Projet 3A HSB Rapport final

CentraleSupélec

Prédiction de maladies cardiaques suite à un traitement de cancer de l'enfant

Thomas Ménard et Clara Cousteix

Sous la direction de Mahmoud Bentriou, Sarah Lemler et Véronique Lechevalier

Avril 2023

Résumé

Ce projet est porté par deux acteurs principaux, le laboratoire MICS de CentraleSupélec et l'équipe INSERM de l'Institut Gustave Roussy, Villejuif. L'enjeu de ce sujet est de prédire l'apparition de maladies cardiaques suite à un traitement ant-cancéreux chez l'enfant ou l'adolescent.

Pour réaliser ces prédictions, nous avons à disposition une base de données de 7670 patients, issus de la cohorte FCCSS, French Childhood Cancer Survivor Study, traités dans plus de 30 centres d'oncologie français. Parmi les patients de la cohorte, nous possédons les matrices de doses issues de traitements en radiothérapie 3D de près de 4000 patients. De plus, nous possédons des données cliniques sur les traitements par chimiothérapie. La quantité importante de données nous a permis d'utiliser des algorithmes de machine learning et deep learning, et de comparer leurs résultats.

Nous avons entraînés les algorithmes de Machine Learning (Random Forest, XGBoost, LightGBM) sur les indicateurs dose-volume issus des matrices de doses. Ces données sont plus succintes et donc adaptées à des algorithmes de Machine Learning. Nous avons entraînés les algorithmes de Deep Learning (réseau linéaire, réseau convolutionnel, réseau à chemins multiples) sur les matrices de doses 3D. Concernant les algorithmes de Machine Learning, nous obtenons des résultats d'environ 67% pour la balanced accuracy et d'environ 50% pour le score de rappel. Concernant les algorithmes de Deep Learning, nous obtenons des résultats d'environ 69% pour la balanced accuracy et d'environ 60% pour le score de rappel, jusqu'à 67% pour le réseau à chemins multiples.

Il apparaît donc que les réseaux de neurones parviennent mieux à prédire les individus positifs que les algorithmes de Deep Learning, ce qui est primordial en prédiction de maladies. Les résultats obtenus permettront d'améliorer le suivi des patients présents dans cette base de données en augmentant ou réduisant le nombre de séance de suivie suite au cancer.

Table des matières

1 Introduction

1.1 Contexte

Chaque année, en France, on enregistre 1 750 nouveaux cas de cancer chez les enfants de moins de 15 ans, et 800 cas chez les adolescents âgés de 15 à 19 ans. En d'autres termes, un enfant sur 440 environ sera atteint d'un cancer avant l'âge de 15 ans. Depuis 50 ans, les progrès thérapeutiques ont permis d'augmenter progressivement les taux de guérison des cancers des enfants et des adolescents. Actuellement, près de 80% des malades guérissent, même s'il y a des différences en fonction des diagnostics. On estime à 50 000 le nombre d'adultes ayant été touché par un cancer avant l'âge de 20 ans.

Au fur et à mesure de l'augmentation des taux de guérison, médecins et chercheurs ont pris progressivement conscience des effets nocifs des traitements. Cette prise de conscience des effets nocifs des traitements a été tardive car il s'agit principalement d'effets à long terme, qui apparaissent parfois plusieurs dizaines d'années après les traitements. Afin d'étudier et de mieux comprendre les effets iatrogènes de ces traitements, une cohorte portant le nom French Childhood Cancer Survivor Study (FCCSS) a été créée.

L'objectif principal de la cohorte FCCSS est d'étudier l'ensemble du devenir à long terme des enfants et adolescents traités pour un cancer. Cette cohorte comprend les adultes traités avant 2000 pour un cancer de l'enfant ou de l'adolescent dans une trentaine de services d'oncologie en France.

L'objectif de notre travail à travers ce projet est de prédire l'apparition de maladies cardiaques grâce à des algorithmes de machine learning, puis de tenter d'améliorer la prédiction grâce à des réseaux de neurones. Pour cela, nous sommes encadrés par trois chercheurs du laboratoire MICS de CentraleSupélec : Mahmoud Bentriou, Véronique Letort Le Chevalier et Sarah Lemler ; et par Rodrigue Allodji, biostatisticien et épidémiologiste à Gustave Roussy.

1.2 Effets iatrogènes à la suite à un traitement anti-cancéreux

Il existe trois traitements principaux pour traiter un cancer. La chirurgie est privilégiée lorsque la tumeur est localisée. Elle consiste donc à retirer les cellules cancéreuses ou la partie du corps touchée. Un des traitements les plus connus du grand public est la chimiothérapie, qui consiste à administrer au patient des médicaments afin de tuer les cellules cancéreuses. Et enfin, la radiothérapie consiste à détruire les cellules cancéreuses par radiation. Les thérapies peuvent être associées entre elles.

La radiothérapie utilise des rayons ionisants pour détruire les cellules cancéreuses. Cela peut être fait en utilisant un équipement externe qui dirige les rayons vers la zone touchée, ou en impliquant l'insertion d'un dispositif radioactif à l'intérieur du corps. La radiothérapie est souvent utilisée en combinaison avec d'autres formes de traitement. Malheureusement, cette thérapie n'est pas ciblée et peut entraîner donc des effets secondaires variés. Les effets secondaires de la radiothérapie découlent de l'exposition aux rayons ionisants, qui peuvent endom-

mager les cellules saines dans la zone traitée, ainsi que les vaisseaux sanguins et les nerfs. Notamment, les maladies cardiaques graves peuvent être déclenchées à la suite d'une exposition du coeur au faisceau de la radiothérapie.

La prise de certains médicaments de chimiothérapie peut aussi avoir un impact sur le risque de déclarer une maladie cardiaque de haut grade.

Nous chercherons à prédire l'apparition d'une maladie cardiaque grave selon les traitements reçus par le patient à l'aide de méthodes de Machine et Deep Learning.

1.3 Livrable attendu

Le livrable attendu est le rapport final du projet, présenté sous la forme d'un article scientifique. Il sera accompagné des scripts utilisés, des poids des modèles entraînés et d'une documentation. Il est attendu que les algorithmes de Deep Learning entraînés sur les matrices de doses 3D soient comparés à la baseline, algorithmes de Machine Learning entraînés sur des données issues d'histogrammes doses-volumes, ainsi qu'à la littérature.

2 Etude bibliographique

2.1 Machine et Deep Learning en Oncologie, Physique Médicale et Radiologie

L'irradiation des tissus normaux est inévitable au cours de la radiothérapie et est le principal facteur limitant l'augmentation de la prescription. Il existe donc un compromis entre la volonté de détruire les tissus cancéreux et la volonté de préserver les tissus sains environnants. L'optimisation de ce compromis est le défi fondamental de la radiothérapie et peut être éclairé à l'aide de données de radiothérapie [8].

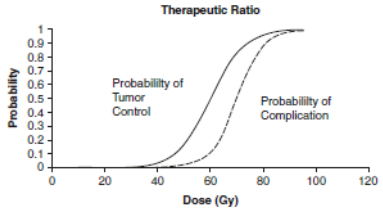

FIGURE 1 – Ratio thérapeutique en radiothérapie [8]

L'optimisation du ratio thérapeutique, présenté sur la figure 1, fait partie des objectifs du chapitre 17 du livre "Machine and Deep Learning in Oncology, Medical Physics and Radiology" : « Modelling Radiotherapy Response : TCP and NTCP ». Dans ce chapitre, on passe en revue des méthodes pour classifier ou quantifier la TCP (Probabilité de contrôle de la tumeur) et la NTCP (la

probabilité de complication de tissus non-cancéreux), afin de mieux prédire la réponse des tissus cancéreux et non-cancéreux à l'irradiation. Le chapitre est très complet sur les méthodes et bonnes pratiques de la littérature. Il traite notamment les méthodes de Machine Learning pour modéliser la NTCP, ce qui se rapproche de notre projet, et référence les papiers qui ont traité le sujet. Il traite aussi les différents travaux de Deep Learning sur des données de radiologie.

Ce chapitre sera une référence pour nous guider dans nos travaux.

2.2 Fonction de coût appropriée aux datasets déséquilibrés

Les algorithmes de Machine et Deep Learning peuvent être très sensibles au caractère déséquilibré d'une base de données. Le papier CAO, Kaidi, WEI, Colin, GAIDON, Adrien, et al. Learning imbalanced datasets with label-distribution-aware margin loss. Advances in neural information processing systems, 2019, vol. 32. a été publié à l'occasion de la 33ème Conférence «Neural Information Processing Systems » (NeurIPS 2019) à Vancouver au Canada. Il présente une revue de plusieurs méthodes pour améliorer la généralisation des modèles entraînés sur des datasets très déséquilibrés, mais aussi propose une nouvelle fonction de coût inspirée des Support Vector Machine (SVM). Ces méthodes sont comparées sur Idivers datasets publics déséquilibrés [4].

Pondération de la fonction de coût Il est possible de pondérer la fonction de coût afin qu'elle prenne en compte le caractère déséquilibré. Mais cela rend l'optimisation plus difficile, surtout dans les schémas de classes fortement déséquilibrées [4]. De plus, les performances sont moins bonnes sur la classe majoritaire. Enfin, selon les sources de l'article, il serait important d'inclure une régularisation dans le réseau.

Resampling Il est possible de faire du resampling dans chaque mini-batch, soit en faisant de l'undersampling de la classe majoritaire soit de l'oversampling de la classe minoritaire, afin d'équilibrer chaque mini-batch. Cela n'est pas faisable pour des classes très déséquilibrées et il y a un risque d'overfitting sur les classes minoritaires [4].

Loss à régularisation personnalisée L'idée dans cette méthode est de faire une margin loss dont les marges dépendent de la distribution de chaque classe, ainsi que présenté figure 2, notée LDAM. Cela revient à faire une régularisation non évidente, en donnant plus de régularisation à la classe minoritaire qu'à la classe majoritaire.De plus, les auteurs proposent un algorithme d'entraînement différé en deux étapes. La première étape s'entraîne d'abord avec la fonction de coût LDAM non pondérée. La seconde étape déploie une perte LDAM repondérée avec un learning rate plus faible. Empiriquement, la première étape conduit à une bonne initialisation pour la deuxième étape avec des pertes repondérées [4].

Dans le tableau 3, les auteurs comparent les différentes méthodes évoquées sur le dataset IMDB review, qui est déséquilibré en faveur des éléments positifs. La méthode LDAM-DRW surpasse les autres méthodes en terme d'erreur

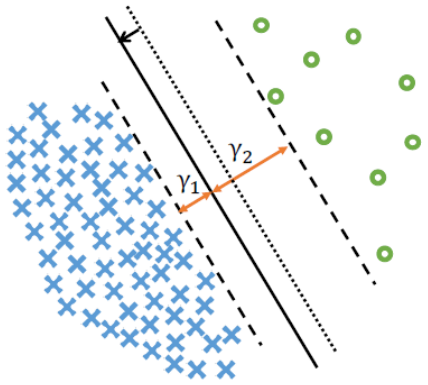

FIGURE 2 – Hinge Margin Loss, avec γ_1 et γ_2 tel que $\gamma_i \propto n_i^{-1/4}$ [4]

moyenne. Elle ne performe pas aussi bien que l'ERM sur les éléments positifs mais est bien meilleure sur les éléments négatifs. On voit qu'il existe toujours un compromis à trouver entre une bonne performance sur la classe minoritaire et sur la classe majoritaire.

Table 1: Top-1 validation errors on imbalanced IMDB review dataset. Our proposed approach LDAM-DRW outperforms the baselines.

Approach	Error on positive reviews	Error on negative reviews	Mean Error
ERM	2.86	70.78	36.82
RS	7.12	45.88	26.50
RW	5.20	42.12	23.66
LDAM-DRW	4.91	30.77	17.84

FIGURE 3 – Synthèse des résultats des différentes méthodes comparées ; ERM = Empiric Risk Minimization, pas de pondération, RS = Resampling, RW = Reweighting , LDAM-DRW = méthode présentée dans le papier [4]

Ces méthodes pourront être des pistes pour notre étude ou pour de futurs travaux. Nous utiliserons notamment les méthodes de pondération et de resampling.

2.3 Deep Learning sur des données de doses radiothérapiques - une revue

L'article "Deep Learning for Radiotherapy Outcome Prediction Using Dose Data" a été publié en 2021 par Appelt et Gilbert, tous deux professeurs chercheurs à l'université de Leeds, ainsi que trois autres chercheurs anglo-saxons : B. Elhaminia , A. Gooya , M. Nix.

L'article propose une revue de 27 papiers, qui utilisent des données de distributions de doses pour de la prédiction de toxicité ou de réponse de la tumeur.

Dans certains papiers, il est ajouté des variables cliniques, les contours d'organes, de l'imagerie médicale CT ou PET scan [2].

Les articles donnent leurs résultats principaux en AUC, qui semble être une métrique usuelle pour ce genre de travaux [2].

Les auteurs de l'étude ont conclu que l'utilisation de l'apprentissage profond pour prédire les résultats de la radiothérapie peut aider les médecins à mieux planifier le traitement et à améliorer les résultats pour les patients. Ils ont également souligné que davantage de recherches sont nécessaires pour évaluer la faisabilité et l'efficacité de cette approche dans un cadre clinique réel.

Il sera intéressant de se référer à ces articles afin de se comparer à la littérature et d'appliquer de nouvelles méthodes.

2.4 Réseau de neurones pour la prédiction de l'issue de la radiothérapie du foie

L'article "Neural networks for deep radiotherapy dose analysis and prediction of liver SBRT outcomes" a été publié en 2020 dans la revue Physics in Medicine and Biology. Cet article faisait partie des articles cités dans le papier présenté section 2.3. A travers cet article, nous avons étudié une méthode utilisant les réseaux de neurones pour prédire les résultats de la radiothérapie stéréotaxique du foie (SBRT).

Les auteurs ont utilisé des données de dose de radiation de 220 patients atteints de cancer du foie pour entraîner un réseau de neurones à prédire la réponse du foie à la SBRT. Le réseau de neurones présenté a été conçu pour analyser les matrices de doses ainsi que des variables cliniques pour prédire la réponse du foie à la SBRT. Ils utilisent une base de données de matrices de doses sur le foie, ainsi que certaines variables cliniques telles que l'âge, le sexe, l'anatomie abdominal, l'historique des comorbidités, autres thérapies sur le foie et des tests fonctionnels du foie. Ainsi, les auteurs ont construit un réseau à plusieurs chemins : un chemin convolutionnel pour les matrices de doses et un chemin linéaire pour les données cliniques, ensuite concaténés pour la prédiction. Le réseau est illustré sur la figure 4 [11].

Les classes de sortie sont appelées PPO (les patients à rémission long terme) et NPO (les patients à rechute ou décès inférieur à 2 ans). Les performances du réseau de neurones sont meilleures que celles d'une Random Forest ou d'une SVM, comme on peut le voir tableau 1 [11].

	SVM	RF	Multipath Network
Survival primary liver cancer	0.639	0.750	0.772
Disease progression for primary liver cancer	0.721	0.653	0.787
Survival Metastatic liver cancer	0.720	0.652	0.671
Disease progression for metatatic liver cancer	0.523	0.566	0.592

TABLE 1 – Synthèse des résultats comparant SVM, Random Forest et Multipath Network [11]

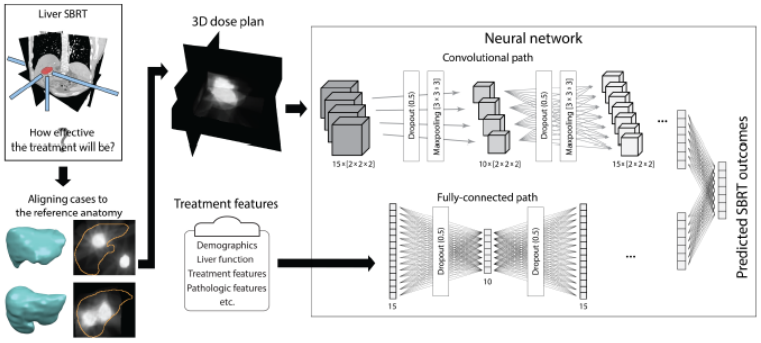

FIGURE 4 – Multipath Network [11]

La population d'étude est de 220 patients. De l'augmentation a été réalisée sur la classe minoritaire et les foies ont été recentrés et normalisés en entrée. De plus, les auteurs ont réalisé un préentrainement sur 2644 images 3D de différents organes [11].

Il est intéressant de noter que les auteurs ont ajouté à leur publication une carte d'activation pour identifier les régions déterminantes pour la décision du réseau. Sur la figure 5, les couleurs rouges correspondent aux zones de risques importants pour une rechute ou un décès sous 2 ans, tandis que les couleurs bleues correspondent à un risque négligeable [11].

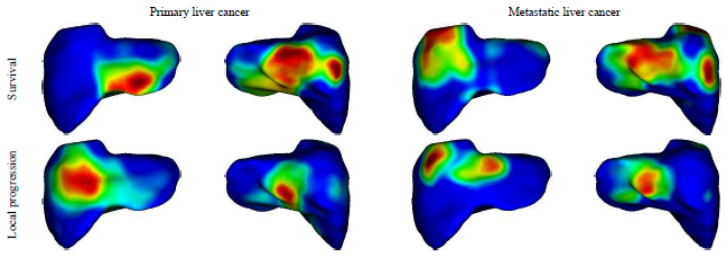

FIGURE 5 – Carte des zones du foie à risque pour la survie après traitement anti-cancéreux et pour la rechute [11]

On pourra utiliser un réseau de neurones similaire à celui proposé dans l'article pour notre problème et comparer nos résultats.

3 Matériels et Méthodes

3.1 Données

3.1.1 Cohorte FCCSS : la population d'étude

La cohorte FCCSS (French Childhood Cancer Survivor Study) est une cohorte regroupant les données de plus de trente centres d'oncopédiatrie français. L'étude est à l'initiative de l'équipe INSERM de l'hôpital Gustave Roussy à Villejuif (France). La cohorte rassemble 7670 patients traités pour des cancers de l'enfant ou de l'adolescence, comme le montre la figure 6. Le but de cette cohorte est d'étudier les effets à long terme des traitements anti-cancéreux, notamment le risque d'apparition de maladies cardiaques graves.

Les patients sont identifiés par une double clé (ctr, numcent) qui sont le numéro du centre et le numéro d'identification du patient dans le centre. A chaque patient sont associées de nombreuses données tirées des dossiers médicaux : données cliniques, administratives ou encore temporelles. On peut y trouver par exemple les divers traitements anti-cancéreux reçus, le siège et type du cancer, les dates des derniers examens. On dénombre ainsi jusqu'à 15000 données par patient. Parmi elles, il existe notamment des dates de suivi du patient et l'apparition d'une maladie cardiaque grave.

On peut voir sur la figure 7 que les patients qui développent une maladie cardiaque grave sont largement minoritaires mais non négligeables. Il sera donc intéressant d'étudier quels peuvent être les facteurs de risque pour l'apparition d'une maladie cardiaque grave à l'aide d'analyses de survie sur les données de la cohorte FCCSS.

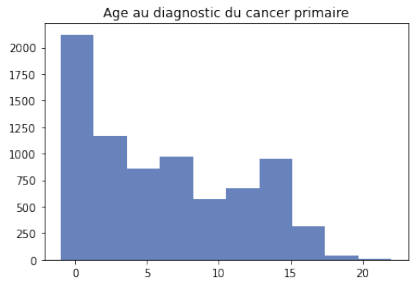

FIGURE 6 – Age du diagnostic du cancer primaire

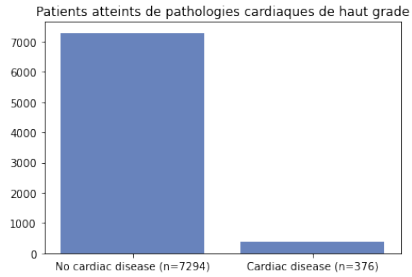

FIGURE 7 – Patients atteints d'une pathologie cardiaque de haut grade

3.1.2 Données dose-volume

Il est usuel en radiothérapie de travailler avec la distribution de dose-volume cumulée, aussi appelé histogramme dose-volume cumulé (DVH). Cela permet de résumer graphiquement la distribution du rayonnement dans un volume d'intérêt d'un patient traité par radiothérapie [7].

Les courbes DVH peuvent avoir des profils différents, en fonction de la dose prescrite, comme on peut le voir sur la figure 8.

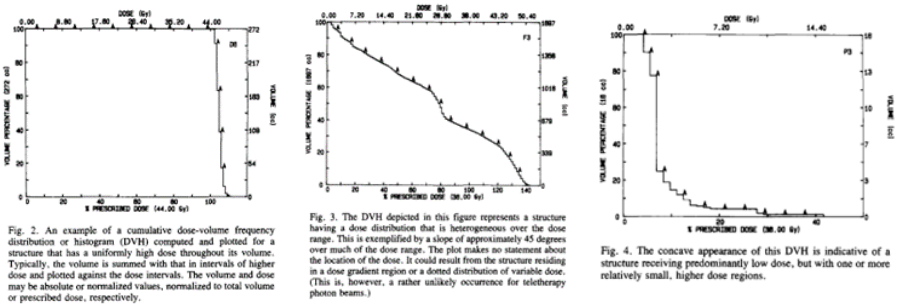

Fig. 2. An example of a cumulative dose-volume frequency distribution or histogram (DVH) computed and plotted for a structure that has a uniformly high dose throughout its volume. Typically, the volume is summed with that in intervals of higher dose and plotted against the dose intervals. The volume and dose may be absolute or normalized values, normalized to total volume or prescribed dose, respectively.

Fig. 3. The DVH depicted in this figure represents a structure having a dose distribution that is heterogeneous over the dose range. This is exemplified by a slope of approximately 45 degrees over much of the dose range. The plot makes no statement about the location of the dose. It could result from the structure residing in a dose gradient region or a dotted distribution of variable dose. (This is, however, a rather unlikely occurrence for teletherapy photon beams.)

Fig. 4. The concave appearance of this DVH is indicative of a structure receiving predominantly low dose, but with one or more relatively small, higher dose regions.

FIGURE 8 – Exemples d'histogramme doses-volumes [7]

Sur le premier DVH, on peut en conclure que la dose prescrite a été uniformément répartie sur le volume d'intérêt et on ne montre pas d'éventuels effets de bords. Sur le deuxième DVH, la répartition de la dose dans le volume est bien plus hétérogène. Sur le troisième DVH, on observe que le volume d'intérêt reçoit globalement une dose faible avec quelques régions exposées à un rayonnement plus fort. Cela pourrait être interpréter comme le traitement de la tumeur avec un rayonnement concentré sur la zone à contrôler et des effets de bords sur l'organe.

Les DVH peuvent également être utilisés comme données d'entrée pour estimer la probabilité de contrôle de la tumeur (TCP) et la probabilité de complication du tissu normal (NTCP) [7].

Dans le cadre de notre projet, on s'intéresse en particulier aux complications du tissu cardiaque à la suite d'un traitement par rayonnement. On possède des données dose-volume sous forme de quantile, qui permettent de reconstruire des histogrammes doses-volumes similaires à ceux qu'on a pu présenter sur la figure 8. Un exemple est représenté sur la figure 9.

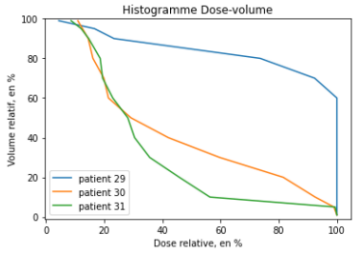

Figure 2 : Histogramme Doses-Volumes issus de nos données

FIGURE 9 – Histogramme Dose-volume avec les données de la cohorte FCCSS

Les données dose-volumes sont liées à la cohorte FCCSS par l'identifiant

unique du patient (ctr, numcent), présenté dans la partie 3.1.1. On possède les données dosimétriques de près 4000 patients de la cohorte FCCSS. Les patients restants sont exclus de l'étude.

Pour chaque patient, nous avons des données de la forme dv_X_N, où X représente un indicateur dose-volume. Par exemple, pour X = D99, cela représente la dose en Gy reçue par au moins 99% du coeur (quantile à 1% de la distribution). D50 est donc la médiane. Pour X = V5, c'est le volume de l'organe qui a reçu au moins 5Gy (donc varie entre 0 et 100%). La lettre N représente sur quelle partie de l'organe les indicateurs sont calculés.

On possède les données dose-volumes pour plusieurs parties du cœur :

- tout le cœur
- oreillette droite
- oreillette gauche
- ventricule droit
- ventricule gauche
- myocarde

Ces données, une fois reliées à la base FCCSS, permettent de faire de premiers algorithmes de prédiction de maladies cardiaques en fonction de la dose reçue sur le cœur. On peut déjà remarquer une corrélation entre les doses reçues par le cœur du patient et la pathologie cardiaque grave sur la figure 10.

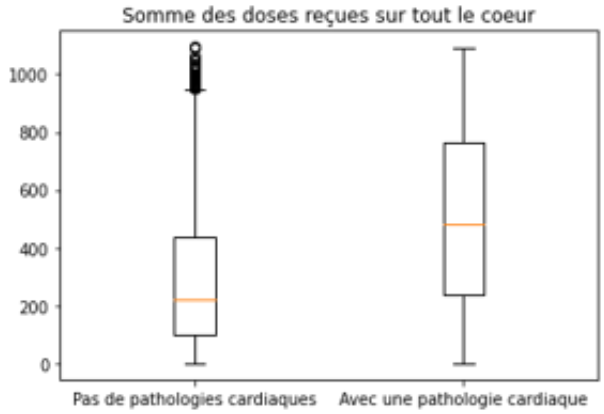

FIGURE 10 – Somme des doses reçues sur tout le coeur

Bien que le DVH soit largement utilisé dans l'évaluation de la qualité du plan de traitement et du pronostic de radiothérapie, la distribution tridimensionnelle de la dose peut décrire les effets du rayonnement de manière plus explicite. En effet, les données issues des DVH n'informent que sur la distribution de la dose et non pas sur la répartition dans l'espace. On travaillera donc aussi avec les matrices de doses tridimensionnelles, que l'on présentera section 3.1.3.

3.1.3 Matrice de dose 3D

La matrice de dose d'un patient traité par radiothérapie est la reconstitution 3D de la dose cumulée reçue sur toutes les séances dans un intervalle de 6 mois. Elle permet de cibler la tumeur et de quantifier l'irradiation des tissus non cancéreux.

Pour constituer la matrice de doses, on utilise le fichier csv de la séance de radiothérapie. Celui-ci indique pour chaque voxel la dose prescrite. Un voxel est un volume unitaire de l'image 3D, à l'image du pixel pour une image 2D. Si le patient a fait plusieurs séances de radiothérapie dans un laps de temps de 6 mois, on somme les csv correspondants à ces séances rapprochées. Ainsi, on pourra transformer le csv résultant en image NIFTI, en croppant autour de la région d'intérêt, ici le cœur. Cela donne une image en dégradé de gris, avec en clair les zones les plus exposées et en sombre les zones les moins exposées, comme on peut le voir sur la figure 11. Dans le cadre du projet, nous n'avons pas travaillé sur l'exploitation des csv, Mahmoud Bentriou l'ayant fait avant nous.

Il est possible de faire le lien entre les matrices de doses 3D et les patients de la cohorte FCCSS grâce à leurs identifiants (ctr, numcent). Au total, on possède les matrices de doses 3D de près de 4000 patients. Les patients restants sont donc exclus de l'étude. Les matrices de doses 3D seront utilisées pour les algorithmes de prédiction de pathologie cardiaque par réseau de neurones.

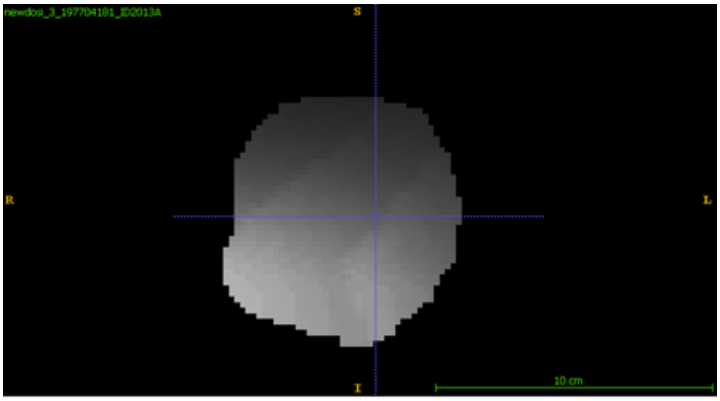

FIGURE 11 – Slice d'une image nii, visualisée à l'aide du logiciel ITK Snap

3.1.4 Données cliniques de chimiothérapie

Nous disposons également d'une base de données comprenant les doses de chimiothérapies reçues par chaque patient. Différents médicaments de chimiothérapie sont présents, on en dénombre près de 70 dans cette base de données. La structure des données est de la forme suivante : do_XXX_cumsum, avec XXX le nom du médicament et cumsum pour la somme cumulée. L'ordre de grandeur

des doses dépend des médicaments et des patients, cependant les doses cumulées sont généralement comprises entre 5 et 50 $000 mg/m^2$.

Cette base de données est particulièrement intéressante puisqu'elle nous permettra d'étudier une possible corrélation entre les doses de chimiothérapie reçues et l'apparition d'une maladie cardiaque. Nous pourrons donc rajouter ces données à notre dataset comprenant les matrices doses volumes de radiothérapie.

3.2 Méthodes

3.2.1 Analyse de survie

L'analyse de survie est une méthode statistique permettant d'estimer le temps restant avant l'occurrence d'un événement d'intérêt [10]. En biostatistique, on considère souvent la mort de l'individu, mais l'analyse de survie peut être utilisée pour d'autres événements. Dans notre cas il s'agit de l'apparition d'une maladie cardiaque grave.

Les données de survie sont des données temporelles : date de dernière nouvelle si l'événement n'est pas advenu ou alors date de l'événement si celui-ci est advenu. On les soustrait à une date de début de l'observation, ce qui nous permet d'avoir une durée de survie pour chaque individu. Certaines données peuvent être dites "censurées" quand un patient sort de l'étude ou bien quand l'événement n'est pas arrivé pendant la période d'observation. Ces données sont tout de même prise en compte dans les modèles de survie [10].

Les modèles d'analyse de survie comme Kaplan-Meier permettent d'estimer la fonction de survie $S(T)$, c'est-à-dire la probabilité que l'événement advienne après un temps t donné.

$$S(t) = \mathbb{P}(T > t)$$

Ainsi, on peut tracer une courbe de survie qui estime à un instant t la proportion de personnes pour qui l'événement d'intérêt n'est pas encore advenu [10].

Les courbes de survie permettent de comparer deux sous-populations A et B sur leurs temps de survie respectifs. Il est usuel d'appuyer cette comparaison d'un test du log-rank, qui permet de quantifier la force statistique de la comparaison et de la séparation des courbes. Ce test statistique est basé sur la statistique du Chi-deux, avec comme hypothèse nulle que les courbes de survie ne sont pas différentes. On peut ensuite extraire la p-value associée au test de valider ou de rejeter l'hypothèse nulle [14].

Dans notre cas, l'événement observé est l'apparition d'une maladie cardiaque. Si un patient est censuré, il n'a donc pas contracté de maladie cardiaque à notre connaissance. Cependant, il arrive qu'on perde certains patients de vue et ceux-ci sont considérés comme censurés. Pour ces patients, il existe un risque non négligeable qu'ils contractent une maladie cardiaque après avoir quitté la cohorte. Il est ainsi souhaitable de les exclure de l'étude si le temps de censure est trop court. En effet, nous n'avons pas la certitude qu'ils n'aient pas déclaré de maladies cardiaque après leur censure. Cependant, pour des temps de censure supérieurs à 40 ou 50 ans, on peut considérer qu'il est peu probable

que les patients contractent une maladie cardiaque suite aux traitements anti-cancéreux. Ainsi, pour un individu donnée i, on observe soit sa durée de survie T_i (s'il a eu une maladie cardiaque) soit sa durée de censure C_i (si l'évènement n'a pas encore eu lieu, ou si l'individu est sorti de l'étude). Plus précisément, la durée observée est $\min(T_i, C_i)$, avec pour condition que $C_i > C_{censure}$, où $C_{censure} = 40$ ou 50 ans.

Cela correspond à une réduction de la cohorte d'étude, tel qu'on peut le voir sur le tableau 2.

	# Patients négatifs	# Patients positifs	# Patients totaux
$C_{censure} > 40$ ans	282	1106	1378
$C_{censure} > 50$ ans	282	297	579

TABLE 2 – Cohorte d'étude

3.2.2 Machine Learning

Dans le but de mieux appréhender notre problème, nous nous sommes essayés au Machine Learning afin d'observer l'apparition d'une maladie cardiaque. Les algorithmes de machine learning sont moins coûteux en temps de calcul, il est donc intéressant d'observer les résultats obtenus avec ce type de modèle. Pour entraîner nos modèles, nous avons réalisé un pré-traitement des données.

Pré-traitement des données

Dans un premier temps, nous avons rassemblé les fichiers contenant les données nécessaires à la prédiction de l'évènement "apparition d'une maladie cardiaque". Pour réaliser les prédictions, nous nous sommes principalement appuyés sur deux bases de données. La première est la base de données clinique de la cohorte FCCSS (French Childhood Cancer Survivor Study). Dans cette base de données, nous avons également ajouté le temps de survie correspondant à chaque patient. Nous utiliserons deux temps de censure minimal, 40 ans et 50 ans, pour prédire l'apparition de l'évènement. La seconde base de données contient les indicateurs doses-volumes pour les patients.

Finalement, nous avons assemblé ces deux bases de données afin d'obtenir la base de données présentée sur la figure 12. Les colonnes 'ctr' et 'numcent', représentant l'identifiant unique du patient, nous permettent de lier les deux bases de données.

Une fois cette base de données, avec l'ensemble des données DVH, créée, nous avons annoté les patients qui appartiendront au dataset d'entraînement, de validation et de test. Le dataset d'entrainement représente 60% du dataset global, les sets de validation et de test correspondent à 20%. le Nous avons également ajouté une colonne "card_age_40" et "card_age_50" en fonction de la date de censure du patient. En effet, les modèles seront entraînés sur deux temps de censure différents, 40ans et 50ans.

	ctr	numcent	dv_V01_1320	dv_V05_1320	dv_V1_1320	dv_V2_1320	dv_V5_1320	dv_V10_1320	dv_V15_1320	dv_V20_1320	...
0	3	197204357	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.0	...
1	3	197608259	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.0	...
2	3	197704050	0.0	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.0	...
3	3	197704181	100.0	44.532788	12.736392	0.000000	0.000000	0.0	0.0	0.0	...
4	3	197704641	100.0	100.000000	94.521714	22.741981	0.017031	0.0	0.0	0.0	...

5 rows × 154 columns

FIGURE 12 – Aperçu du dataset de la cohorte FCCSS

Constructions des datasets

Comme évoqué précédemment, nous allons créer deux datasets différents en fonction du temps de censure connu des patients. Le dataset composé des patients avec le temps de censure de 50 ans comprend moins de patients. Ce dataset est plutôt équilibré puisqu'il contient 56 patients avec une maladie cardiaque et 60 patients sans maladie cardiaque. Le second dataset, avec le temps de censure égal à 40 ans est beaucoup moins équilibré (219 patients sans maladie cardiaque, 57 patients avec).

Pour le second dataset, avec le temps de censure supérieur à 40 ans, nous avons dû rééquilibrer notre dataset. Un jeu de données déséquilibré peut causer des problèmes lors de l'entraînement d'un modèle de machine Learning. En effet, le modèle peut avoir une performance plus élevée pour les classes surreprésentées et une performance plus faible pour les classes sous-représentées ; ou alors il peut manquer d'apprendre les caractéristiques distinctives des classes sous-représentées, ce qui peut réduire sa capacité à les classer correctement. Ainsi, il est important de prendre en compte l'équilibre des classes lors de la préparation des données pour éviter ces problèmes potentiels dans les modèles de machine Learning. Nous avons donc utilisé les méthodes under-sampling et d'over-sampling pour réduire le déséquilibre [4].

L'under-sampling permet de résoudre le problème de déséquilibre des classes dans les données d'apprentissage en réduisant la taille de la classe surreprésentée. Nous avons réalisé under-sampling en choisissant aléatoirement un sous-ensemble de la classe surreprésentée pour équilibrer les tailles des classes [4].

Dans un second temps nous avons mis en place une méthode d'over-sampling afin d'augmenter le nombre d'individus présent dans la classe sous-représentée. Nous avons donc généré des données synthétiques à l'aide de la technique SMOTE. Cette méthode génère de nouvelles instances de la classe sous-représentée en interpolant les points entre les exemples existants de cette classe [5].

Nous avons utilisé la méthode bootstrap pour équilibrer nos datasets. Celle-ci est illustrée sur le schéma figure 13. Il s'agit de générer n échantillons de la classe majoritaire par tirage aléatoire de taille prédéfini. Alors, chaque échantillon sera associé à la classe minoritaire et cela générera n nouveaux datasets d'entraînements. A partir de ces datasets, on peut entraîner n modèles légèrement différents, puisqu'une partie de leur ensemble d'entraînement est différente. On peut choisir différents rapports entre chaque classe dans ces sous-datasets

ainsi générés : par exemple 1 :1, mais aussi 2 :1, 1 :3,... Nous utiliserons cette méthode pour construire des sous-datasets équilibrés (de rapport 1 :1) [9].

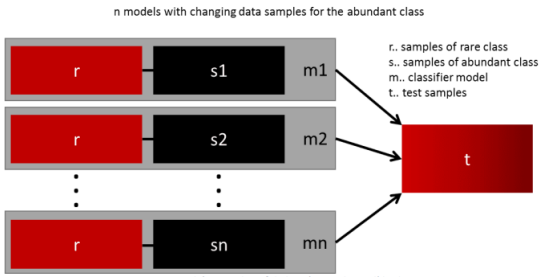

FIGURE 13 – Schéma Bootstap

Algorithmes de Machine Learning utilisés

Nous avons utilisé les algorithmes suivants sur les données dose-volume : Random Forest, le XGBoost, et le LightGBM [3][6][12]. Ces algorithmes ont été entraîné sur le dataset comprenant les patients avec le temps de censure égale à 40.

Random Forest

Cet algorithme d'apprentissage automatique est basé sur des arbres de décision. L'idée derrière Random Forest est de construire un grand nombre d'arbres de décision dépendants et de les combiner pour obtenir une prédiction plus fiable [3]. Chaque arbre est formé à partir d'un sous-ensemble aléatoire de données d'entraînement et d'un sous-ensemble aléatoire de caractéristiques, ce qui permet de réduire la variabilité et de minimiser l'overfitting [3].

XGBoost

L'algorithme XGBoost est également basé sur les arbres de décision, et utilise la technique de gradient boosting. Un des avantages de XGB par rapport à un Random Forest est sa capacité à minimiser l'overfitting en utilisant des algorithmes de régularisation [6].

LightGBM

Le LightGBM, qui est un algorithme d'apprentissage automatique développé pour la classification et la régression. Il est basé sur les arbres de décision et utilise une technique d'optimisation, le gradient boosting. Cet algorithme se distingue notamment du XGBoost en utilisant une technique d'arbre de décision appelée arbre de décision découplé, permettant une construction plus rapide. De

plus, LightGBM utilise une stratégie de division de feuille dite de gain de gain, qui permet de sélectionner les caractéristiques les plus utiles pour la division des feuilles [12].

3.2.3 Deep Learning

Dans cette section, nous allons présenter l'architecture de nos algorithmes de Deep Learning.

Traitement des données et construction du dataset

Les données utilisées pour les algorithmes de Deep Learning sont les matrices de doses, présentées section 3.1.3, et les données cliniques de chimiothérapie, présentées section 3.1.4. La même sélection des patients est effectuée que dans la section Machine Learning 3.2.2, avec des temps de censure supérieurs au égaux à 40 ans ou à 50 ans.

Aucune transformation des données de dose n'est appliquée aux matrices de doses, celles-ci étant déjà normalisées.

Concernant les données cliniques de chimiothérapie, celles-ci sont sommées selon leur famille de médicaments : agents alkylants, anthracycline ou vinca-alcaloïdes.

Architecture des réseaux de neurones

Nous avons comparé plusieurs architectures de réseaux de neurones. Le détail de l'architecture de chaque réseau est disponible sur le repository Github.

Réseau fully connected Premier réseau implémenté avec 3 couches linéaires et des fonctions d'activations ReLU. Il est noté FirstNet sur le git.

Réseau de convolutions On a implémenté un second réseau, avec 6 couches de convolutions, avec chacune une activation ReLU et une batch-normalisation. Le réseau se conclut sur trois couches linéaires. Il est noté CNN_tho sur le git.

Réseau de convolutions avec les données de chimiothérapie Inspiré de l'article [4], on fait un réseau à deux chemins. Le premier chemin est pour les matrices de dose, avec 3 couches de convolutions. Chacune est suivie d'une activation ReLU, de Dropout et de batch-normalisation. Le second chemin prend en entrée les variables cliniques, avec une couches linéaire, suivie d'une activation ReLU. On fusionne ensuite les sorties de ces deux chemins, qui passeront à nouveau par 2 couches linéaires. Il est noté NewNet sur le git.

Autres

Loss On utilise la cross-entropy Loss avec des poids associés aux proportions de chaque classe [15]

Optimizer On utilise l'optimizer Adam avec un learning rate de 10^3 [13]

Régularisation On applique une régularisation L2 avec un wieght_decay de 10^4 [16]

3.3 Moyens humains, techniques, infrastructures

Dans le cadre de ce projet, nous avons travaillé sous la direction de Mahmoud Bentriou, Sarah Lemler et Véronique Lechevalier. Nous avons utilisé des travaux préliminaires réalisés par Mahmoud Bentriou. Les données ont été préparées par l'équipe Inserm à l'Institut Gustave Roussy, représentée par Rodrigue S. ALLODJI.

Concernant les moyens techniques et en infrastructure, nous avons travaillé sur nos ordinateurs sans GPU, sur Google Colab ainsi que le Mésocentre de l'Université Paris-Saclay (GPU NVIDIA gpua100). Nous avons aussi été équipé par le MICS de deux disques durs.

4 Résultats

4.1 Courbes de survie Kaplan-Meier

A l'aide des données de la base FCCSS, on a pu construire les données de survie sur la probabilité d'avoir une pathologie cardiaque de grade 3 ou plus. Ainsi, on a pu faire des études en comparant plusieurs types de populations basées sur diverses variables cliniques.

Tout d'abord, on constate sur la figure 14 l'impact que la radiothérapie peut avoir sur la probabilité d'avoir une maladie cardiaque. On observe sur la figure 15 que la prise d'agents alkylants ou d'anthracyclines est un facteur de risque pour le risque de maladies cardiaques.

De plus, parmi les patients traités par radiothérapie, la position du cancer primaire semble avoir un impact sur la probabilité d'avoir une maladie cardiaque. En effet, les cancers primaires situé proche du coeur, labelisé "thorax" sur la figure 16, ont une moins bonne survie que les cancers primaires éloignés du coeur, par exemple tête, jambe ou même région abdominale.

4.2 Analyse en Composantes Principales

Nous avons réalisé une analyse en composante principale (ACP) sur les données issues de la base de données doses-volumes. L'ACP consiste à projeter les données sur un plan à 2 dimensions en choisissant les deux composantes principales qui capturent la plus grande variabilité possible des données d'origine [1].

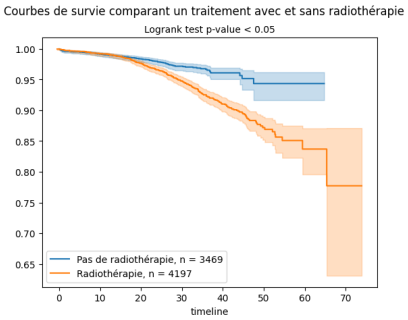

FIGURE 14 – Courbes de survie comparant traitement avec ou sans radiothérapie

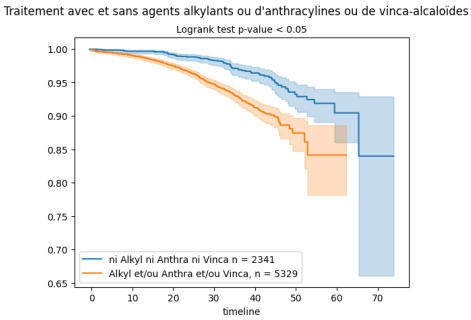

FIGURE 15 – Courbes de survie comparant traitement avec ou sans chimiothérapie

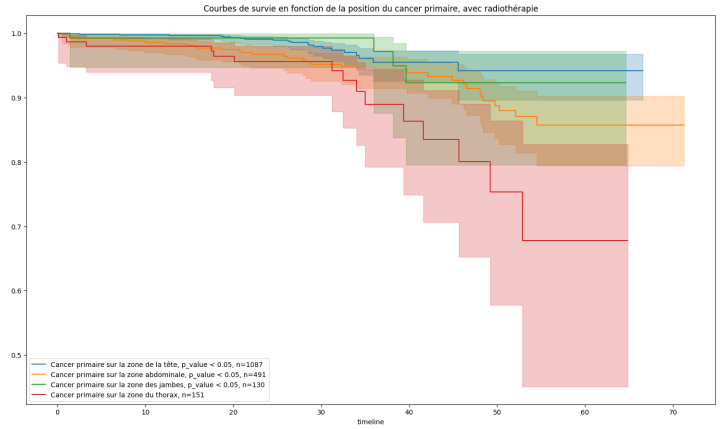

FIGURE 16 – Courbes de survie comparant les zones de cancers primaires

Cela permet de visualiser graphiquement les relations entre les différents individus ou variables dans les données. Nous obtenons un graphe où se superposent des points rouges (patients avec maladie cardiaque) et des triangle bleus (patients sans maladie cardiaque), visible figure 17. Le point rouge plus grand que les autres, situé au niveau de l'origine, ainsi que le triangle bleu plus grand que les autres, correspondent à la moyenne de chaque classe.

Du fait de la répartition des données, nous pouvons voir que les individus ne sont pas séparables facilement. Nous ne pourrons donc pas nous tourner vers une simple SVM pour classifier nos individus. Nous nous tournons donc vers des arbres de décision, en bagging ou boosting.

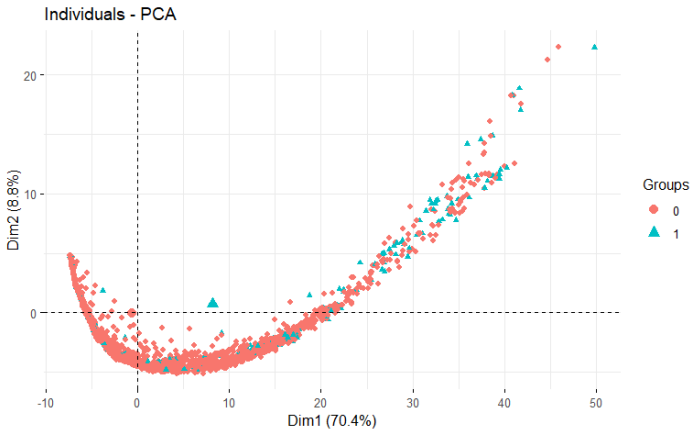

FIGURE 17 – Analyse en composantes principales des données doses-volumes

4.3 Machine learning à partir des données doses volume

Nous allons donc opter pour des algorithmes plus fins détaillés section 3.2.2. Les figures 18, 19 et 20 montrent la matrices de confusion associée à chaque algorithme, sur le set de validation.

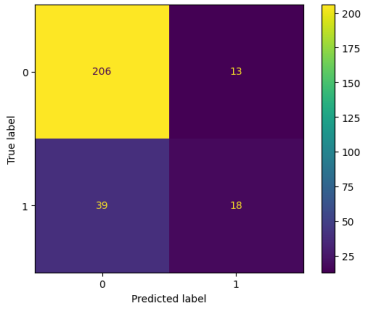

FIGURE 18 – Matrice de confusion Random Forest, avec un temps de censure supérieur à 40 ans

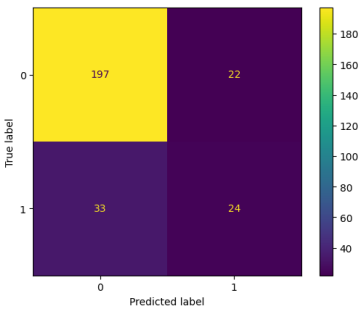

FIGURE 19 – Matrice de confusion XGBoost, avec un temps de censure supérieur à 40 ans

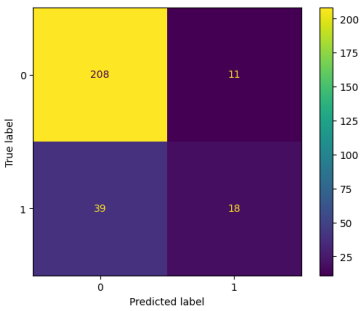

FIGURE 20 – Matrice de confusion LightGBM, avec un temps de censure supérieur à 40 ans

Le tableau 3 regroupe les résultats obtenus grâce aux trois modèles mentionnés précédemment. Le dataset de validation et de test sont de même taille. Ils comprennent 276 individus.

Temps de censure : 40 ans

Les résultats présentés dans le tableau 3 ont été calculés avec un temps de censure égale à 40. On peut remarquer que les valeurs obtenues sur le test set sont très proches des valeurs obtenues sur le validation set, ce qui confirme la

	Random Forest			XBG			LGBM		
	BA	Rappel	AUC	BA	Rappel	AUC	BA	Rappel	AUC
validation set	65%	35%	**0.73**	**69%**	**52%**	0.71	68%	49%	0.72
test set	**68%**	**88%**	**0.76**	63%	86%	0.75	62%	86%	0.73

TABLE 3 – Synthèse des résultats des algorithmes de Machine Learning pour un temps de censure supérieur à 40 ans

fiabilité de nos résultats sur nos données.

Temps de censure : 50 ans

	Random Forest			XBG			LGBM		
	BA	Rappel	AUC	BA	Rappel	AUC	BA	Rappel	AUC
validation set	**81%**	**70%**	**0.88**	78%	70%	0.83	78%	70%	0.84
test set	54%	42%	0.54	**57%**	**47%**	**0.54**	56%	47%	0.54

TABLE 4 – Synthèse des résultats des algorithmes de Machine Learning pour un temps de censure supérieur à 50 ans

Les résultats présentés dans le tableau 4 ont été calculés avec un temps de censure égale à 50. Nous pouvons observé que les résultats sur le dataset de validation sont meilleurs que ceux obtenus sur le set de test, ce qui est alarmant pour la fiabilité de nos algorithmes avec un temps de censure égal à 50 ans. Il faudrait investiguer la raison d'une telle différence.

Ainsi nous obtenons des résultats relativement bons pour prédire l'absence de maladie cardiaque chez un patient. Cependant, en utilisant les algorithmes de Machine Learning nous avons du mal à prédire la présence d'une maladie cardiaque. De plus, les données DVH sur lesquelles sont entraînés les algorithmes de Machine Learning ne donnent pas d'information de la répartition spatiale de la dose. Nous pouvons donc nous tourner vers des algorithmes de Deep Learning, entraînés sur des matrices de doses 3D, dont les résultats sont présentés dans la section suivante.

4.4 Deep Learning

Dans le tableau 5, respectivement le tableau 6, on présente les résultats des trois architectures décrites section 3.2.3 pour un temps de censure supérieur à 40 ans, respectivement supérieur à 50 ans. Les meilleurs résultats sont **en gras**. On remarque que pour le validation set des deux populations, le réseau Multipath Network performe mieux que le réseau linéraire et que le réseau convolutionnel si on compare à la balanced accuracy (BA) et au score de rappel. Il faut noté qu'il manque les valeurs de l'AUC ROC pour le validation set.

Concernant les performances sur le test set, un dataset réservé à cet usage, qui n'a jamais servi à l'entraînement des modèles, on remarque que les perfor-

mances sont très différentes de celle du validation set. Le Multipath Network n'est plus le réseau qui performe le mieux. Il semble donc assez mauvais à généraliser. Il serait intéressant d'investiguer une telle différence de performance.

	Fully Connected Network			Convolutional Network			Multipath Network		
	BA	Rappel	AUC	BA	Rappel	AUC	BA	Rappel	AUC
validation set	67%	53%	NA	69%	56%	NA	**71%**	**67%**	NA
test set	**66%**	**54%**	**0.69**	65%	43%	0.56	54%	52%	0.50

TABLE 5 – Synthèse des résultats des algorithmes de Deep Learning pour un temps de censure supérieur à 40 ans

	Fully Connected Network			Convolutional Network			Multipath Network		
	BA	Rappel	AUC	BA	Rappel	AUC	BA	Rappel	AUC
validation set	80%	70%	NA	85%	73%	NA	**86%**	**79%**	NA
test set	66%	54%	0.69	**76%**	**67%**	**0.58**	60%	65%	0.55

TABLE 6 – Synthèse des résultats des algorithmes de Deep Learning pour un temps de censure supérieur à 50 ans

Les matrices de confusion sur le validation set pour un temps de censure supérieur à 40 ans sont comparables aux figures 21, 22 et 23. On remarque que le Multipath Network est meilleur que les autres pour distinguer les positifs, au détriment d'une bonne perfromance sur les négatifs. Dans le cadre du médical, il est plus grave de manquer des positifs (alors faux négatifs) que de prédire des faux positifs. Dans cette perspective, le Multipath Network est meilleur que le réseau linéaire ou que le réseau convolutionnel sur le set de validation.

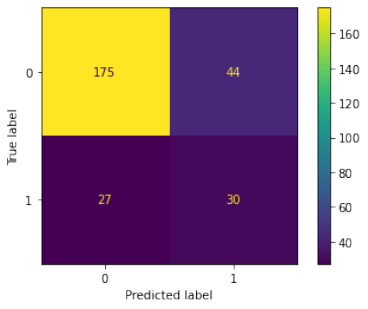

FIGURE 21 – Matrice de confusion FCN, avec un temps de censure supérieur à 40 ans

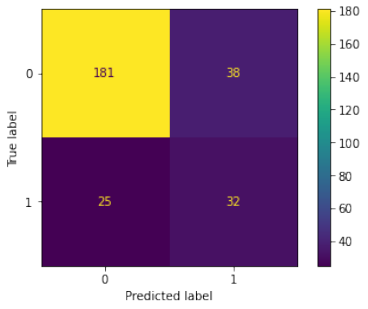

FIGURE 22 – Matrice de confusion CNN, avec un temps de censure supérieur à 40 ans

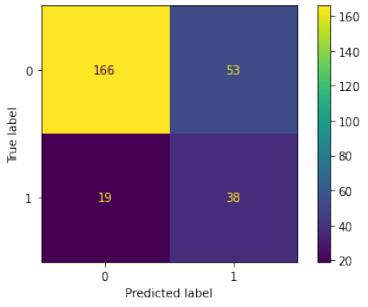

FIGURE 23 – Matrice de confusion Multipath Network, avec un temps de censure supérieur à 40 ans

On peut comparer nos résultats en Deep Learning avec les résultats obte-

nus en Machine Learning, présenté dans les tableaux 3 et 4. On remarque que l'amélioration est de l'ordre de 2 ou 3 % sur la Balanced Accuracy et de l'ordre de 20% sur le score de Rappel. Ainsi, si on ne compare que la Balanced Accuracy, on peut douter de la plus-value des algorithmes de Deep Learning par rapport à ceux de Machine Learning. Cependant, au regard du score de rappel, on constate que les algrithmes de Deep Learning sont meilleurs à la détection des patients positifs, c'est-à-dire qui pourront déclarer une maladie cardiaque grave.

Les performances sont encore à améliorer. Nous présenterons nos pistes d'amélioration dans la section 5.

5 Conclusion et Perspectives

Conclusion

Les analyses de survie nous ont aidé à intuiter quelles pouvaient être les variables cliniques influentes pour la prédiction des maladies cardiaques. On peut retenir par exemple le traitement par radiothérapie, le traitement en chimiothérapie des agents alkylants ou d'anthracylines ou bien encore le siège du cancer primaire.

Par la suite, on a construit des algorithmes de Machine Learning (XGBoost, LightGBM, BalancedRandomForest) sur les indicateurs dose-volumes. Ces algorithmes nous permettent d'établir une baseline, dont les performances seront comparées aux résultats ultérieurs. Le principal frein à l'amélioration de nos modèles est le manque de données avec pathologie cardiaque. Pour cela, nous avons essayé plusieurs techniques d'échantillonnage pour agrandir artificiellement et pour équilibrer le dataset d'entraînement. De plus, les indicateurs dose-volumes ne fournissent qu'un résumé d'une représentation 3D, nous perdons donc de l'information qui pourrait s'avérer précieuse.

Pour finir, nous avons développé différents réseaux de neurones : un réseau linéaire, un réseau convolutionnel et un réseau multi chemin inspirée de la bibliographie. Ces réseaux ont été entraînés avec Adam régularisé et une Cross Entropy Loss weighted, sur la base de matrices de doses 3D. Les résultats obtenus sont comparables aux résultats obtenus en Machine Learning. Toutefois, les réseaux de neurones semblent meilleurs pour prédire les individus positifs que les algorithmes de Machine Learning. Les performances sur le test set sont malheureusement décorellées aux performances sur le set de validation.

Perspectives

Tout d'abord, il serait intéressant de comprendre d'où vient la différence de performance entre le set de test et le set de validation. Il faudrait peut-être augmenter la régularisation, en ajoutant du Dropout par exemple.

Pour améliorer nos résultats en Deep Learning, nous pourrions créer une architecture qui, en plus de prendre en entrée les matrices de doses 3D, pourrait

utiliser les données cliniques mises à notre disposition. Cette architecture serait proche de celle présentée dans l'article [11]. Intuitivement, il semble y avoir une corrélation entre l'apparition d'une maladie cardiaque et le parcours de soins du patient. Afin d'améliorer nos résultats, nous souhaiterions donc implémenter un réseau de neurones capable de prendre en considération les différents traitements suivis par le patient. Il sera sûrement nécessaire de faire une normalisation des données cliniques, voire une sélection de variables. Enfin, on pourrait essayer d'augmenter la taille du réseau de convolutions.

Il pourrait aussi être intéressant de construire des cartes d'activation, comme dans l'article [11], afin de voir quelles sont les zones du coeur les plus à risques.

Remerciements

Nous tenons à remercier chaleureusement chaque personne nous ayant permis de réaliser ce projet.

Tout d'abord, nous tenons à remercier vivement DR Mahmoud BENTRIOU, chercheur au MICS, qui nous a suivi durant ces six mois de travail. Nous le remercions chaleureusement pour tout ce qu'il a pu nous apporter pendant ce temps à ses côtés. Ses conseils avisés nous ont permis d'apprendre énormément de choses et d'acquérir de nouvelles connaissances et compétences. Nous souhaitons aussi le remercier pour la confiance qu'il a su nous accorder dès les premiers instants.

Nous remercions vivement DR Veronique LETORT et DR Sarah LEMLER qui ont su se rendre disponibles quand cela été nécessaire et ont toujours pris le temps pour nous conseiller de façon pédagogique.

Nous souhaitons remercier Emmanuel ODIC, responsable de la mention Healthcare et Services en Biomédicales à CentraleSupélec, pour cette année très enrichissante et les cours de qualités qu'il a pu dispenser durant cette année, nous permettant ainsi de développer nos connaissances et compétences dans le secteur de la santé. Nous tenons aussi à le remercier pour son écoute et sa disponibilité.

Nous remercions aussi l'équipe du Mésocentre pour leur confiance et leur sérieux dans la gestion du super-ordinateur de l'Université Paris-Saclay.

Références

[1] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews : computational statistics*, 2(4) :433–459, 2010.

[2] AL Appelt, B Elhaminia, A Gooya, A Gilbert, and M Nix. Deep learning for radiotherapy outcome prediction using dose data–a review. *Clinical Oncology*, 34(2) :e87–e96, 2022.

[3] Leo Breiman. Random forests. *Machine learning*, 45 :5–32, 2001.

[4] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

[5] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote : synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16 :321–357, 2002.

[6] Tianqi Chen and Carlos Guestrin. Xgboost : A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[7] RE Drzymala, Radhe Mohan, L Brewster, J Chu, Michael Goitein, W Harms, and M Urie. Dose-volume histograms. *International Journal of Radiation Oncology* Biology* Physics*, 21(1) :71–78, 1991.

[8] Issam El Naqa and Martin J Murphy. *Machine and Deep Learning in Oncology, Medical Physics and Radiology*. Springer, 2022.

[9] Ravi Garg, Shu Dong, Sanjiv Shah, and Siddhartha R Jonnalagadda. A bootstrap machine learning approach to identify rare disease patients from electronic health records. *arXiv preprint arXiv :1609.01586*, 2016.

[10] Manish Kumar Goel, Pardeep Khanna, and Jugal Kishore. Understanding survival analysis : Kaplan-meier estimate. *International journal of Ayurveda research*, 1(4) :274, 2010.

[11] Bulat Ibragimov, Diego AS Toesca, Yixuan Yuan, Albert C Koong, Daniel T Chang, and Lei Xing. Neural networks for deep radiotherapy dose analysis and prediction of liver sbrt outcomes. *IEEE journal of biomedical and health informatics*, 23(5) :1821–1833, 2019.

[12] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm : A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[13] Diederik P Kingma and Jimmy Ba. Adam : A method for stochastic optimization. *arXiv preprint arXiv :1412.6980*, 2014.

[14] David G Kleinbaum, Mitchel Klein, David G Kleinbaum, and Mitchel Klein. Kaplan-meier survival curves and the log-rank test. *Survival analysis : a self-learning text*, pages 55–96, 2012.

[15] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[16] Zijun Zhang. Improved adam optimizer for deep neural networks. In *2018 IEEE/ACM 26th international symposium on quality of service (IWQoS)*, pages 1–2. Ieee, 2018.