National College of Ireland

Food Authentication Using Dimensionality Reduction techniques and Ensemble Algorithms on Spectroscopic Datasets

MSc Research Project

Data Analytics

Tushar Patil

Student ID: x18182020

School of Computing

National College of Ireland

Supervisor: Manaz Kaleel

National College of Ireland

MSc Project Submission Sheet

Student Name:	Tushar Santosh Patil
Student ID:	X18182020
Programme:	Data Analytics **Year:** 2020
Module:	MSc Research Project
Supervisor:	Manaz Kaleel
Submission Due Date:	17/08/2020
Project Title:	Food Authentication Using Dimensionality Reduction techniques and Ensemble Algorithms on Spectroscopic Datasets
Word Count:	5299 **Page Count:** 23

School of Computing

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: .....

Date: 14/08/2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Food Authentication Using Dimensionality Reduction techniques and Ensemble Algorithms on Spectroscopic Datasets

Tushar Patil
x18182020

Abstract

The objective of the studies in food authentication domain is to correctly label the unknown food samples. In this research study three different food authenticity datasets of different types : meat, olive oil and honey are studied. The samples collected using Near-infrared spectroscopy method pose major challenges : the resulting datasets are high dimensional data, i.e. number of predictor variables(p) are much more than the number of observations (n),(n<<p) and the datapoints suffer from inherent collinearity problems. This research study proposes to apply three different dimensionality reduction algorithms to determine the principal components and then feed these embedding spaces to AdaBoost classifiers with DCT and SVM as base estimator. In addition, Random Forests classifier is also applied on the datasets. The aim of this research is to find the optimal combination of the dimensionality reduction algorithms and the classification algorithms that yields optimum level of accuracy. From the results of the study, it is observed that in case of meat data, LDA-AdaBoost Svm approach outperforms other approaches, whereas in case of honey dataset Random Forest classifier outperforms other approaches. In case of olive oil dataset AdaBoostDct with original dataset without any transformation outperforms other approaches.

Keywords *:* Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA), Uniform Manifold Approximation and Projection (UMAP), t-SNE (Stochastic Neighbour Embedding), Adaptive Boosting (AdaBoost), Decision tree (DCT), Support Vector Machines (SVM), Random Forest (RF).

Contents

1 Introduction

Food authentication is the process of verifying the compliance of the food products with their labels. Application of the modern techniques to capture food data, such as near-infrared spectroscopy (NIR) and microscopy (NIRM), Fourier transform (FT) and hyperspectral imaging, results in thousands of spectral data points that are suffering from inherent collinearity problems. Worldwide there are several studies aiming to determine the geographic origin, species, type of food and adulteration type (Song, et al., 2020; Parastar, et al., 2020; Pérez-Rodríguez, et al., 2019; Jiang, et al., 2020; Joswiak, et al., 2019; Kessler, et al., 2015; Bisgin, et al., 2018; Pérez-Rodríguez, et al., 2019; Downey, et al., 2003; Singh & Domijan, 2019). These research studies commonly used Principal Component Analysis (PCA) as dimensionality reduction technique and for classification Support Vector Machines (SVM) is used.

The aim of this research is to use dimensionality reduction techniques coupled with ensemble learning algorithms for the classification of food samples in food authentication studies. The purpose of this study is to find the optimal combination of embedding space and ensemble classifier by trying different combinations that yields optimum accuracy. The study classifies food samples based on their origin, species, type (such as types of wine) and adulteration type. In this study apart from employing the PCA technique that is based on finding the axis or hyperplane of maximum variance, Linear Discriminant Analysis (LDA) that is based on finding the hyperplane of largest class separation, is also used for dimensionality reduction. In addition the novel manifold learning algorithms namely Uniform Manifold Approximation and Projection (UMAP) and t-Stochastic Neighbour Embedding (t-SNE) are used to further explore the domain area of food authentication studies. Hence, this research study aims to achieve optimal accuracy, recall and precision and therefore proposes a model that is efficient. The model is implemented using Scikit-Learn python libraries along with supporting libraries, where in the first phase of the research project, higher dimensional data is reduced to lower embedding space using techniques like PCA, LDA and UMAP. In the second and final phase, boosting algorithm AdaBoost with two different base estimators SVM and decision tree are trained and then tested on embedding

space of datasets - honey, meat and olive oil. In addition, Random Forests classifier is also trained on the three datasets. The performance of all the classifiers is measured using the performance metrics – Accuracy, Auc-roc score, precision, recall and f1- score. The results obtained from this study are also compared_to the study using similar datasets. From the results it is observed that in case of meat dataset LDA-AdaBoost SVM outperforms the PLS approach taken in the study (Singh & Domijan, 2019) using similar dataset .

This paper is structured as below:
Section 2: It summarizes the literature review that describes dimension reduction techniques and classification methods that are common in food authentication studies.
Section 3: This section describes system architecture, methodology and the data used in this research.
Section 4: This section thoroughly discusses and analyses the results obtained by the proposed model. It compares the results of this study with another similar study using same datasets.
Section 5: It concludes the research project and discusses the future work.
Section 6: Expresses gratitude.

2 Literature Review

In this section the research studies are reviewed and summarized in matrix form mentioned below. All these studies are based on food authentication and most of them use PCA and SVM, to generate the embedding space and for classification purpose respectively. Motivated by this literature survey, the research study proposes to use SVM as a base estimator in Adaboost ensemble, instead of using it as a direct classifier and also explore other dimensionality reduction techniques apart from PCA.

Work name	Year	Algorithm used	Dataset	Dataset split	Performance evaluation metrics	Performance achieved
Organic apple authentication based on diffraction grating and image processing **(Song, et al., 2020)**.	2020	SVM, Locally weighted partial least squares classifier (LW-PLS), Logistic Regression and k-NN. PCA is used for dimensionality reduction.	Dataset comprises of images of 150 organic and conventional apples of 3 different varieties – Pink lady, Gala and Braeburn	Training and test set are split into the ratio of 2:1	Accuracy.	classification accuracies of 93-100%.
Use of smartphone videos and pattern recognition for Authentication of milk and olive	2020	PLS-DA , LW-PLSC and PCA for dimensionatlity reduction.	A total of 160 unadulterated and adulterated olive oil samples are used.	Leave-one-out-cross-validation method is used.	Accuracy.	96.2% for Olive oil and 100% for milk.

oil **(Jiang, et al., 2020)**.			The milk dataset has a total of 138 data samples			
Comparing SVM and ANN models for Species Identification of Food Contaminating Beetles. **(Bisgin, et al., 2018)**	2018	SVM with radial basis function kernel and ANN.	Dataset consists of 6900 images of 15 species of bettle.	Dataset split into 80:20 for training and testing respectively.	Accuracy, Sensitivity, Precision, Specificity and Matthew's correlation coefficient (MCC).	**Sensitivity** - Almost near perfect . **Precision** - Excellent for some species and low for others. **Specificity**- Excellent. **MCC** -not very high. **Accuracy** – 87%.
Classify organic and conventional wheat using the MeltDB 2.0 metabolomics analysis platform **(Kessler, et al., 2015)**.	2015	SVM, t-SNE and PCA for dimensionality reduction.	Dataset comprises od total 313 wheat samples of 11 different cultivars produced for 3 years.	The algorithm is trained and tested on 80% of the data. The remaining 20% out-of-the-bag data is used for validation.	Accuracy, Sensitivity and Specificity.	Accuracy- 90.32%, Sensitivity- 90.32%, and Specificity- 90.32% (when trained and tested on 3 years data).
Authentication of the geographical origin and the botanical variety of avocados using liquid chromatography fingerprinting and deep learning methods. **(Martín-Torres, et al., 2020)**	2020	PLD-DA and SVM. PCA.	108 avocados samples from different origin, varieties and ripeness characteristis.	30% of the samples from each class for external validation set while the remaining samples constitutes the training set.	Sensitivity, specificity, precision, accuracy, AUC, Mathews correlation coefficient and Kappa coefficient.	SVM performed better than PLS-DA. Sensitivity – 1.00, Specificity – 0.80, Precision – 0.87, accuracy – 0.91, AUC- .90, and Kappa coefficient – 0.82

Brown rice Authenticity evaluation by spark discharge laser-induced breakdown spectroscopy **(Pérez-Rodríguez, et al., 2019)**.	2019	Four different classification methods LDA, random forests (RF), SVM and k-NN are used. XGBoost for feature selection based on the feature importance matrix.	Dataset comprises of total 66 spectral samples of brown rice collected from 2 different regions Corrientes and Mercedes (both in Argentina).	Dataset split into the ratio of 70:30 for training and testing set respectively.	Accuracy, Sensitivity and Specificity.	k-NN outperformed other classifiers with accuracy of 84%, sensitivity of 100% and specificity of 78%.
Integration of handheld NIR and machine learning to "Measure & Monitor"chicken meat authenticity **(Parastar, et al., 2020)**.	2020	Random subspace discriminant ensemble (RSDE).	Dataset comprises of total 153 fresh chicken samples and total 133 thawed samples.	Dataset split into the ratio of 70:30 for training and testing respectively.	Accuracy, Precision, Sensitivity, Specificity and Error rate.	(RSDE) method significantly outperformed other common classification methods such PLS-DA, ANN and SVM with classification accuracy of >95%, precision >96% and sensitivity >95%.
Comparison Of Machine Learning Models in Food Authentication Studies **(Singh & Domijan, 2019)**.	2019	k-NN, DCT, Logit Boost, SVM,PLS, Bayesian Kernel Projection Classifier. DR techniques – Marginal Relevance, Genetical Algorithm, PCA and FPCA.	Dataset comprises of total 153 fresh chicken samples and total 133 thawed samples.	Dataset split into the ratio of 50:50 for training and testing respectively.	Accuracy and Standard deviation.	PLS yields classification accuracy of 94% on meat dataset, 90% accuracy on Olive Oil dataset and 95% on Honey dataset.

3 Methodology

3.1 Data Description

This research study uses three datasets comprising of the different types of food samples - meat (McElhinney et al., 1999), olive oil (Downey et al., 2003) and honey (Fouratier et al., 2003) and for complete sample collection process these papers can be referred. All the three datasets are high dimensional i.e. p>>n, as seen from the below Table 1 and sample measurements are taken using near-infrared spectroscopy method.

Table 1 – Data Description

Dataset	Features (p)	Samples (n)	Labels	Labelled by
Olive Oil	1051	65	Crete, Peloponese and Other.	origin
Meat	1051	231	Chicken, Lamb, Pork, Beef, Turkey.	species
Honey	1051	478	Bi, Hfs, fg (adulterated) and pure.	adulteration

Bi - beet invert syrup , Hfs - high fructose corn syrup and fg - fructose-glucose.

3.2 Data Splitting and tuning

All the three raw datasets -meat, honey and olive oil, are randomly split into the training set (70 %), validation set (15%) and testing set (15%). The high dimensional - 70% training set, 15% validation set, and 15% testing set is then transformed to low embedding space using dimensionality techniques such as PCA, LDA and UMAP. This transformed 15% of the validation set is then used to fine tune the classifiers – AdaBoost and Random Forests. Furthermore, the transformed training set is then fed to the classifiers AdaBoost - DCT, AdaBoost – SVM and Random Forests for training purpose. Finally, the transformed 15% testing set is used to perform the predictions to calculate the performance of the classifiers.

3.3 Data Transformation and Pre-processing

This study applies dimensionality reduction techniques such as LDA, PCA, t-SNE and UMAP in the pre-processing step. Figure 1, Figure 2 and Figure 3 shows the visualization of the three data sets -meat, honey and olive oil comprising of 1051 variables using dimensionality reduction techniques stated above. As shown in all the three figures components of PCA, UMAP and t-SNE contain some part of the related information but they are not able to clearly distinguish the classes. On the other hand LDA clearly separates each of the class. In Figure 1 (UMAP vs t-SNE part and PCA) local structures of data or local clusters are preserved well for each of the separate class, but LDA performs far better than other techniques. In case of meat dataset, Chicken and Turkey both belong to the same class as they have same taxonomy. In this case UMAP and t-SNE algorithms has done better job in clustering the local structures and cluster

the similar classes together, but LDA has done far better job by clearly identifying and discriminating these groups of similar classes from each other. As seen in the Figure 2 (honey) and Figure 3 (olive oil) – LDA has outperformed other techniques and has clearly separated the classes – Crete, Peloponnese and Other.

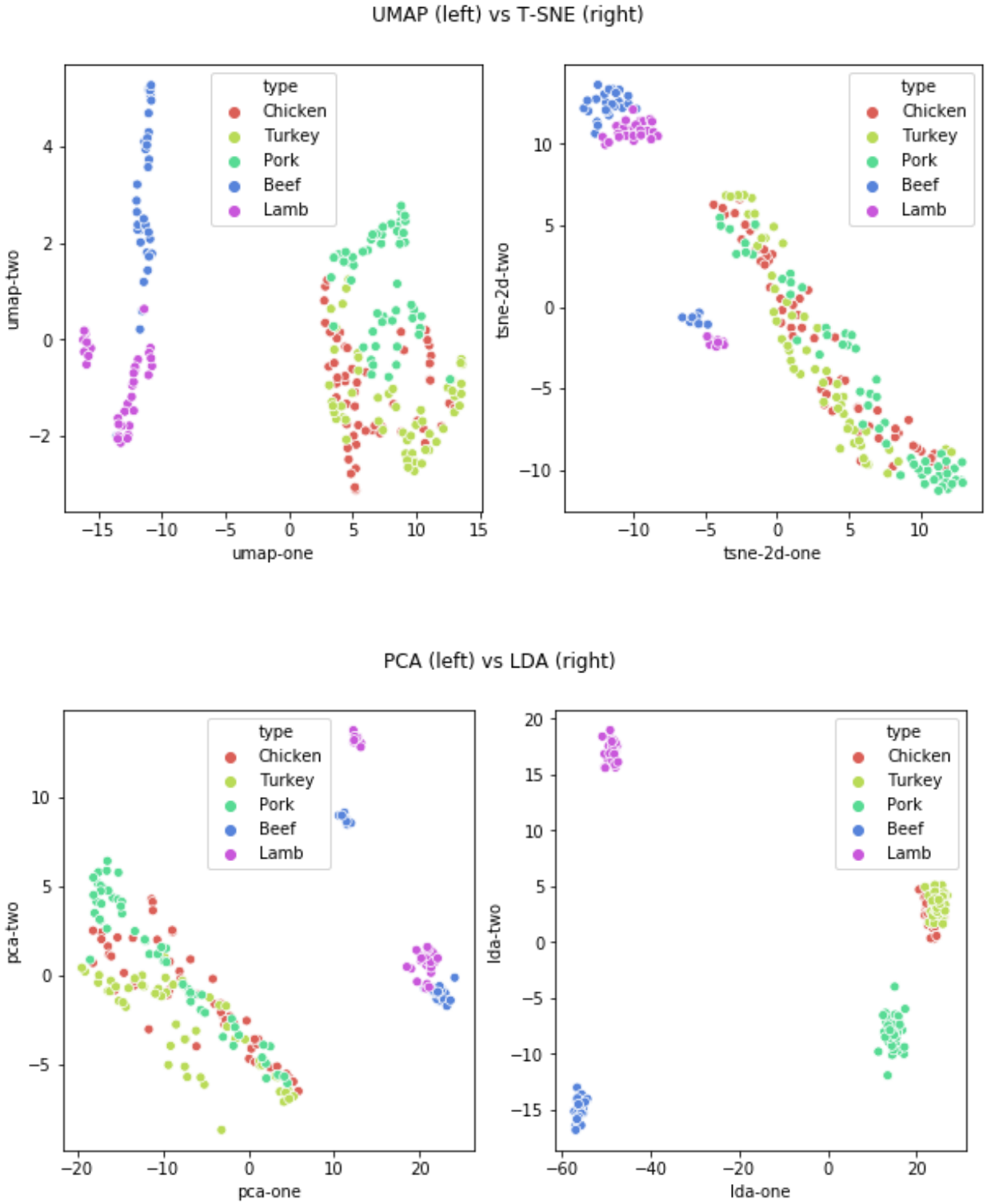

Figure 1 – UMAP vs t-SNE and PCA vs LDA on Meat dataset

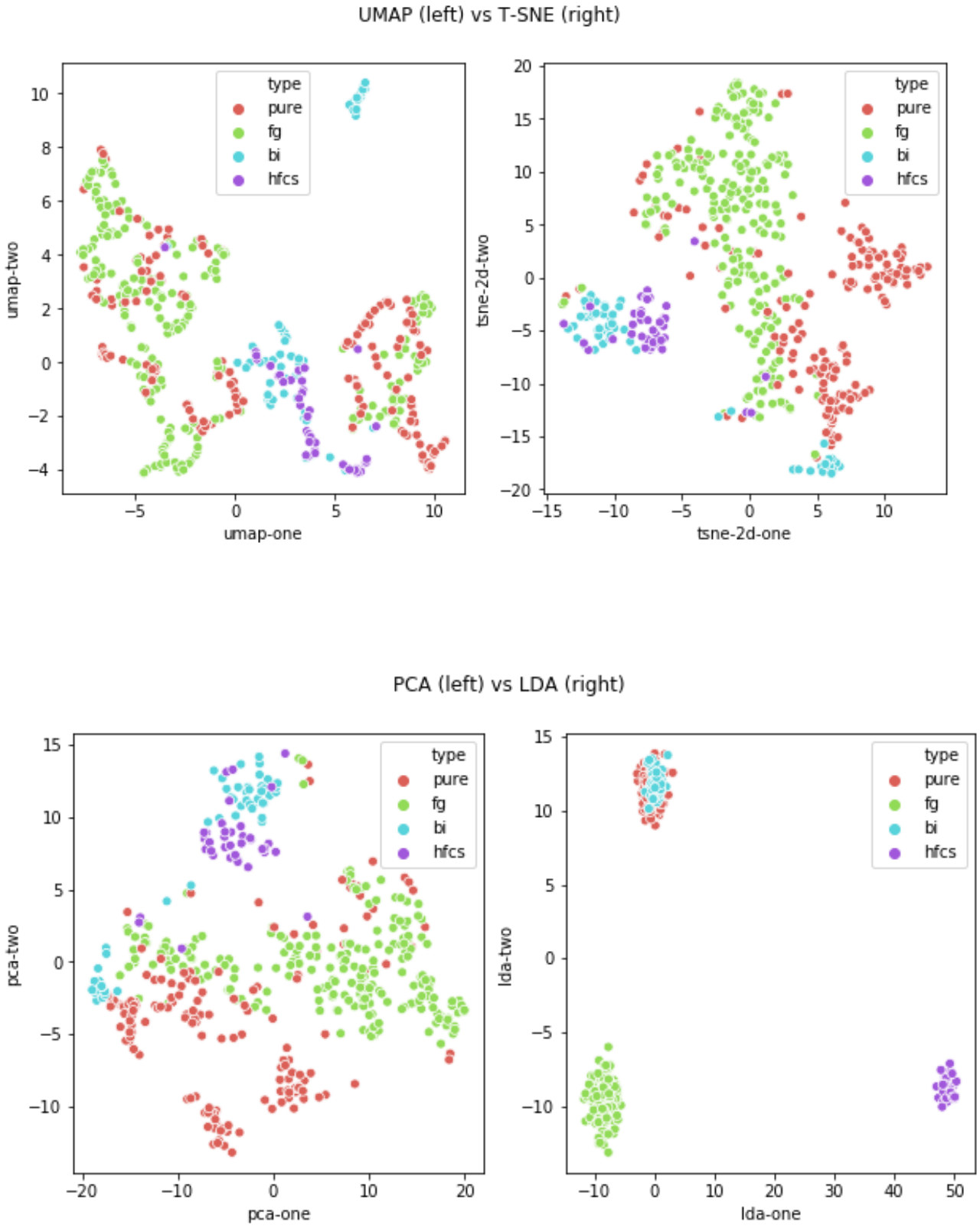

Figure 2- UMAP vs t-SNE and PCA vs LDA on Honey dataset

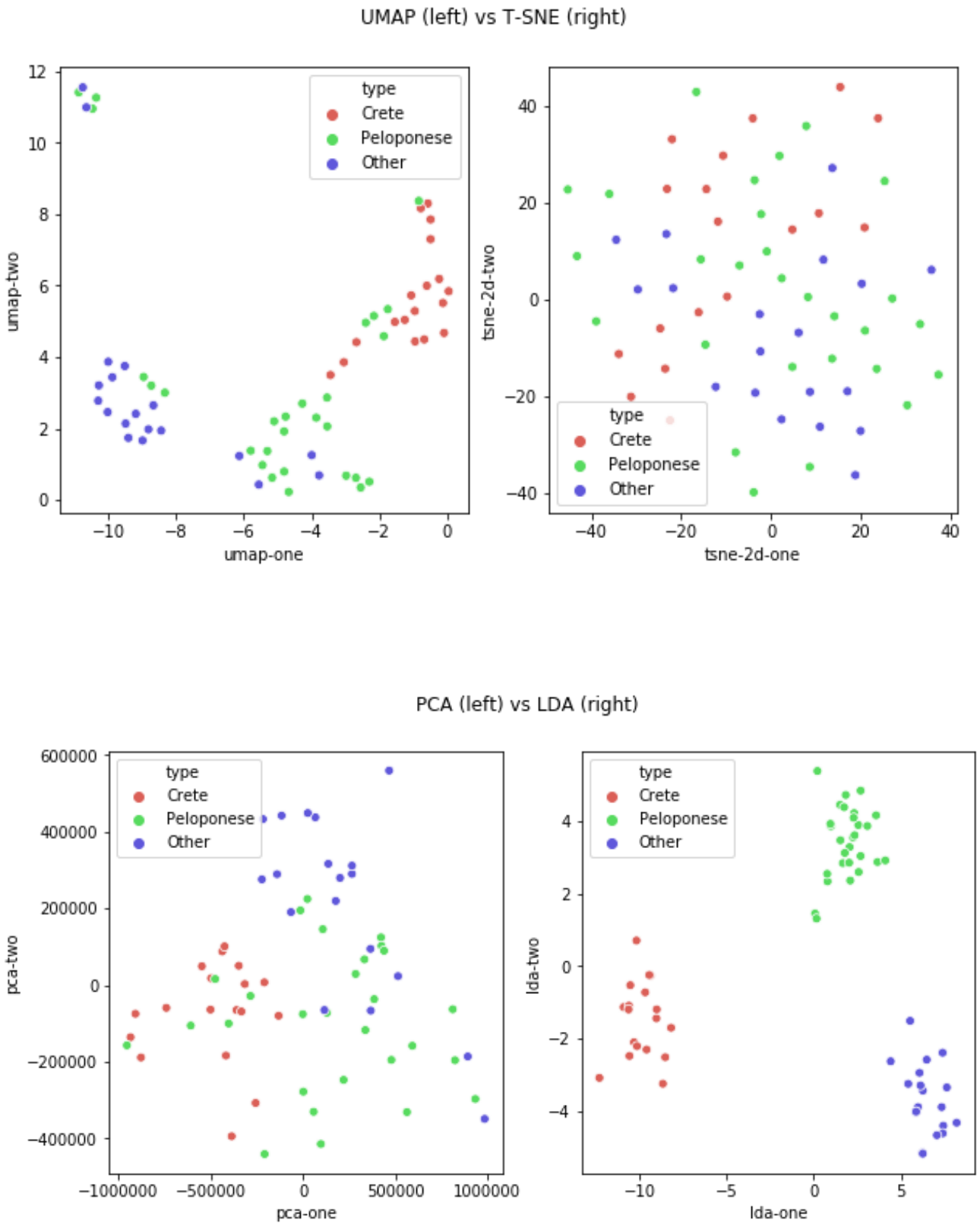

Figure 3 - UMAP vs t-SNE and PCA vs LDA on Olive oil dataset

3.4 System Architecture Design

3.4.1 Dimensionality Reduction

There are two main approaches to the dimensionality reduction algorithm : projection based and the one with manifold learning. This study uses both the approaches to all the three datasets and applies LDA, PCA, UMAP dimensionality reduction techniques on the meat, honey and olive oil datasets to generate lower embedding space from high dimensional input data. However, it is important to note that, the t- SNE is used for visualization purpose only as its output cannot be fed to any classifier.

3.4.1.1 Linear Discriminant Analysis:

LDA is supervised and it finds the hyperplane of maximum class separation as opposite to PCA. LDA identifies the axis that separate the classes with the largest margin of separation.

3.4.1.2 Principal Component Analysis:

PCA is the most popular and extensively used dimensionality reduction algorithm. PCA finds the axis that accounts for the largest amount of variance in the training set. It also finds a second axis orthogonal to the first one, that accounts for the largest amount of remaining variance. In case of high dimensional data, PCA finds a third axis, orthogonal to both the previous axes, and a fourth, and a fifth and so on – as many as the number of dimensions in the dataset. (Geron, 2019).

3.4.1.3 Uniform Manifold Approximation and Projection :

UMAP works on the notion that local distances are of more importance than the global distances. UMAP, a novel manifold learning technique introduced by Becht, et al.,(2018), is theoretically based on the notion of topological data analysis. UMAP algorithm uses Riemannian geometry in order to bridge the gap between the assumptions of topological data analysis and real word data. UMAP is similar to t-SNE except that it uses cross-entropy loss function unlike of Kullback-Leibler divergence used by t-SNE, which enables UMAP to retain the global structure. UMAP works in 2 stages, in the first stage it constructs weighted k-neighbour graph while in the second stage low dimensional representation of this graph is constructed.

3.4.1.4 t- Stochastic Neighbour Embedding (t-SNE) :

t-SNE was introduced by Hinton & Maaten, (2008) to visualize high dimensionality data though the embedding generated it cannot be used as input to the classifier. t- SNE operates in two stages, in first stage it generates probability distribution of high dimensional data and in second stage it generates similar probability distribution of lower dimensional representation of the datapoints. It then minimizes the distance between the two-probability distribution using KL divergence. t-SNE overcomes the shortcoming of crowding problem suffered by other manifold learning methods such as Sammon Mapping and Stochastic Neighbour Embedding.

3.4.2 Classification Algorithms

In case if complex question is posed to thousands of random people and then their answers are aggregated, it is observed that aggregated answer is better than the individual specialist's answer. Similarly, it is observed that the collection of the predictors produces the predictions that are much better than the prediction of the individual predictor. A cluster of predictors is known as ensemble and this method or technique is known as Ensemble learning algorithm (Geron, 2019). This study presents the methodology that uses ensemble method called as Boosting and Bagging for classification task. ***The five classifiers presented in this project are***

- AdaBoost Dct (with no embedding),
- Umap - AdaBoost Dct (with umap embedding as input),
- PCA – AdaBoost Dct (with pca embedding as input),
- LDA - AdaBoost Svm (with lda embedding as input) and
- Random Forests.

3.4.2.1 Boosting

Boosting is an ensemble learning algorithm that combines several weak learners into the strong learner. The general notion on which the boosting method is based, is to sequentially train the predictors wherein each predictor is trying to correct its predecessor. This study uses the most popular boosting method - AdaBoost. One way for a new predictor to correct its predecessor is to pay a bit more attention to the training instance that the predecessor underfitted. This is the technique used by AdaBoost. In AdaBoost classifier, first the base classifier namely Dct or SVM is trained and then using this trained object the predictions are made on training set. The relative weights of the misclassified training instances are updated and increased. Subsequently, the algorithm further trains the second predictor on these updated weights to make predictions on the training set and update the instance weights and so on (Geron, 2019). Once all the predictors are trained, ensemble makes the final prediction.

In this study the low dimensional representation of the meat dataset, generated during the pre-processing stage using the techniques such as UMAP, and PCA, is fed to the classifiers AdaBoost DCT and the embedding space generated using LDA is fed to the classifier AdaBoost SVM. Similar approach is taken for the honey and oil olive datasets as shown the Figure 4 and Figure 5 .

3.4.2.1.1 ADABOOST based on DCT

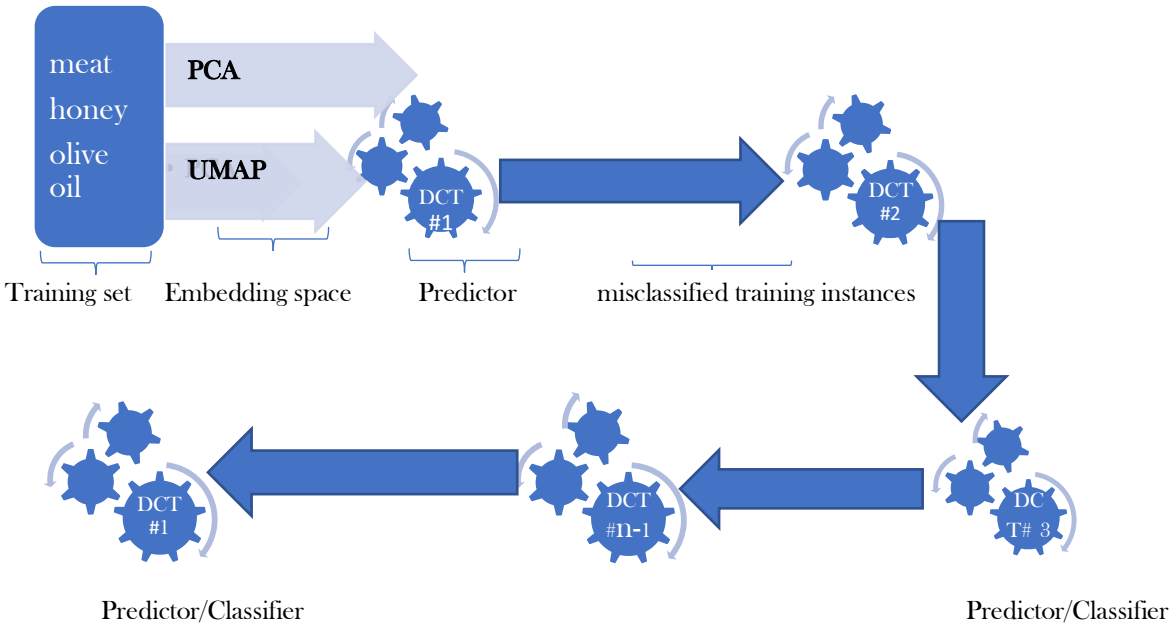

Figure 4 – AdaBoost Sequential Training with DCT as base estimator on umap & pca embedding.

3.4.2.1.2 ADABOOST based on SVM

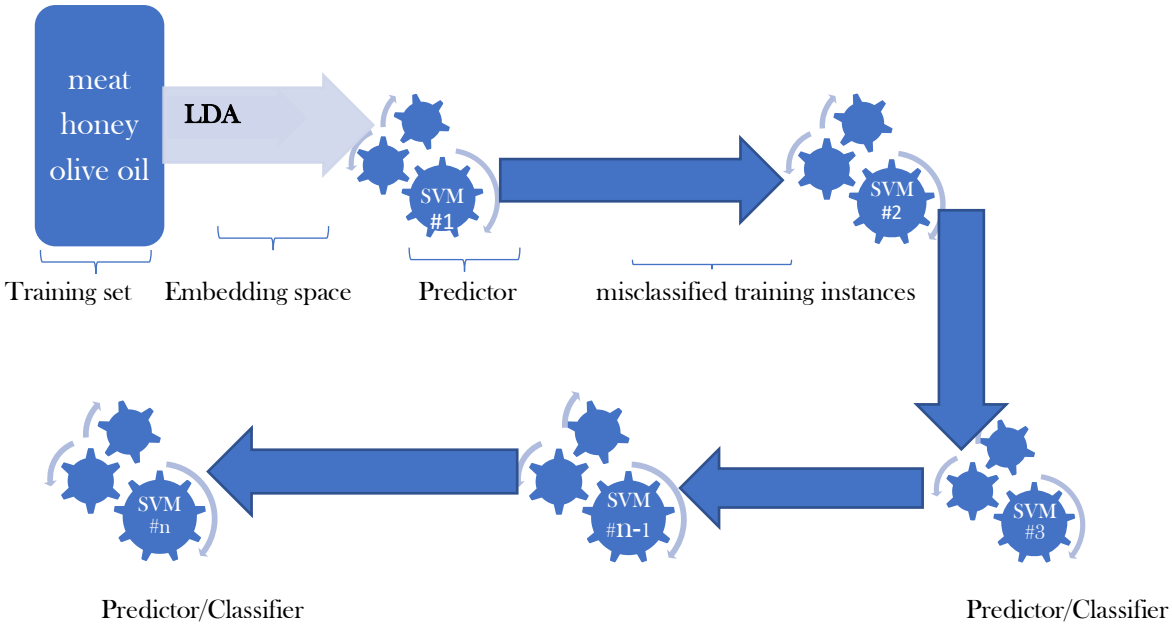

Figure 5 – AdaBoost Sequential Training with SVM as base estimator on lda embedding.

3.4.2.2 Bagging

In case of Bagging technique, every predictor is trained using the same algorithm and different random subsets of the training data is used to train each predictor. Finally, the individual predictors results are aggregated to produce the final prediction. *There is no explicit dimensionality reduction method used in case of Random Forests classifier as the algorithm itself has the quality to rank the features by their relative importance.*

3.4.2.2.1 Random Forests

Random forests are an ensemble of decision trees trained using bagging method. Once all the decision trees (no of trees can be controlled using parameter) are trained, then a prediction for a new instance can be made by an ensemble just by simple aggregation of the predictions of the decision trees. Another important quality of Random forests is that it can measure the relative importance of each feature and hence can be used as dimensionality reduction technique. Scikit-Learn measure a feature's importance by looking the ability of the tree node to reduce impurity on average that use that feature (Geron, 2019).

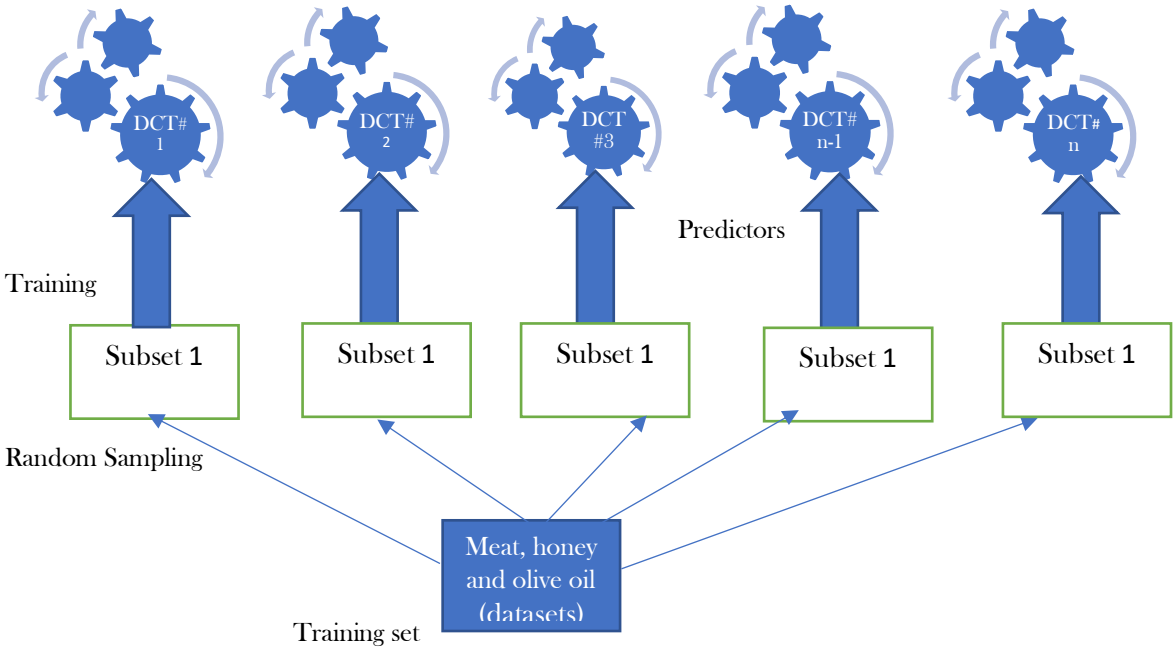

Figure 6 – Random Forests involves training several Decision tree predictors on different random samples of the training set. Random Forests also identifies important features.

3.5 System Architecture Implementation

The implementation of the architecture is done using the Jupyter using python 3 programming language. The architecture of this research project consists of 6 modules developed using Scikit-learn (Pedregosa, et al., 2011), Pandas (McKinney & others, 2010), Matplotlib (Hunter, 2007), UMAP and supporting libraries, each performing the distinct tasks. All the functions in the

modules are generic and applies to all the three datasets – meat, honey and olive oil. These modules should work and produce the results even if another spectroscopic dataset is used, with very little change ONLY in the implementation.py file. Below is the description of each of the module :

3.5.1 datasetsSplitTrainTestVal.py

- This module takes any dataset (of the 3 datasets used) as input and generates train, test and validation split into the ratio of 70:15:15.
- The important function in this module is get_train_val_test_set() that returns the separate data portions mentioned above.

3.5.2 createEmbeddings.py

- The three important functions in this module are get_umap_embedding(), get_pca_embedding() and get_lda_embeddings() that returns the embedding space for meat, honey and olive oil datasets.
- This module is responsible to take split dataset as input (output of *datasetsSplitTrainTestVal.py*) and transform it in the low dimensional embeddings using the functions stated above.

3.5.3 options.py

- It loads the hyperparameters from the *hyperparam*.csv file for the classifiers - AdaBoost and Random Forests .

3.5.4 trainClassifiers.py

This module contains 5 important functions- get_trained_adaboost_dct(), get_trained_adaboost_dct_umap(),get_trained_adaboost_dct_pca(),get_trained_adaboost_svm_lda () and get_trained_random_forest () that returns the trained classifiers for all the three datasets.

- The module gets the hyperparameters from *options.py* and does the fine tuning of the hyperparameters using GridSearchCV method on validation set (15%) portion of the input dataset. The two more important functions of this module that returns the best estimators are - get_best_adaboost_estimator() and get_best_randomfor_estimator().
- It then retrieves the embeddings generated by the module *createEmbeddings*.py by calling the functions - get_umap_embedding(), get_pca_embedding() and get_lda_embedding().
- Finally, after fine tuning and getting the embeddings, it trains the best estimators on the transformed training set (70% of the dataset) using the functions stated above.

3.5.5 datasetsVisualizations.py

- The two important functions in this module are prepare_data_for_visualizations() and show_umap_tsne_lda_pca_components() .
- It is responsible to visualize the clusters for all the three datasets using Umap, t-SNE, LDA and PCA techniques. These visualizations are shown in section 3.4.

3.5.6 Implementation.py

This is the key module of the project and is responsible to perform the prediction on new instances and serialize the model objects. All this operations/processing is done on three datasets used in this research project.

- It calls function show_umap_tsne_lda_pca_components() mentioned in section 3.2.5 to plot the extracted principal components.
- The module fetches the trained classifier objects by calling the five get_trained_....() functions mentioned in section 3.2.4 on each of the datasets.
- For each of the datasets, the module then iterates through each of the classifiers and perform the predictions on unseen testing sets (15%).
- It creates the data frame to store the accuracy score, auc-roc score and then plot the accuracy score bar graph and auc-roc bar graph. It also prints the classification report for each classifier.
- Finally, it saves the model objects.

3.6 Model Evaluation:

The performance of all the five classifiers, AdaBoost Dct, UMAP – AdaBoost Dct, LDA – AdaBoost SVM, PCA – AdaBoost Dct and Random forests, is measured using the performance metrics such as *Accuracy score*, *Auc - Roc score* and the classification report detailing the *Precision, Recall and F1-score* by label for meat, honey and olive oil datasets.

4 Results

The results obtained from the five classifiers of this research project can be seen in the Table 2. LDA – AdaBoost SVM classifier has outperformed the other classifiers with accuracy score of 97% and Auc- roc score of 98% on meat dataset, followed by AdaBoost-Dct and UMAP AdaBoost-Dct with accuracy scores of 83%. In case of olive oil dataset AdaBoost-Dct outperformed other classifiers with accuracy score of 90%. The olive oil dataset being the smallest of the 3 datasets (65 samples) used in this project, applying the dimensionality reduction techniques has resulted in the significant loss of information, that is the reason all the other model using techniques such as umap, pca and lda has lower accuracy scores. In case of honey dataset, Random forest scores the highest accuracy and auc-roc score of 90%. **Furthermore, the result obtained from this research project is also compared with another study using similar datasets**

but using different approaches. It is observed that the LDA -AdaBoost SVM model presented in this research project has outperformed the PLS approach used by another study (Singh & Domijan, 2019). In case of meat dataset LDA – AdaBoost SVM model has accuracy score of 97% whereas the PLS approach has the accuracy score of 94%, so there is an improvement in the accuracy by 3%. In case of Olive oil dataset, accuracy score of the both the approaches is same of 90%. In case of honey dataset accuracy score of PLS approach is 95% whereas LDA – AdaBoost SVM approach yields 90% accuracy score.

Table 2 – Accuracy score and Auc – roc score for meat, olive oil and honey datasets.

	Meat		Olive Oil		Honey	
Model	Accuracy	Auc-roc	Accuracy	Auc-roc	Accuracy	Auc-roc
1. AdaBoost - Dct	83	90	90	88	83	86
2. UMAP - AdaBoost-Dct	83	90	60	70	67	73
3. PCA – AdaBoost-Dct	74	85	80	87	65	73
4. LDA – AdaBoost -SVM	97	98	80	82	83	88
5. Random Forest	80	88	80	83	90	91
Comparison with another study using similar dataset (Singh & Domijan, 2019)						
PLS approach	94	NA	90	NA	95	NA

Similar results described above and as seen in the table 2 are shown graphically in the form of bar graph plot in Figure 7 and Figure 8 for better comparison of the accuracy score and auc-roc score results.

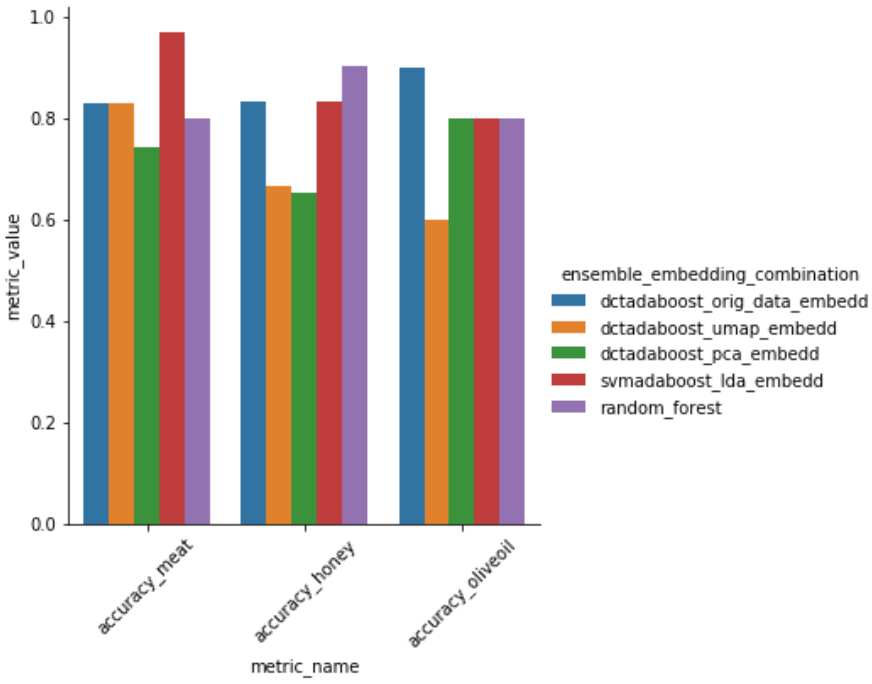

Figure 7 – Accuracy Score bar graph for meat, honey and olive oil datasets by combination of DR technique and the ensemble model.

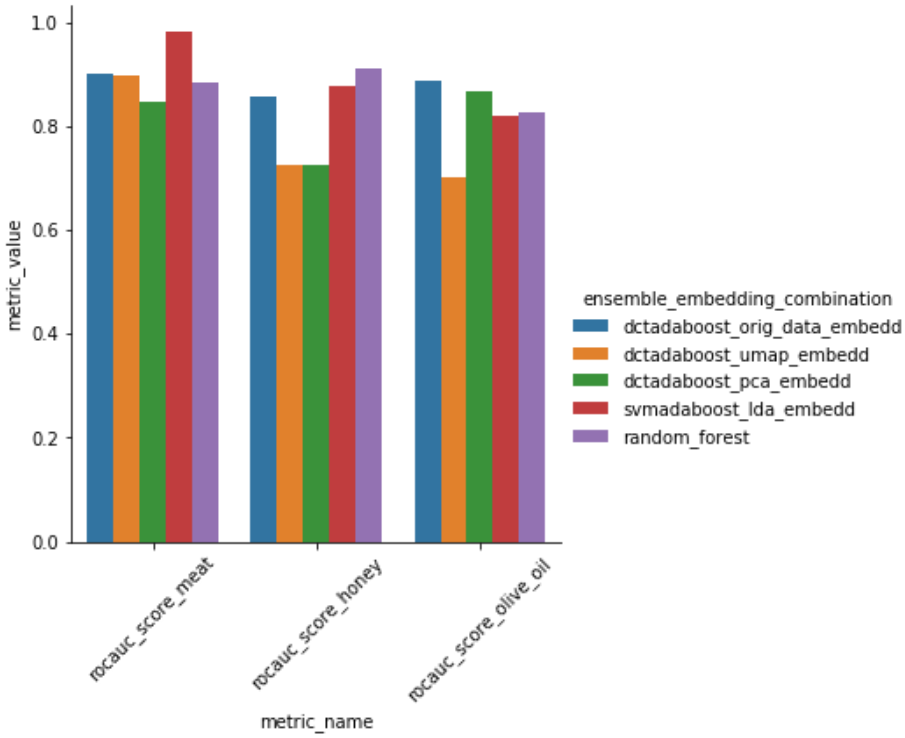

Figure 8 – ROC-AUC Score bar graph for meat, honey and olive oil datasets by combination of DR technique and the ensemble model.

The Table 3 shows the detailed classification report of meat dataset with other performance metrices such as precision, recall and f1-score. The report shows the precision, recall and f1-scores by labels for each of the five classifiers and these scores are highest for the LDA – AdaBoostSVM classifier for all the labels in case of meat dataset. Similar classification reports can be seen in the Table 4 for honey dataset and in Table 5 for Olive oil dataset. In case of honey dataset, the scores are highest for the Random Forest classifier for all the labels whereas in case of the olive oil dataset, AdaBoost-Dct model has the highest scores and it is observed the dimensionality reduction techniques are not very effective in this case.

Table 3 - Classification report of meat dataset by labels

Classifier	Labels	Precision	Recall	F1-score
AdaBoost - Dct	Chicken	1.00	1.00	1.00
	Turkey	1.00	0.38	0.55
	Pork	1.00	1.00	1.00
	Beef	0.62	1.00	0.76
	Lamb	0.86	0.86	0.86
UMAP - AdaBoost-Dct	Chicken	1.00	1.00	1.00
	Turkey	0.75	0.75	0.75
	Pork	1.00	0.83	0.91
	Beef	0.86	0.75	0.80
	Lamb	0.62	0.86	0.75
PCA – AdaBoost-Dct	Chicken	0.71	0.83	0.77

	Turkey	0.71	0.62	0.67
	Pork	0.86	1.00	0.92
	Beef	0.62	0.62	0.62
	Lamb	0.83	0.71	0.77
LDA – AdaBoost - SVM	Chicken	1.00	1.00	1.00
	Turkey	1.00	0.88	0.93
	Pork	1.00	1.00	1.00
	Beef	1.00	1.00	1.00
	Lamb	0.88	1.00	0.93
Random Forest	Chicken	1.00	1.00	1.00
	Turkey	0.80	0.50	0.62
	Pork	1.00	1.00	1.00
	Beef	0.78	0.88	0.82
	Lamb	0.56	0.71	0.63

Table 4 - Classification report of honey dataset by labels

Classifier	Labels	Precision	Recall	F1-score
AdaBoost - Dct	Pure	0.71	0.62	0.67
	Fg	0.89	0.86	0.87
	Bi	0.62	0.71	0.67
	Hfscs	0.86	0.90	0.88
UMAP - AdaBoost-Dct	Pure	0.62	0.62	0.62
	Fg	0.72	0.72	0.72
	Bi	0.50	0.29	0.36
	Hfscs	0.62	0.71	0.67
PCA – AdaBoost-Dct	Pure	0.56	0.62	0.59
	Fg	0.71	0.75	0.73
	Bi	0.75	0.43	0.55
	Hfscs	0.57	0.57	0.57
LDA – AdaBoost - SVM	Pure	0.42	0.62	0.50
	Fg	1.00	0.94	0.97
	Bi	0.88	1.00	0.93
	Hfscs	0.78	0.67	0.72
Random forest	Pure	0.78	0.88	0.82
	Fg	0.92	0.94	0.93
	Bi	0.71	0.71	0.71
	Hfscs	1.00	0.90	0.95

Table 5 - Classification report for olive oil dataset by labels

Classifier	Labels	Precision	Recall	F1-score
AdaBoost - Dct	Crete	1.00	1.00	1.00
	Peloponese	0.80	1.00	0.89
	Other	1.00	0.50	0.67

UMAP - AdaBoost-Dct	Crete	1.00	0.75	0.86
	Peloponese	0.67	0.50	0.57
	Other	0.25	0.50	0.33
PCA – AdaBoost-Dct	Crete	0.80	1.00	0.89
	Peloponese	1.00	0.50	0.67
	Other	0.67	1.00	0.80
LDA – AdaBoost - SVM	Crete	0.80	1.00	0.89
	Peloponese	0.75	0.75	0.75
	Other	1.00	0.50	0.67
Random forest	Crete	1.00	0.75	0.86
	Peloponese	0.80	1.00	0.89
	Other	0.50	0.50	0.50

5 Conclusion and Future Work

This research project applies not only the different approaches of dimensionality reduction techniques such as linear methods (Feature projection) and Manifold learning but also used different approaches of ensemble models such as Boosting and Bagging. This study is successful in finding the best optimal combination of the ensemble model and dimensionality reduction technique – LDA AdaBoost-Svm with accuracy score of 97% in case of meat dataset that outperformed not only other ensembles used in this study but also PLS approach of another study mentioned in the results section. In case of olive oil dataset, none of the dimensionality reduction techniques performed better. As the olive oil dataset in extremely small and further applying the reduction techniques on it resulted in the signification loss of information. Instead applying the Adaboost-Dct model outperformed other ensemble models that are employing reduction techniques. In case of honey dataset, Random Forest classifier outperforms other approaches, while it is important to note that the approach using LDA technique is second best here and also in olive oil dataset. Furthermore, key aspect of the framework developed in this study is that it is generic and can be used with any high dimensional spectroscopic (NIR) datasets of size - small and medium with very little modifications. In future, the Deep Neural Network approach along with Stacked Auto-Encoder can be experimented to further improve the accuracy. Also, the behaviour of the existing architecture of this research project on larger and very large datasets will be interesting thing to explore and experiment.

6 Acknowledgments

I would like to express my sincere gratitude to my project supervisor Manaz Kaleel for his continuous and relentless support throughout the duration of the project work. He always welcomed open discussions and was generous enough to allow us to ask any trivial questions right

from the first lecture itself. Such kind of support was truly inspiring and motivating. Lastly, I would like to thank my friends and family for their motivation and support.

References

Jiang, N., Song, W., Wang, H. & Vincent, J., 2020. Use of smartphone videos and pattern recognition for food authentication. *Sensors and Actuators B: Chemical,* Volume 304.

Kessler, N. et al., 2015. Learning to Classify Organic and Conventional Wheat â€“ A Machine Learning Driven Approach Using the MeltDB 2.0 Metabolomics Analysis Platform. *Frontiers in Bioengineering and Biotechnology,* Volume 3.

Bisgin, H. et al., 2018. Comparing SVM and ANN based Machine Learning Methods for Species Identification of Food Contaminating Beetles. *Scientific Reports,* 8(1).

Parastar, H. et al., 2020. Integration of handheld NIR and machine learning to "Measure & Monitor" chicken meat authenticity. *Food Control,* Volume 112.

Pérez-Rodríguez, M. et al., 2019. Brown rice authenticity evaluation by spark discharge-laser-induced breakdown spectroscopy. *Food Chemistry,* Volume 297.

Singh, M. & Domijan, K., 2019. Comparison of Machine Learning Models in Food Authentication Studies. *30th Irish Signals and Systems Conference (ISSC).*

Song, W., Jiang, N., Wang, H. & Wang, G., 2020. Evaluation of machine learning methods for organic apple authentication based on diffraction grating and image processing. *Journal of Food Composition and Analysis,* Volume 88.

Hinton, G. E. & Maaten, L., 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research,* pp. 1-48.

Becht, E. et al., 2018. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology,* 37(1), pp. 38-44.

Geron, A., 2019. *Hands-On Machine Learning with Scikit-Learn, Keras and Tensorflow.* 2nd ed. s.l.:O'Reilly.

Hunter, J. D., 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering,* Volume 9, pp. 90--95.

McKinney, W. & others, 2010. Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference,* Volume 445, pp. 51--56.

Downey, G., McIntyre, P. & Davies, A., 2003. Geographic classification of extra virgin olive oils from the eastern mediterranean by chemometric analysis of visible and near-infrared spectroscopic data. *Applied spectroscopy,* Volume 57, pp. 158-163.

Fouratier, V., Kelly, J. & Downey, G., 2003. Detection of honey adulteration by addition of fructose and glucose using near infrared transflectance spectroscopy. *Journal of Near Infrared Spectroscopy,* Volume 11, pp. 447-456.

McElhinney, J., Downey, G. & Fearn, T., 1999. Chemometric processing of visible and near infrared reflectance spectra for species identification in selected raw homogenised meats. *Journal of Near Infrared Spectroscopy,* Volume 7, pp. 145-154.