CentraleSupélec – EI Mondor

Soutenance de projet

Reda Hamama,Adiel Van Hecke, Alexis Le Parco, Amram El Bazis, Marilou Bernard de Courville

Méthodologie

1: préparer nos données	2: recherche des composantes	3: entraîner et tester des modèles
On va chercher à imputer des données et à les normaliser → besoin de reshape notre dataframe	On fait passer à nos données une ACP/un lasso afin de conserver les 15 composantes les plus importantes	On considère les modèles de Random Forest et Régression Logistique, et on évalue leur performance avec l'AUC

Traitement des données - global

Problème: il manque des données liées à des phases précises
Solution: imputer des données

Méthode:

1. on convertit le fichier .csv pour avoir un fichier “large”
2. on impute les données manquantes en groupant par type de cancer avec la formule *mean + np.random.uniform(-1, 1) * std* (dans un second temps: lignes supprimées)
3. on reconvertit le fichier .csv pour retrouver un fichier “long”

```
patient_num | classe_name | temps_inj | mesure1 | ...
```

```
patient_num | classe_name | VEIN_mesure1 | ART_mesure1 | ...
```

*

Traitement des données - multislice

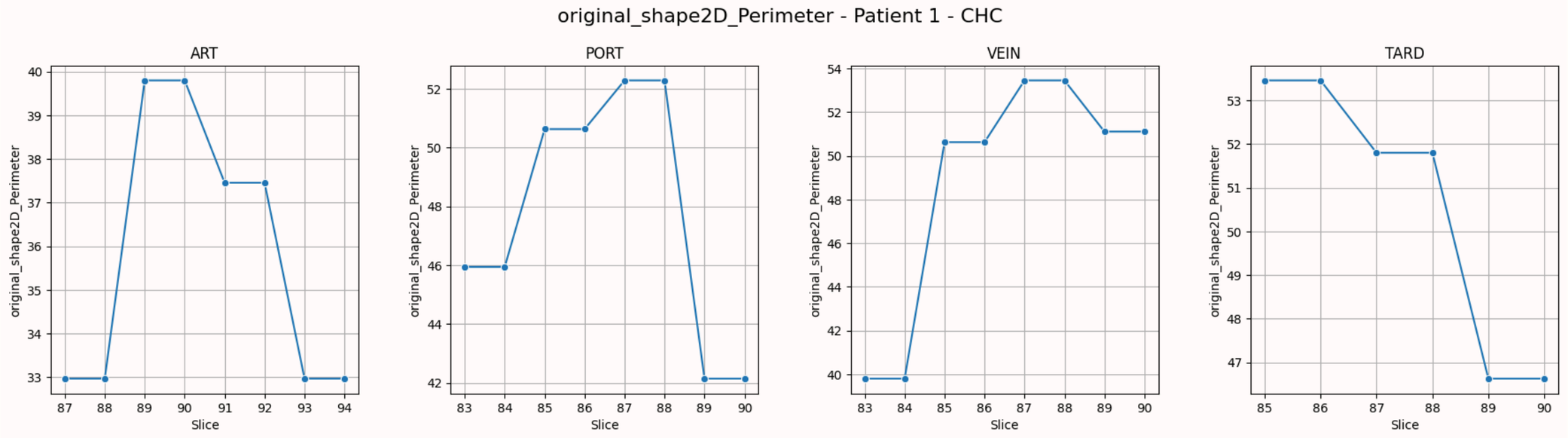
Problème: Les slices ne correspondent pas à une région anatomique fixe du foie, variant à la fois intra- et inter-patient
Solution: Calculer des nouvelles métriques synthétiques

Méthode:

1. Extraction des variables informatives
2. Calcul de nouvelles métriques

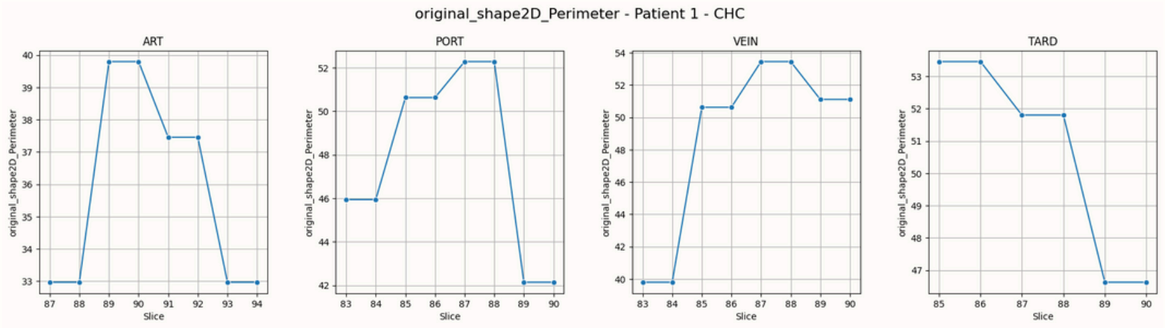
*

Traitement des données - multislice



*

Traitement des données - multislice



Nouvelles Variables :
Extraite de la dérivée

- MOYENNE
- ÉCART-TYPE
- AMPLITUDE
- ASYMÉTRIE
- ÉNERGIE

*

Traitement des données - Donnée d'observation

- Remplissage des cases vides
- Transformation des données globales en préparation à l'assemblage

*

Première méthode : 4 datasets

multi, global_multi, global, global_visu

Modèles de classification :

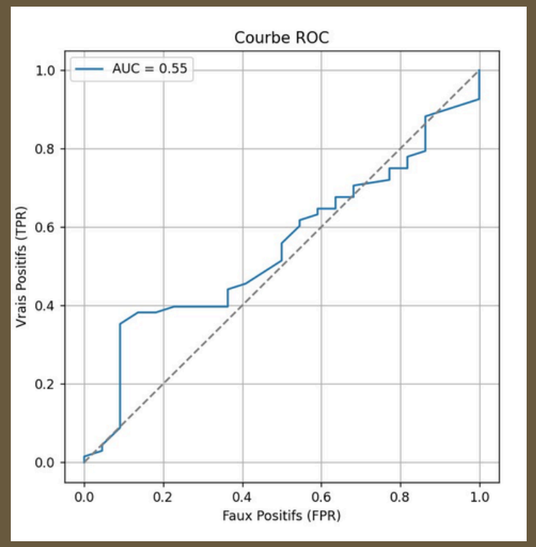
2 modèles de classification :

- RandomForest
- Régression logistique

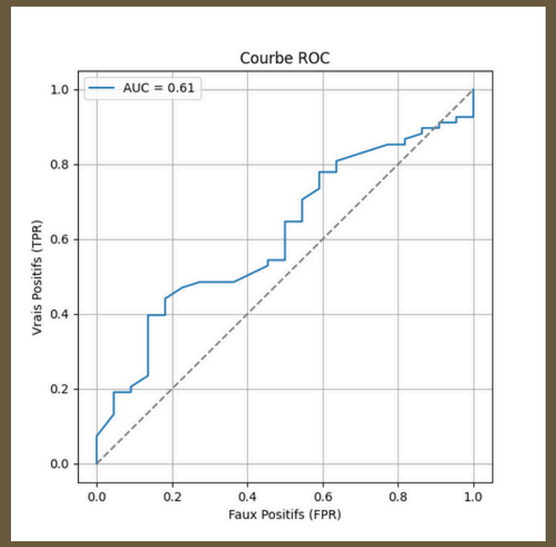
Particularités des modèles :

- Smote
- GroupShuffleSplit

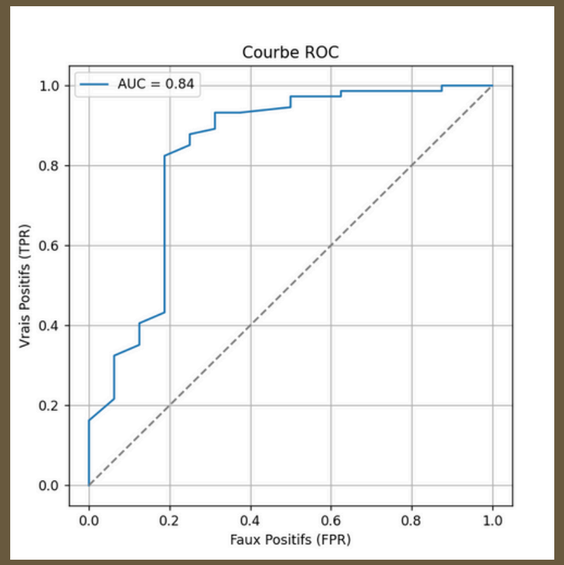
Analyse d'efficacité : Courbes AUC-ROC



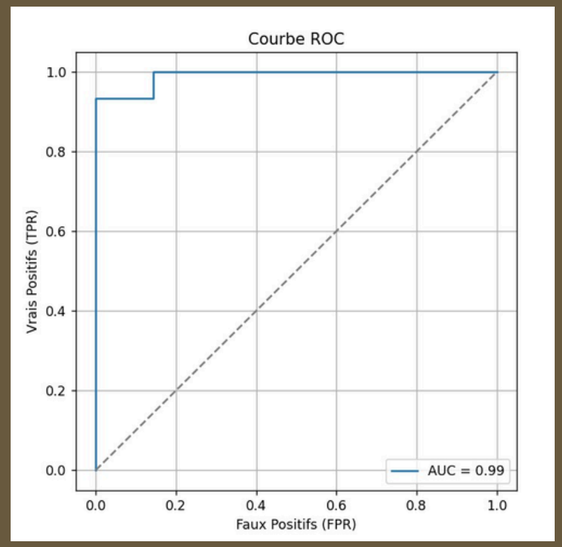
Courbe pour le multi



Courbe pour le global_multi

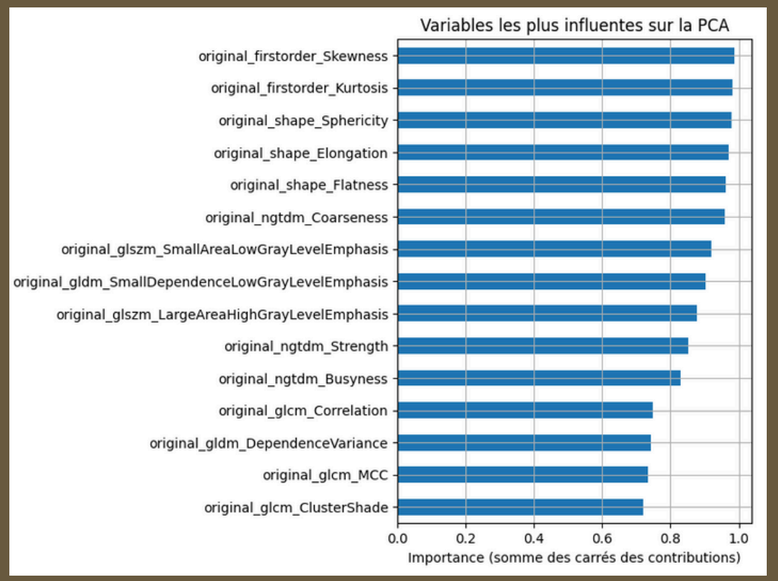


Courbe pour le global

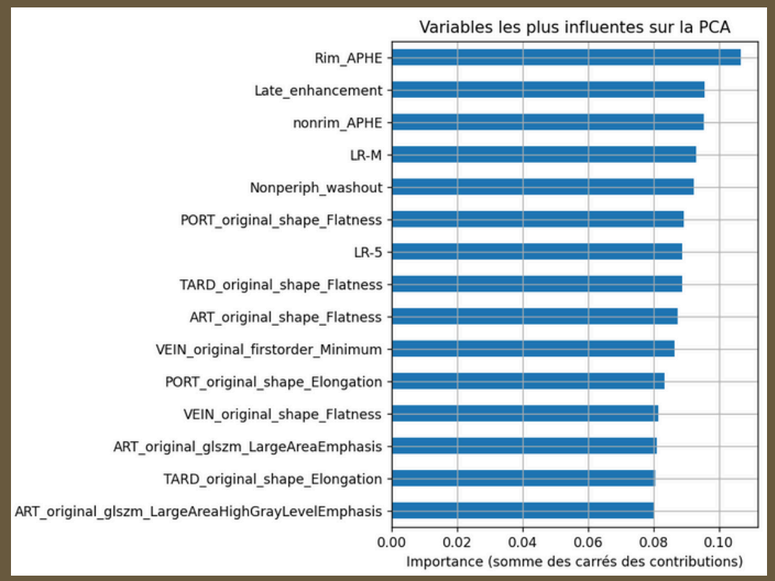


Courbe pour le global_visu

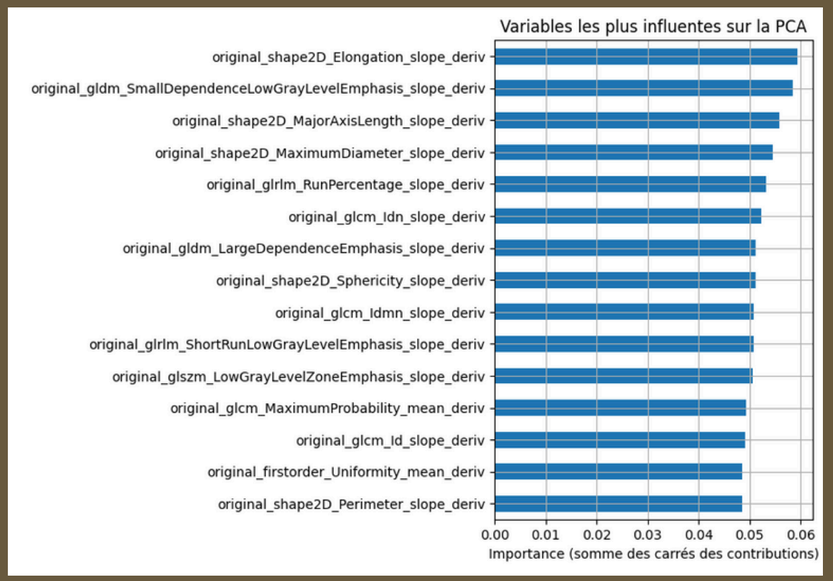
Analyse des variables les plus discriminantes :



Variance pour le global



Variance pour le global_visuel



Variance pour le multi

*

Deuxième méthode : Régression Lasso

Un nouveau modèle plus interprétable : la régression lasso

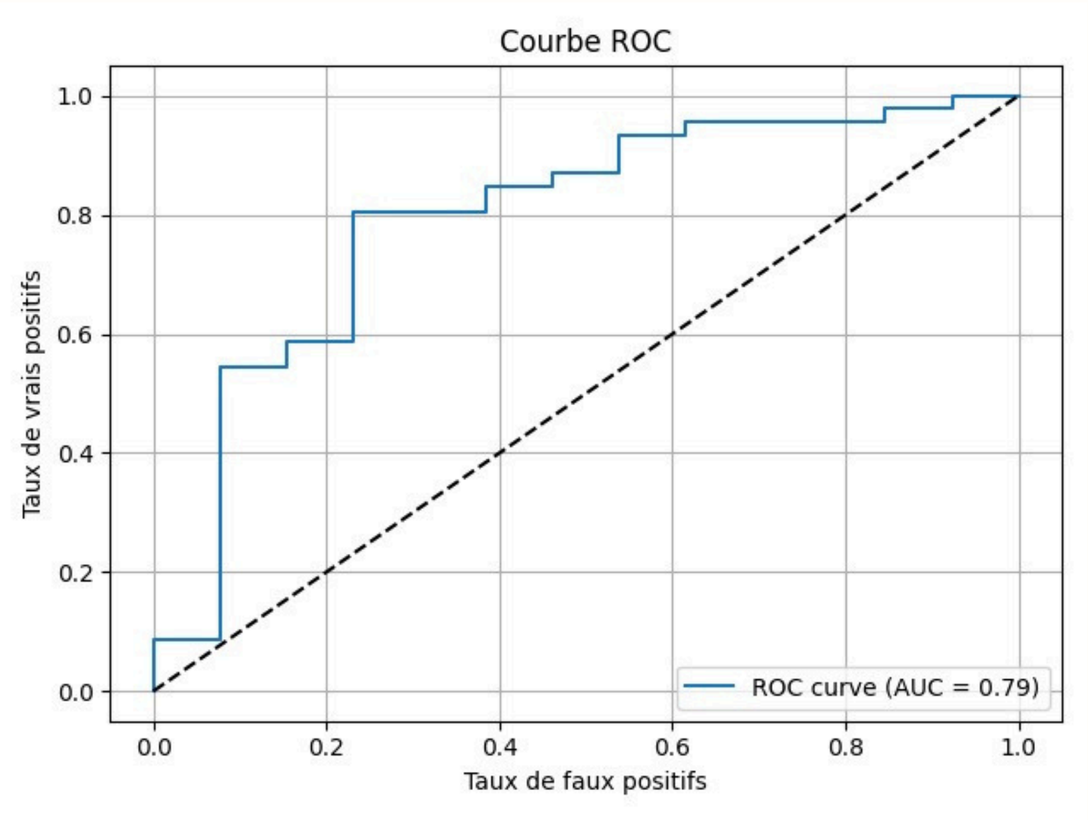
Précision pondérée:0.72

Sensibilité:0.57

Spécificité:0.86

Précision:0.801

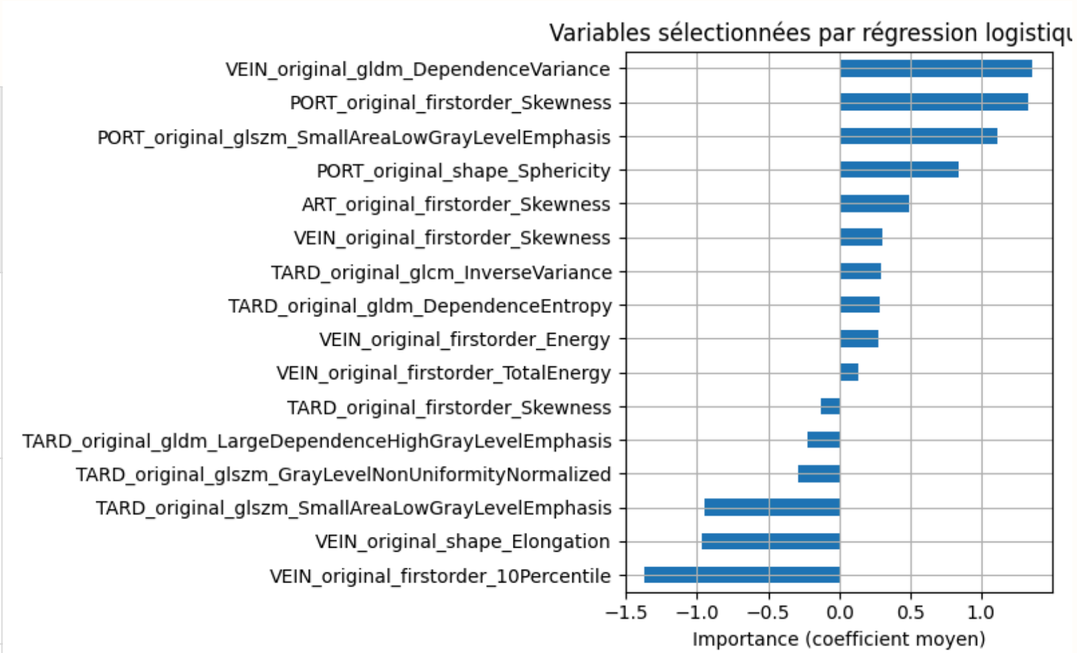
Classe positive = ‘CCK’



Interprétation :

Variables déterminantes:

Niveau de gris	Forme	Texture
original_firstorder_10Percentile	original_shape_Elongation	original_gldm_DependenceVariance ,DependenceEntropy ,LargeDependenceHighGrayLevelEmphasis
original_firstorder_Energy	original_shape_Sphericity	original_glcm_InverseVariance
original_firstorder_Skewness		original_glszm_GrayNonUniformityNormalized ,SmallAreaLowGrayLevelEmphasis
original_firstorder_TotalEnergy		



- Classe 0 :CCK
- Classe 1: CHC

*

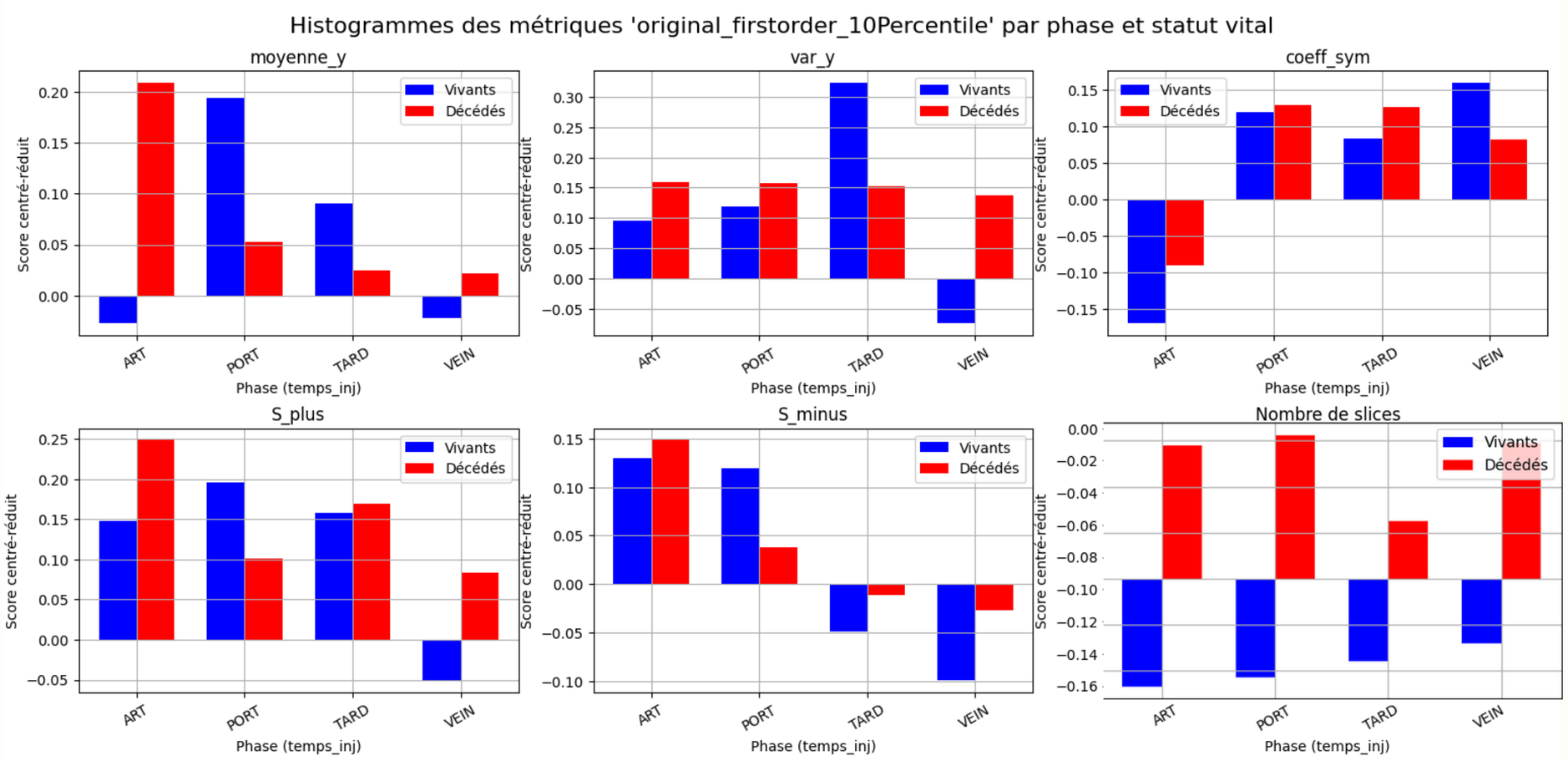
Nouveau défi: Analyse de survie

Différentes métriques pour le multislice

- Moyenne
- Variance
- Skewness
- $S+ = \Sigma [y_k - y_{k-1}]+$
- $S- = \Sigma [y_k - y_{k-1}]-$
- Nombre de Slices

Pour chaque couple (variable,métrique) , on calcule leur score défini comme:

$$\text{score}(\text{var}, \text{metri}) = |\text{médiane}_{\text{vivants}}(\text{metri}(\text{var})) - \text{médiane}_{\text{morts}}(\text{metri}(\text{var}))|$$



Conclusion

*

- 2 types de modèles complémentaires pour déterminer le type de cancer
- score pour discriminer les variables en survie

Pistes d'améliorations :

- Utilisation d'un DataSet plus équilibré
- Réalisation d'un modèle de classification pour l'analyse de survie

*

Avez-vous des *questions ?*

Merci pour votre attention