Methods for Big Data in Audiology

Marta Campi [1], Mareike Buhl[1], Gareth W. Peters[2]
Perrine Morvan[1,3], Catherine Boiteux[3], Hung Thai-Van [1]

[1] Hearing Institute, CERIAH - Paris, France

[2]University of California, Santa Barbara - Santa Barbara, US

[3]Amplifon - Paris, France

EFAS, Zagreb, May 24, 2024

Center for Research and Innovation in Human Audiology

Outline

1. Introduction to Machine Learning
2. Motivation & Research Questions
3. Methods
4. Results
5. Conclusions

Outline

What is Machine Learning?

Machine Learning (ML) is the field focusing on the development of algorithms, able to achieve a certain task (such as recognition, prediction, etc.). The algorithm implements a **mathematical model with unknown parameters, which should be learnt on the data**.

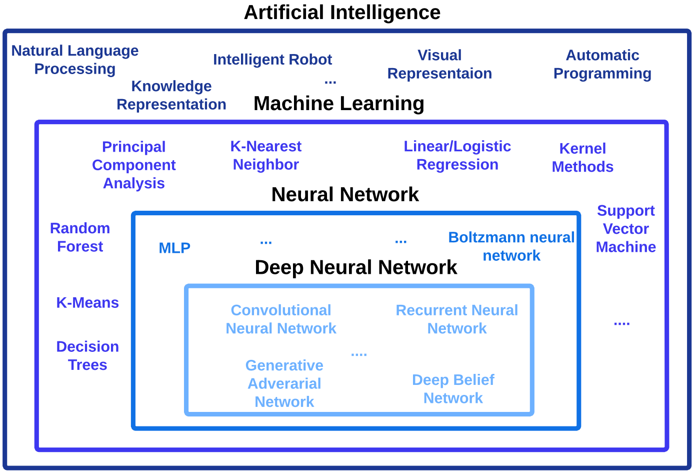
Formal definition by **Tom Mitchell**:
A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Examples

- Search engines (e.g. Google)
- Recommender systems (e.g. Netflix)
- Automatic translation (e.g. Google Translate)
- Speech understanding (e.g. Siri, Alexa)
- Game playing (e.g. AlphaGo)
- Personalized medicine

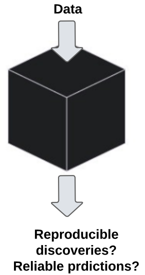
Introduction to Machine Learning

By considering the whole Artificial Intelligence world:



How to turn these ML models into reliable tool for audiological care?

- **Reproducibility**. Conclusion I draw today need to hold up tomorrow.
- **Reliability**. Users need to understand how the model make prediction.
- **Transparency**. Users needs to check for validity of the results given the assumptions.
- **Avoid Implicit Bias**. Users need to be able to check whether the model does not learn biases.
- **Interpretability**. Users need to interpret model decisions on a local and global level.
- **Coverage**. Users need to be able to compute their predictions confidence.
- **Discovery**. Users needs to distil insights/new knowledge learnt.
- **Parsimony**. Users need to ensure that the model adheres to the principle of parsimony, maintaining simplicity with a minimal number of parameters.
- **Expert Opinion**. Users need to validate results based on experts' opinion.



How to turn these ML models into reliable tool for audiological care?

- **Reproducibility**. Conclusion I draw today need to hold up tomorrow.
- **Reliability**. Users need to understand how the model make prediction.
- **Transparency**. Users needs to check for validity of the results given the assumptions.
- **Avoid Implicit Bias**. Users need to be able to check whether the model does not learn biases.
- **Interpretability**. Users need to interpret model decisions on a local and global level.
- **Coverage**. Users need to be able to compute their predictions confidence.
- **Discovery**. Users needs to distil insights/new knowledge learnt.
- **Parsimony**. Users need to ensure that the model adheres to the principle of parsimony, maintaining simplicity with a minimal number of parameters.
- **Expert Opinion**. Users need to validate results based on experts' opinion.

Outline

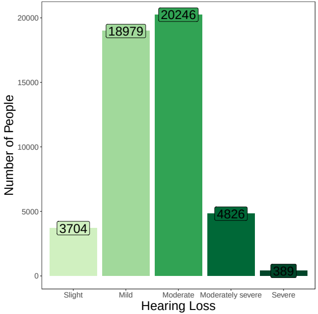
Motivation & Research Questions

Data: 24,072 adults, with symmetric hearing loss , age range between 40 to 90 (French Amplifon Database) for which we have: **Audiogram**, **Speech-in-quiet**, **Speech-in-noise**.

We partition this data set according to the Pure Tone Average (PTA) categories

Hearing Loss Category Definition

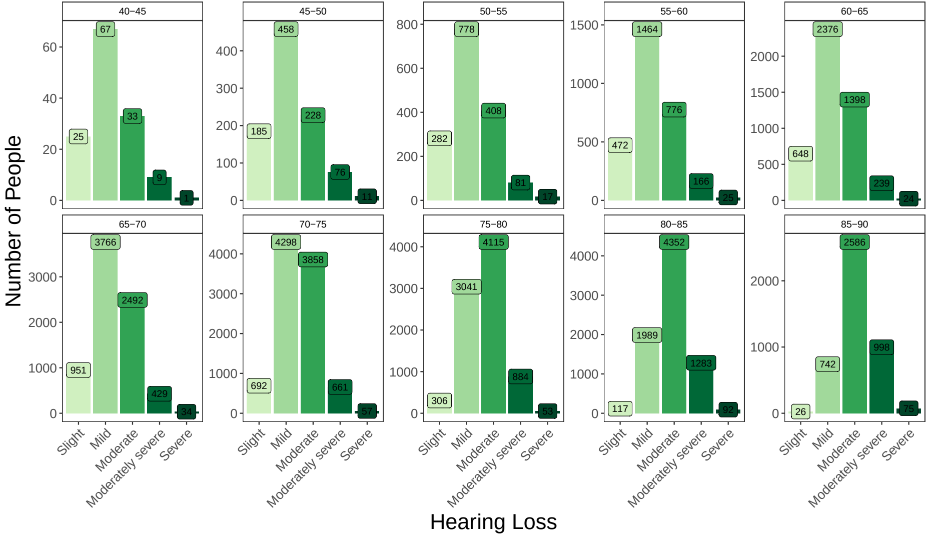
Degree of hearing loss	PTA range (dB HL)
Slight	16 to 25
Mild	26 to 40
Moderate	41 to 55
Moderately severe	56 to 70
Severe	71 to 90



Research Question 1) By considering the PTA categories, can we quantify how the audiogram and the speech tests characterise these hearing loss categories with ML solutions?

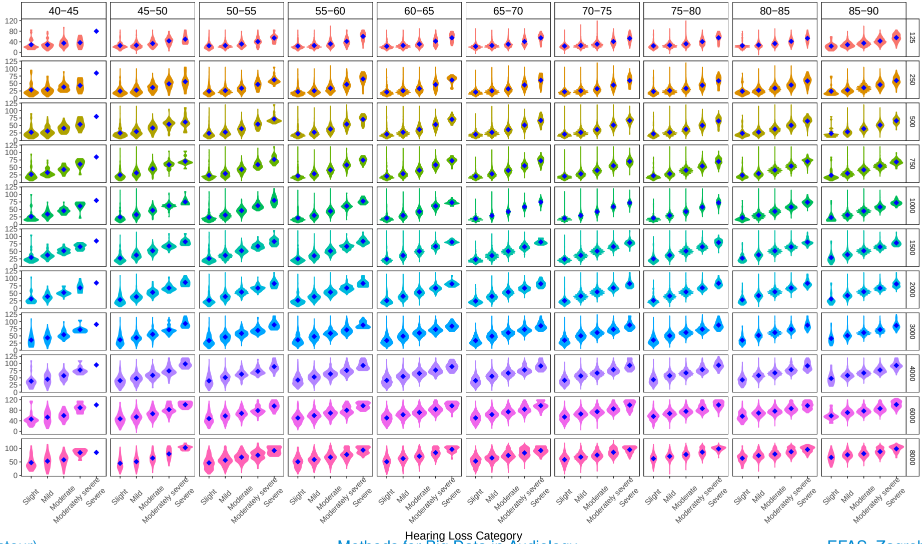
Motivation & Research Questions

If you one adds age grouping, then:



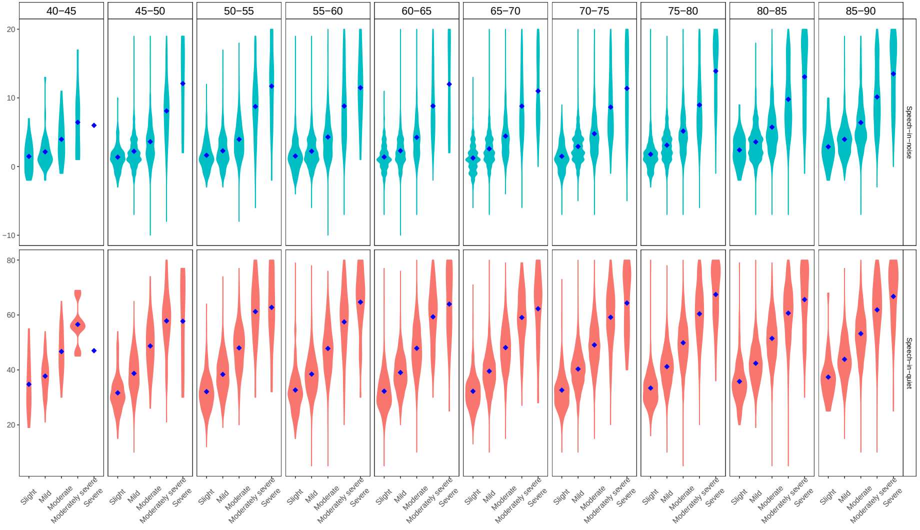
Motivation & Research Questions

If you one looks ad the distribution over age of the left audiogram, then:



Motivation & Research Questions

If you one looks ad the distribution over age of the speech tests, then:



Motivation & Research Questions

If we observe these plots, these represent **a complex dataset** characterised by **several parameters** as age, individual response to pure tone, individual response to speech (in quiet and in noise) corresponding to a more challenging scenario.

We put ourself in the perspective of classification, opposite to the one of regression.

What are standard practices in ML when such a complex dataset is analysed? Talk Goals.

1. Data Visualisation. **Talk Goal 1**. Understanding how to visualise high dimensional data in lower dimensional spaces.
2. Feature Design. **Talk Goal 2**. Understanding the concept of feature map and how to design features that are 1) **interpretable**, 2) **parsimonious**, 3) **in univariate and multivariate spaces**.
3. Feature Selection. **Talk Goal 3**. Understanding how to significantly select features carrying a statistical meaning without overfitting.

Research Question 2) Is there a value in analysing data using (non-linear) feature maps, especially in the context of statistical or machine learning methods, as opposed to working directly on raw data?

Outline

❶ Data Visualisation. **Talk Goal 1**

A standard practice in ML is to first look at the data set, particularly when the number of dimensions (i.e. the number of attributes/input variables available) is big.

Which tools are available for data visualisation?

This task corresponds to applying a **dimensionality reduction** technique such as Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), Uniform Manifold Approximation and Projection, Multidimensional Scaling, kernel PCA, Linear Discriminant Analysis, Factor Analysis, ...

These techniques vary in terms of

- assumptions on the underlying data
- computational complexity
- interpretability
- ability to capture different types of data structures
- captured information and output

The choice of dimensionality reduction technique depends on the specific characteristics of the dataset and the objectives of the analysis.

Methods

We selected the **t-Distributed Stochastic Neighbor Embedding (t-SNE)**, which converts a high-dimensional data set $\mathcal{S} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ into a two- or three-dimensional data set $\tilde{\mathcal{S}} = \{\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_n\}$ that is easier to observe. It is particularly effective when data are affected by complex structures such as non-stationary and non-linear contents.

The algorithm

1. t-SNE models the Euclidean distance between two high-dimensional $\mathbf{x}_i$ and $\mathbf{x}_j$ as the joint probabilities p_{ij}

$$p_{j|i} = \frac{\exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-\|\mathbf{x}_i - \mathbf{x}_k\|^2/2\sigma_i^2\right)} \qquad p_{i|i} = 0 \qquad p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

2. t-SNE measures the similarity between two low-dimensional $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{x}}_j$ as:

$$q_{ij} = \frac{\left(1 + \|\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j\|^2\right)^{-1}}{\sum_{k\neq l}\left(1 + \|\tilde{\mathbf{x}}_k - \tilde{\mathbf{x}}_l\|^2\right)^{-1}} \qquad q_{ii} = 0.$$

3. The identification of the points in the low dimension $\tilde{\mathcal{S}}$ is given by minimising the Kullback-Leibler divergence between the two joint distributions P and Q:

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

❷ Feature Design. Talk Goal 2

A second standard practice in ML corresponds to **Feature Design** or **Feature Engineering** or **Feature Extraction**.

This is process of transforming the input data $\mathcal{S} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ to minimise **Type I Error** and **Type II Error** tp **provide clear discrimination between classes**. It enables one to

- capture domain knowledge (e.g., periodicity or relationships between features).
- express non-linear relationships using linear models.
- encode non-numeric features to be used as inputs to models.

A Classification Example

Let y be the true class label of an instance, with $y = 1$ for positive $y = 0$ for negative classes. Our model predicts $\hat{y}$, where $\hat{y} = 1$ is a positive prediction and $\hat{y} = 0$ a negative prediction. **The confusion matrix** is given as

	Predicted Positive ($\hat{y} = 1$)	Predicted Negative ($\hat{y} = 0$)
Actual Positive ($y = 1$)	True Positive (TP)	False Negative (FN)
Actual Negative ($y = 0$)	False Positive (FP)	True Negative (TN)

Type I Error: or false positive, occurs when the model incorrectly predicts a positive class (*FP*).
Type II Error: or false negative, occurs when the model incorrectly predicts a negative class (*FN*).

Methods

Our initial suggestions

Features Design by Statistical Tests					
Feature	Test	H_0	H_1	Test Statistic	Distribution
Mean	T-test	$\mu_d^{(g)} = \mu_d^{(h)}$	$\mu_d^{(g)} \neq \mu_d^{(h)}$	$T = \frac{\left(\bar{X}_d^{(g)} - \bar{X}_d^{(h)}\right)}{S_p^2 \sqrt{\frac{1}{n_g} + \frac{1}{n_h}}}$	Student's t
Mean	Welch T-test	$\mu_d^{(g)} = \mu_d^{(h)}$	$\mu_d^{(g)} \neq \mu_d^{(h)}$	$T = \frac{\left(\bar{X}_d^{(g)} - \bar{X}_d^{(h)}\right)}{\sqrt{\frac{S_d^{2(g)}}{n_g} + \frac{S_d^{2(h)}}{n_h}}}$	Student's t
Variance	Variance Ratio	$\sigma_d^{2(g)} = \sigma_d^{2(h)}$	$\sigma_d^{2(g)} \neq \sigma_d^{2(h)}$	$F = \frac{S_d^{2(g)}}{S_d^{2(h)}}$	Fisher–Snedecor
Distr.	Kolmogorov Smirnov	$F_d^{(g)}(x) = F_d^{(h)}(x)$	$F_d^{(g)}(x) \neq F_d^{(h)}(x)$	$D = \sup_x \left\lvert \hat{F}_d^{(g)}(x) - \hat{F}_d^{(h)}(x) \right\rvert$	Free
Copula		$C_g = C_h$	$C_g \neq C_h$	$E_{n_g, n_h} = \frac{\hat{C}_g - \hat{C}_h}{\sqrt{\frac{1}{n_g} + \frac{1}{n_h}}}$	Free

3 Feature Selection. Talk Goal 3

Feature Selection Procedure.

1. Consider groups $g =$ Slight and $h =$ Mild and select the input data attribute $d = f_{125}$.
2. Take the feature **Mean** and perform **the T-test** following the **distribution Student's t**.
3. Is the p-value significant?
4. **Yes**. Retain this feature and compute **the sample mean estimate** of input data attribute $d = f_{125}$ for each of the groups (i.e. Slight, Mild, Moderate, Moderately severe, Severe) so to have each group well represented in the feature space. **No**. Discard this feature for the input data attribute $d = f_{125}$.
5. Repeat this procedure for each input data attribute d, each statistical test and each pairwise contrasts of the PTA categories.

This process is equivalent to formulating a map $\varphi(\cdot)$, which can be defined as follows:

$$\varphi : \mathbb{R}^d \to \mathbb{R}^{d'}, \quad \varphi(\boldsymbol{x}_i) = \boldsymbol{z}_i$$

where $\boldsymbol{x}_i \in \mathbb{R}^d$ is the original input vector with $d = 23$ features and $\boldsymbol{z}_i \in \mathbb{R}^{d'}$ is the transformed feature vector in the new feature space of dimension d'.

We formulated $\varphi(\cdot)$ through **pairwise contrasts** of each audiological test, between the PTA categories.

Outline

Our Application:

In our dataset, the input data matrix denoted as $\boldsymbol{X}_{N \times d}$ corresponds do the audiogram for the two ears and the two speech tests, therefore $\boldsymbol{x}_i \in \mathbb{R}^d$, with $d = 23$ and $N = 24,072$.

We formulated $\varphi(\cdot)$ through **pairwise contrasts** of each audiological tests, between the PTA categories.

The contrasts are performed through the use of several **statistical tests** which are **interpretable**, **parsimonious**, **robust to unbalanced datasets**, **transparent** and provide a direct **ranking of the feature based on the p-values**.

We are going to observe steps 1, 2 and 3 for each of the extracted features and compare them to the same steps applied to the raw data.

Results

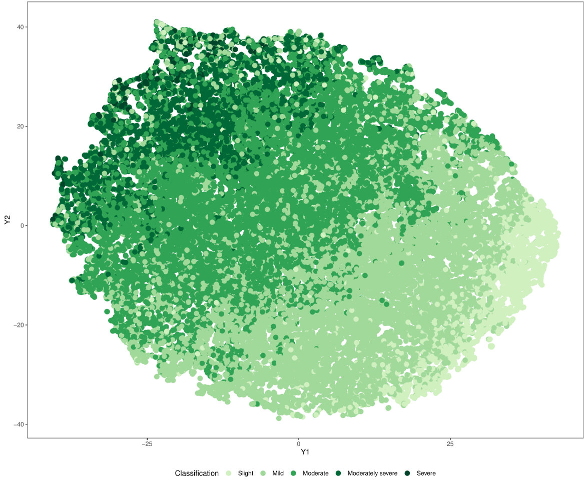
1 Data Visualisation. **Talk Goal 1**

Applying t-SNE to the left audiogram:

Left Audiogram

ID	L_125	...	L_8000
1	20	...	40
2	40	...	50
...	...	...	...
...	...	...	...
24072	30	...	30

t-SNE



Results

1 Data Visualisation. **Talk Goal 1**

Applying t-SNE to the left audiogram:

K-means results:

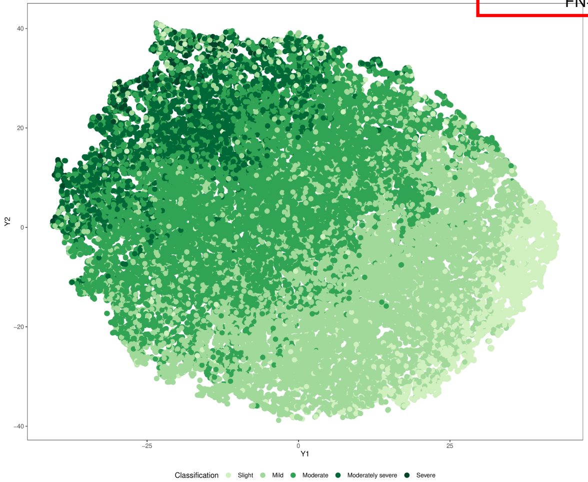
$$\textbf{FPR} = \frac{FP}{FP+TN} = 0.4$$

$$\textbf{FNR} = \frac{FN}{FN+TP} = 0.7$$

Left Audiogram

ID	L_125	...	L_8000
1	20	...	40
2	40	...	50
...	...	...	...
...	...	...	...
24072	30	...	30

t-SNE



Results

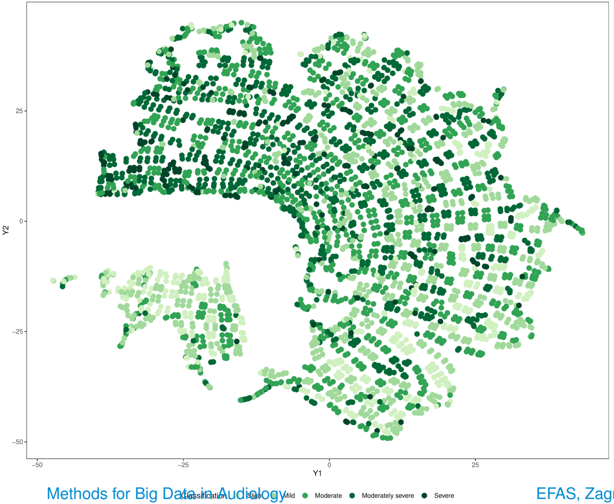
1 Data Visualisation. **Talk Goal 1**

Applying t-SNE to the speech tests:

Speech Tests

ID	SRT_Q	SRT_N
1	3.6	4.7
2	7.8	5.3
...	...	...
...	...	...
24072	3.4	9.4

t-SNE



Results

1 Data Visualisation. **Talk Goal 1**

Applying t-SNE to the speech tests:

K-means results:

$$\textbf{FPR} = \frac{FP}{FP+TN} = 0.37$$

$$\textbf{FNR} = \frac{FN}{FN+TP} = 0.58$$

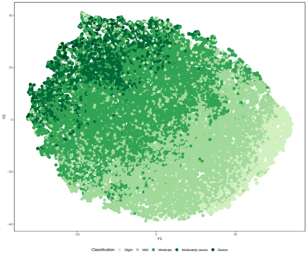
Speech Tests

ID	SRT_Q	SRT_N
1	3.6	4.7
2	7.8	5.3
...	...	...
...	...	...
24072	3.4	9.4

t-SNE

Results

1. Data Visualisation. **Talk Goal 1**
2. Feature Design. **Talk Goal 2**
3. Feature Selection. **Talk Goal 3**



K-means results:

$$\textbf{FPR} = \frac{FP}{FP+TN} = 0.4$$

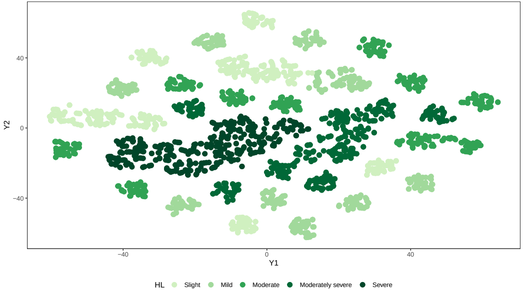
$$\textbf{FNR} = \frac{FN}{FN+TP} = 0.7$$

Applying t-SNE to Feature Mean engineered on the left audiogram:

- **Engineered Feature**:

$$\boldsymbol{z} = \bar{x}_d^{(g)} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_{d,i}^{(g)})$$

- **Selected Features**:
  $f_{500}, f_{500}, f_{2000}, f_{6000}$



Results

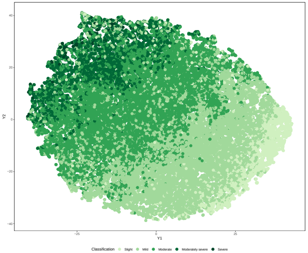
1. Data Visualisation. **Talk Goal 1**
2. Feature Design. **Talk Goal 2**
3. Feature Selection. **Talk Goal 3**

Applying t-SNE to Feature Mean engineered on the left audiogram:

- **Engineered Feature**:

$$\boldsymbol{z} = \bar{x}_d^{(g)} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_{d,i}^{(g)})$$

- **Selected Features**: $f_{500}, f_{500}, f_{2000}, f_{6000}$



K-means results:

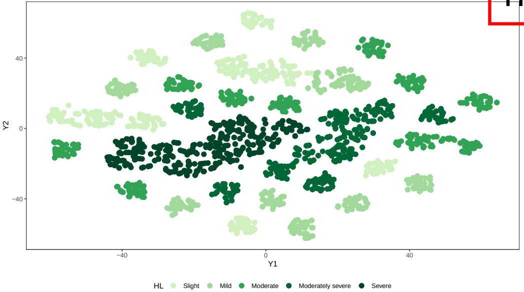
$\textbf{FPR} = \frac{FP}{FP+TN} = 0.4$

$\textbf{FNR} = \frac{FN}{FN+TP} = 0.7$

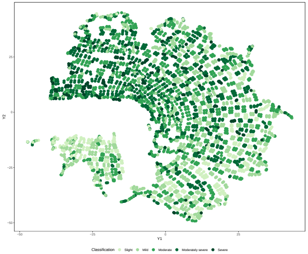
K-means results:

$\textbf{FPR} = \frac{FP}{FP+TN} = 0.1$

$\textbf{FNR} = \frac{FN}{FN+TP} = 0.07$



Results



K-means results:

$$\textbf{FPR} = \frac{\text{FP}}{\text{FP+TN}} = 0.37$$

$$\textbf{FNR} = \frac{\text{FN}}{\text{FN+TP}} = 0.58$$

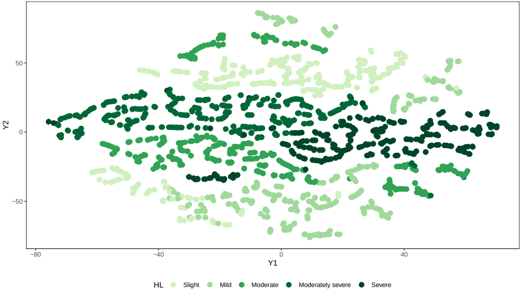
1. Data Visualisation. **Talk Goal 1**
2. Feature Design. **Talk Goal 2**
3. Feature Selection. **Talk Goal 3**

Applying t-SNE to Feature Mean engineered on the speech tests:

- **Engineered Feature**:

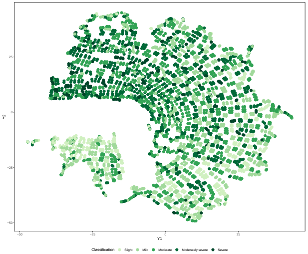
$$\boldsymbol{z} = \bar{x}_d^{(g)} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_{d,i}^{(g)})$$

- **Selected Features**: SRT in quiet, SRT in noise (SNR)



Results

1. Data Visualisation. **Talk Goal 1**
2. Feature Design. **Talk Goal 2**
3. Feature Selection. **Talk Goal 3**



K-means results:

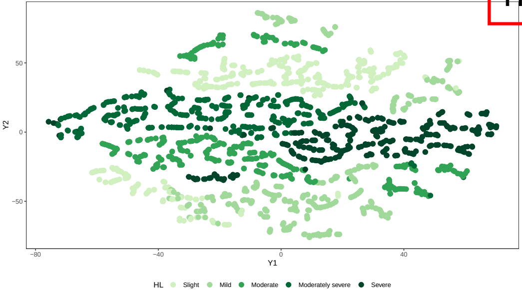
$$\textbf{FPR} = \frac{FP}{FP+TN} = 0.37$$
$$\textbf{FNR} = \frac{FN}{FN+TP} = 0.58$$

Applying t-SNE to Feature Mean engineered on the speech tests:

- **Engineered Feature**:

$$\boldsymbol{z} = \bar{x}_d^{(g)} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_{d,i}^{(g)})$$

- **Selected Features**: SRT in quiet, SRT in noise (SNR)



K-means results:

$$\textbf{FPR} = \frac{FP}{FP+TN} = 0.12$$
$$\textbf{FNR} = \frac{FN}{FN+TP} = 0.13$$

Outline

Conclusions

- ML tools must be reliable tools in audiological care practices
- Such reliability property is induced through the model formulation and its properties which we have above discussed
- Complex datasets should be carefully analysed and explored through several standard ML practices
- The concept of feature engineering is highly precious and should be further explored in audiology for the purpose of auditory profiling
- Interpretation, parsimony and expert opinion should always be sought when ML is applied in this area to provide a better patient care