SEVENTH FRAMEWORK PROGRAMME
FP7-ICT-2011-C

Grant agreement for: Collaborative project

Deliverable D2.3

Documentation and tool development for global classifier

Project acronym: PLEASED
Project full title: " PLants Employed As SEnsor Devices "
Grant agreement no: 296582

Responsible Partner: University of Southampton
Version date: 12-02-2015

Contents

1 Introduction

The goal of the PLEASED project is to lay the foundation of a system that can use electrophysiological signals from plants to identify the environmental constituents they are in. Toward this goal, the PLEASED approach is first to explore the possibility of classifying external stimuli using the electrophysiological signal from a single plant species and then combine the response of different plant species to more accurately identify the external stimuli. The latter part stems from the fact that not all the species may have same sensitivity to an arbitrary external stimulus and the may show individualized response to a stimulus. The idea of combining the responses from multiple plat species is to exploit this individualized response to identify the external stimulus more accurately. This document presents an overview of the findings as a result of the analysis carried out on four plant species (cabbage, rosemary, sage, mint) to determine their discriminatory abilities when two different types of stimuli (acid and salt) are applied to them. The experimental protocol, signal categories, analysis methods adopted and the results obtained are listed in the following sections.
The primary motivation was to determine the behavior of a group of plants (cabbage, rosemary, sage and mint) to external stimuli and to recognize the stimulus applied. In view of this, the research question addressed in this exploration can be summarised as:

1. Can the background (before the application of stimulus) and post-stimulus (after the application of stimulus) signals for each plant species be distinguished?
2. Can the post-stimulus signals for different stimuli (acid, salt) for each plant species be distinguished?

2 Experimental protocol

The details of the experiments performed on four plant species is presented below:

Plant types: Cabbage, Rosemery, Sage, Mint

Electrodes: 2 (1 reference, 1 on the plant)

Stimulus: Acid (H_2So_4) and Salt (NaCl)

Duration: Day1 – Acid stimulus on 4 plants, Salt stimulus on 4 plants, for three hours each.
Day2 – Acid stimulus on 4 plants, Salt stimulus on 4 plants, for three hours each. Four new plants were used for experimental set up.

Signal categories: The signals from each plant prior to the application of the stimuli were recorded, *viz.* background and the signals after the application of the stimuli, *viz.* post-stimulus were also recorded on both days of the experiments. The signals for each of the four plant categories are:

1. Background for different experimental setup

- $BG_{Acid1\_cabbage}$, $BG_{Salt1\_cabbage}$, $BG_{Acid2\_cabbage}$, $BG_{Salt2\_cabbage}$
- $BG_{Acid1\_rosemery}$, $BG_{Salt1\_rosemery}$, $BG_{Acid2\_rosemery}$, $BG_{Salt2\_rosemery}$
- BG_{Acid1_sage}, BG_{Salt1_sage}, BG_{Acid2_sage}, BG_{Salt2_sage}
- BG_{Acid1_mint}, BG_{Salt1_mint}, BG_{Acid2_mint}, BG_{Salt2_mint}

2. Post-stimulus for different experimental setup

- $PS_{Acid1\_cabbage}$, $PS_{Salt1\_cabbage}$, $PS_{Acid2\_cabbage}$, $PS_{Salt2\_cabbage}$
- $PS_{Acid1\_rosemery}$, $PS_{Salt1\_rosemery}$, $PS_{Acid2\_rosemery}$, $PS_{Salt2\_rosemery}$
- PS_{Acid1_sage}, PS_{Salt1_sage}, PS_{Acid2_sage}, PS_{Salt2_sage}
- PS_{Acid1_mint}, PS_{Salt1_mint}, PS_{Acid2_mint}, PS_{Salt2_mint}

3 Approach

The approach followed to solve the above mentioned research problems can be highlighted as:

1. Signal pre-processing

2. Feature extraction – time domain, frequency domain and time-frequency domain features

3. Statistical tests – Wilcoxon ranksum test and ANOVA analysis to find the most discriminant features between background signals and post-stimulus signals and also between the two stimuli (acid, salt).

4. Pattern recognition – classification and clustering to differentiate between both categories of post-stimulus signals (acid/salt) for each plant species.

3.1 Signal pre-processing

All the Background and post-stimulus signals were pre-processed by using a 6th order, Chebyshev type II filter, which provided the best optimization of the cost function using a cut-off frequency of 0.77Hz and stop-band ripple of 100. This filter parameter was used to on the raw signals. Once filtered the signals were divided into non-overlapping blocks of 1024 samples.
.

3.2 Feature extraction

Table 1 lists the details of the features and their description. The following features were extracted from each window of the filtered signals as a result of time domain and time-frequency domain analysis.

No	Features	Description
		Time Domain analysis
1	Mean	Average of the signal
2	Std	Standard deviation of samples
3	Kurtosis	measure of the 'peakedness' of a signal
4	Skewness	measure of the symmetry of the signal
5	Signal Energy	Energy content of the signal
		Frequency Domain analysis
6	PSD_max	Maximum amplitude of the power spectral density (fft)
7	PSD_min	Minimum amplitude of the power spectral density (fft)
8	FFT_power	Energy content of the fft signal
		Time-Frequency analysis
9	Energy_haar_l1	Energy content of the haar level1 DWT coefficient
10	Energy_haar_l2	Energy content of the haar level2 DWT coefficient
11	Energy_haar_l3	Energy content of the haar level3 DWT coefficient
12	ZCR_haar_l1	Zero crossing rate of the haar level1 DWT coefficient
13	ZCR_haar_l2	Zero crossing rate of the haar level2 DWT coefficient
14	ZCR_haar_l3	Zero crossing rate of the haar level3 DWT coefficient
15	Energy_db3_l1	Energy content of the db3 level1 DWT coefficient
16	Energy_db3_l2	Energy content of the db3 level2 DWT coefficient
17	Energy_db3_l3	Energy content of the db3 level3 DWT coefficient
18	ZCR_db3_l1	Zero crossing rate of the db3 level1 DWT coefficient
19	ZCR_db3_l2	Zero crossing rate of the db3 level2 DWT coefficient
20	ZCR_db3_l3	Zero crossing rate of the db3 level3 DWT coefficient
21	Energy_coif3_l1	Energy content of the coif3 level1 DWT coefficient
22	Energy_coif3_l2	Energy content of the coif3 level2 DWT coefficient
23	Energy_coif3_l3	Energy content of the coif3 level3 DWT coefficient
24	ZCR_coif3_l1	Zero crossing rate of the coif3 level1 DWT coefficient
25	ZCR_coif3_l2	Zero crossing rate of the coif3 level2 DWT coefficient
26	ZCR_coif3_l3	Zero crossing rate of the coif3 level3 DWT coefficient
27	Contrast_cgau3	Image generated by CWT using complex Gaussian - measure of local level variations which takes high values for image of high contrast
28	Correlation_cgau3	Image generated by CWT using complex Gaussian - measure of correlation between pixels in two different directions
29	Energy_cgau3	Image generated by CWT using complex Gaussian - Measure of signal energy
30	Homogeneity_cgau3	Image generated by CWT using complex Gaussian - measure that takes high values for low-contrast images
31	Entropy_cgau3	Image generated by CWT using complex Gaussian - measure of randomness and takes low values for smooth images
32	Contrast_cmor	Image generated by CWT using complex Morlet - measure of local level variations which takes high values for image of high contrast
33	Correlation_cmor	Image generated by CWT using complex Morlet - measure of correlation between pixels in two different directions
34	Energy_cmor	Image generated by CWT using complex Gaussian - Measure of signal energy
35	Homogeneity_cmor	Image generated by CWT using complex Morlet - measure that takes high values for low-contrast images
36	Entropy_cmor	Image generated by CWT using complex Morlet - Measure of randomness and takes low values for smooth images

Table 1: List of features extracted from pre-processed Background and post-stimulus signals.

3.3 Discrimination between Background and Post-stimulus signals

This exploration enlists the steps undertaken to fulfil the first research objective to distinguish between background and post-stimulus signals for each plant species. A statistical approach based on ANOVA and Wilcoxon ranksum test was performed to determine the distinguishing features. Analysis of variance (ANOVA) is a collection of statistical models used in order to analyze the differences between group means and their associated procedures (such as variation among and between groups). It provides a statistical test of whether or not the means of several

groups are equal, and therefore generalizes the commonly used *t*-test to more than two groups. In general, the purpose of analysis of variance (ANOVA) is to test for significant differences between means.

The Wilcoxon ranksum test is a non-parametric statistical hypothesis test used to compare two independent random samples taken from two populations whose distributions are identical. The Wilcoxon rankusm does not require that populations have normal distribution which is the primary difference from the paired *t*-test. The key steps involved in this exploration are:

Step1: The null hypothesis is that there is no difference between the background signals of a particular species collected during the four experiments. This is because prior to the application of the stimulus (acid or salt), the background signal of each plant species used for different experimental protocols should have the same morphology.

Step2: Hence, all the 36 features extracted from each of the 4 Background signals undergo the ANOVA test to determine the features which prove the null hypothesis to be false.

Step3: The features which prove the null hypothesis to be true (the one's which find no difference between the background signals of the 4 experimental conditions) are grouped together and are used to determine those features which produce the required discrimination between the Background and the post-stimulus (acid, salt) signals. This can be represented as:

Group1: $[BG_{Acid1\_cabbage}, BG_{Salt1\_cabbage}, BG_{Acid2\_cabbage}, BG_{Salt2\_cabbage}] - [PS_{Acid1\_cabbage}, PS_{Acid2\_cabbage}]$
Group2: $[BG_{Acid1\_cabbage}, BG_{Salt1\_cabbage}, BG_{Acid2\_cabbage}, BG_{Salt2\_cabbage}] - [PS_{Salt1\_cabbage}, PS_{Salt2\_cabbage}]$

Step4: The discrimination between the background and the stimulus groups (acid and salt) is determined through the Wilcoxon ranksum hypothesis test and the commonly used *t*-test. The null hypothesis is that there is a difference between the features of both the groups (Group1, Group2).

Step5: The common intersecting features which produce the discrimination in Group1 and Group2 are selected and a histogram plot of each of the discriminating features is plotted to observe the difference between the distributions.

Step6: Steps1 – 5 are repeated for each of the four plant species.

3.4 Results

The discriminating features between the background and post-stimuli signals for each species (as a result of Step 5) are listed below:

Cabbage	Rosemery	Sage	Mint
Energy_cmor	*Homogeneity_cmor*	*Homogeneity_cmor*	*Homogeneity_cmor*
Homogeneity_cmor	*Energy_cmor*	*Energy_cmor*	*Energy_cmor*
Energy_cgau3	*Entropy_cmor*	*Entropy_cgau3*	*Contrast_cgau3*
Entropy_cgau3	*Energy_cgau3*	*Entropy_cmor*	*Entropy_cgau3*
Entropy_cmor	*Entropy_cgau3*	*Contrast_cmor*	*Energy_cgau3*
Correlation_cgau3	*Contrast_cmor*		*Homogeneity_cgau3*
Contrast_cgau3	*Energy_coif3_l1*		*Entropy_cmor*
Energy_db3_l1	*Energy_db3_l1*		*Correlation_cgau3*
Std	*Energy_haar_l1*		*Mean*
Energy_haar_l1	*Energy*		*Contrast_cmor*
Energy	*FFT_power*		*ZCR_coif3_l3*
FFT_power	*Std*		*Energy_db3_l3*
Energy_coif3_l2			
Energy_coif3_l1			
Energy_haar_l2			
Energy_db3_l3			
Energy_haar_l3			
Energy_coif3_l3			
Mean			
Contrast_cmor			

Table 2: List of features determined through Wilcoxon ranksum and Ttest to distinguish between background and post-stimulus signals for each plant species.

The features pertaining to each plant species as listed in Table 2, are further plotted in a histogram to determine the most distinguishing features. The histogram plot for cabbage is shown in Figure 1:

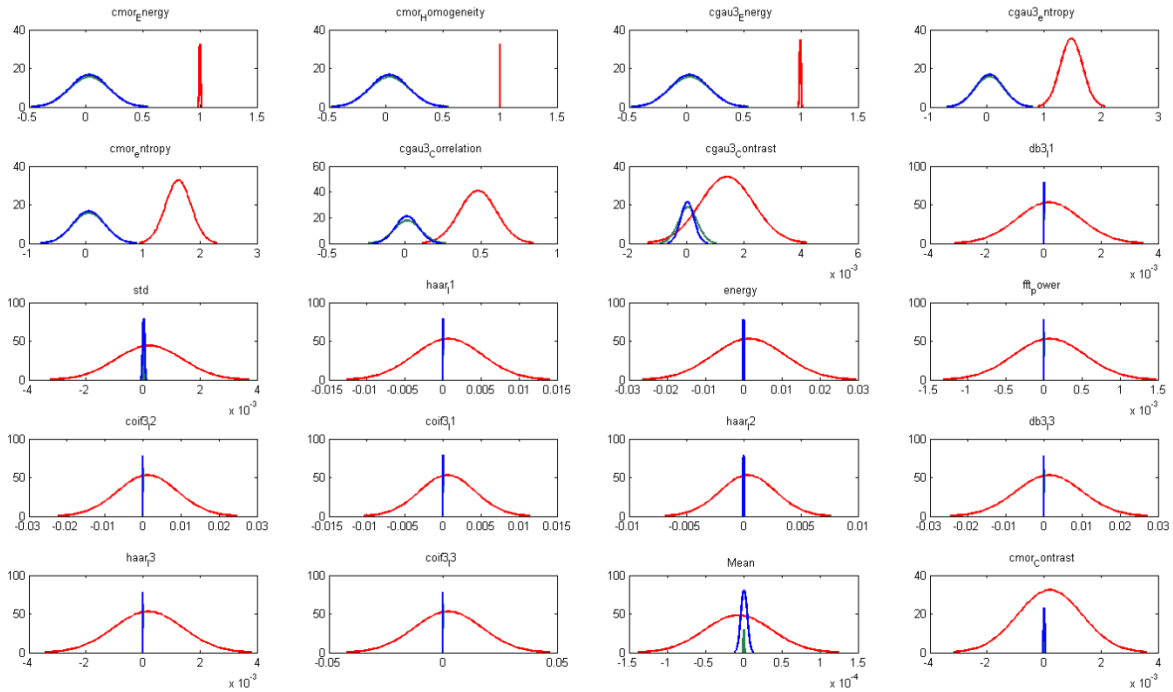

Figure 1: Histogram plot for Cabbage, showing discrimination between background ('blue') and post-stimulus signals ('red') for each of the 20 features as determined in Table 1.

Therefore, from Figure 1, it is clear that although the statistical tests help us to infer a list of 20 features (cf. Table 1) that discriminate between the background and post-stimuli signals, but the

histogram plot of the pre-processed signals show that the features - *Energy_cmor*, *Homogeneity_cmor*, *Energy_cgau3*, *Entropy_cgau3*, *Entropy_cmor* and *Correlation_cgau3* are the most discriminative.

Similarly, as seen from Figure 2, for rosemery, the features - *Homogeneity_cmor*, *Energy_cmor*, *Entropy_cmor*, *Energy_cgau3*, *Entropy_cgau3*, turn out to be the most discriminating features.

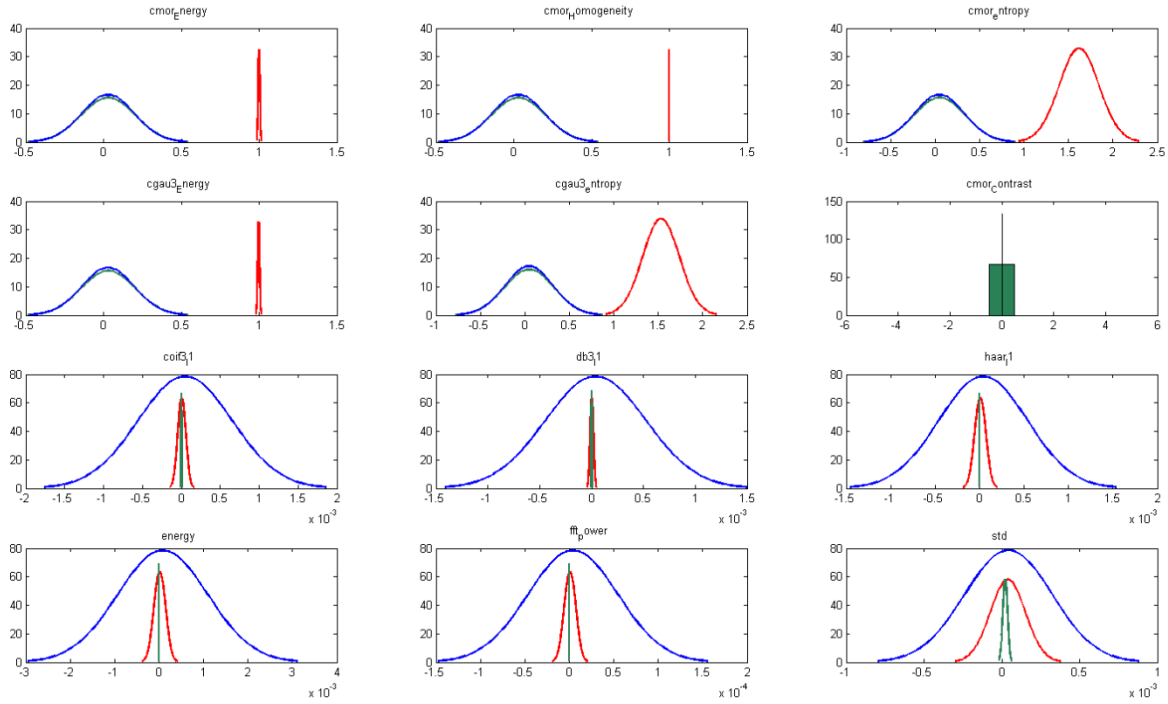

Figure 2: Histogram plot for Rosemery, showing discrimination between background ('blue') and post-stimulus signals ('red') for each of the 12 features as determined in Table 1.

For sage (cf. Figure 3), the features - *Homogeneity_cmor*, *Energy_cmor*, *Entropy_cgau3*, *Entropy_cmor*, reflect a discrimination between the background and post-stimulus signals.

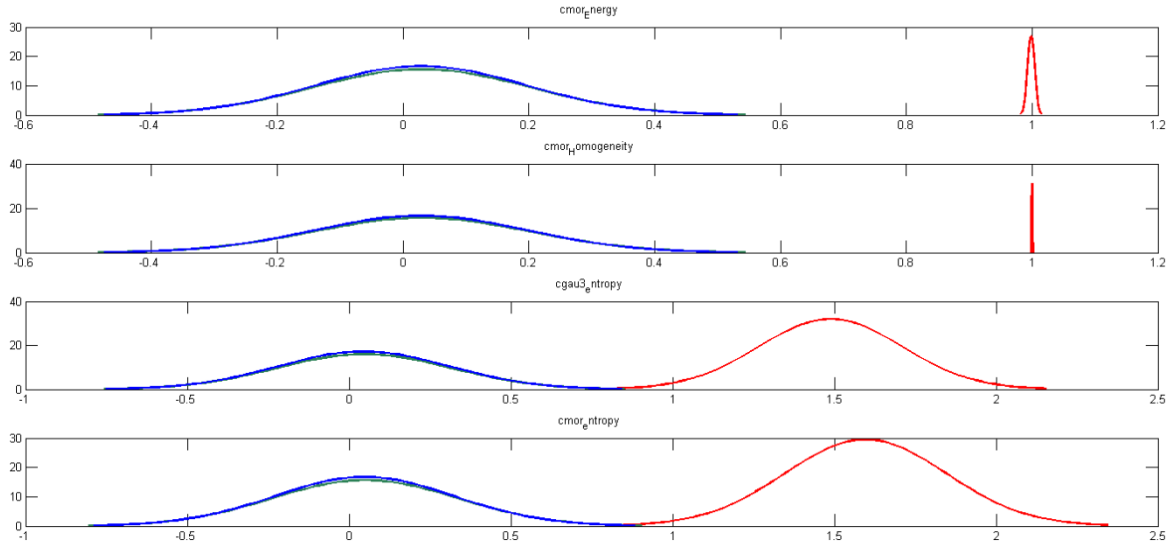

Figure 3: Histogram plot for Sage, showing discrimination between background ('blue') and post-stimulus signals ('red') for each of the 4 features as determined in Table 1.

For mint the seven discriminating features are - *Homogeneity_cmor*, *Energy_cmor*, *Contrast_cgau3*, *Entropy_cgau3*, *Energy_cgau3*, *Homogeneity_cgau3*, *Entropy_cmor* which is further evident from Figure 4.

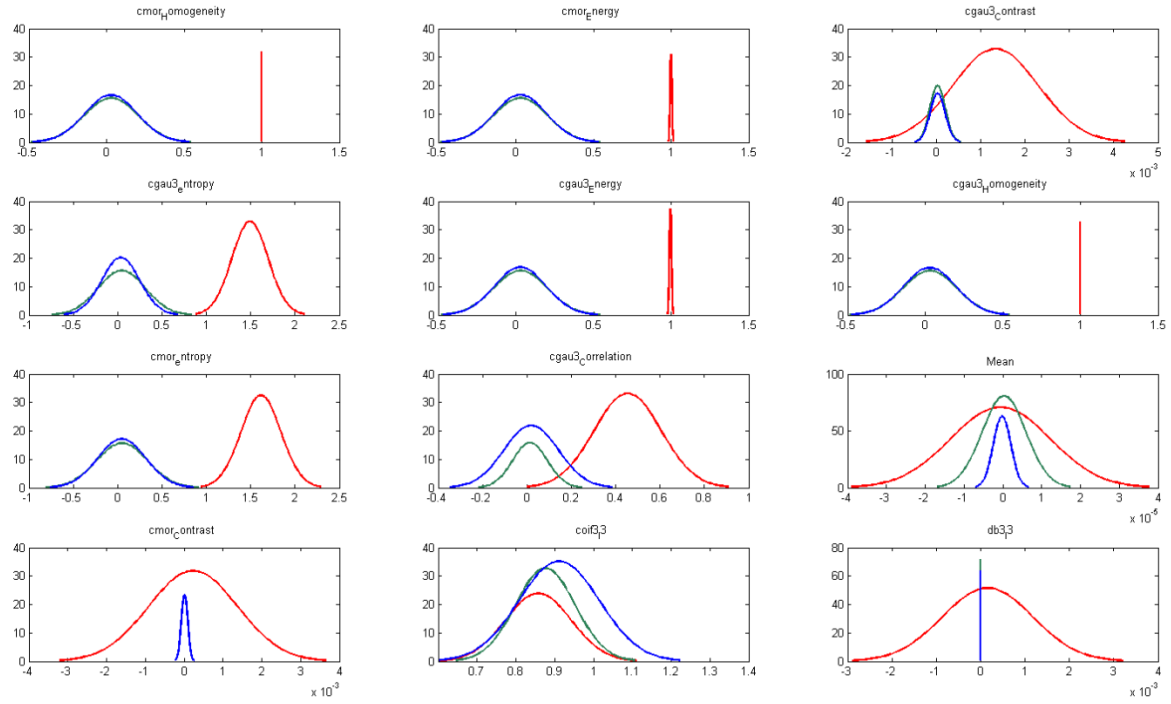

Figure 4: Histogram plot for Mint, showing discrimination between background ('blue') and post-stimulus signals ('red') for each of the 12 features as determined in Table 1.

Therefore, the final list of features which produce considerable discrimination between the background and post-stimulus signals for each plant species have been listed in Table 2.

Cabbage	Rosemery	Sage	Mint
Energy_cmor	*Homogeneity_cmor*	*Homogeneity_cmor*	*Homogeneity_cmor*
Homogeneity_cmor	*Energy_cmor*	*Energy_cmor*	*Energy_cmor*
Energy_cgau3	*Entropy_cmor*	*Entropy_cgau3*	*Contrast_cgau3*
Entropy_cgau3	*Energy_cgau3*	*Entropy_cmor*	*Entropy_cgau3*
Entropy_cmor	*Entropy_cgau3*		*Energy_cgau3*
Correlation_cgau3			*Homogeneity_cgau3*
			Entropy_cmor

Table 2: List of discriminating features as determined from the histogram plot.

3.5 Discrimination between the post-stimulus signals for acid and salt stimuli

Having determined the features which produce discrimination between the background and post-stimulus signals (generated as a result of acid and salt stimuli), in this section the methodology adopted to discriminate between the two post-stimulus signals for each plant species has been described. This exploration also helps to ascertain the behaviour of each of the four plant species in discriminating between the two different stimuli.

Step1: The null hypothesis is that there is no difference in signal characteristics between the post-stimulus signals using acid between the two experiments performed over two days for each category of plants, i.e. [$PS_{Acid1\_cabbage}$, $PS_{Acid2\_cabbage}$]. This hypothesis is tested using Wilcoxon ranksum test and a list of features are selected which satisfy the hypothesis.

Step2: Similarly there is no difference between the post-stimulus signals generated as a result of salt stimulus between the two experiments performed over different days for each category of plants, i.e. [$PS_{Salt1\_cabbage}$, $PS_{Salt2\_cabbage}$]. A list of features having no difference between the two post-stimulus salt groups is selected.

Step3: The features selected in Step1 and Step2 are used to determine the most discriminating features between the post-stimulus signals for acid and salt. A similar kind of null hypothesis using Wilcoxon ranksum test is performed to determine the discriminating feature sets.

Step4: Steps 1 – 3 are repeated for each of the 4 plant species.

3.6 Results

The discriminating features between the post-stimuli signals for acid and salt for each plant species are listed below:

Cabbage	Rosemery	Sage	Mint
ZCR_db3_l2 *ZCR_coif3_l2*			*Energy_db3_l2* *Energy_coif3_l2* *Energy_haar_l2* *Skewness*

Table 3: List of features determined through Wilcoxon ranksum to distinguish between the post-stimulus signals for acid and salt for each plant species.

Therefore the features *ZCR_db3_l2* and *ZCR_coif3_l2* computed from the post-stimulus signals of the cabbage plant are supposed to be responsive towards discriminating between acid and salt. Similarly, the discriminating features for the mint plant are listed in Table 3. However, it is interesting to note that the plant species rosemery and sage did not respond to the differences in the applied stimuli (acid and salt) using the investigated features.

A further analysis from the histogram plot reveals that there is hardly any separation between the two categories of post-stimulus signals – acid and salt. Here for the plot, features other than the one's mentioned in Table 3 were also considered. Figures 5 – 8 show the histogram plots for cabbage, rosemery, sage and mint respectively.

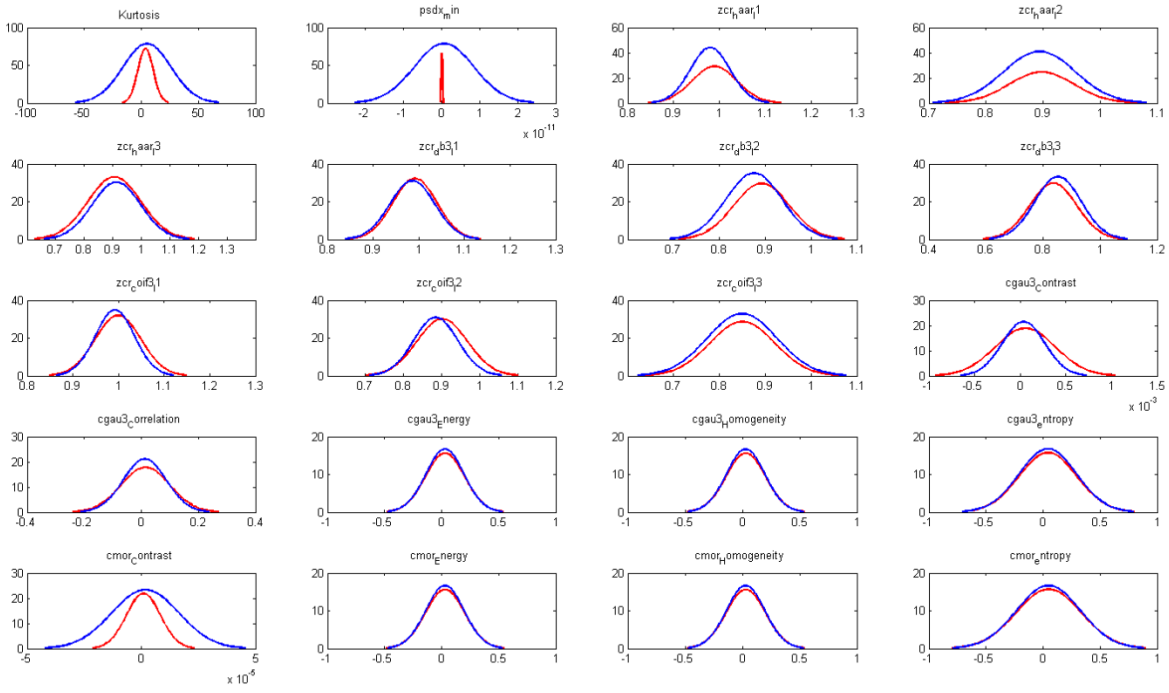

Figure 5: Histogram plot for features extracted from post-stimulus signals generated as a result of acid ('red') and salt ('blue') stimulus for the Cabbage plant.

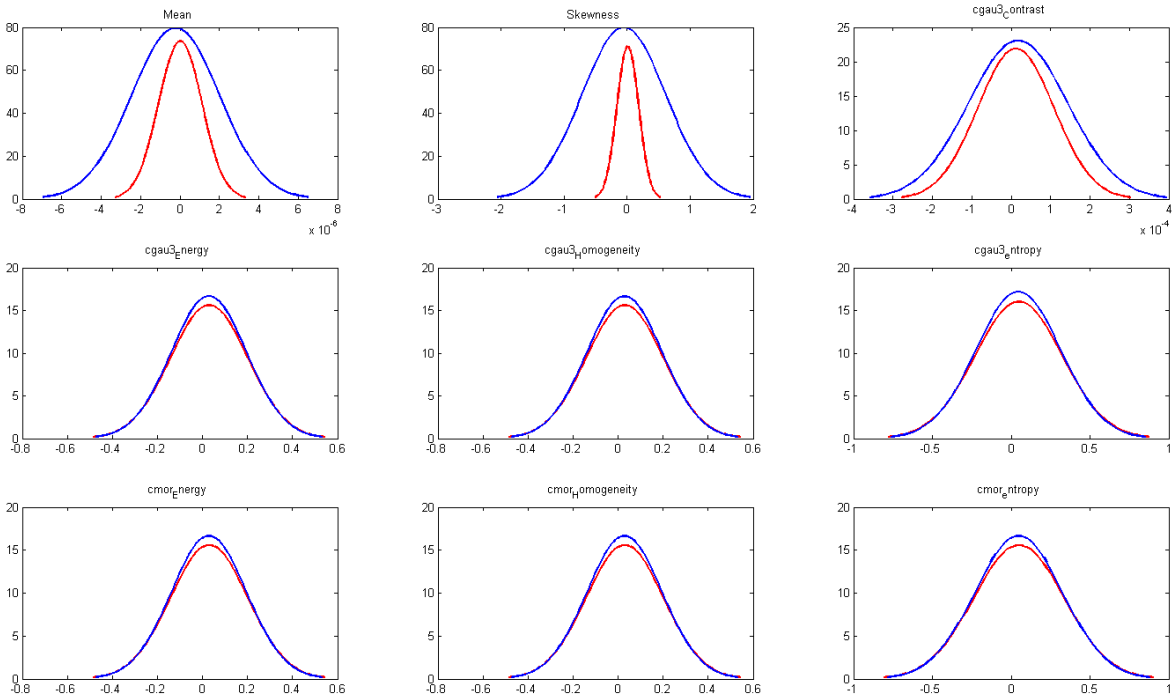

Figure 6: Histogram plot for features extracted from post-stimulus signals generated as a result of acid ('red') and salt ('blue') stimulus for the Rosemery plant.

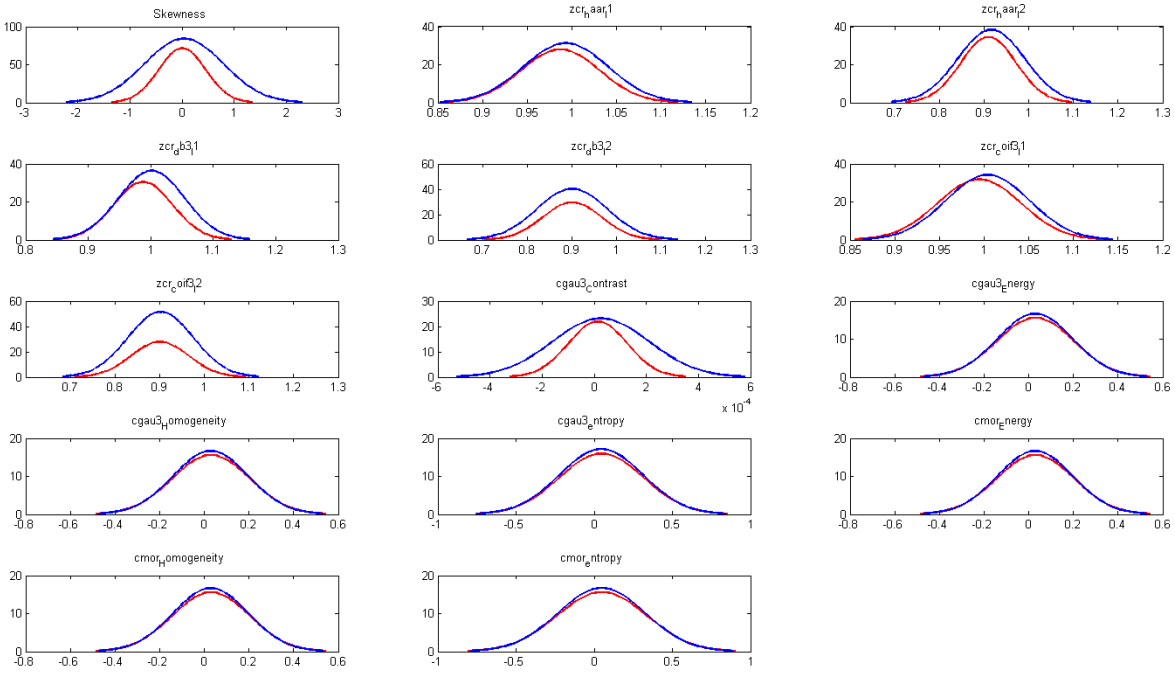

Figure 7: Histogram plot for features extracted from post-stimulus signals generated as a result of acid ('red') and salt ('blue') stimulus for the Sage plant.

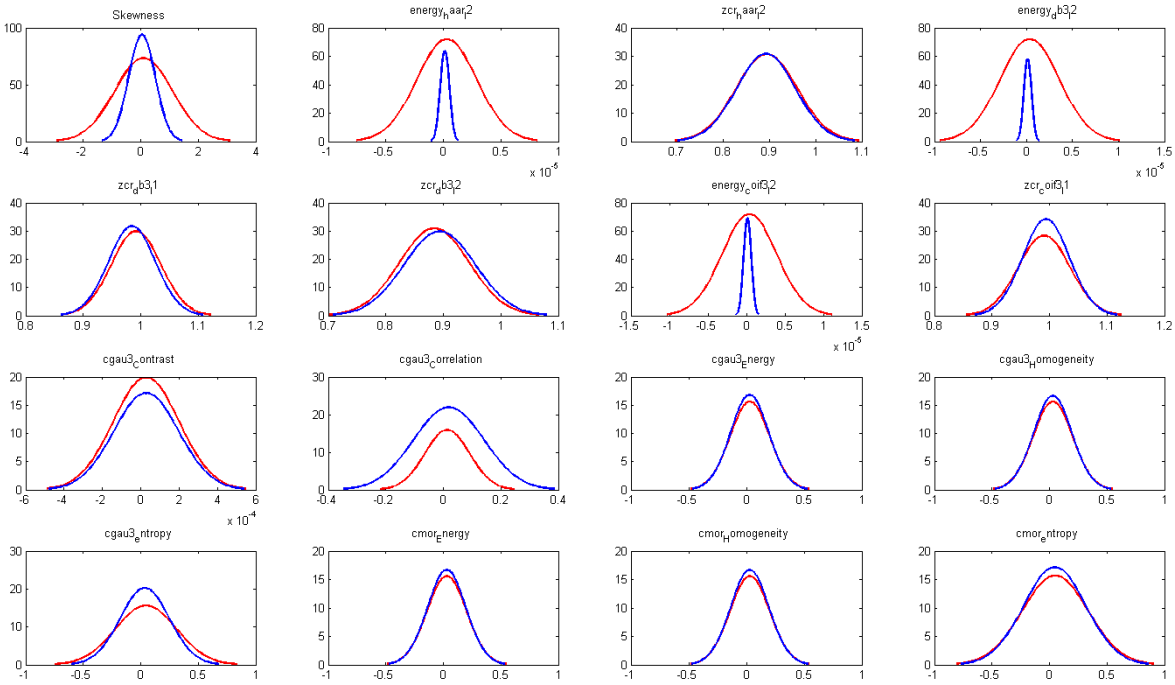

Figure 8: Histogram plot for features extracted from post-stimulus signals generated as a result of acid ('red') and salt ('blue') stimulus for the Mint plant.

4 Methodology based on pattern recognition

It is evident from the plots (Figures 5 – 8) that there is minimal separation between the post-stimulus signals for acid and salt stimuli for each of the plant species. Therefore, a further exploration to classify the signals based on pattern recognition based approach was undertaken. There were two approaches that were undertaken – 1) use of supervised learning algorithms to classify the post-stimulus signals and 2) use of *k*-means clustering and minimum distance computation. These are described in detail in the following sections. The basic approach of this methodology has been depicted in Figure 9.

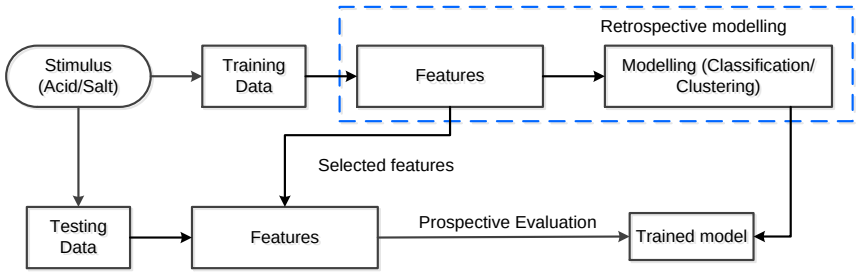

Figure 9: Retrospective modelling and prospective evaluation to classify post-stimulus signals for acid and salt for each plant species.

In this exploration the features pertaining to the second order statistics, generated as a result of continuous wavelet transform (CWT) were not considered because of reduction in the sample length. Therefore, only 26 features were considered for this analysis which is listed in Table 4.

No	Features
1	Mean
2	Std
3	Kurtosis
4	Skewness
5	Signal Energy
6	PSD_max
7	PSD_min
8	FFT_power
9	Energy_haar_l1
10	Energy_haar_l2
11	Energy_haar_l3
12	ZCR_haar_l1
13	ZCR_haar_l2
14	ZCR_haar_l3
15	Energy_db3_l1
16	Energy_db3_l2
17	Energy_db3_l3
18	ZCR_db3_l1
19	ZCR_db3_l2
20	ZCR_db3_l3
21	Energy_coif3_l1
22	Energy_coif3_l2
23	Energy_coif3_l3
24	ZCR_coif3_l1
25	ZCR_coif3_l2
26	ZCR_coif3_l3

Table 4: List of extracted features considered for classification and clustering based analysis

4.1 Classification based on decision boundary

Supervised classification techniques involve two phases – training a model with a given set of observations and evaluating the trained model with new set of observations (testing). Here, the post-stimulus data for each plant species is used to develop a model retrospectively. The trained model is cross-validated in association with three supervised learning algorithms independently – linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and support vector machines (SVM) [1]. The trained model (classifier) is then prospectively evaluated on new set of data in association with the best performing learning algorithm to classify the acid and salt stimuli for each individual species.

Taking an example, considering cabbage,

$[PS_{Acid1\_cabbage}, PS_{Salt1\_cabbage}]$ are used for retrospective modelling,

the developed model was used for prospectively evaluating on $[PS_{Acid2\_cabbage}, PS_{Salt2\_cabbage}]$.

In view of this, the procedure is divided in to three parts – feature selection, retrospective modelling and prospective evaluation, each of which is discussed in the following sections.

4.1.1 Feature selection

Feature selection helps to select the optimal number of features thereby reducing the computational load and helps in achieving the best possible classification accuracy. The extracted features were normalised and the Wrapper approach was followed using the sequential forward selection (*sfs*) searching technique. It selects various feature vector combinations to test for the minimal classification error probability and is computationally simple. Here, the selection of the optimal number of features depends strongly on the employed classification algorithm.

The *sfs* technique can be explained with a working example by considering a feature vector comprising of four different features [X_1, X_2, X_3, X_4]. First, the best ranked feature is computed, say X_2, and the classification performance is evaluated with X_2. Secondly, all two-dimensional feature vector combinations with X_2 are computed: [X_1, X_2], [X_2, X_3], [X_2, X_4] and the classification performance for each of the combinations is evaluated. Thirdly, all three-dimensional feature vector combination with X_2 are computed: [X_1, X_2, X_3], [X_1, X_2, X_4] and the classification performance is evaluated with both the combinations. Finally, the features forming the best feature vector combination are selected [2].

The number of features selected in each of the cases is highlighted in the corresponding results.

4.1.2 Retrospective modelling

The recognition model was developed retrospectively using the data collected on Day 1, in association with the learning algorithms LDA, QDA and SVM. The model is verified through 10 runs of 10-fold cross-validation whereby out of the total *N* available data samples, 9*N/10* data samples are used to train the classifier whilst the remaining sample size (*N/10*) is used to test the classifier. This is repeated such that each individual data segment takes the role of the classifier test data. The true error could then be calculated as the mean of the error over all *10* runs performed [3].

Therefore for retrospective modelling the following data are considered for each of the plant species:

- $PS_{Acid1\_cabbage}$, $PS_{Salt1\_cabbage}$
- $PS_{Acid1\_rosemery}$, $PS_{Salt1\_rosemery}$
- PS_{Acid1_sage}, PS_{Salt1_sage}
- PS_{Acid1_mint}, PS_{Salt1_mint}

4.1.3 Results for the retrospective modelling

The classification results for the cross-validation stage for each of the three learning algorithms - LDA, QDA and SVM are presented in Table 5.

Classifier	Plant Species	Acid (%)	Salt (%)	Features
LDA	Cabbage	97.1	98.6	Std, Energy_haar_l1, Energy_db3_l2, ZCR_coif3_l1
	Rosemery	100	85.8	Kurtosis, PSD_min, Energy_db3_l3
	Sage	74.3	61.4	ZCR_db3_l1, ZCR_coif3_l3
	Mint	82.8	75.8	Mean, PSD_max, ZCR_haar_l3, ZCR_db3_l3
QDA	Cabbage	98.6	97.1	Energy_haar_l1
	Rosemery	95.8	97.1	Mean, Kurtosis, Energy_db3_l3
	Sage	98.6	57.1	Mean, Energy_db3_l2, ZCR_db3_l1, ZCR_coif3_l3
	Mint	92.9	74.1	Mean, PSD_min, ZCR_haar_l1, ZCR_db3_l1, ZCR_db3_l2, ZCR_db3_l3, ZCR_coif3_l3
SVM	Cabbage	54.3	68.1	ZCR_db3_l2, ZCR_coif3_l1, ZCR_coif3_l2
	Rosemery	65.7	54.1	ZCR_db3_l3
	Sage	77.5	51.4	ZCR_coif3_l3
	Mint	84.3	74.5	ZCR_haar_l3, ZCR_db3_l2, ZCR_db3_l3

Table 5: Cross-validation of the retrospective model for each plant species.

4.1.4 Prospective evaluation

The model was prospectively evaluated on the data from Day 2 for each species as follows:

- $PS_{Acid2\_cabbage}$, $PS_{Salt2\_cabbage}$
- $PS_{Acid2\_rosemery}$, $PS_{Salt2\_rosemery}$
- PS_{Acid2_sage}, PS_{Salt2_sage}
- PS_{Acid2_mint}, PS_{Salt2_mint}

The results for prospective exploration are presented in Table 6. The table lists the sensitivity for acid and salt recognition for each of the four species and for the three learning algorithms. The table also lists the features that are selected as a result of feature selection.

Classifier	Plant Species	Acid (%)	Salt (%)	Features
LDA	Cabbage	37.1	100	Std, Energy_haar_l1, Energy_db3_l2, ZCR_coif3_l1
	Rosemery	10	65.7	Kurtosis, PSD_min, Energy_db3_l3
	Sage	74.3	27.2	ZCR_db3_l1, ZCR_coif3_l3
	Mint	67.1	18.5	Mean, PSD_max, ZCR_haar_l3, ZCR_db3_l3
QDA	Cabbage	50	100	Energy_haar_l1
	Rosemery	7.1	80	Mean, Kurtosis, Energy_db3_l3
	Sage	24.3	20	Mean, Energy_db3_l2, ZCR_db3_l1, ZCR_coif3_l3
	Mint	71.3	24.3	Mean, PSD_min, ZCR_haar_l1, ZCR_db3_l1, ZCR_db3_l2, ZCR_db3_l3, ZCR_coif3_l3
SVM	Cabbage	47.3	62.9	ZCR_db3_l2, ZCR_coif3_l1, ZCR_coif3_l2
	Rosemery	40	51.2	ZCR_db3_l3
	Sage	78.5	15.8	ZCR_coif3_l3
	Mint	70	18.6	ZCR_haar_l3, ZCR_db3_l2, ZCR_db3_l3

Table 6: Prospective evaluation for each plant species for the three learning algorithms.

The results clearly illustrate that for none of the plant species, across all the algorithms, both acid and salt are classified up to a satisfactory level (> 60%). Although the model was successful in distinguishing between acid and salt signals in the retrospective phase, it fails to generalize for the prospective study on the data that it has not seen during modelling. This is a typical problem with many classifiers, where they perform well on the *training* dataset but perform poorly on the *testing* dataset (data on which it has not been trained). Hence, it can be inferred that the learnt model or the classifier is poorly generalized because it cannot perform well on new set of data.

One of the reasons for this failure could also be the less number of data points. The employed classification algorithms here namely LDA, QDA and SVM primarily work based on a decision boundary based system. Data points lying on either side of the decision boundary are classified accordingly to the competing classes. LDA and QDA are also affected by outlier data points which might lead to a complicated decision boundary which caters well for the variations of the *training* set but fails to generalize for the data points not used for the modelling (*testing* set). Although SVM caters to outliers by concentrating only on the support vectors that lie proximal to the decision boundary rather than all the data points, none of these methods can effectively model a sparse data distribution in the respective feature space.

4.1.5 Further exploration using classification

Hence, for further evaluation, experimental data was collected from the same four plant species using the same experimental protocol as earlier. Experiments were conducted on two consecutive days with acid and salt stimuli to new set of plants for each experimental session. Hence, this leads to a new data structure in addition to the one's mentioned before:

- $PS_{Acid3\_cabbage}$, $PS_{Salt3\_cabbage}$, $PS_{Acid4\_cabbage}$, $PS_{Salt4\_cabbage}$
- $PS_{Acid3\_rosemery}$, $PS_{Salt3\_rosemery}$, $PS_{Acid4\_rosemery}$, $PS_{Salt4\_rosemery}$
- PS_{Acid3_sage}, PS_{Salt3_sage}, PS_{Acid4_sage}, PS_{Salt4_sage}
- PS_{Acid3_mint}, PS_{Salt3_mint}, PS_{Acid4_mint}, PS_{Salt4_mint}

Here, only the post-stimulus signals have been mentioned. Based on this dataset two new explorations were made. These explorations have been described with the cabbage plant as an illustrative example. The same methodology is followed for the other three plant species.

1). Forming a retrospective model (training) based on:

$[PS_{Acid1\_cabbage}, PS_{Salt1\_cabbage}, PS_{Acid2\_cabbage}, PS_{Salt2\_cabbage}, PS_{Acid3\_cabbage}, PS_{Salt3\_cabbage}]$

Evaluating the model (testing) on:

$[PS_{Acid4\_cabbage}, PS_{Salt4\_cabbage}]$.

4.1.6 Results

The results for the prospective evaluation for each plant species are presented in Table 7:

Classifier	Plant Species	Acid (%)	Salt (%)	Features
LDA	Cabbage	100	34.8	PSD_max, Energy_db3_l3, ZCR_db3_l1, ZCR_db3_l2, Energy_coif3_l3,
	Rosemery	100	55.1	ZCR_haar_l2, ZCR_db3_l1, ZCR_db3_l3, ZCR_coif3_l3
	Sage	95.7	2.9	Mean, Skewness, Energy_haar_l3
	Mint	92.75	14.49	Mean, Std, Energy_haar_l1, Energy_haar_l2, Energy_db3_l1, Energy_db3_l2, ZCR_db3_l3, Energy_coif3_l1, Energy_coif3_l3, ZCR_coif3_l3
SVM	Cabbage	92.8	47.8	Energy_db3_l1, ZCR_db3_l1, ZCR_coif3_l2
	Rosemery	100	33.3	ZCR_db3_l1, ZCR_db3_l3, ZCR_coif3_l1
	Sage	91.3	0	Std
	Mint	88.4	27.5	Energy_haar_l1, ZCR_haar_l1, ZCR_db3_l3, ZCR_coif3_l3

Table 7: Prospective evaluation for each plant species for the learning algorithms LDA, SVM with data from Day 1, day 2 and Day 3 used for forming the model and evaluating the model on Day 4 data.

Table 7 indicates that acid detection improves but salt detection is poor and inconsistent across plant species and learning algorithms. In this exploration, the QDA algorithm was not used since its results (cf. Table 5 and 6) were on similar lines to LDA. In the following section a clustering based approach has been discussed which was adopted in view of its ability to cater to the underlying data distribution and search for a unique feature space where the data can be represented in compact clusters having a minimal within-class variance.

4.2 Clustering based exploration

Having explored the classification algorithms, it is imperative to explore a different algorithmic formulation that can classify acid and salt stimuli for each plant species. In this exploration, a *k*-means clustering based methodology is used to form compact clusters in a multi-dimensional feature space representing the training data. The test data is associated with each respective cluster based on a Euclidean or Mahalanobis distance based minimum distance classifier. A major advantage of the *k*-means algorithm is its computational simplicity making it an attractive choice for a wide variety of applications [2]. It is a well-perceived fact in the research community that cluster analysis is primarily used for unsupervised learning where the class labels for the training data are not available. However, the *k*-means algorithm can also be used for supervised learning where the class labels of the training data are known a priori. In this proposed methodology, the class labels for the training data pertaining to the stimulus signals are

known. This helps to have a definitive estimate of the underlying cluster structure to be formed on the data (three clusters), thereby facilitating a faster convergence during cluster formation for reduced time complexity [1,4].

The basic philosophy of the methodology has been illustrated in Figure 10, where two clusters Acid and Salt are formed on the *training* dataset corresponding to the post-stimulus data of Day 1 and Day 2 grouped together, in a 2-dimensional feature space (Feature 1 (f_1) and Feature 2 (f_2)). The distance of the *test* vector *Test* from each of the two cluster centroids are represented by the distances d_{Acid}, d_{Salt}. These two distance measures are compared to estimate the proximity of the *Test* dataset to each cluster and assigned to the nearest one. This methodology can be further scaled up by forming more clusters corresponding to new categories of stimulus and associating a new dataset (corresponding to a test dataset) to the proximal cluster. The formation of unique clusters corresponding to each type of stimulus can be achieved by selecting the optimum number of features which help to discriminate stimulus patterns in the respective feature space.

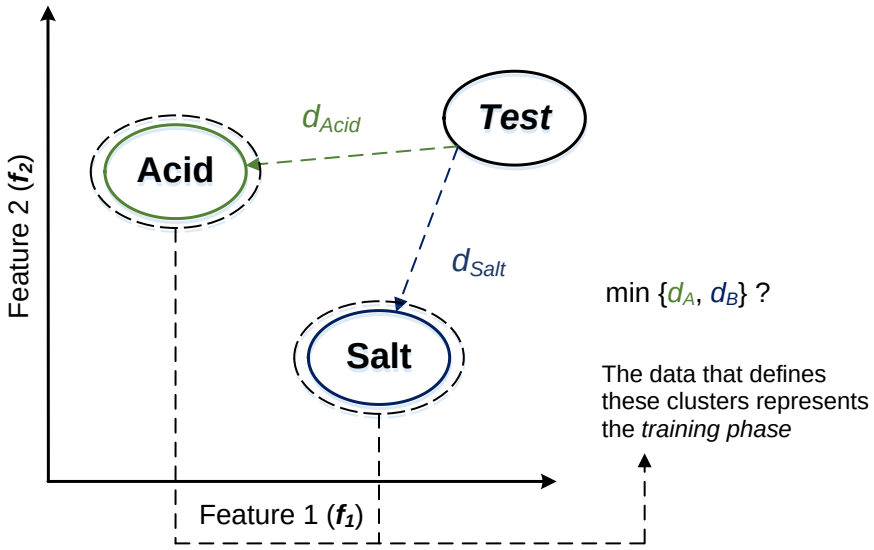

Figure 10: Illustration of the clustering and minimum distance classifier based methodology.

The process proceeds with the selection of the optimum features which helps in forming a feature space where compact clusters are formed. This process is explained in further details.

4.2.1 Feature selection

The extracted features are linearly normalized and the best features for each plant species are selected by using the low-complexity class-separability measure based on scatter matrices which ranks the 26 features for each plant-stimulus combination [2]. The scatter matrices quantify the scatter of feature vectors in the feature space. The rank of each individual feature for a multiple-class scenario is determined by the *R* value calculated as:

$$R = \frac{S_m}{S_b} \quad (1)$$

$$S_m = S_w + S_b \quad (2)$$

where S_w and S_b are the within-class and between-class scatter matrices respectively and S_m is the mixture scatter matrix.

$$S_w = \sum_{i=1}^{c} P_i S_i \quad (3)$$

where P_i denotes the priori probability of a given class i = 1, 2,…c and S_i is the respective covariance matrix of class i.

$$S_b = \sum_{i=1}^{c} P_i (m_i - m_0)(m_i - m_0)^T \quad (4)$$

$$m_0 = \sum_{i=1}^{c} P_i m_i \quad (5)$$

where m_0 is the global mean vector. A high value of R represents a small within-class variance and a large between-class distance among the data points in the respective feature space [2]. The ranked features are sorted in descending order with respect to their R values. A sequential forward selection (*sfs*) technique is employed, selecting the first i features of the ranked feature set in each iteration (i = 2…26) and it is checked if the data from the *training* phase can be correctly clustered in a multi-dimensional feature space. This has been described in detail in the following section.

4.2.2 Cluster formation on the training dataset

The fundamental concept of cluster analysis is to form groups of similar objects as a means of distinguishing them from each other and can be applied in any discipline involving multivariate data. With a given dataset X = {x_i}, i = 1,...,n to be clustered into a set of k clusters, the k-means algorithm iterates to minimize the squared error between the empirical mean of a cluster and the individual data points, defined as the cost function, J:

$$J(\theta, u) = \sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij} \| x_i - \theta_j \|^2 \quad (6)$$

where θ_j is the cluster center and u_{ij} = 1 if x_i lies close to θ_j, or 0 if otherwise. Initially k centroids are defined and the data vectors are assigned to a cluster label depending on how close they are to each centroid. The k centroids are recalculated from the newly defined clusters and the process

of reassignment of each data vector to each new centroid is repeated. The algorithm iterates over this loop until the data vectors from the dataset X form clusters and the cost function *J* is minimized [5].

Here, as described the clusters are formed using data from the training set while the testing data is checked for its proximity to the formed clusters. Two explorations were made, details of which along with the data structure considered and the results obtained are mentioned here.

1). Forming the clusters using the data from the first two days:

Training - [$PS_{Acid1\_cabbage}$, $PS_{Salt1\_cabbage}$, $PS_{Acid2\_cabbage}$, $PS_{Salt2\_cabbage}$]

and evaluating the proximity of the test data (Day 3 and 4) to the formed clusters:

Test 1 - [$PS_{Acid3\_cabbage}$, $PS_{Salt3\_cabbage}$] and

Test 2 - [$PS_{Acid4\_cabbage}$, $PS_{Salt4\_cabbage}$].

Here, the minimum distance classifier (based on Euclidean and the Mahalanobis distance) was used to determine the proximity of the test data collected on Day3 and Day 4 to the clusters formed on Day 1 and Day 2.

- The cluster formation using *k*-means runs on the *training* dataset for each subject comprising of feature vectors (26 features) extracted from each post-stimulus signal.
- The algorithm runs in conjunction with the *sfs* algorithm sequentially selecting a combination of 2 to 26 ranked features in each step (*i*).
- A threshold of 25% of the expected number of data points is set for each of the two clusters formed (i.e. for Acid: 140 ± 35 and for Salt: 140 ± 35). This threshold value was experimentally selected since it produced the best results. If the number of data points in each cluster is within the threshold, it is considered as correctly clustered for that particular combination (*i*) of features selected (where $i = 2...26$).
- The distance of the mean of the *training* dataset for each class label from the cluster centroids is computed and thereby each cluster is assigned with the class label that has its closest proximity to that particular class of the *training* dataset.

A minimum distance classifier is used to compute the distance of the *test* dataset from the centroid of each cluster in a multi-dimensional feature space (considering the feature combination of the current step, *i*) based upon: a). Euclidean distance and b). Mahalanobis distance. The Mahalanobis distance is used to measure the distance of a point from a data distribution. The data distribution is characterized by the mean and the covariance matrix which defines the shape of how the data is distributed in the feature space and is generally hypothesized as a multivariate Gaussian distribution. Here, the Mahalanobis distance takes into consideration the covariance of the clusters along with their mean for the maximum likelihood estimation of

the covariance matrix and hence is effective for clusters with larger variance along one or many directions and in general having an ellipsoidal shape [2].

- The *test* dataset is assigned to a particular cluster depending on the minimum distance computed for each of the two measures (Euclidean or Mahalanobis).
- The predicted label is verified with respect to the known annotations thereby ascertaining the accuracy of the prediction.

The results are summarised in Table 8, for each plant species and both the distance measures. The features are selected based on the methodology described above. The results clearly show the effectiveness of the methodology over the classification technique used earlier with higher sensitivities for acid and salt detection.

Distance Measure	Plant Species	Acid (%)	Salt (%)	Acid (%)	Salt (%)	Features
		Day 3		Day 4		
Euclidean	Cabbage	46.4	95.7	63.8	79.7	Mean, ZCR_haar_l1, ZCR_coif3_l1, ZCR_db3_l1
	Rosemery	55.1	95.7	100.0	73.9	Mean, Skewness, ZCR_db3_l1, ZCR_coif3_l1, ZCR_haar_l1
	Sage	78.3	34.8	21.7	88.4	Mean, ZCR_haar_l1, ZCR_coif3_l1, Skewness, ZCR_db3_l1, ZCR_haar_l3, ZCR_haar_l2, ZCR_coif3_l2, ZCR_coif3_l3, ZCR_db3_l3, ZCR_db3_l2, Std, Energy_db3_l1
	Mint	65.2	82.6	81.2	59.4	ZCR_haar_l1, ZCR_db3_l1, ZCR_coif3_l1, ZCR_db3_l2, ZCR_haar_l2, ZCR_coif3_l3
Mahalanobis	Cabbage	44.9	94.2	72.5	85.5	Mean, ZCR_haar_l1, ZCR_coif3_l1, ZCR_db3_l1, ZCR_haar_l2, ZCR_db3_l2, ZCR_haar_l3, ZCR_db3_l3, ZCR_coif3_l2, ZCR_coif3_l3
	Rosemery	55.1	94.2	91.3	71.0	Mean, Skewness, ZCR_db3_l1
	Sage	78.3	49.3	21.7	87.0	Mean, ZCR_haar_l1, ZCR_coif3_l1, skewness, ZCR_db3_l1, ZCR_haar_l3, ZCR_haar_l2, ZCR_coif3_l2, ZCR_coif3_l3, ZCR_db3_l3, ZCR_db3_l2, std, mean
	Mint	62.3	89.9	88.4	60.9	ZCR_haar_l1, ZCR_db3_l1, ZCR_coif3_l1, ZCR_db3_l2, ZCR_haar_l2, ZCR_coif3_l3

Table 8: Clustering results, where clusters are formed on Day 1 and Day 2 and prospectively evaluated on data from Day 3 and Day 4 using a Euclidean and Mahalanobis distance measure.

However, the sensitivities achieved are not uniformly on the higher side for each plant species for both acid and salt detection over the two days (Day 3 and 4). Hence a new method was tried by considering three day's data as training and evaluating the system to detect the presence of acid and salt on the data of Day 4.

2). Forming the clusters using the data from Day 1, 2 and 3:

$[PS_{Acid1\_cabbage}, PS_{Salt1\_cabbage}, PS_{Acid2\_cabbage}, PS_{Salt2\_cabbage}, PS_{Acid3\_cabbage}, PS_{Salt3\_cabbage}]$

and evaluating the proximity of the test data (Day 4) to the formed clusters:

$[PS_{Acid4\_cabbage}, PS_{Salt4\_cabbage}]$.

Similarly, a minimum distance classifier based on both the Euclidean and the Mahalanobis distance were used to determine the proximity of the test data collected on Day4 to the clusters formed by considering data of Day1, Day2 and Day3. The results for the prospective evaluation of Day 4 data are presented in Table 9.

Distance Measure	Plant Species	Acid (%)	Salt (%)	Features
Euclidean	Cabbage	98.6	73.9	Mean, ZCR_haar_l1, ZCR_coif3_l1, ZCR_db3_l1, ZCR_haar_l2, ZCR_db3_l2, ZCR_haar_l3, ZCR_db3_l3, ZCR_coif3_l2, ZCR_coif3_l3, Skewness, Energy_haar_l1, Energy_coif3_l1
	Rosemery	100	52.2	Mean, Skewness, ZCR_db3_l1, ZCR_coif3_l1, ZCR_haar_l1, ZCR_coif3_l3, ZCR_coif3_l2, ZCR_haar_l3, ZCR_haar_l2, ZCR_db3_l3, ZCR_db3_l2
	Sage	0	0	
	Mint	0	0	
Mahalanobis	Cabbage	95.7	76.8	Mean, ZCR_haar_l1, ZCR_coif3_l1, ZCR_db3_l1, ZCR_haar_l2, ZCR_db3_l2, ZCR_haar_l3, ZCR_db3_l3, ZCR_coif3_l2, ZCR_coif3_l3, Skewness, Energy_haar_l1, Energy_coif3_l1
	Rosemery	100	54	Mean, Skewness, ZCR_db3_l1, ZCR_coif3_l1, ZCR_haar_l1, ZCR_coif3_l3, ZCR_coif3_l2, ZCR_haar_l3, ZCR_haar_l2
	Sage	0	0	
	Mint	0	0	

Table 9: Clustering results, where clusters are formed on Day 1, Day 2 and Day 3 data and prospectively evaluated on data from Day 4 using a Euclidean and Mahalanobis distance measure.

It clearly shows that although for cabbage and rosemery the results are better than the previous occasion, for the plant species sage and mint, the cluster formation itself was not successful. This indicates that the required number of data points could not be clustered together in the respective feature space and hence the fields in Table 9 for these two plants have been left vacant. However, from the last approach (cf. Table 8), detection of acid and salt stimuli for sage and mint were successful (although inconsistent in the sensitivity values), which implies that adding the third day's data leads to some erroneous characteristics in the cluster formation. This is also testified by the fact that results in Table 8 show variations in the results of Day 3 and Day 4 although the experimental conditions (concentration of stimulus, duration of experiments, etc.) were kept constant. This reflects on the variability inherent in the characteristic signals of a same plant species which is used for experiments on multiple occasions.

5 Conclusions and Future works

In this exploration, there was a two-fold target – first, to see how different plant species react to external stimuli applied over multiple occasions and secondly once the background signal of a plant is separated from a post-stimulus signal, is it possible to discriminate between the applied stimuli.

Throughout this exploration, each plant species was considered individually, primarily aimed at understanding their individual behaviours to external stimuli. Therefore as a first step, a few statistical tests like analysis of variance (ANOVA), Wilcoxon ranksum and *t*-test were performed and it was established that background signals of a plant could be clearly distinguished from post-stimulus signals. Once this was established, the next task was to discriminate the plant signals after the application of the stimulus (post-stimulus). The statistical tests and histogram plots of the features considered for the post-stimulus signals showed that the responses of all the four types of plants to either of the two stimuli were not distinguishable. Hence a pattern recognition based methodology was adopted to classify these two post-stimulus signals (as a result of acid and salt application) for each individual plant species.

The results from the clustering based exploration were more promising than the linear decision boundary based classifiers, however the results reflected the variability inherent within the plant response to the same stimuli over multiple occasions.

Hence as a future work, a rigorous experimental protocol involving different plants performed in a controlled environment is required to ensure minimum external variations. A large dataset from a group of plants could be studied using the techniques mentioned here considering individual plants. Once a successful recognition model is determined for each plant species which can recognise the application of different stimulus to the plants, this information can be clubbed together to formulate a global classifier which shares a common feature library and standardized classification methodology to distinguish external stimulus being applied to a network of plants.

References

[1] J. L. Semmlow, *Biomedical and Medical Image Processing*, 2nd ed. CRC Press, 2008.

[2] S. Theodoridis, A. Pikrakis, K. Koutroumbas, and D. Cavouras, *Pattern Recognition*, Fourth. Academic Press, 2009.

[3] R. Gutierrez-Osuna, "Lecture 13: Validation," *Retrieved February*, vol. 28, p. 2007, 2006.

[4] R. T. T. Hastie and J. Friedman, *The Elements of Statistical Learning*. Springer, 2009.

[5] D. Biswas, A. Cranny, N. Gupta, K. Maharatna, and S. Ortmann, "Recognition of Elementary Upper Limb Movements in an Activity of Daily Living using Data from Wrist Mounted Accelerometers," in *Health Informatics (ICHI), 2014 IEEE-Computer Society International Conference on*, 2014, pp. 232–237.