CentraleSupélec

université PARIS-SACLAY

Rapport ST4 :

EI Mondor

Détection du type de cancer du foie et prédiction de la survie

Réalisé par le groupe :
Marilou Bernard de Courville
Amram El Bazis
Reda Hamama
Alexis Le Parco
Adiel Van Hecke

6 juin 2025

1 Introduction

Dans le cadre de notre ST4 *Big Data et Santé*, nous avons choisi l'enseignement d'intégration en partenariat avec le service de radiologie de l'hôpital Henri Mondor. Notre projet consiste à développer un outil d'aide à la décision pour la détection du type de cancer du foie et la prédiction de la survie des patients.

1.1 Données

Notre approche repose donc sur des données anonymisées fournies par l'hôpital Henri Mondor, qui comprennent des données radiomiques, cliniques, et d'observations médicales. Ces données sont réparties dans plusieurs tableaux. Les patients sont identifiés par le couple `(patient_num, classe_name)`, où `patient_num` est le numéro du patient et `classe_name` est le type de cancer du foie. Les types de cancer du foie présents dans le dataset sont les suivants : CCK, CHC et Mixtes. De plus, on prend en compte différentes phases d'injections, dans `temps_inj`, qui sont : artérielle, portale et veineuse et tardive, notées respectivement `ART, PORT, VEIN, TARD`. De plus, nous avons un set de données contenant des coupes de la tumeur, nous donnant des informations sur la taille et l'intérieur de la tumeur.

Voici donc des statistiques sur le dataset que nous avons utilisé :

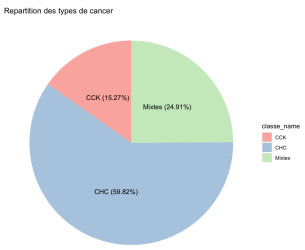

FIGURE 1.1 – Répartition des types de cancer du foie dans le dataset

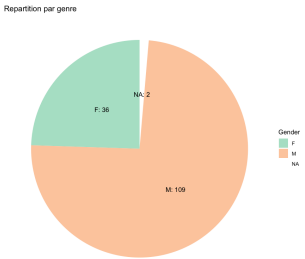

FIGURE 1.2 – Répartition des genres dans le dataset

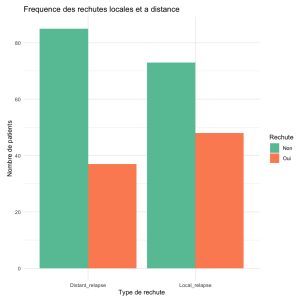

FIGURE 1.3 – Fréquence de rechute des patients, par type de rechute

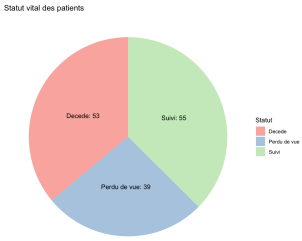

FIGURE 1.4 – Statut vital des patients

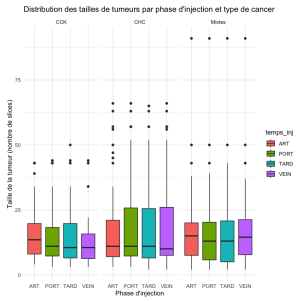

FIGURE 1.5 – Distribution de la taille des cancers par type de cancer et par phase

En tout, nous avons un dataset de 148 patients.

On remarque que le dataset est déséquilibré, contenant une majorité de patients atteints de CHC, et une minorité de patients atteints de CCK. De plus, on remarque que la majorité des patients sont des hommes.

2 Démarches

2.1 Traitement des données

2.1.1 Contexte

Plusieurs dataset ont été utilisés pour cette étude, des données sur le descriptif des patients, sur les résultats radiomiques globaux et multislice des patients sur leur tumeur, et enfin un jeu de données sur les observations visuelles des radiologues. Beaucoup de données

non analysables ont été retirés des dataset pour ne garder que les informations sur la forme, le niveau de gris et les textures.

2.1.2 Multislicing

L'avantage du jeu de données multislice est d'avoir des métriques sur chaque coupe de la tumeur. Ceci dit, la difficulté derrière est qu'un numéro de slice ne définisse pas une région fixe du foie entre les patients ni entre les phases d'un même patient.
Aussi faut-il traiter les données pour les rendre exploitables et comparables entre les patients. Pour ce faire, la démarche suivie est de déterminer des nouvelles métriques sur les courbes des variables en fonction des slices (Voir Figure 2.1.)

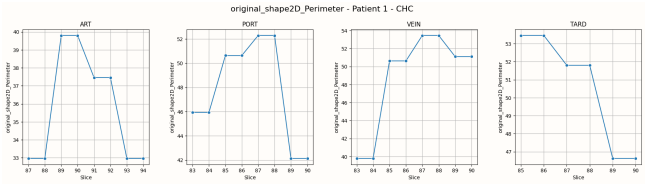

FIGURE 2.1 – Courbe représentant le périmètre en fonction des slices sur les différentes phases d'un patient

Plusieurs variables, décrivant les variations locales, ont été extraites notamment la moyenne, l'écart-type, la symétrie et l'énergie.

2.1.3 Sélection des variables

Une fois les données traitées, nous avons utilisé une analyse des composantes principales pour sélectionner les variables les plus informatives.

3 Résultats

Durant cet enseignement d'intégration, nous avons obtenu de nombreux résultats, en particulier en combinant différents jeux de données : global, multislice et visuel.

3.1 Modèles

Dans un premier temps, notre travail s'est concentré sur deux modèles de classification : la forêt aléatoire (*Random Forest*) et la régression logistique.

Pour assurer le bon fonctionnement de nos programmes, ces modèles reposent sur deux fonctions essentielles : `SMOTE` et `GroupShuffleSplit`.

La méthode `SMOTE` permet de rééquilibrer les classes dans le jeu d'entraînement. La fonction `GroupShuffleSplit`, quant à elle, garantit que les différentes phases d'un même patient ne soient pas séparées entre le jeu d'entraînement et le jeu de test.

3.2 AUC pour le dataset multislice

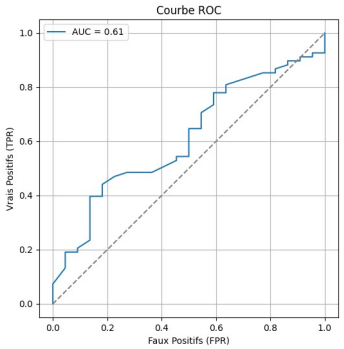

FIGURE 3.1 – AUC pour le dataset multislice

On observe une valeur d'AUC de 0,61, ce qui correspond à un modèle globalement faible, à peine supérieur au hasard.

3.3 AUC pour le dataset fusionné multislice et global

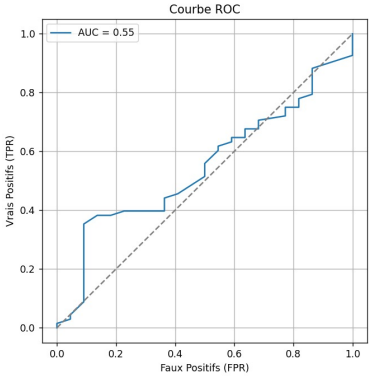

FIGURE 3.2 – AUC pour le dataset fusionné multislice et global

On observe une valeur d'AUC de 0,55, indiquant également un modèle peu performant. Ainsi, les deux jeux de données intégrant des données multislice donnent des résultats globalement décevants.

3.4 AUC pour le dataset global

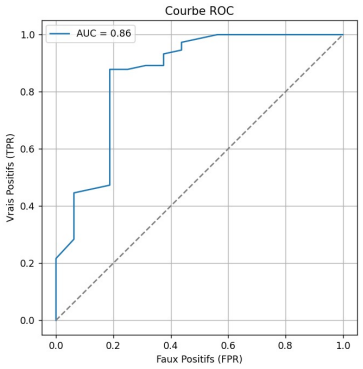

FIGURE 3.3 – AUC pour le dataset global

On obtient ici une valeur d'AUC de 0,86, ce qui est très satisfaisant. Le modèle est à la fois fiable et stable, le score étant obtenu en moyenne sur plusieurs tests.

3.5 AUC pour le dataset global fusionné avec le dataset visuel

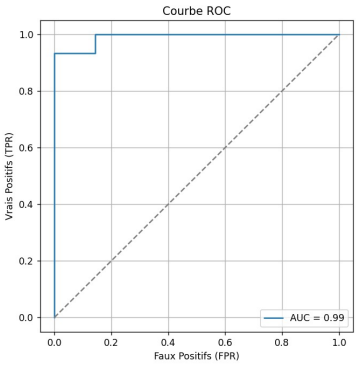

FIGURE 3.4 – AUC pour le dataset global fusionné avec le dataset visuel

On observe une valeur d'AUC de 0,99, ce qui est exceptionnellement élevé. Cela témoigne de la présence de variables très discriminantes dans le dataset visuel, issues des observations des radiologues. Ce résultat est conforme à nos attentes.

En résumé :

Le modèle basé uniquement sur le dataset global obtient de très bons résultats.

Le modèle enrichi avec les observations visuelles des radiologues permet une prédiction quasiment parfaite.

Les jeux de données contenant des informations multislice n'ont malheureusement pas permis d'améliorer la performance obtenue avec le dataset global.

3.6 Régression LASSO

La méthode précédente produisant des variables d'importance relativement homogène, elle rend l'interprétation difficile. Afin de permettre aux radiologues d'identifier plus facilement les variables pertinentes ainsi que leur mode d'influence, nous avons exploré une approche alternative plus interprétable.

Nous avons ainsi réalisé une régression logistique en utilisant l'ensemble de nos paramètres numériques, identifiés par leur phase d'injection, comme variables explicatives, et la variable `Classe_name` (valant 1 pour CHC et 0 pour CKC) comme variable cible. Pour l'entraînement du modèle, nous avons utilisé 80 % des données disponibles.

Les données de train et de test sont créées avec de l'échantillonnage stratifié pour avoir la même proportion de classe dans les deux ensembles,

La régression aboutit à une équation de la forme :

$$Y = \beta X, \quad \text{avec } \beta = (\beta_1, \ldots, \beta_{528}),\ X = (X_1, \ldots, X_{528}),$$

où les coefficients β_j sont estimés par maximisation de la vraisemblance.

Afin de sélectionner uniquement les variables les plus pertinentes, nous avons appliqué une pénalisation de type *Lasso* (L1). Cette méthode force certains coefficients β_j à s'annuler, ne conservant que ceux dont l'impact est significatif sur la prédiction. Cela permet une meilleure interprétation du rôle de chaque variable dans la classification.

Nous avons ensuite tracé la courbe ROC du modèle (représentant le taux de vrais positifs en fonction du taux de faux positifs pour chaque seuil de classification). Le point optimal, situé le plus près du coin supérieur gauche du graphique, correspond à un seuil de classification de 0,21. À ce seuil, le modèle atteint une précision de 80 %, mais la sensibilité (c'est-à-dire la capacité du modèle à détecter correctement les cas positifs, ici les CKC) n'est que de 57 %.

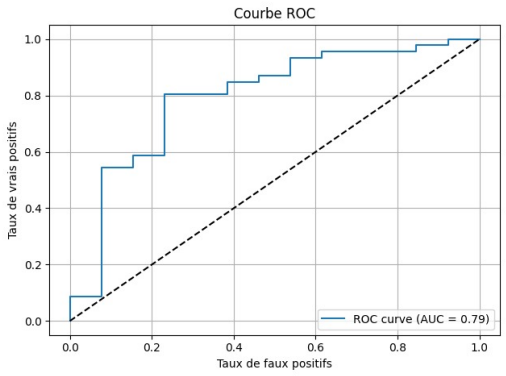

FIGURE 3.5 – Figure : Courbe ROC

Cette faible sensibilité s'explique en grande partie par le déséquilibre entre les classes. C'est pourquoi nous nous intéressons également à la précision pondérée, qui atteint 72 % et prend en compte le déséquilibre des classes dans l'évaluation globale du modèle.

Une analyse du modèle donne comme variables prépondérantes les variables :

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		

TABLE 3.1 – Tableau des variables radiomiques sélectionnées

Les variables les plus influentes sont celles avec un coefficient absolu le plus élevé, les négatives indiquant une augmentation de la probabilité d'être classé comme CKC et les positives d'être CHC.La distriubtion des coefficients est représentée dans la figure suivante.

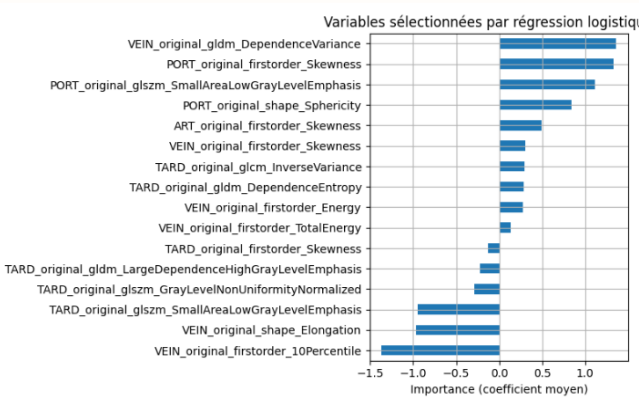

FIGURE 3.6 – Influence des variables sélectionnées

4 Survie

4.1 Analyse des variables multi-slices pour la prédiction de la survie

Dans cette section, nous avons entrepris une analyse approfondie des données `radiomiques_multislice.csv` des patients atteints de cancer, dans le but d'identifier les variables les plus pertinentes pour prédire la survie des patients. Notre étude s'est concentrée sur différents types d'injection, à savoir l'ART, le PORT, le TARD et le VEIN, en prenant en considération plusieurs indicateurs mesurés sur une période d'au moins un an. L'objectif principal était de déterminer quelles caractéristiques étaient les plus significatives pour distinguer les patients décédés des patients vivants.

Pour garantir la cohérence de notre échantillon, nous avons inclus uniquement les patients ayant été suivis pendant au moins un an. Cette mesure visait à garantir que les données recueillies étaient représentatives d'une période significative de l'évolution de la maladie.

Pour visualiser les données, nous avons construit des courbes où les numéros de *slice* étaient placés sur l'axe des abscisses et les valeurs des variables sur l'axe des ordonnées. Chaque courbe représente les valeurs d'une variable spécifique pour un patient donné, avec des points reliés pour illustrer l'évolution de cette variable à travers les *slices*.

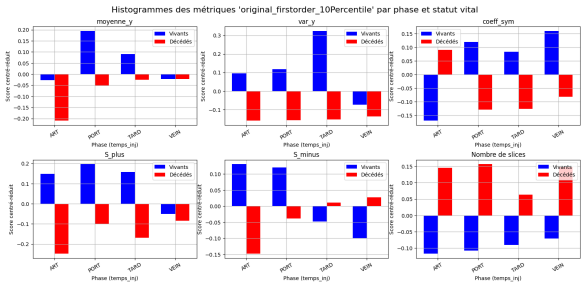

FIGURE 4.1 – Représentation des données de la variable "original_firstorder_10Percentile" en fonction du numéro de slices, avec chaque patient représenté par une couleur différente

En utilisant ces courbes, nous avons calculé diverses métriques pour chaque patient, notamment :

- La moyenne des valeurs
- La variance
- Le nombre de *slices* par patient
- L'analyse de la symétrie à l'aide du coefficient d'asymétrie (*skewness*)
- $S^+ = \sum [y_k - y_{k-1}]^+$
- $S^- = \sum [y_k - y_{k-1}]^-$

Après avoir normalisé ces métriques en les centrant et les réduisant, nous avons séparé les patients en deux groupes distincts : les patients décédés et les patients vivants. En calculant les médianes pour chaque métrique et chaque variable dans chaque groupe, nous avons pu évaluer les différences entre les deux populations. Cette approche nous a permis de définir un score pour chaque couple (variable, métrique), représentant la différence absolue entre la médiane des patients décédés et celle des patients vivants :

$$\text{Score}(\text{var},\text{met}) = |\text{médiane}_{\text{vivant}}(\text{met}(\text{var})) - \text{médiane}_{\text{mort}}(\text{met}(\text{var}))|$$

Les couples ayant les scores les plus élevés sont considérés comme les plus discriminants.

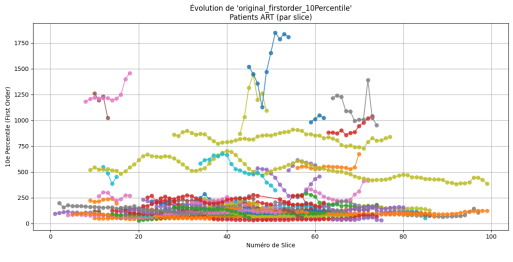

FIGURE 4.2 – Représentation des données de la variable `"original_firstorder_10Percentile"` en fonction du numéro de slices, avec chaque patient représenté par une couleur différente

Comme attendu, le nombre de *slices* est une métrique discriminante pour toutes les variables. En effet, un patient ayant une tumeur plus grande a plus de chances d'en décéder qu'un autre.

En ce qui concerne les autres métriques, nous obtenons, comme visible dans l'histogramme ci-contre, des résultats spécifiques à chaque phase d'injection :

Phase	Métriques	Variables discriminantes
ART	S^+, variance	original_glcm_Idn, original_glcm_Idm, original_glcm_SumEntropy
PORT	variance	original_glszm_SmallAreaEmphasis, original_glrlm_RunPercentage, original_gldm_SmallDependenceEmphasis, original_glcm_Idn
TARD	skewness	original_firstorder_Energy, original_firstorder_TotalEnergy, original_firstorder_Kurtosis
VEIN	skewness S^+	original_firstorder_Energy, original_firstorder_Median original_firstorder_Uniformity

TABLE 4.1 – Couples (métrique, variable) les plus discriminants selon la phase d'injection

5 Conclusion

En conclusion, au cours de notre projet, nous avons pu établir deux types de modèles complémentaires permettant de déterminer le type de cancer du foie. Par ailleurs, l'utilisation du score dans l'analyse de survie nous a permis d'identifier des variables pertinentes associées à la survie des patients.

En perspective, il serait pertinent d'exploiter une base de données plus équilibrée en termes de répartition des classes afin de construire un modèle de classification dédié à l'analyse de survie. De plus, il est intéressant de noter que l'on aurait pu étudier la durée de la survie des patient, alors que l'on a simplement étudier binairement si un patient survivait ou non.