d3*+ $l_1$ | KCDA-*d4*+ $l_1$ | KCDA-*d5*+ $l_1$ |
| 91.46 | 74.17 | 85.21 | 89.79 | 82.71 |
| CSDA+ $l_2$ | KCDA-*d2*+ $l_2$ | KCDA-*d3*+ $l_2$ | KCDA-*d4*+ $l_2$ | KCDA-*d5*+ $l_2$ |
| **93.75** | 85.63 | 93.13 | **94.38** | 86.25 |

only show the best results from a larger analysis using the AD where among the classical kernels, exponential kernels have not exceeded the performance attained by the polynomial kernels used. The polynomial kernel of degree *deg* can be indicated by the annotation *-ddeg* in such tables which also exhibit the performance of the classifiers when combined with one of the representations in (32)–(34). In most classification cases in Table 7, a better performance is obtained when a low polynomial degree kernel is used with the quadratic kernel attaining the best performance in Table 7 and where a regression based representation can only achieve competitive results. That is, in Table 7, KCDA-*d2* is slightly better than CSDA and KCDA-*d2*+ $l_1$ with the latter classifier attaining competitive but not better results than the linear classifier. The other point one can take from Tables 7 and is that when KCDA-*d2* is combined with any of the regression-based representations, it persistently tends to do better than any of the classifiers in the table when combined with the same regression-based representation which really underlines the performance of a kernel-based sub class LDA. The matte can also be shown in Table 8 with a combination of KCDA and CRC representation edging slightly better than the linear CSDA in terms of classification performance. Taking the investigation further and looking deeper into the matter of using regression-based representations, a comparison of the results of fusion of multiple classifiers is shown in Table 9. Of the fusion techniques employed by the authors, only the results of the Weighted Majority Voting (WMV) and DST are reported. The weights associated with the WMV rule are the classification rates of the individual classifiers involved. Other fusion techniques have not yielded competitive res [55]ults when compared with DST and WMV; and that includes the simple fusion rules used in Table 6. Further, the analysis extended to include a polynomial kernel of up to degree 5 only; and as in the analysis in Tables 6 and 7, exponential kernels attained lower performance and have been omitted from Table 9.

Compared to the performance of the CSDA in Table 8, a combination of CSDA and 4 KCDAs combined with a CRC representation using a WMV rule can attain better results than CSDA in the extent of the experiment carried out using the AD. That is, the kernelization of the proposed sub-class LDA has yielded competitive results in Tables 6 and 7; and is edging with over a margin of 2 % better than CSDA in the experiments of Table 9. This shows a relatively good improvement to the extent of the experiments in Tables 6–8.

Nevertheless, regression based representations when combined with classifiers did not achieve an overall improvement in classification performance comparable to the success enjoyed in face and palmprint recognition in Refs. [49,51,54]. Furthermore, sub-class classification may have by its own merit lessened the impact of non-linearity by dividing data into clusters in which samples share the same statistical attributes. Moreover, in the literature, kernelization and classical kernels used have been reported to not always guarantee improvement of results [17]. That is, if samples are grouped in clusters, the kernelization that traditionally has been used to solve non-linearity may not yield improvements when globally applied. Rather, a multi-kernel approach may be more useful as one is not certain of the data non-linearity attributes and whether clustering would take account of such attributes. That is, to the extent of the experiments carried out, the results in Table 9 would favour the use of multiple kernels with fusion techniques.

## 5. Conclusions

Food quality analysis using miniature spectrometers augurs a new citizens' realm where the war against food fraud may enlist ordinary citizens in the forefront of the fight instead of having them as mere victims. This has been made possible because of technology miniaturization, its cost effectiveness, ease of use and short processing time. The technology however may not yield top shelf performance, and coupled with the amount of data that it may generate due mainly to its possible wide deployment, the matter would lend itself very nicely to being a machine learning application. That is, the addressed aim in the paper is to bring food analysis under the general umbrella of ML and to tackle it in a pure data driven approach. The authors have collected data in

**Table 9**
Combination with five polynomial kernels (up to *d5*; CSDA uses kernel *d1*).

| DST + $l_0$ *d1-5* | DST + $l_1$ *d1-5* | DST + $l_2$ *d1-5* | WMV + $l_0$ *d1-5* | WMV + $l_1$ *d1-5* | WMV + $l_2$ *d1-5* |
|---|---|---|---|---|---|
| 93.13 | 90.63 | 93.13 | 92.71 | 91.25 | **95.83** |
| DST + $l_0$ *d1,3,4* | DST + $l_1$ *d1,3,4* | DST + $l_2$ *d1,3,4* | WMV + $l_0$ *d1,3,4* | WMV + $l_1$ *d1,3,4* | WMV + $l_2$ *d1,3,4* |
| 93.96 | 94.17 | **94.58** | 93.54 | 90.62 | 93.75 |
| DST + $l_0$ *d1,2,4* | DST + $l_1$ *d1,2,4* | DST + $l_2$ *d1,2,4* | WMV + $l_0$ *d1,2,4* | WMV + $l_1$ *d1,2,4* | WMV + $l_2$ *d1,2,4* |
| 94.58 | 89.38 | **95.0** | 91.67 | 92.29 | **96.67** |

multiple sessions and attempted to mimic data collection exercises as they may be carried out by ordinary users who would not have a particular training to appreciate the spectroscopy technology complexity. Further, as it is common in ML, models built using collected data would be tested using samples collected in different sessions. Although it is common in ML, this matter is deemed necessary to establish the capabilities of miniature devices in a real-world scenario which is a matter usually overlooked in research papers. It is within this context that the paper introduces a new sub-class LDA for spectral based data analysis. The choice is well justified by the fact that samples collected in different sessions may exhibit different statistical features, as such forming subclasses within the same class. The proposed classifier takes account of the distances between class averages and subclass averages to build a measure of scatterness and attempts to minimise the distance between samples in a subclass cluster and its average. The suggested subclass LDA achieves competitive or better results than traditional ML algorithms and similar subclass classifiers when applied on general ML datasets. This clearly demonstrates the benefit of using the adopted approach which is a point that confirms what has already been reported in the literature, and the merit of the suggested algorithm over similar solutions. The matter has been demonstrated in olive oil data collected in multiple sessions which is the exact point for which the suggested algorithm was devised. Further, the subclass classification approach exhibits more flexibility in classification than similar ML methods when it comes to using clusters, which is a matter that has been clearly shown in the relatively extensive set of conducted experiments and in which the authors' algorithm has improved the results. It is also a flexibility in which the scatterness metric can include both the distance of class averages to data average and the distance of subclass averages to class averages. A further extension shown in the paper is the kernel version of the suggested algorithm. For a detailed analysis, the authors have also deployed regression-based representations for their reported ability to classify distorted data. Although the kernel-based algorithm has shown high competitiveness with the linear classifier, the clear improvement in results could only be achieved when multiple kernels were used. This may suggest that subclass classification may be addressing data nonlinearity as a multimodal data distribution; that nonlinearity may be related to models' degeneration and only becomes obvious when multiple data collection sessions are used. The general assumption is that spectrometers, food samples and spectral samples collection may be affected by background and ambient conditions. Hence, inter-session classification may be exhibiting some data nonlinearity which otherwise would be almost linear per collection session. This may explain why not a single classical and globally applied kernel can achieve overall improvement.

## CRediT authorship contribution statement

**Omar Nibouche:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Fayas Asharindavida:** Data curation, Software, Validation, Visualization. **Hui Wang:** Data curation, Formal analysis, Funding acquisition, Supervision. **Jordan Vincent:** Data curation, Methodology, Validation. **Jun Liu:** Formal analysis, Funding acquisition, Methodology, Project administration, Supervision. **Saskia van Ruth:** Conceptualization, Data curation, Visualization. **Paul Maguire:** Conceptualization, Data curation, Formal analysis, Methodology. **Enayet Rahman:** Conceptualization, Data curation, Investigation, Methodology.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests

Omar Nibouche reports equipment, drugs, or supplies was provided by Ulster University. Omar Nibouche reports a relationship with Ulster University that includes: employment.

## Data availability

Data will be made available on request.

## References

[1] R. Tauler, B. Walczak, Steven Brown, in: R.T.B.W. Steven Brown (Ed.), Comprehensive Chemometrics: Chemical and Biochemical Data Analysis, second ed., Elsevier, Oxford, 2020.
[2] R. Duda, P. Hart, D. Stork, Pattern Classification, second ed., Wiley, 2000.
[3] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, Cambridge, MA, 2013.
[4] G.J. McLachlan, Discriminant Analysis and Statistical Pattern Recognition, John Wiley & Sons, 2004.
[5] R. Fisher, The statistical utilization of multiple measurements, Annals of Eugenics 8 (1938) 376–386.
[6] Y.A. Ghassabeh, F. Rudzicz, H.A. Moghaddam, Fast incremental lda feature extraction, Pattern Recogn. 48 (2015) 1999–2012.
[7] T. Hastie, R. Tibshirani, Discriminant analysis by Gaussian mixtures, J. Roy. Stat. Soc. B (1996) 155–176.
[8] W. Song, H. Wang, U. Power, E. Rahman, J. Barabas, J. Huang, J. McLaughlin, C. Nugent, P. Maguire, Classification of respiratory syncytial virus and sendai virus using portable near-infrared spectroscopy and chemometrics, IEEE Sensor. J. (2022).
[9] D. Chu, L.-Z. Liao, M.P. Ng, X. Wang, Incremental linear discriminant analysis: a fast algorithm and comparisons, IEEE Transact. Neural Networks Learn. Syst. 26 (11) (Nov. 2015) 2716–2735.
[10] K. Chumachenko, J. Raitoharju, M. Gabbouj, A. Losifidis, Incremental fast subclass discriminant analysis, in: 2020 IEEE International Conference on Image Processing (ICIP), 2020.
[11] I. Vranckx, J. Raymaekers, B. De Ketelaere, P.J. Rousseeuw, M. Hubert, Real-time discriminant analysis in the presence of label and measurement noise, Chemometr. Intell. Lab. Syst. 208 (2021).
[12] H. Wan, Cluster-Based Supervised Classification, Ulster University, UK, 2020. PhD Thesis.
[13] H. Wan, G. Guo, H. Wang, X. Wei, A New Linear Discriminant Analysis Method to Address the Over-reducing Problem, Springer International Publishing, 2015.
[14] K. Chumachenko, A. Losifidis, M. Gabbouj, Robust fast subclass discriminant analysis, in: 28th European Signal Processing Conference (EUSIPCO), 2020.
[15] W. Song, H. Wang, P. Maguire, O. Nibouche, Local Partial Least Square classifier in high dimensionality classification, Neurocomputing 234 (2017) 126–136.
[16] B. Chen, L. Yuan, H. Liu, Z. Bao, Kernel subclass discriminant analysis, Neurocomputing 71 (2007) 455–458.
[17] N. Gkalelis, V. Mezaris, I. Kompatsiaris, T. Stathaki, Mixture subclass discriminant analysis link to restricted Gaussian model and other generalizations, IEEE Transact. Neural Networks Learn. Syst. 24 (1) (Jan. 2013) 8–21.
[18] N. Gkalelis, V. Mezaris, I. Kompatsiaris, Mixture subclass discriminant analysis, IEEE Signal Process. Lett. 18 (5) (May 2011) 319–322.
[19] A. Maronidis, A. Tefas, I. Pitas, Subclass graph embedding and a marginal Fisher analysis paradigm, Pattern Recogn. 48 (2015) 4024–4035.
[20] H. Wan, H. Wang, G. Guo, X. Wei, Separability-oriented subclass discriminant analysis, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2) (Feb. 2018) 409–422.
[21] M. Zhu, A.M. Martinez, Subclass discriminant analysis, IEEE Trans. Pattern Anal. Mach. Intell. 28 (8) (Aug. 2006) 1274–1286.
[22] M. Martinez, M. Zhu, Where are linear feature extraction methods applicable? IEEE Trans. Pattern Anal. Mach. Intell. 27 (12) (Dec. 2005) 1934–1944.
[23] H. Wan, H. Wang, B. Scotney, J. Liu, A novel Gaussian mixture model for classification, in: 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), 2019.
[24] M. Esteki, Z. Shahsavari, J. Simal-Gandara, Use of spectroscopic methods in combination with linear discriminant analysis for authentication of food products, Food Control 91 (2018) 100–112.
[25] S. van Ruth, N. Liu, How organic is organic milk? Can we have a quick check? NIR News 30 (2) (2019) 18–21.
[26] J. Yan, L. van Stuijvenberg, S.M. van Ruth, Handheld near-infrared spectroscopy for distinction of extra virgin olive oil from other olive oil grades substantiated by compositional data, Eur. J. Lipid Sci. Technol. 121 (12) (2019).
[27] J. Riu, G. Gorla, B. Giussani, Miniaturized near-infrared instruments in dairy products or dairy industry: first steps in a long-distance race? NIR News 32 (2021) 17–19.
[28] R.H. Furlanetto, T. Moriwaki, R. Falcioni, M. Pattaro, A. Vollmann, A.C. Sturion Junior, W.C. Antunes, M.R. Nanni, Hyperspectral reflectance imaging to classify lettuce varieties by optimum selected wavelengths and linear discriminant analysis, Remote Sens. Appl.: Society and Environment 20 (2020).
[29] Y. Zhou, S. Yan, Y. Ren, S. Liu, Rolling bearing fault diagnosis using transient-extracting transform and linear discriminant analysis, Measurement 178 (2021).

[30] W. Lin, Q. Gao, M. Du, W. Chen, T. Tong, Multiclass diagnosis of stages of Alzheimer's disease using linear discriminant analysis scoring for multimodal data, Comput. Biol. Med. 134 (2021).
[31] W. Song, H. Wang, P. Maguire, O. Nibouche, Nearest clusters based partial least squares discriminant analysis for the classification of spectral data, Anal. Chim. Acta 1009 (7 June 2018) 27–38.
[32] C.H. Li, B. Kuo, C. Lin, LDA-based clustering algorithm and its application to an unsupervised feature extraction, IEEE Trans. Fuzzy Syst. 19 (1) (2011) 152–163.
[33] K. Chumachenko, J. Raitoharju, A. Iosifidis, M. Gabbouj, Speed-up and multi-view extensions to subclass discriminant analysis, Pattern Recogn. (2021).
[34] S.W. Kim, A pre-clustering technique for optimizing subclass discriminant analysis, Pattern Recogn. Lett. 31 (6) (2010) 462–468.
[35] G. Anouar, F. BaudAT, Generalized discriminant analysis using a kernel approach, Neural Comput. 12 (10) (2000) 2385–2404. Neural Comput.
[36] S.W. Kim, R. Duin, On using a pre-clustering technique to optimize LDA-based classifiers for appearance-based face recognition, in: L. Rueda, D. Mery, J. Kittler (Eds.), Lecture Notes in Computer Science, Progress in Pattern Recognition, Image Analysis and Applications. CIARP, 4756, Springer, Berlin, Heidelberg, 2007, 2007.
[37] Y. Pang, S. Wang and Y. Yuan, "Learning regularized LDA by clustering," IEEE Transact. Neural Networks Learn. Syst., vol. 25, no. 12, pp. 2191-2201, 201.
[38] H. Park, J. Choo, B.L. Drake, K. Jinwoo, Linear discriminant analysis for data with subcluster, in: 19th International Conference on Pattern Recognition, Tampa, 2008.
[39] F. Kosmowski, T. Worku, Evaluation of a miniaturized NIR spectrometer for cultivar identification: the case of barley, chickpea and sorghum in Ethiopia, PLoS One (2018).
[40] S. Chang, Y. Shihong, L. Qi, Clustering characteristics of UCI dataset, in: 39th Chinese Control Conference, Shenyang, 2020.
[41] Integrated Light Source Reflection/Backscatter Probes," Ocean Insight, [Online]. Available: https://www.oceaninsight.com/products/fibers-and-probes/probes/reflectionbackscatter-probes/integrated-light-source-reflectionbackscatter-probes/. [Accessed 23 January. 2023].
[42] Microspectrometers," Ocean Insight, [Online]. Available: https://www.oceaninsight.com/products/spectrometers/microspectrometer/. [Accessed 23 January. 2023].
[43] M.R. Rana, M. Babor, A.A. Sabuz, Traceability of sweeteners in soy yogurt using linear discriminant analysis of physicochemical and sensory parameters, Journal of Agriculture and Food Research 5 (2021).
[44] J. Müller-Maatsch, S.M. van Ruth, Handheld devices for food authentication and their applications: a review, Foods 10 (12) (2021).
[45] N. Liu, H. Parra, A. Pustjens, K. Hettinga, P. Mongondry, S.v. Ruth, Evaluation of portable near-infrared spectroscopy for organic milk authentication, Talanta 184 (2018) 128–135.
[46] K. Zhao, L. Li, Z. Chen, R. Sun, G. Yuan, J. Li, A survey: optimization and applications of evidence fusion algorithm based on Dempster–Shafer theory, Appl. Soft Comput. 124 (2022).
[47] Q.D. Zhan, M. Yang, X. Feng, Sparse representation or collaborative representation: which helps face recognition?, in: IEEE International Conference on Computer Vision (ICCV'11)., Barcelona, Spain, 2011.
[48] E. Candes, J. Romberg, T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information, IEEE Trans. Inf. Theor. 2 (52) (2006) 489–509.
[49] J. Wright, A. Yang, A. Ganesh, S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 2 (31) (2009) 210–227.
[50] I. Rida, S. Al-Maadeed, A. Mahmood, A. Bouridane, S. Bakshi, Palmprint identification using an ensemble of sparse representation, IEEE Access 6 (2018) 3241–3248.
[51] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. Huang, S. Yan, Sparse representation for computer vision and pattern recognition, Proc. IEEE 6 (98) (2010) 1031–1044.
[52] J. Xu, W. An, L. Zhang, D. Zhang, Sparse, collaborative, or nonnegative representation: which helps pattern classification? Pattern Recogn. 88 (2019) 679–688.
[53] S. J. Kim, K. Koh, M. Lustig, S. Boyd and D. Gorinevsky, "An interior-point method for large-scale $\ell$1-regularized least squares," IEEE Journal of Selected Topics in Signal Processing, vol. vol. 1, no. no. 4, pp. pp. 606-617, Dec. 200.
[54] O. Nibouche, J. Jiang and P. Trundle, "Analysis of performance of palmprint matching with enforced sparsity," Digit. Signal Process., vol. Volume 22, no. Issue 2, pp. Pages 348-355, 2012.
[55] UC Irvine Machine Learning Repository," [Online]. Available: https://archive.ics.uci.edu/datasets. [Accessed 15 March. 2024].