



eNOSE calibration

FOR LUNG CANCER DETECTION

Andrea Merlina, Valerio Bruschi

eNose

1. Statistics
2. Objectives
3. Data contextualization
 1. Outliers
4. Data Analysis
 - a) Intro
 - b) LDA
 - c) Mahalanobis
 - d) PCA-DA
 - e) Feature Selection
 - f) MLR
5. Conclusions

Statistics

In 2008 approximately 12.7 million cancers were diagnosed

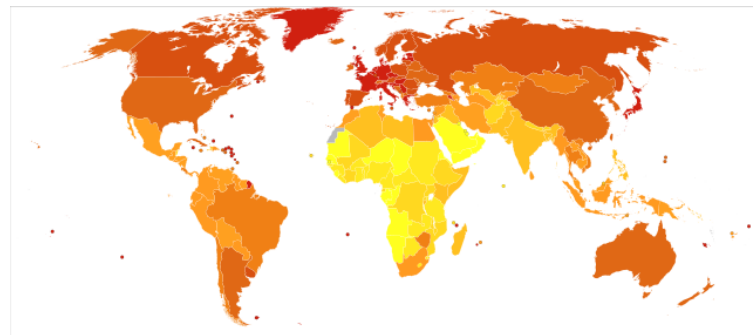
In 2010 nearly 7.98 million people died

Cancers account for approximately **13% of deaths**

Most common type:

- lung cancer (1,4 million deaths)
- stomach cancer (740.000)
- liver cancer (700.000)

Invasive cancer are the leading cause of death in the developed world and the second leading in the developing world



eNose

1. Statistics
2. Objectives
3. Data contextualization
 1. Outliers
4. Data Analysis
 - a) Intro
 - b) LDA
 - c) Mahalanobis
 - d) PCA-DA
 - e) Feature Selection
 - f) MLR
5. Conclusions

Objectives

Composition of the breath of patients with lung cancer contains information that could be used to detect the disease

Breath samples were collected and analyzed by two electronic noses

Two goals:

- ✓ Instrument calibration using a set of key compounds



distinguish **healthy** and **ill** patients

- ✓ Quality assurance of eNose data in the medical setting



Mapping between instruments through the calculation of a **model**

eNose

1. Statistics
2. Objectives
3. Data contextualization
 1. Outliers
4. Data Analysis
 - a) Intro
 - b) LDA
 - c) Mahalanobis
 - d) PCA-DA
 - e) Feature Selection
 - f) MLR
5. Conclusions

Data contextualization

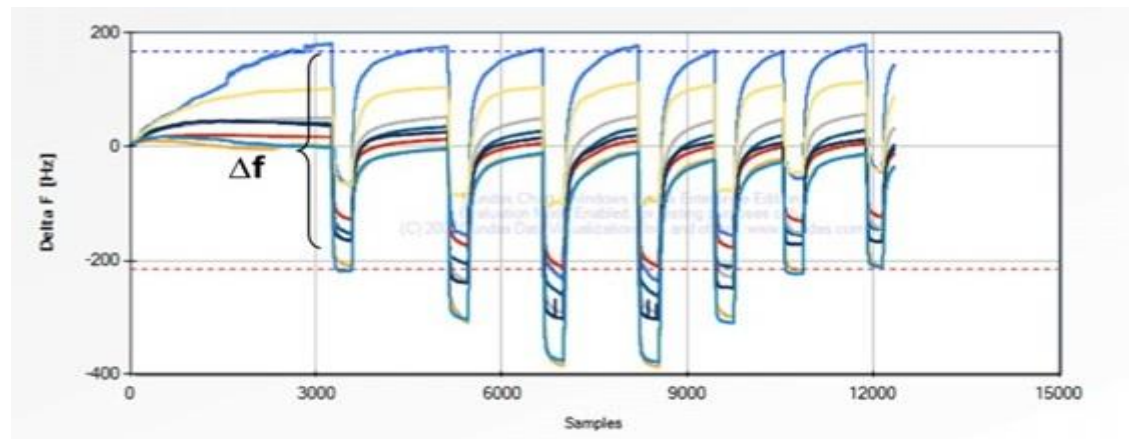
Two datasets obtained as measures from two electronic noses

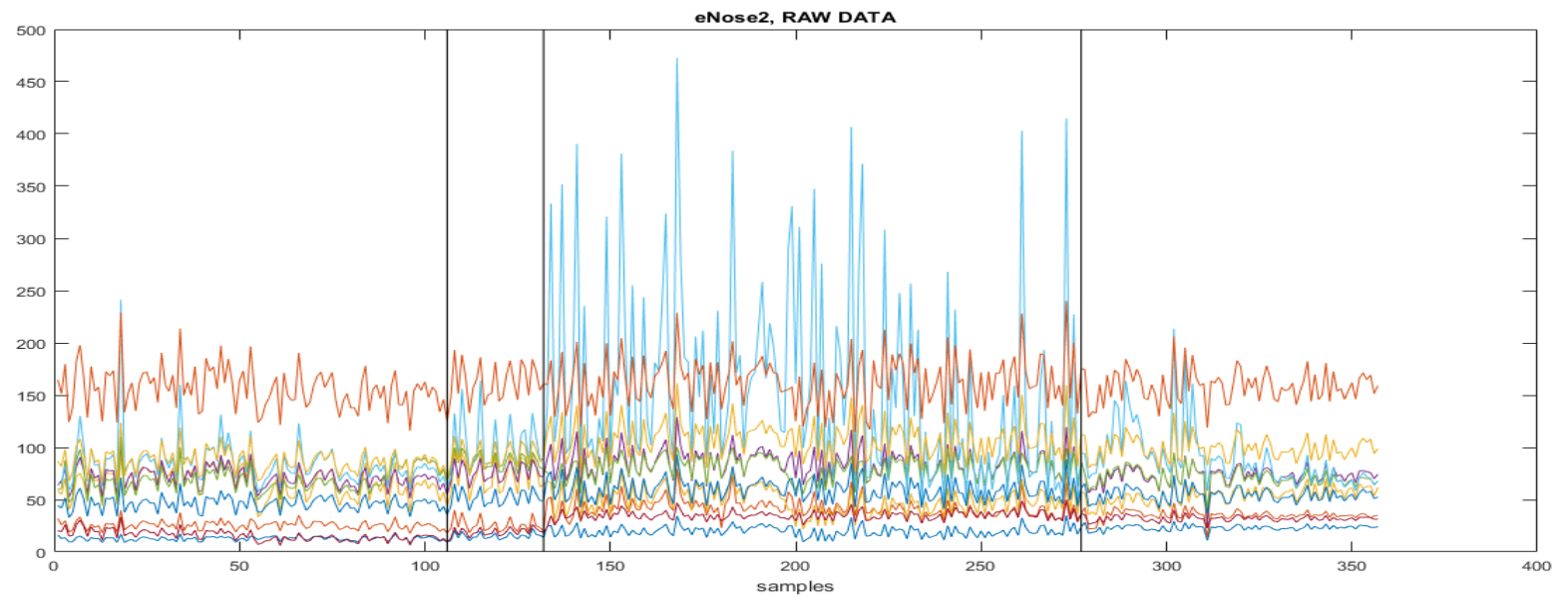
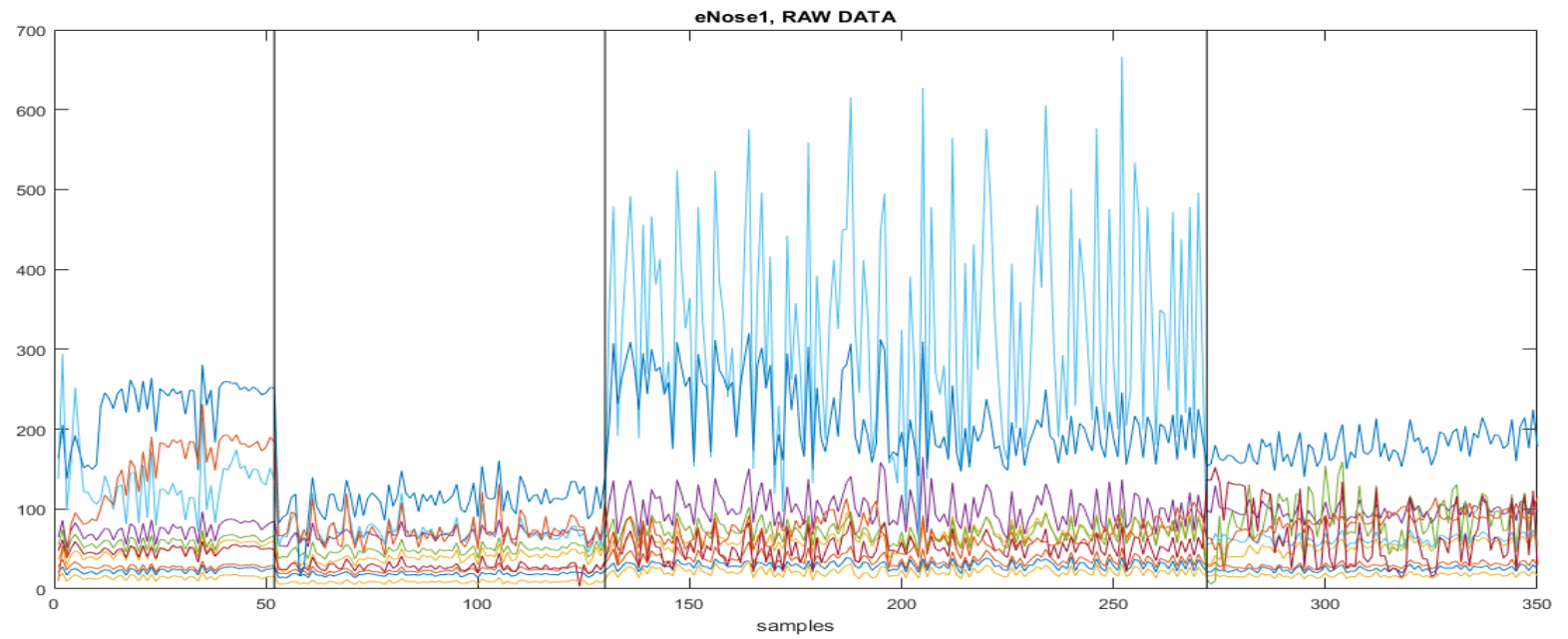
- Cyranose
- ROTV

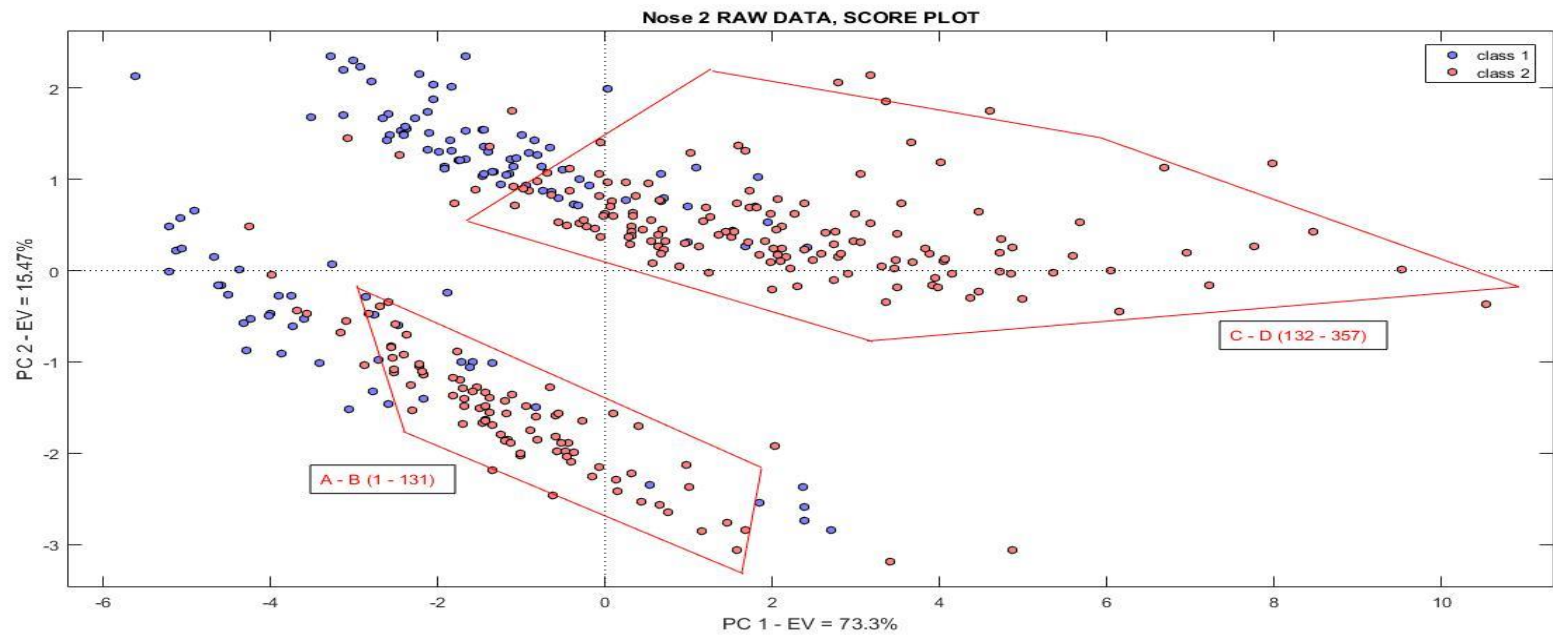
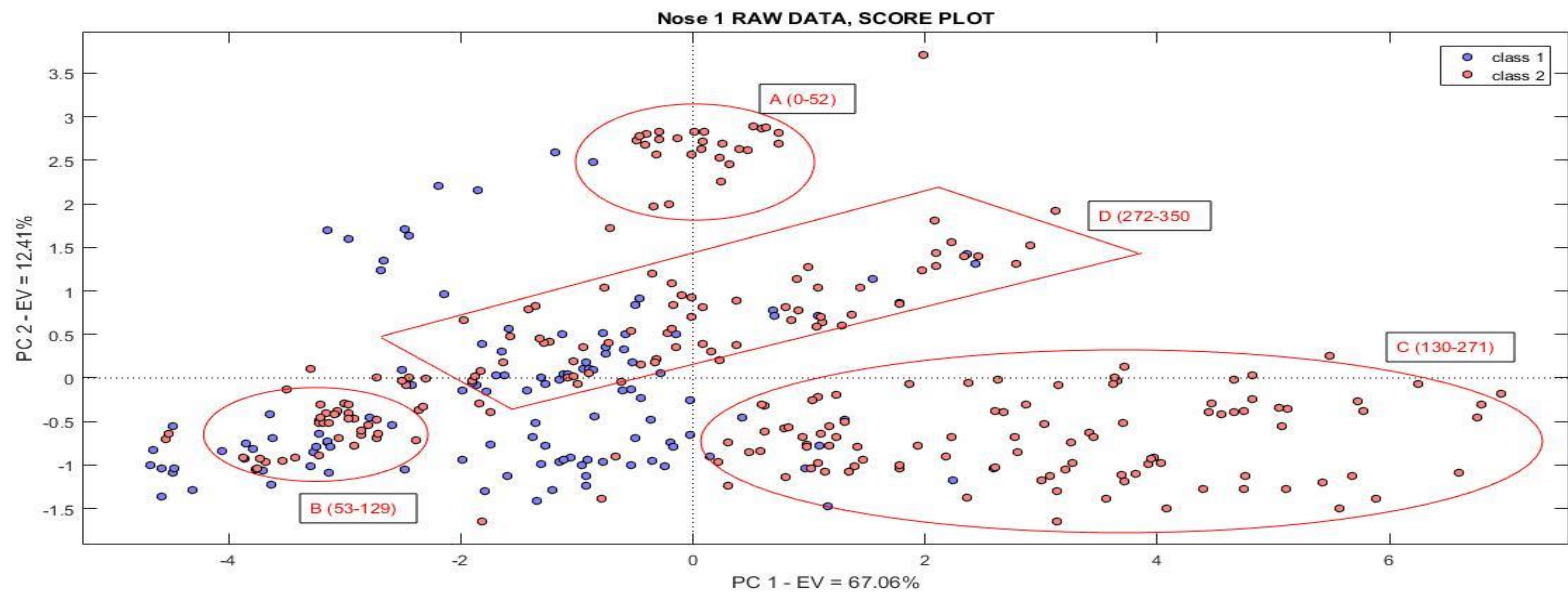
In a medical environment

≈ 350 samples/dataset

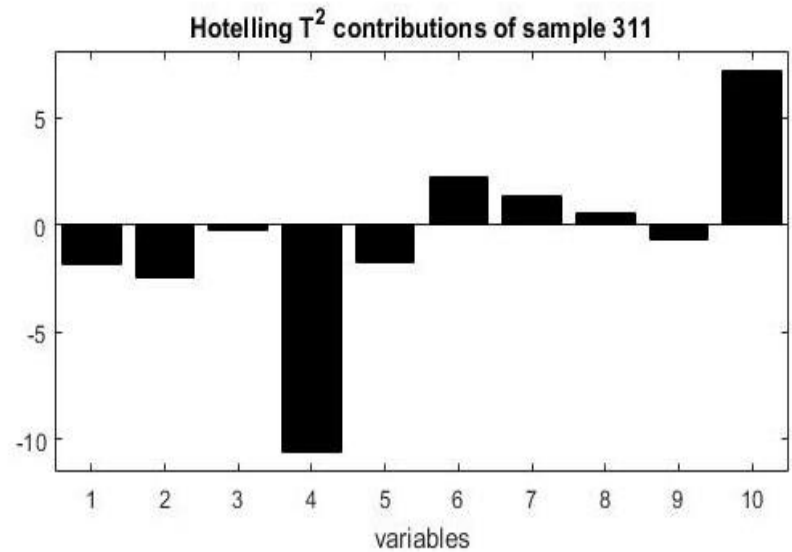
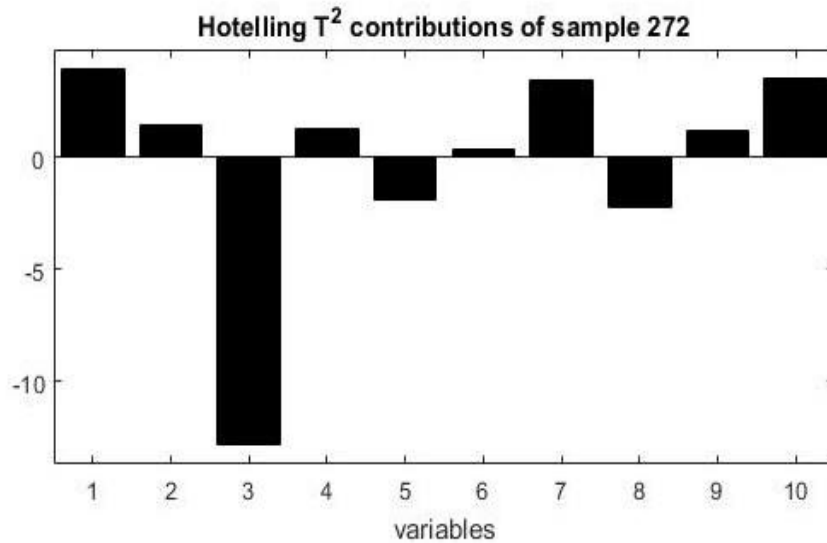
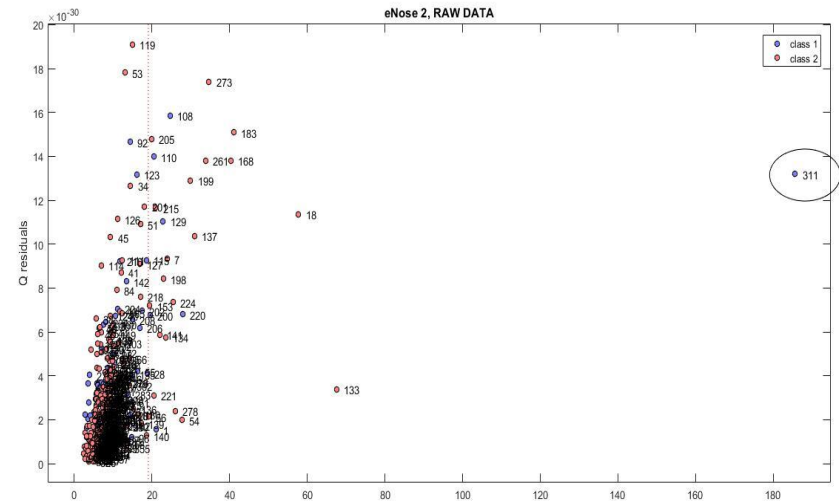
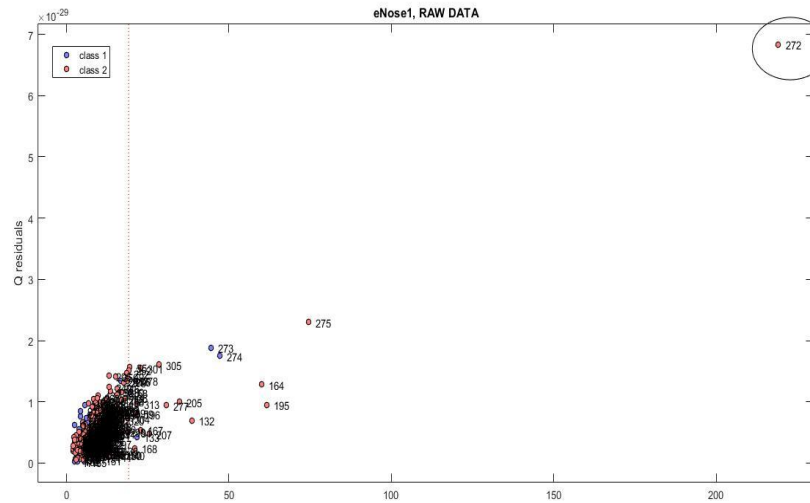
10 features ↔ 10 gas non-selective sensors







Outliers



eNose

1. Statistics
2. Objectives
3. Data contextualization
 1. Outliers
4. Data Analysis
 - a) Intro
 - b) LDA
 - c) Mahalanobis
 - d) PCA-DA
 - e) Feature Selection
 - f) MLR
5. Conclusions

Data Analysis

Using raw data to develop an inter instrumental model:

eNose 1 as training and eNose 2 as test results:

- **LDA:** everybody is classified as **ill**
- **PCA – DA:** everybody is classified as **healthy**
- **Mahalanobis:**

Classes	Classification			Classification
		Class #1	Class #2	Error (%)
	Class #1	45	74	62,18
	Class #2	123	115	51,68
		Accuracy		44,82



eNose

1. Statistics
2. Objectives
3. Data contextualization
 1. Outliers
4. Data Analysis
 - a) Intro
 - b) LDA
 - c) Mahalanobis
 - d) PCA-DA
 - e) Feature Selection
 - f) MLR
5. Conclusions

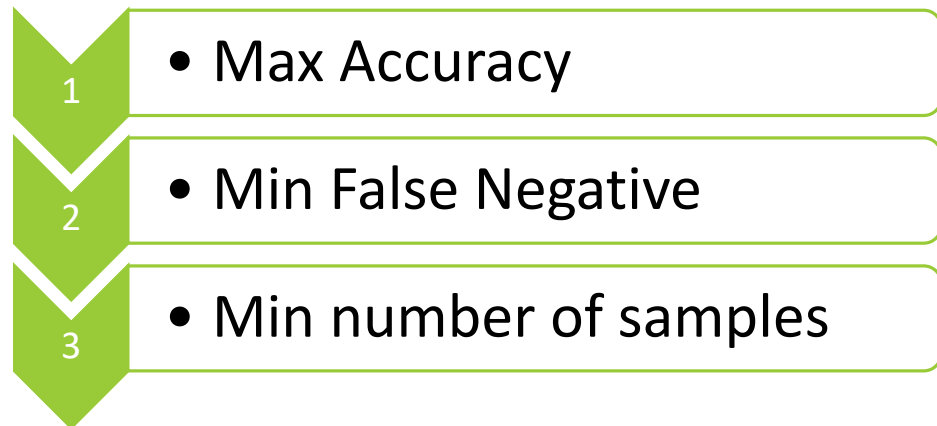
Data Analysis: Intro

Hypotesis:

Normalization of a trunk using a **baseline** of:

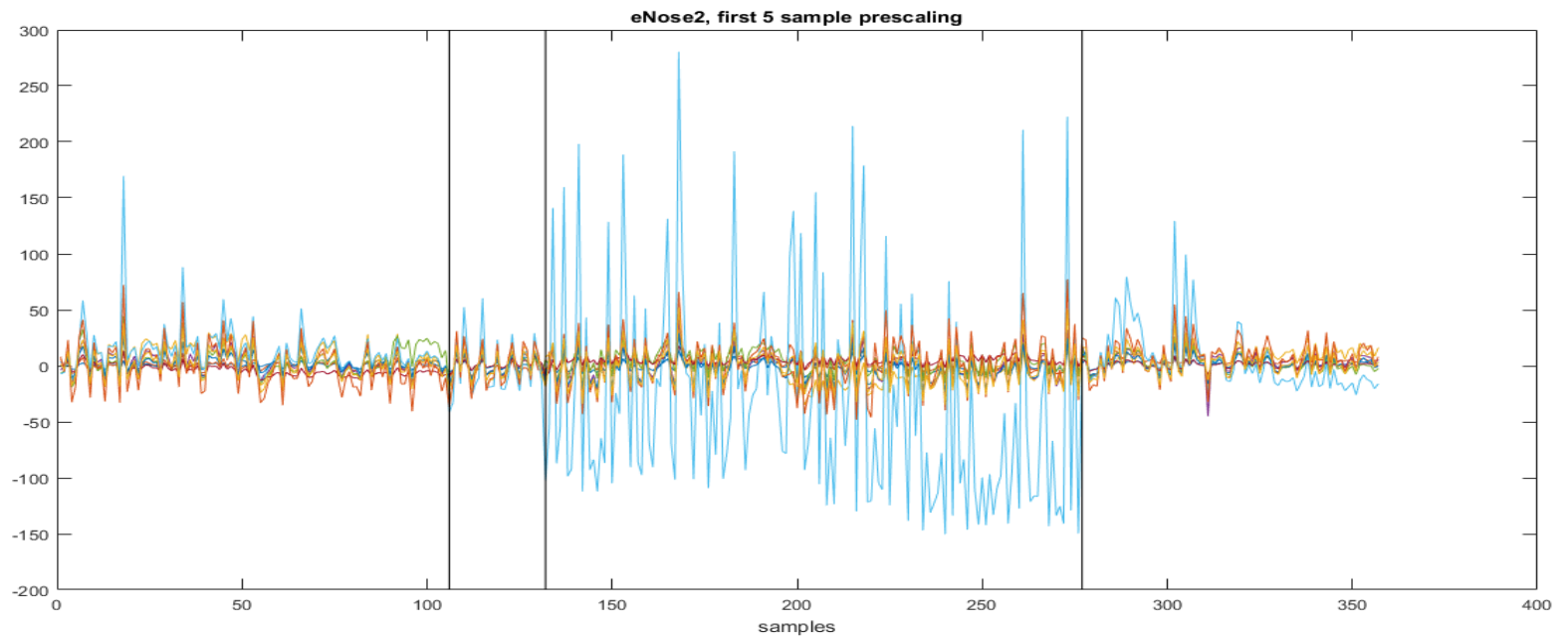
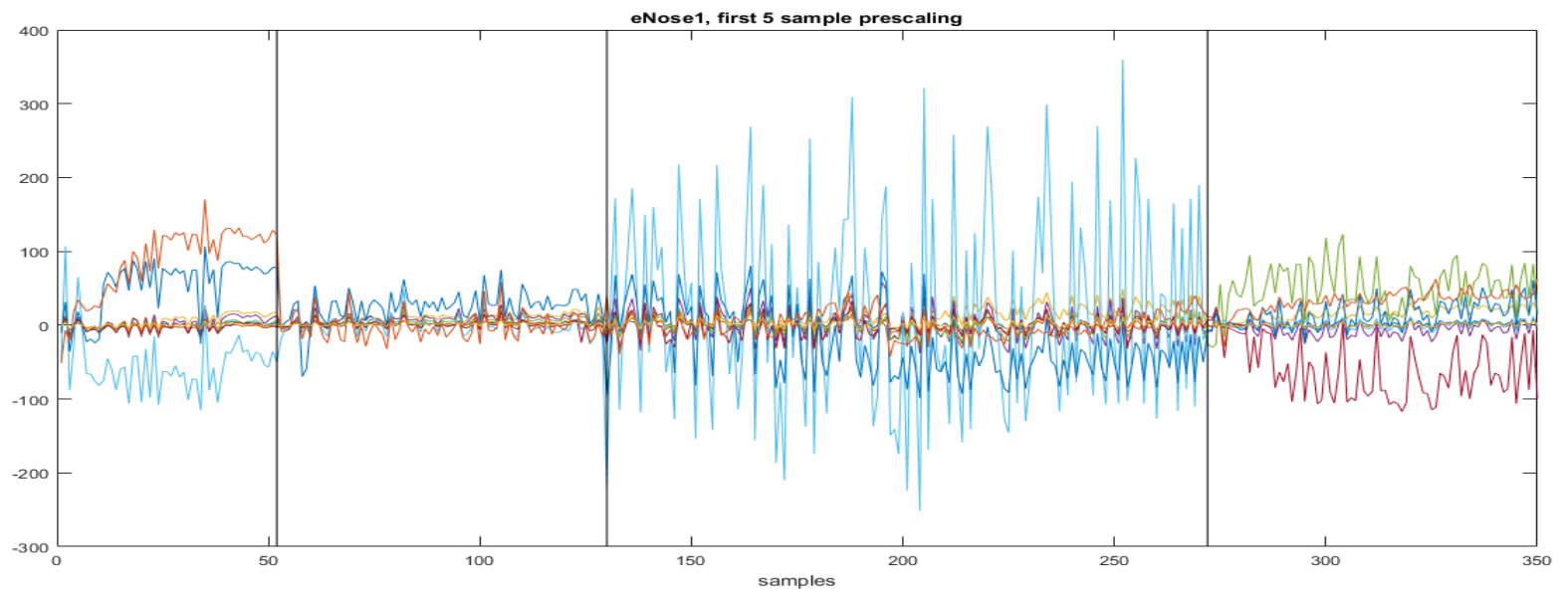
- 1 sample
- 5 samples (about 15%* error for class 1)
- 10 samples (about 15%* error for class 1)
- entire chunk (25%*)

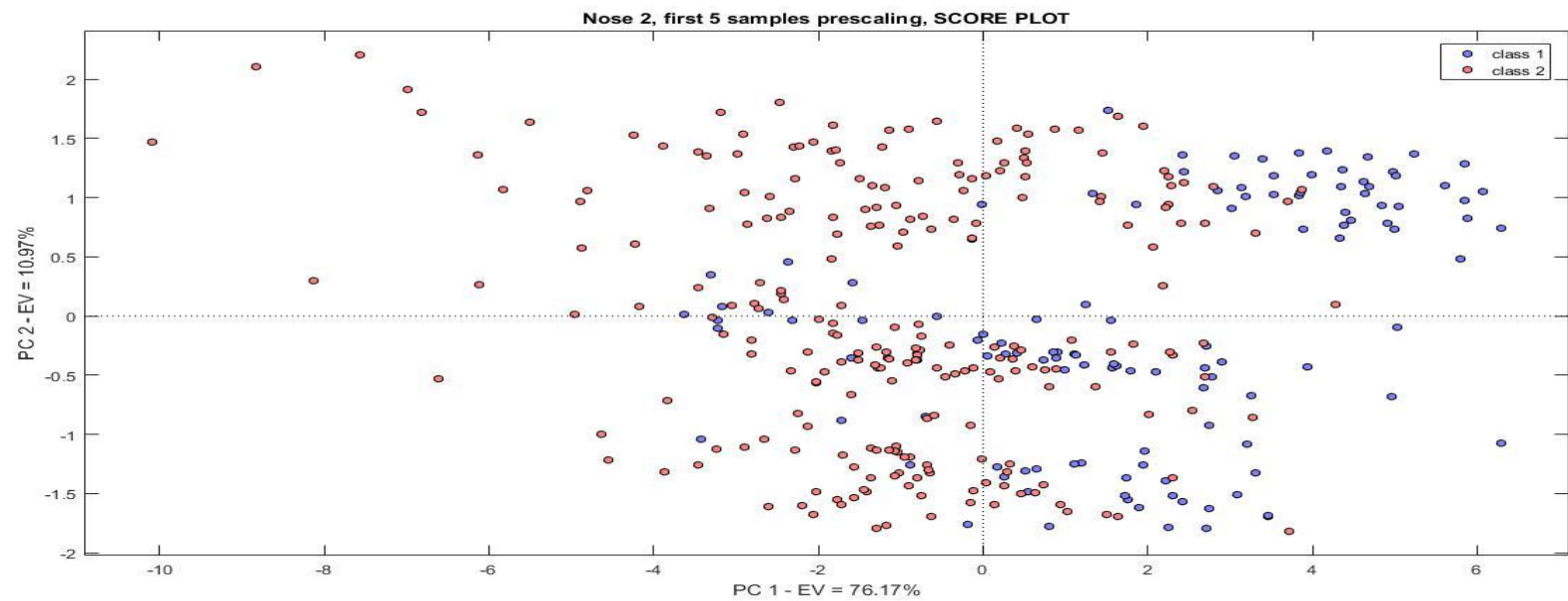
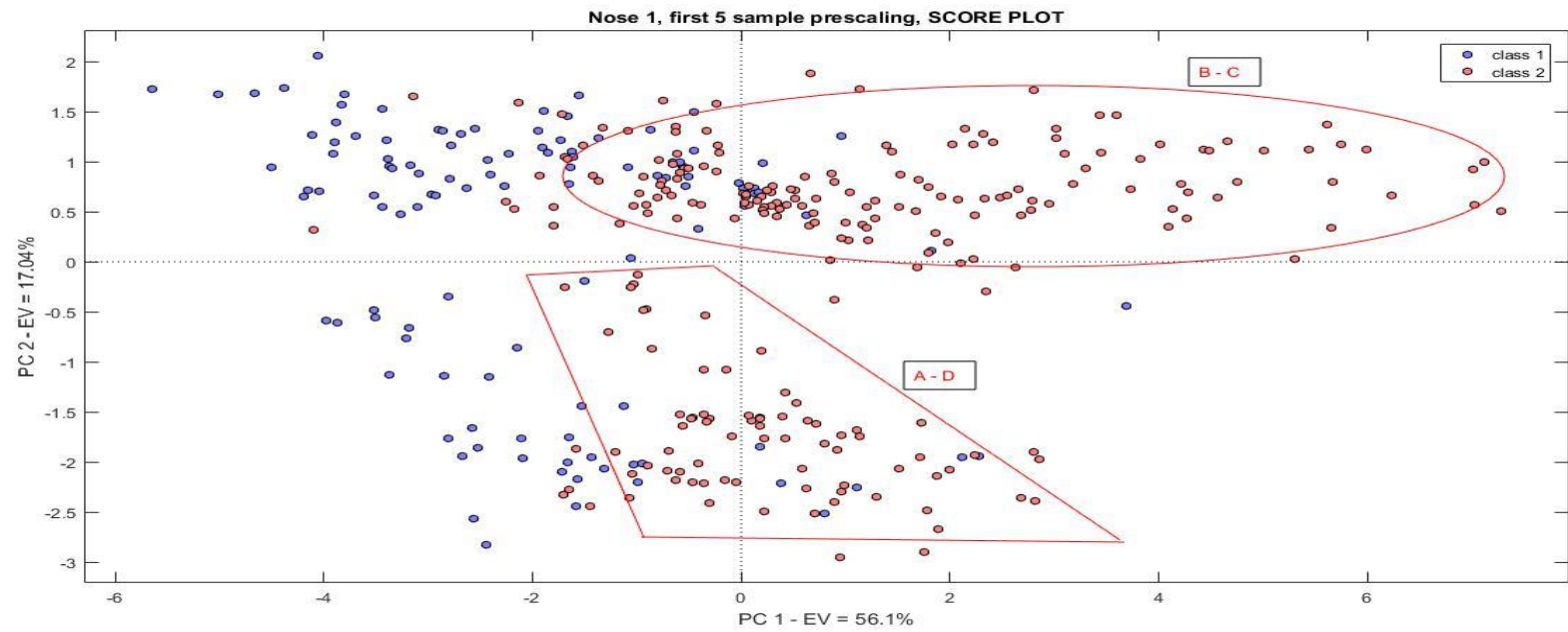
In order to minimize memory effects and maximize reproducibility



Best tradeoff performances/easy of calibration

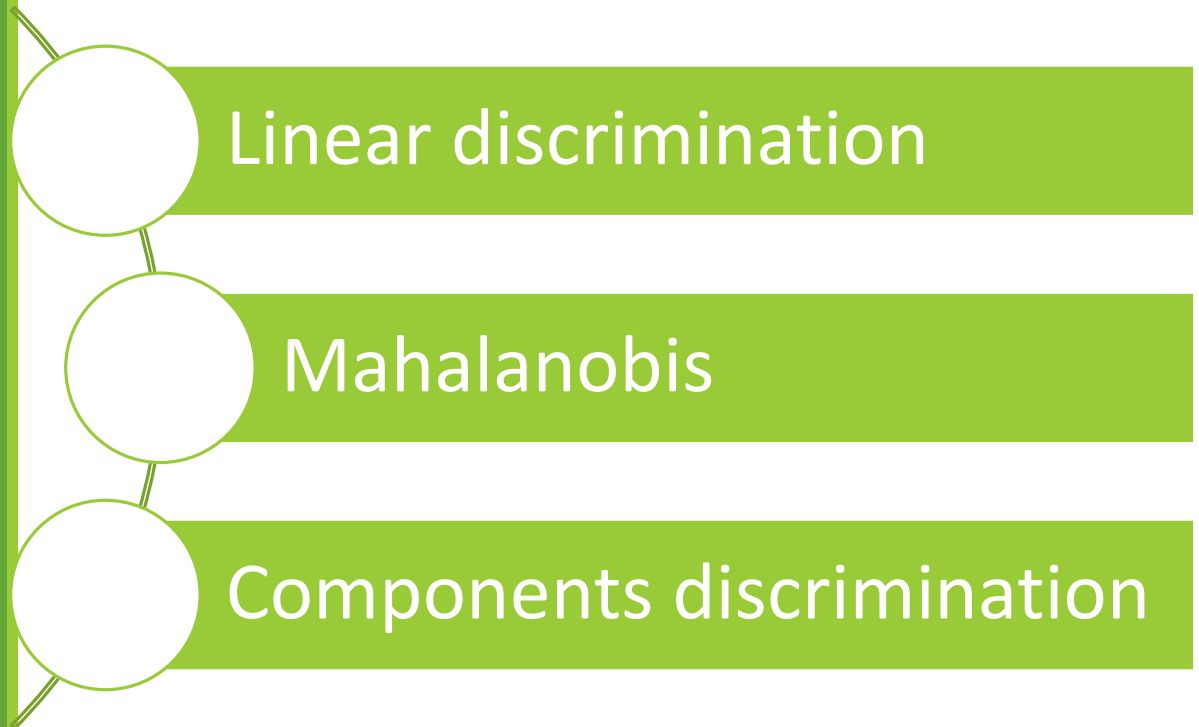
* With our best classification method





1. Statistics
2. Objectives
3. Data contextualization
 1. Outliers
4. Data Analysis
 - a) Intro
 - b) LDA
 - c) Mahalanobis
 - d) PCA-DA
 - e) Feature Selection
 - f) MLR
5. Conclusions

Data Analysis: Intro

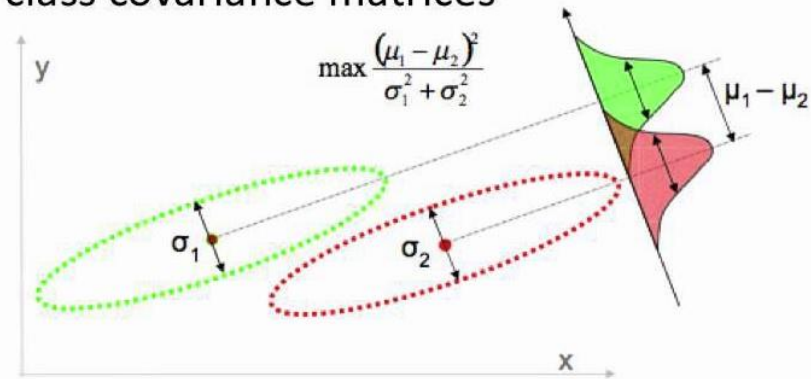


eNose

1. Statistics
2. Objectives
3. Data contextualization
 1. Outliers
4. Data Analysis
 - a) Intro
 - b) LDA**
 - c) Mahalanobis
 - d) PCA-DA
 - e) Feature Selection
 - f) MLR
5. Conclusions

Data Analysis: LDA

- LDA: pick a new dimension that gives:
 - maximum separation between means of projected classes
 - minimum variance within each projected class
- Solution: eigenvectors based on between-class and within-class covariance matrices



Resulting confusion matrix:

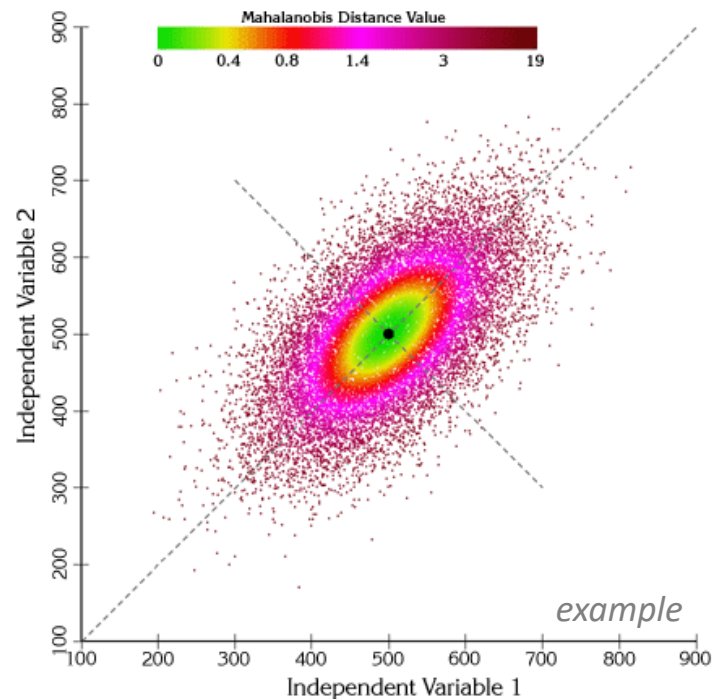
Classes	Classification			Classification Error (%)
		Class #1	Class #2	
	Class #1	89	29	24,58
	Class #2	99	139	41,60
		Accuracy		64,04

eNose

1. Statistics
2. Objectives
3. Data contextualization
 1. Outliers
4. Data Analysis
 - a) Intro
 - b) LDA
 - c) Mahalanobis
 - d) PCA-DA
 - e) Feature Selection
 - f) MLR
5. Conclusions

Data Analysis: Mahalanobis

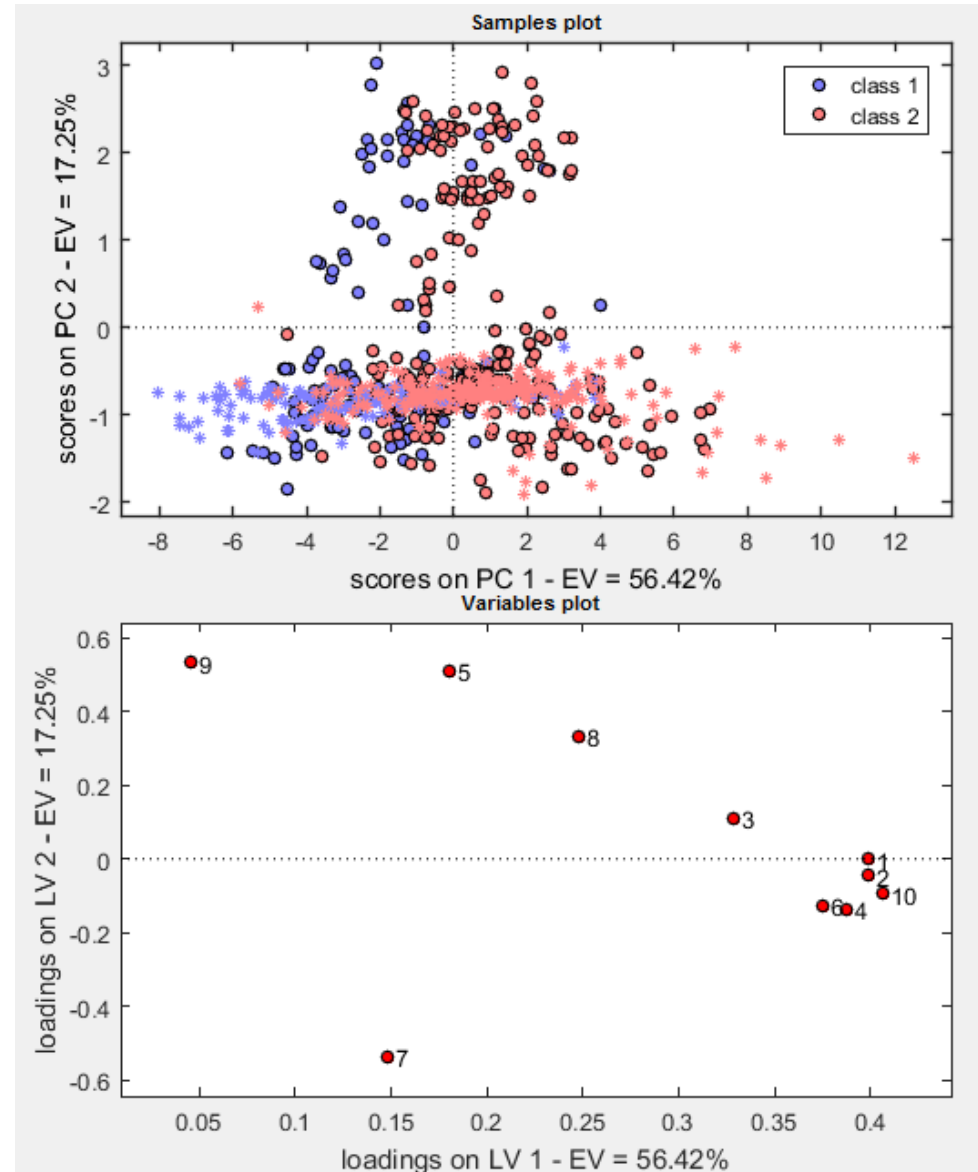
- Sample points are distributed about the center of mass in an ellipsoidal manner.
- the probability of the test point to belong to the set depends not only on the **distance** from the center of mass, but also on the **direction**.
- ***Result: everybody is classified as healthy***



eNose

1. Statistics
2. Objectives
3. Data contextualization
 1. Outliers
4. Data Analysis
 - a) Intro
 - b) LDA
 - c) Mahalanobis
 - d) PCA-DA
 - e) Feature Selection
 - f) MLR
5. Conclusions

Data Analysis: PCA - DA



eNose

1. Statistics
2. Objectives
3. Data contextualization
 1. Outliers
4. Data Analysis
 - a) Intro
 - b) LDA
 - c) Mahalanobis
 - d) PCA-DA
 - e) Feature Selection
 - f) MLR
5. Conclusions

Data Analysis: PCA - DA



- ***Optimum solution***



- ***Maximum accuracy***



- ***Minimum False Negative***

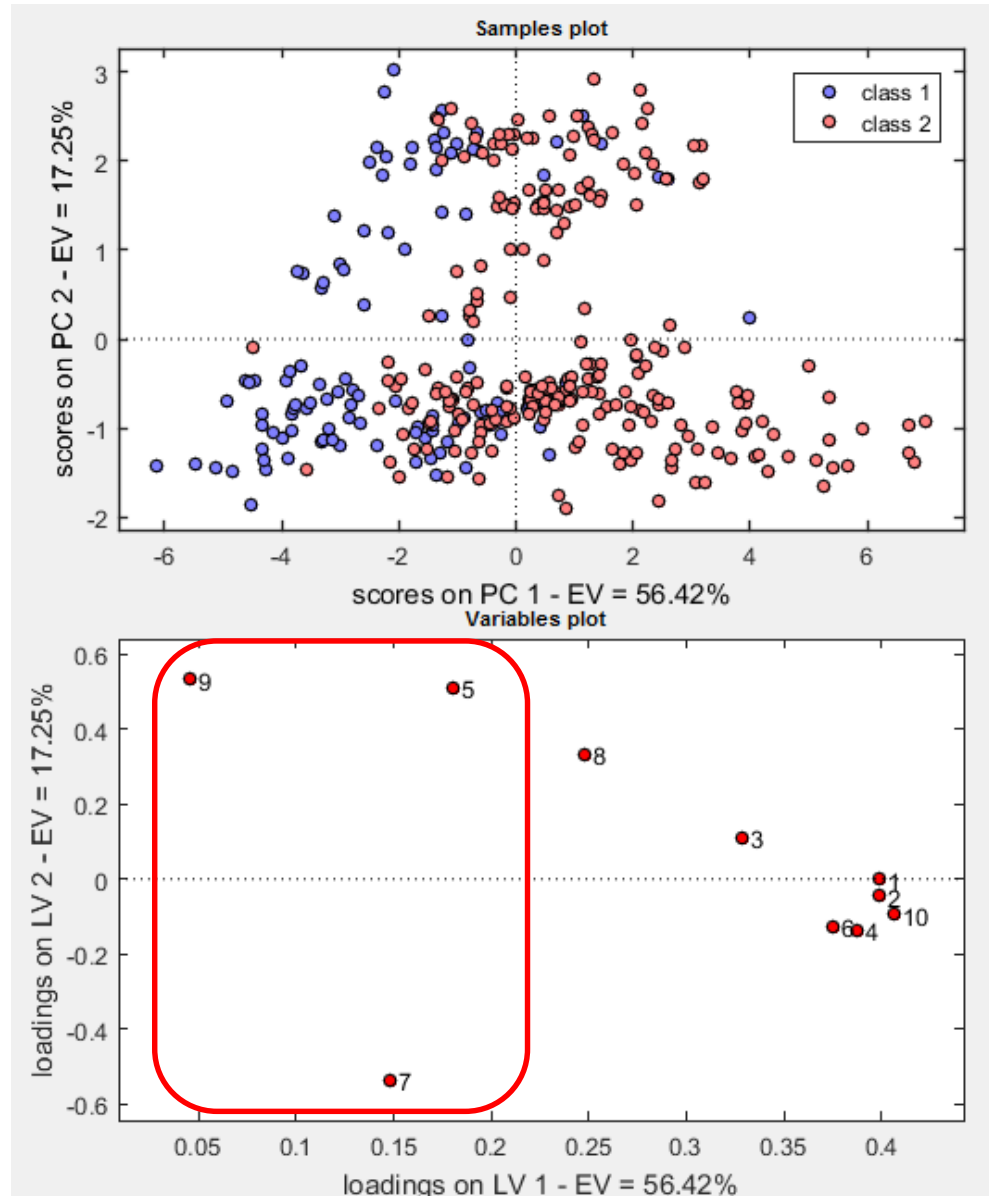
	Classification			Classification
		Class #1	Class #2	Error (%)
	Classes			
	Class #1	100	18	15,25
	Class #2	59	179	24,79
		Accuracy		78,37

eNose

1. Statistics
2. Objectives
3. Data contextualization
 1. Outliers
4. Data Analysis
 - a) Intro
 - b) LDA
 - c) Mahalanobis
 - d) PCA-DA
 - e) Feature Selection
 - f) MLR
5. Conclusions

Data Analysis: Feature Selection

Based on loadings of PCA



eNose

1. Statistics
2. Objectives
3. Data contextualization
 1. Outliers
4. Data Analysis
 - a) Intro
 - b) LDA
 - c) Mahalanobis
 - d) PCA-DA
 - e) Feature Selection
 - f) MLR
5. Conclusions

Data Analysis: Feature Selection

Based on loadings of PCA

PCA - DA

*Close to results
without feature
selection !!*

	Classification			Classification
Classes		Class #1	Class #2	Error (%)
	Class #1	101	17	14,41
	Class #2	65	173	27,31
		Accuracy		76,97

Mahalanobis

	Classification			Classification
Classes		Class #1	Class #2	Error (%)
	Class #1	32	86	72,88
	Class #2	22	216	9,24
		Accuracy		69,66

LDA

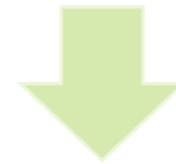
	Classification			Classification
Classes		Class #1	Class #2	Error (%)
	Class #1	90	28	23,73
	Class #2	95	143	39,92
		Accuracy		65,45

eNose

1. Statistics
2. Objectives
3. Data contextualization
 1. Outliers
4. Data Analysis
 - a) Intro
 - b) LDA
 - c) Mahalanobis
 - d) PCA-DA
 - e) Feature Selection
 - f) MLR
5. Conclusions

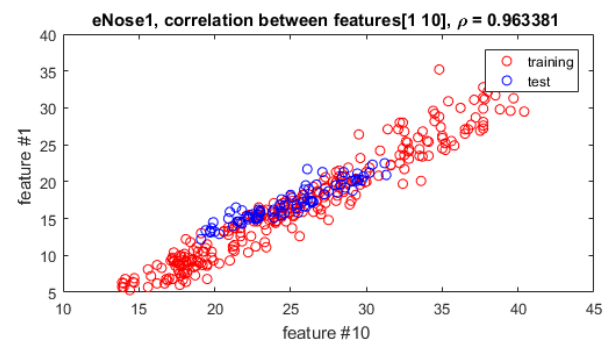
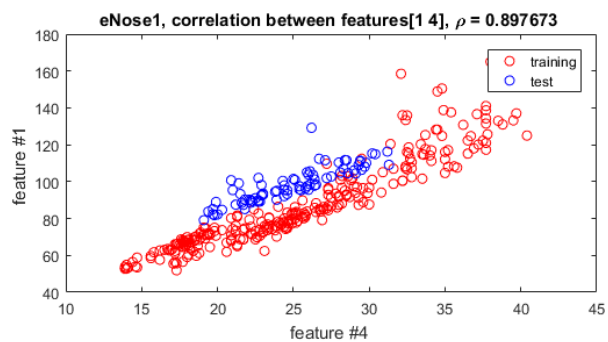
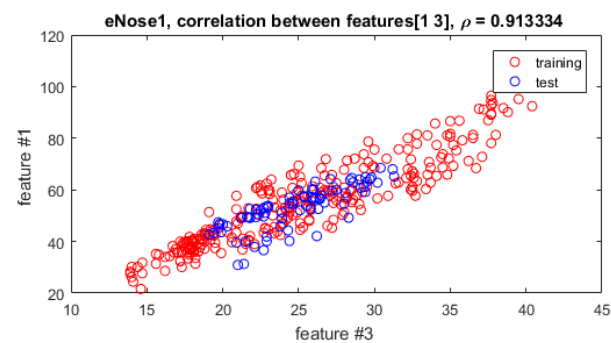
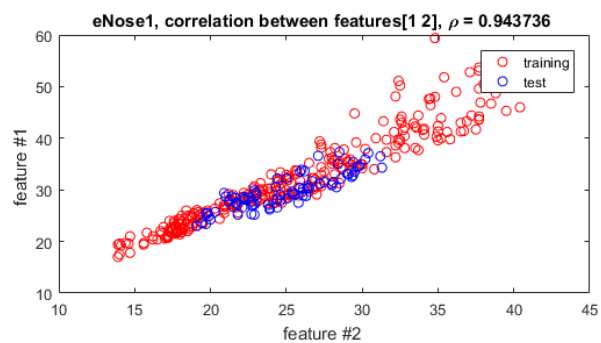
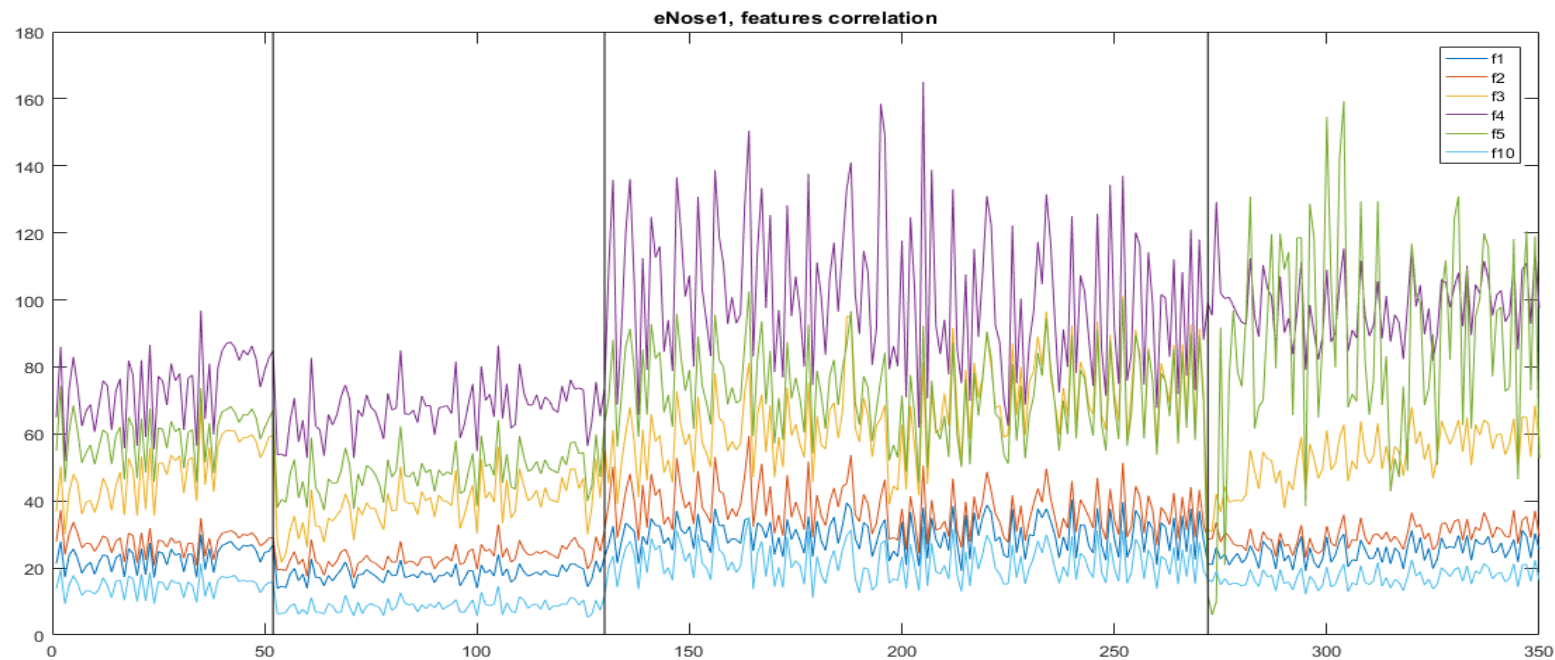
Data Analysis: MLR (intro)

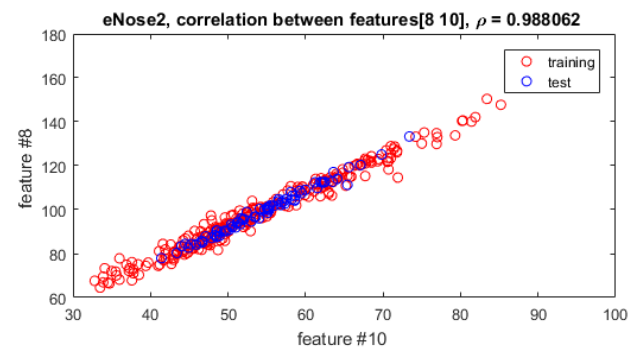
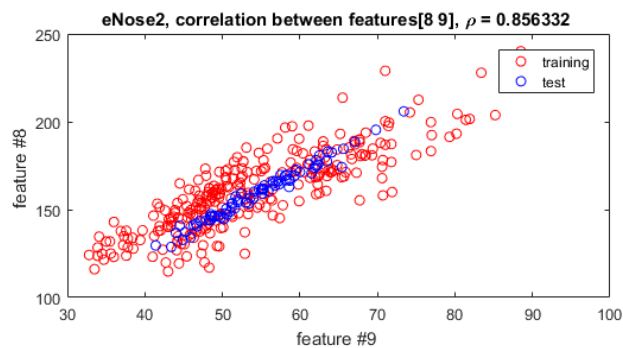
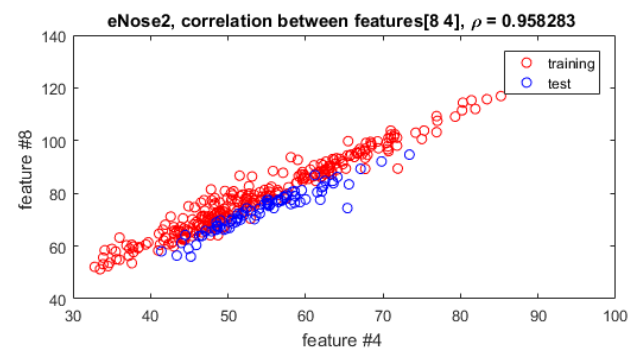
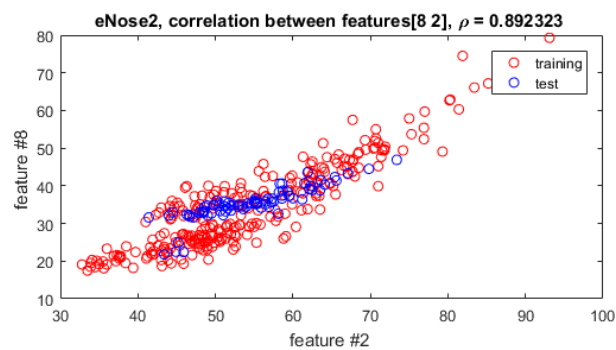
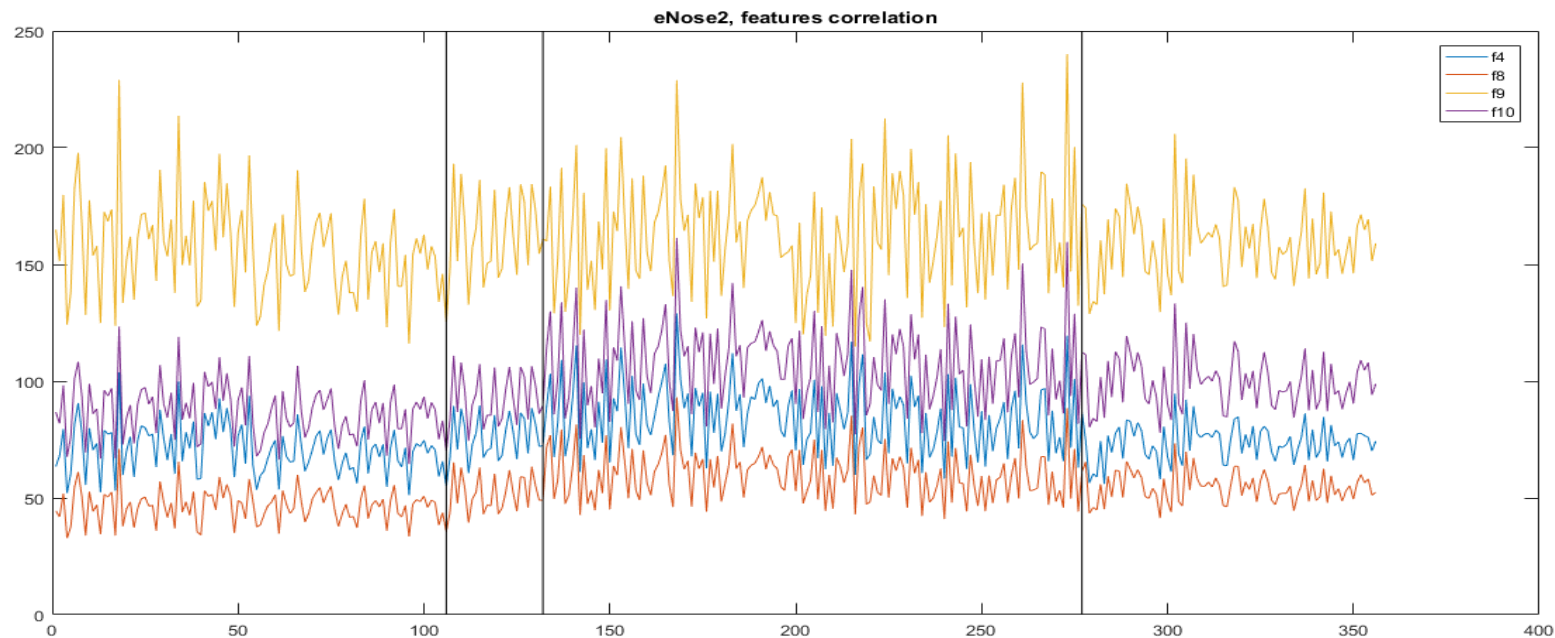
We noticed correlation between features on the same eNose.



❖ *Can we find a regression model in order to reduce the number of features (sensors) used in the experiment?*

❖ *Is the model reliable during time?*





eNose

1. Statistics
2. Objectives
3. Data contextualization
 1. Outliers
4. Data Analysis
 - a) Intro
 - b) LDA
 - c) Mahalanobis
 - d) PCA-DA
 - e) Feature Selection
 - f) MLR
5. Conclusions

Data Analysis: MLR

Multiple linear regression attempts to model the relationship between variables by fitting a **linear** equation to observed data.

$$Y = XB + E$$

- Dataset was divided in training and test sets in a temporal way.

(first 3 days as training, the last one as testing)

$$B = X^+ \cdot Y$$

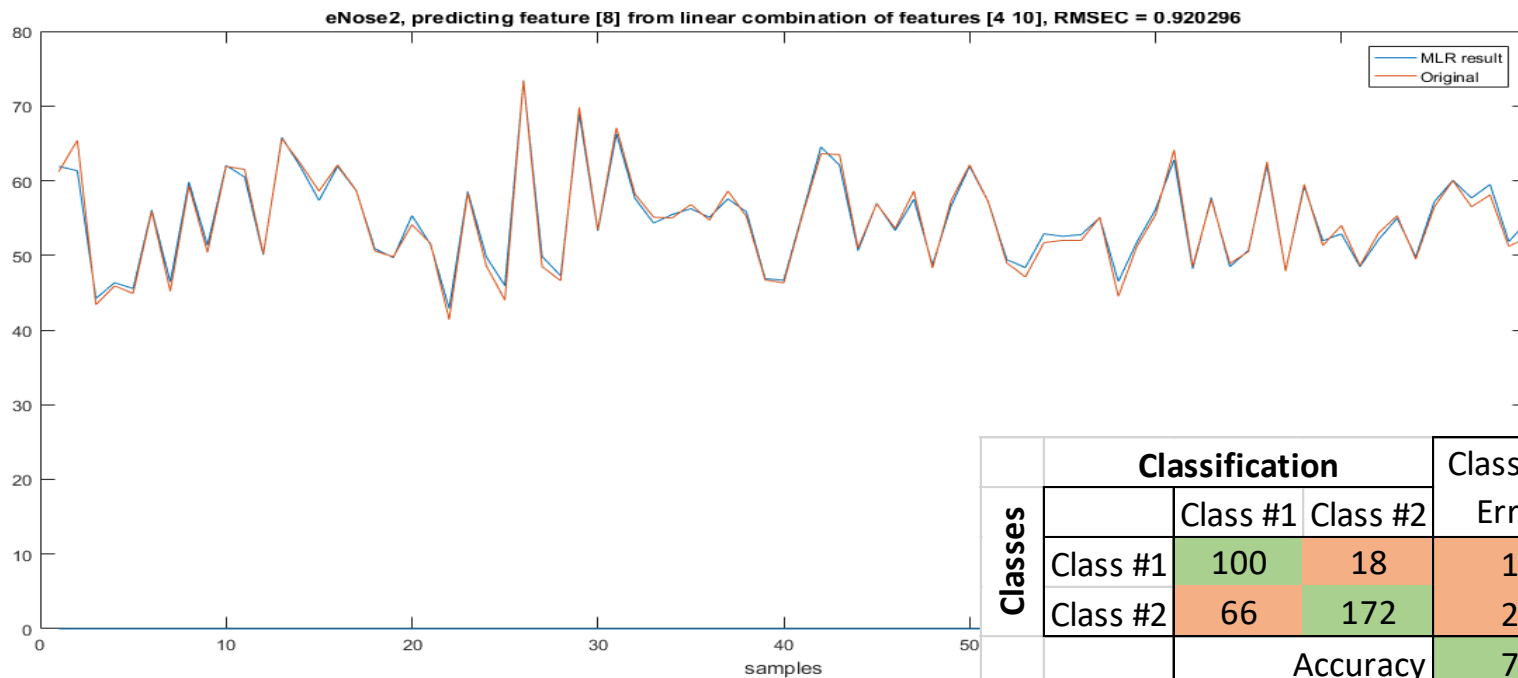
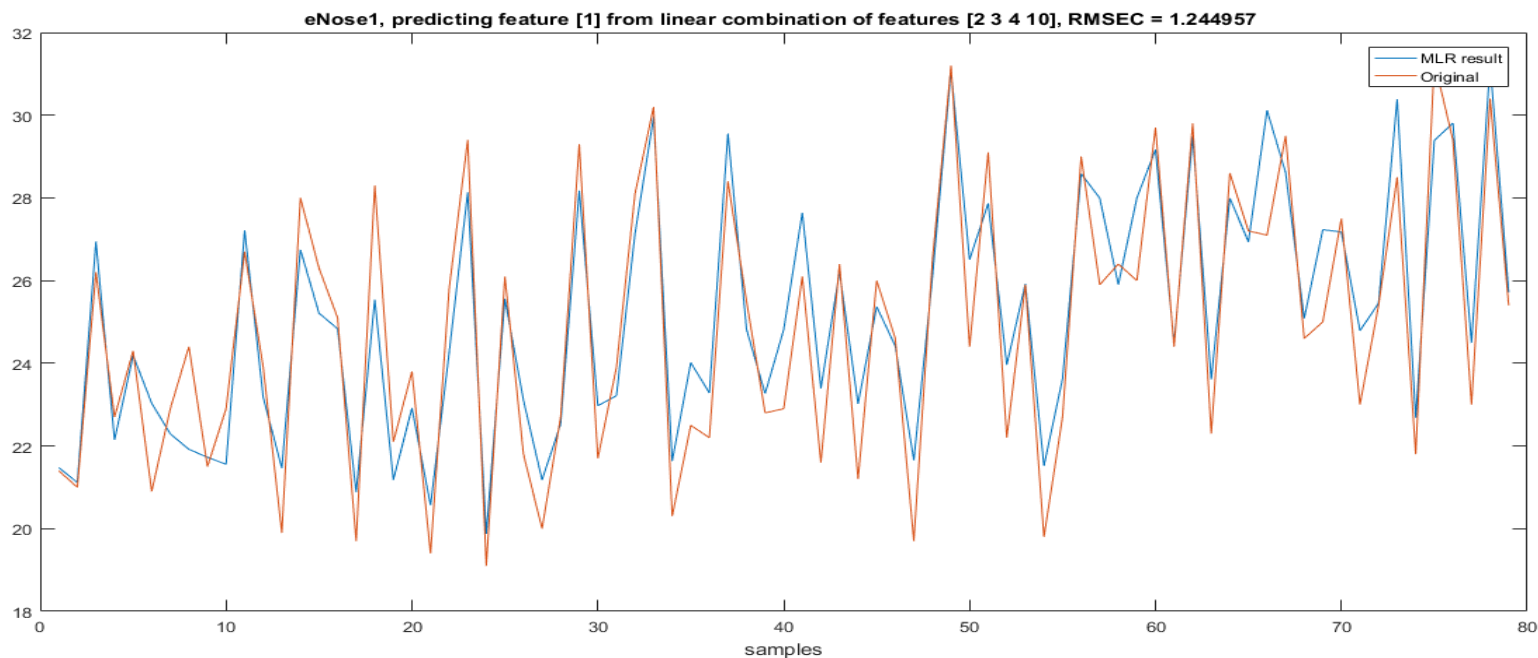
- RMSECV was computed for each feature predicted

$$RMSECV = \sqrt{\frac{PRESS}{N}}$$

$$\text{with } PRESS = \sum_i (y_i^{LS} - y_i)^2$$

MLR hypothesis: $e_y \approx e_x$.

Regression less robust

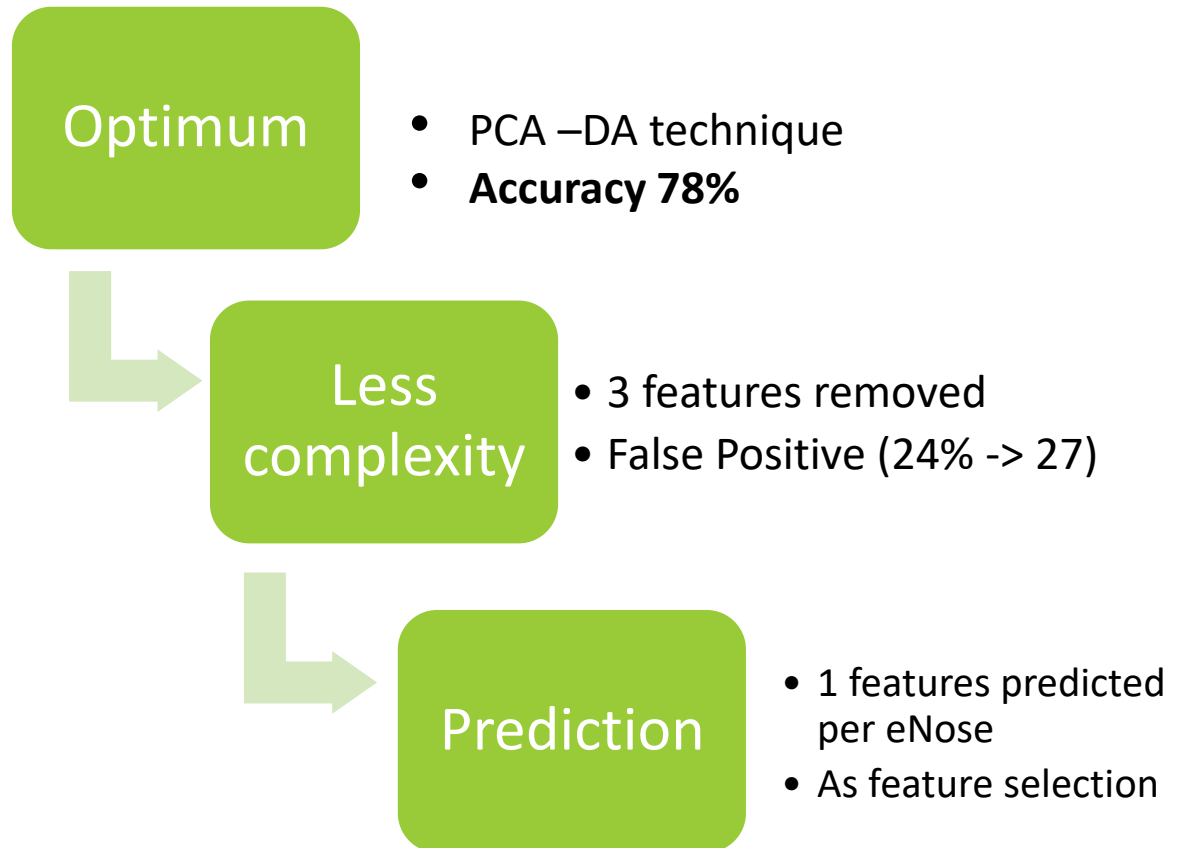


Classes	Classification			Classification
		Class #1	Class #2	Error (%)
	Class #1	100	18	15,25
	Class #2	66	172	27,73
	Accuracy			76,40

eNose

1. Statistics
2. Objectives
3. Data contextualization
 1. Outliers
4. Data Analysis
 - a) Intro
 - b) LDA
 - c) Mahalanobis
 - d) PCA-DA
 - e) Feature Selection
 - f) MLR
5. Conclusions

Conclusions



Thanks for the attention
