SOUTENANCE DE PROJET
Fairness en imagerie médicale
Chest X-Ray NIH14

Alexis VO ▪ Tianqi [redacted] ▪ Rakul N[redacted]

LDD IM, fairness in AI

Vendredi 10 avril 2026

université
PARIS-SACLAY

Objectif du projet

université PARIS-SACLAY

Question de recherche

Peut-on réduire les disparités de **fairness** dans un modèle de classification de radiographies thoraciques, sans trop dégrader la **balanced accuracy** ?

- Identifier les principaux biais présents dans les **métadonnées** du dataset
- Comparer deux stratégies de mitigation :
  - **pre-processing** par reweighing
  - **post-processing** par ajustement de seuils
- Évaluer le compromis entre **équité** et **performance** pour un modèle **ResNet18**

Idée clé : le modèle n'utilise que l'image en entrée, mais les métadonnées restent indispensables pour **auditer** et **corriger** les biais.

Données et contraintes du projet

Dataset étudié

NIH Chest X-ray 14, avec des images compressées en **256**×**256**. Le projet s'appuie sur **trois sous-datasets** du groupe : [redacted] et **Vo**.

Données et contraintes du projet

Dataset étudié

NIH Chest X-ray 14, avec des images compressées en **256**×**256**. Le projet s'appuie sur **trois sous-datasets** du groupe : [redacted] **N**[redacted] et **Vo**.

Comparaison des sous-datasets

Sous-dataset	Images	Patients	% malade
[redacted]	5401	1500	45.6
[redacted]	5708	1500	48.0
Vo	5336	1500	45.3

Données et contraintes du projet

Dataset étudié

NIH Chest X-ray 14, avec des images compressées en **256**×**256**. Le projet s'appuie sur **trois sous-datasets** du groupe : [redacted] et **Vo**.

Comparaison des sous-datasets

Sous-dataset	Images	Patients	% malade
[redacted]	5401	1500	45.6
	5708	1500	48.0
o	5336	1500	45.3

Contraintes méthodologiques

- Code d'entraînement fourni
- Modèle utilisé : **ResNet18**
- Le modèle prend **uniquement l'image** en entrée
- Les métadonnées servent à :
  - auditer les biais
  - construire des corrections

Données et contraintes du projet

Dataset étudié

NIH Chest X-ray 14, avec des images compressées en **256**×**256**. Le projet s'appuie sur **trois sous-datasets** du groupe : [redacted] et **Vo**.

Comparaison des sous-datasets

Sous-dataset	Images	Patients	% malade
[redacted]	5401	1500	45.6
	5708	1500	48.0
Vo	5336	1500	45.3

Contraintes méthodologiques

- Code d'entraînement fourni
- Modèle utilisé : **ResNet18**
- Le modèle prend **uniquement l'image** en entrée
- Les métadonnées servent à :
  - auditer les biais
  - construire des corrections

Question centrale

Jusqu'où peut-on améliorer la fairness sans trop perdre en **balanced accuracy** ?

Pipeline global du projet

université PARIS-SACLAY

1. Audit des biais

Analyse des métadonnées pour détecter les disparités selon `View Position`, `Patient Age` et `Patient Gender`.

Pipeline global du projet

université PARIS-SACLAY

1. Audit des biais

Analyse des métadonnées pour détecter les disparités selon `View Position`, `Patient Age` et `Patient Gender`.

2. Correction avant et après apprentissage

- **Pre-processing** : reweighing des couples groupe $\times$ label
- **Entraînement / prédiction** : pipeline **ResNet18** fourni
- **Post-processing** : ajustement de seuils distincts pour `PA` et `AP`

Pipeline global du projet

université PARIS-SACLAY

1. Audit des biais

Analyse des métadonnées pour détecter les disparités selon `View Position`, `Patient Age` et `Patient Gender`.

2. Correction avant et après apprentissage

- **Pre-processing** : reweighing des couples groupe $\times$ label
- **Entraînement / prédiction** : pipeline **ResNet18** fourni
- **Post-processing** : ajustement de seuils distincts pour `PA` et `AP`

3. Évaluation comparative

Comparaison des méthodes avec les métriques : **Balanced Accuracy**, **DIR**, **SPD**, **EOD** et **AOD**.

Pipeline global du projet

université PARIS-SACLAY

1. Audit des biais

Analyse des métadonnées pour détecter les disparités selon `View Position`, `Patient Age` et `Patient Gender`.

2. Correction avant et après apprentissage

- **Pre-processing** : reweighing des couples groupe $\times$ label
- **Entraînement / prédiction** : pipeline **ResNet18** fourni
- **Post-processing** : ajustement de seuils distincts pour `PA` et `AP`

3. Évaluation comparative

Comparaison des méthodes avec les métriques : **Balanced Accuracy**, **DIR**, **SPD**, **EOD** et **AOD**.

Audit $\rightarrow$ Reweighing $\rightarrow$ ResNet18 $\rightarrow$ Post-processing $\rightarrow$ Évaluation

Audit initial : biais observés

Constat principal

Dans les trois sous-datasets, la variable `View Position` est celle qui présente la disparité la plus forte avant modélisation.

Audit initial : biais observés

Constat principal

Dans les trois sous-datasets, la variable `View Position` est celle qui présente la disparité la plus forte avant modélisation.

Comparaison des biais initiaux

Sous-dataset	DIR Genre	DIR Position	DIR Âge
	0.966	1.371	1.153
	1.013	1.327	1.174
o	1.150	1.343	1.162

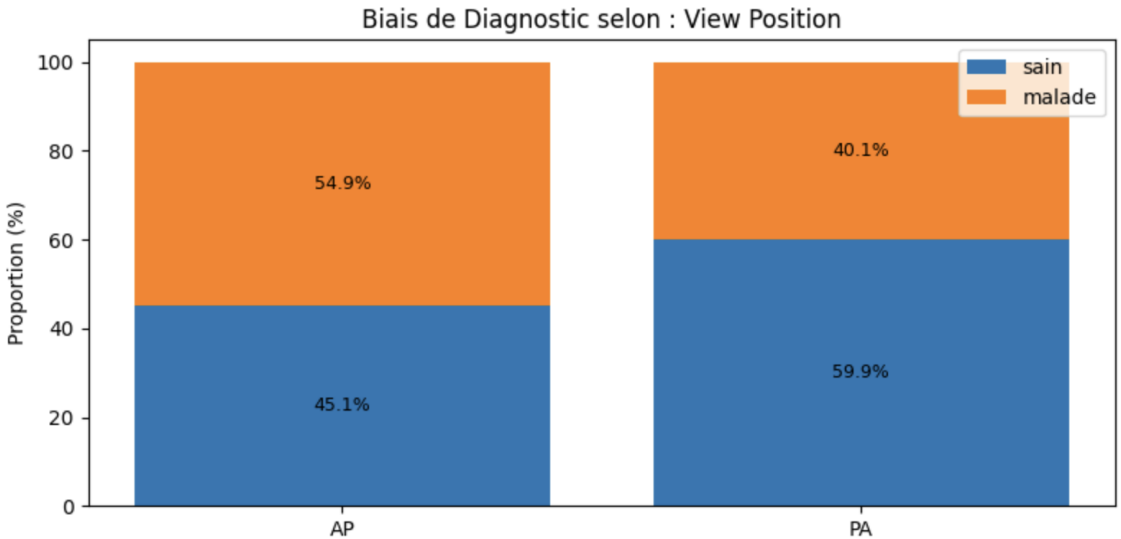
Audit initial : biais observés

Constat principal

Dans les trois sous-datasets, la variable `View Position` est celle qui présente la disparité la plus forte avant modélisation.

Comparaison des biais initiaux

Sous-dataset	DIR Genre	DIR Position	DIR Âge
[redacted]	0.966	1.371	1.153
	1.013	1.327	1.174
o	1.150	1.343	1.162



Exemple illustratif sur le sous-dataset [redacted]

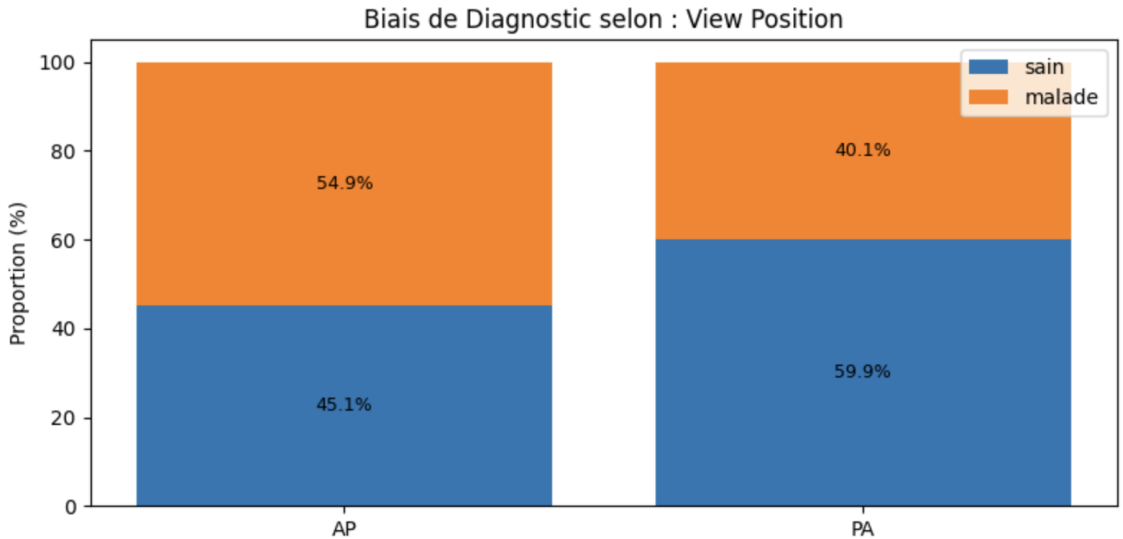
Audit initial : biais observés

Constat principal

Dans les trois sous-datasets, la variable `View Position` est celle qui présente la disparité la plus forte avant modélisation.

Comparaison des biais initiaux

Sous-dataset	DIR Genre	DIR Position	DIR Âge
[redacted]	0.966	1.371	1.153
	1.013	1.327	1.174
Vo	1.150	1.343	1.162



Exemple illustratif sur le sous-dataset Lu

Lecture : la disparité liée à `View Position` est la plus forte et surtout la plus **stable** sur les trois sous-datasets.

Pourquoi `View Position` est le biais principal ?

université PARIS-SACLAY

Interprétation

`AP` / `PA` correspond à une **condition technique d'acquisition**, pas à la pathologie elle-même.

Si le modèle exploite cette différence, il peut apprendre un **shortcut technique** au lieu de se concentrer sur les signes cliniques.

Indice empirique

Avant modélisation, les cas `malade` sont plus fréquents en `AP` qu'en `PA`, et les métriques de fairness montrent une disparité forte sur cette variable.

Exemple représentatif :

- DIR Position élevé
- SPD Position éloigné de 0

Pourquoi `View Position` est le biais principal ?

université PARIS-SACLAY

Interprétation

`AP` / `PA` correspond à une **condition technique d'acquisition**, pas à la pathologie elle-même.

Si le modèle exploite cette différence, il peut apprendre un **shortcut technique** au lieu de se concentrer sur les signes cliniques.

Indice empirique

Avant modélisation, les cas `malade` sont plus fréquents en `AP` qu'en `PA`, et les métriques de fairness montrent une disparité forte sur cette variable.

Exemple représentatif :

- DIR Position élevé
- SPD Position éloigné de 0

Conséquence pour le projet

Nous choisissons donc `View Position` comme **première cible de mitigation**, avant d'examiner d'autres attributs comme l'âge.

Pre-processing : logique du reweighing

université PARIS-SACLAY

Principe

Le reweighing modifie les **poids d'entraînement** pour rééquilibrer les couples **groupe protégé** $\times$ **label**.

L'idée est de :

- augmenter le poids des combinaisons sous-représentées ;
- réduire le poids des combinaisons sur-représentées ;
- corriger les déséquilibres **avant** l'apprentissage.

Pre-processing : logique du reweighing

université PARIS-SACLAY

Principe

Le reweighing modifie les **poids d'entraînement** pour rééquilibrer les couples **groupe protégé** $\times$ **label**.

L'idée est de :

- augmenter le poids des combinaisons sous-représentées ;
- réduire le poids des combinaisons sur-représentées ;
- corriger les déséquilibres **avant** l'apprentissage.

Implémentation dans le notebook

Fonction centrale :

- `compute_reweighing_factor`

Poids construits :

- `W_POSITION`
- `W_AGE`
- `W_POSITION_AGE`

Pre-processing : logique du reweighing

université PARIS-SACLAY

Principe

Le reweighing modifie les **poids d'entraînement** pour rééquilibrer les couples **groupe protégé** × **label**.

L'idée est de :

- augmenter le poids des combinaisons sous-représentées ;
- réduire le poids des combinaisons sur-représentées ;
- corriger les déséquilibres **avant** l'apprentissage.

Implémentation dans le notebook

Fonction centrale :

- `compute_reweighing_factor`

Poids construits :

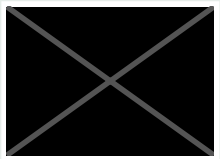
- `W_POSITION`
- `W_AGE`
- `W_POSITION_AGE`

But

Tester si une correction **avant entraînement** améliore la fairness sans modifier l'architecture du modèle.

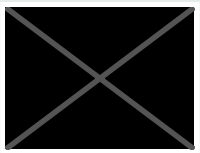
Résultats du pre-processing

Exemples représentatifs : Baseline vs W_POSITION

Sous-dataset	Méthode	BA	DIR Position	SPD Position
	Baseline	0.700	1.848	0.242
	W_POSITION	0.691	1.383	0.175
	Baseline	0.693	1.916	0.314
	W_POSITION	0.684	1.250	0.123
Vo	Baseline	0.649	2.384	0.349
Vo	W_POSITION	0.619	1.173	0.054

Résultats du pre-processing

Exemples représentatifs : Baseline vs `W_POSITION`

Sous-dataset	Méthode	BA	DIR Position	SPD Position
	Baseline	0.700	1.848	0.242
	W_POSITION	0.691	1.383	0.175
	Baseline	0.693	1.916	0.314
	W_POSITION	0.684	1.250	0.123
Vo	Baseline	0.649	2.384	0.349
Vo	W_POSITION	0.619	1.173	0.054

Lecture

Le reweighing améliore déjà nettement la fairness sur `View Position`, avec une baisse généralement **modérée** de la balanced accuracy.

Post-processing : logique de l'ajustement de seuils

Principe

Le post-processing agit **après l'entraînement** : on ne modifie ni le dataset, ni l'architecture du modèle, mais uniquement la **règle de décision finale**.

Post-processing : logique de l'ajustement de seuils

Principe

Le post-processing agit **après l'entraînement** : on ne modifie ni le dataset, ni l'architecture du modèle, mais uniquement la **règle de décision finale**.

Règle de décision

Au lieu d'utiliser un seuil unique, nous appliquons deux seuils distincts :

- un seuil pour les images `PA`
- un seuil pour les images `AP`

$$\hat{y} = 1 \quad \text{si} \quad score_{\text{malade}} \geq \begin{cases} thr_{PA} & \text{si View Position} = \text{PA} \\ thr_{AP} & \text{si View Position} = \text{AP} \end{cases}$$

Post-processing : logique de l'ajustement de seuils

Principe

Le post-processing agit **après l'entraînement** : on ne modifie ni le dataset, ni l'architecture du modèle, mais uniquement la **règle de décision finale**.

Règle de décision

Au lieu d'utiliser un seuil unique, nous appliquons deux seuils distincts :

- un seuil pour les images `PA`
- un seuil pour les images `AP`

$$\hat{y} = 1 \quad \text{si} \quad score_{\text{malade}} \geq \begin{cases} thr_{PA} & \text{si View Position} = \text{PA} \\ thr_{AP} & \text{si View Position} = \text{AP} \end{cases}$$

Recherche des seuils

Nous testons plusieurs couples de seuils sur une grille simple :

$$\{0.3,\ 0.4,\ 0.5,\ 0.6,\ 0.7\}$$

afin d'identifier une correction efficace sur `View Position`.

Objectif du post-processing

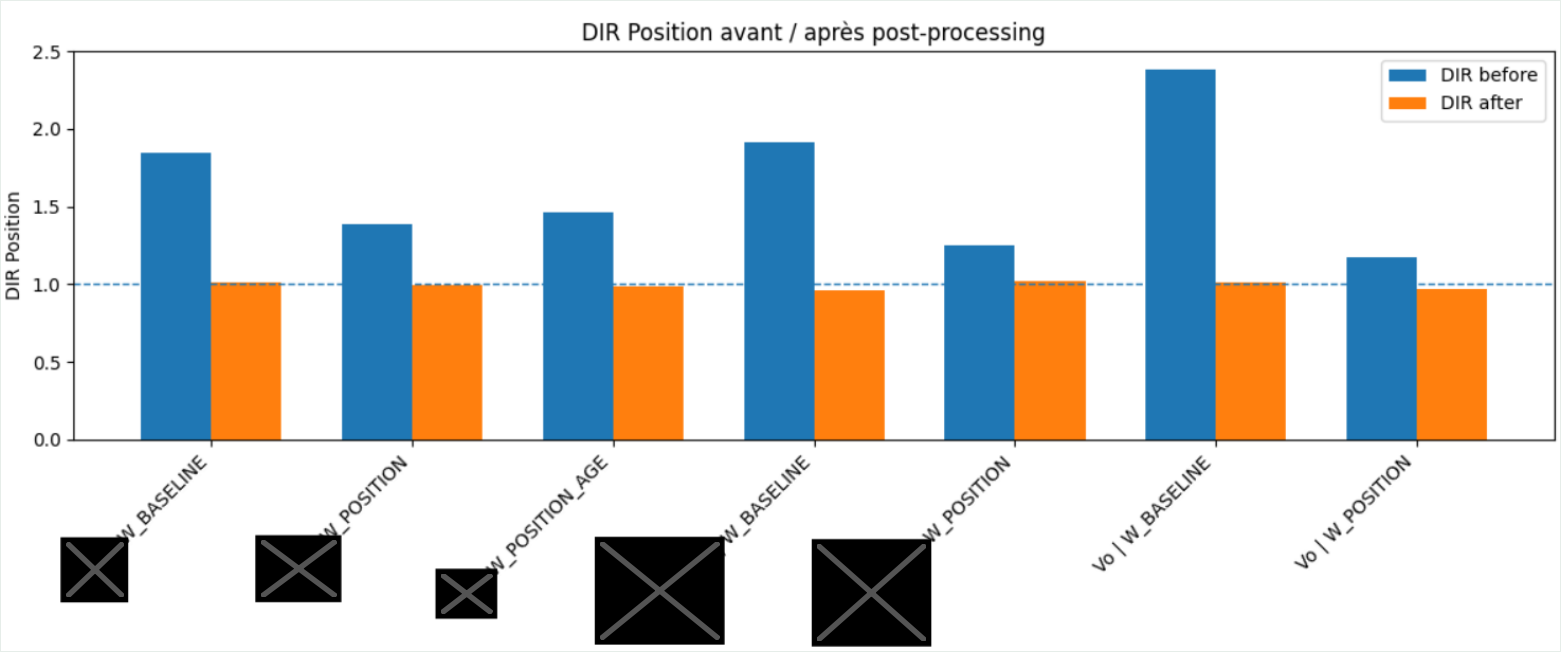
université PARIS-SACLAY

Objectif

Améliorer les métriques de fairness (**DIR**, **SPD**) tout en limitant la perte de **balanced accuracy**.

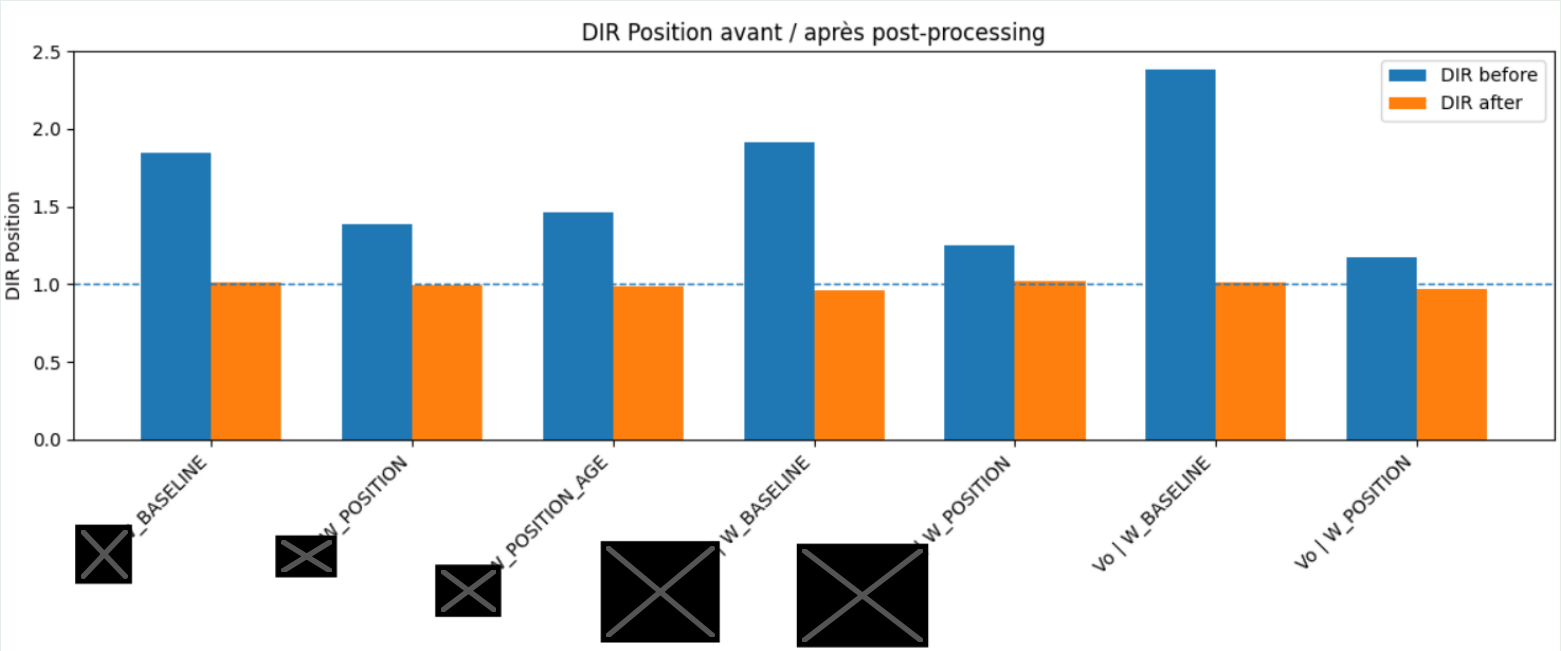
Résultats du post-processing

DIR Position avant / après post-processing



Résultats du post-processing

DIR Position avant / après post-processing



Lecture

Ce graphique compare deux situations :

- le modèle **baseline** ;
- le modèle déjà corrigé par `W_POSITION`.

Dans les deux cas, le post-processing rapproche fortement le **DIR Position** de **1**, ce qui confirme que l'ajustement de seuils reste efficace, y compris après reweighing.

Comparaison finale et conclusion

université PARIS-SACLAY

Pre-processing

- agit **avant** l'entraînement
- améliore déjà la fairness
- baisse généralement **modérée** de BA
- correction plus structurelle

Comparaison finale et conclusion

université PARIS-SACLAY

Pre-processing

- agit **avant** l'entraînement
- améliore déjà la fairness
- baisse généralement **modérée** de BA
- correction plus structurelle

Post-processing

- agit **après** l'entraînement
- correction la plus forte sur `View Position`
- baisse de BA parfois plus marquée selon le dataset
- méthode simple et directe

Comparaison finale et conclusion

Pre-processing

- agit **avant** l'entraînement
- améliore déjà la fairness
- baisse généralement **modérée** de BA
- correction plus structurelle

Post-processing

- agit **après** l'entraînement
- correction la plus forte sur `View Position`
- baisse de BA parfois plus marquée selon le dataset
- méthode simple et directe

Message principal

La fairness doit être pensée comme un problème de **pipeline complet** : audit des biais, correction avant ou après apprentissage, puis arbitrage entre **équité** et **performance**.

Comparaison finale et conclusion

Pre-processing

- agit **avant** l'entraînement
- améliore déjà la fairness
- baisse généralement **modérée** de BA
- correction plus structurelle

Post-processing

- agit **après** l'entraînement
- correction la plus forte sur `View Position`
- baisse de BA parfois plus marquée selon le dataset
- méthode simple et directe

Message principal

La fairness doit être pensée comme un problème de **pipeline complet** : audit des biais, correction avant ou après apprentissage, puis arbitrage entre **équité** et **performance**.

Limites et perspectives

- nombre d'entraînements limité ;
- seuils choisis sur validation ;
- analyse surtout centrée sur `View Position` ;
- perspectives : groupes intersectionnels et seuils appris automatiquement.