UNIVERSITÉ ALIOUNE DIOP DE BAMBEY

N° :.....

UFR : Sciences Appliquées et Technologies de l'Information et de la Communication (SATIC)

Département : Mathématiques

Spécialité : Statistique et Informatique Décisionnelle

Mémoire présenté par :

AHMED SEFDINE

Pour l'obtention du diplôme de :

Master en Statistique et Informatique Décisionnelle

Sujet du mémoire

ANALYSE DE SURVIE POUR LE CANCER DE L'ESTOMAC AU SÉNÉGAL

Soutenu publiquement le 26 avril 2025, devant le jury ci-dessous :

Président	Pr. Gaoussou Camara	UADB
Examinateur	Dr. Mouhamed Amine Niang	UADB
	Dr. Ibrahima Faye	UADB
	Dr. Papa Ibrahima Ndiaye	UADB
Directeur de mémoire	Pr. Aba Diop	UADB
Co-directeur de mémoire	Dr. Idrissa Sy	HALD

Année universitaire 2023-2024

Dédicace

À la lumière de ma vie,
À ceux dont la présence discrète m'a porté plus loin que les mots,
À mes parents, pour leur amour sans bornes et leurs innombrables sacrifices.

À mes enseignants, amis, compagnons de route et d'espoir, qui ont su me faire sourire même dans les jours les plus sombres.

Ce travail est le résultat de toutes ces âmes généreuses.

Remerciements

Au terme de ce travail, je tiens tout d'abord à exprimer ma gratitude la plus sincère à Allah, le Tout-Puissant, pour m'avoir donné la force, la patience et la persévérance nécessaires tout au long de ce parcours. C'est grâce à Sa grâce et Sa miséricorde que j'ai pu mener à bien ce mémoire, malgré les nombreuses difficultés rencontrées.

Je souhaite ensuite remercier profondément le Pr. Aba DIOP et Dr. Idrissa SY, mes encadreurs, pour leur accompagnement exemplaire tout au long de cette étude. Leur rigueur scientifique, leur grande disponibilité, leurs conseils avisés et leur bienveillance ont été des sources constantes de motivation et d'inspiration pour moi. Grâce à leur encadrement de qualité, j'ai pu progresser dans mes réflexions, approfondir mes compétences en analyse de données et m'initier concrètement à la recherche scientifique.

Je veux également remercier, toutes les personnes qui ont contribué, de près ou de loin, à la réalisation de ce mémoire. Il serait difficile de tous les nommer, mais je pense notamment aux enseignants, collègues, amis, personnels administratifs et professionnels de santé rencontrés durant ce parcours. Chacun, à sa manière, a enrichi cette expérience par un mot d'encouragement, un conseil technique, ou un soutien moral précieux.

Résumé

L'analyse des données de survie, une branche spécialisée de l'analyse statistique, se concentre sur l'étude du temps jusqu'à la survenue d'un événement spécifique, tel que le décès de patients atteints de cancer gastrique au Sénégal. Cette étude a pour objectif de comparer l'efficacité de diverses approches, tant classiques que modernes, pour modéliser ce phénomène. Elle commence par l'application de méthodes statistiques traditionnelles, comme l'estimateur de Kaplan-Meier, utilisé pour comparer les courbes de survie entre groupes de variables qualitatives, et le modèle de régression de Cox, qui sert à identifier les facteurs significativement associés au temps de survie. Par la suite, des modèles d'apprentissage automatique, tels que le Random Survival Forest, le Gradient Boosting Survival Tree et un modèle de Deep Survival, ont été explorés pour améliorer la précision des prédictions. Cette étude met en lumière l'importance de combiner des approches classiques et modernes pour modéliser efficacement le temps de survie et ouvre la voie à de nouvelles stratégies pour la gestion des patients atteints de cancer gastrique au Sénégal.

Mots-clés : Analyse de survie, Cancer gastrique, Kaplan-Meier, Modèle de Cox, Apprentissage automatique, Random Survival Forest, Gradient Boosting, Deep Survival, Oncologie

Abstrat : Survival data analysis, a specialized branch of statistical analysis, focuses on the study of the time to occurrence of a specific event, such as the death of gastric cancer patients in Senegal. The aim of this study is to compare the effectiveness of various approaches, both classical and modern, in modeling this phenomenon. It begins with the application of traditional statistical methods, such as the Kaplan-Meier estimator, used to compare survival curves between groups of categorical variables, and the Cox regression model, used to identify factors significantly associated with survival time. Subsequently, machine learning models such as the Random Survival Forest, the Gradient Boosting Survival Tree and a Deep Survival model were explored to improve prediction accuracy. This study highlights the importance of combining classi-

cal and modern approaches to effectively model survival time, and opens the way to new strategies for the management of gastric cancer patients in Senegal.

Keywords : Survival analysis, Gastric cancer, Kaplan-Meier, Cox model, Machine learning, Random Survival Forest, Gradient Boosting, Deep Survival, Oncology

Table des matières

Table des figures

Liste des tableaux

Abréviations et Notations

RSF : Random Survival Forest
GBST : Gradient Boosting Survival Tree
IA : Intelligence artificielle
HALD : Hopital Aristide Le Dantec
ES : Erreur Standard
IBS : Brier Score Intégré
C-index : Indice de concordance de Harell
Deep Surv : Deep Survival
Cox PH : Risque (Hasard) proportionnelle de Cox
HR : Hasard Ratio
UADB : Université Alioune Diop de Bambey
UFR : Unité de Formation et de Recherche
SHAP : SHapley Additive exPlanations
SATIC : Sciences Appliquées et Technologie de l'information et de la Communication
HER2 : Human Epidermal Growth Factor Receptor 2 (ou Récepteur 2 du Facteur de Croissance Épidermique Humain)

Introduction Générale

L'analyse de données, à l'intersection des statistiques, des mathématiques et des probabilités, est devenue un domaine incontournable dans la résolution de problématiques complexes et multidimensionnelles. Ce domaine dépasse largement les frontières des techniques de traitement de données : il représente aujourd'hui une compétence stratégique essentielle pour transformer les organisations, les industries et les individus en acteurs informés et proactifs dans un monde de plus en plus gouverné par la donnée.

Dans ce contexte, l'analyse de données joue un rôle fondamental dans la prise de décision éclairée, l'innovation et la compétitivité. Cette importance est particulièrement manifeste dans un environnement où la quantité de données disponibles ne cesse de croître, rendant nécessaires des outils de plus en plus sophistiqués pour en extraire de la valeur.

Les récentes avancées en science des données, notamment grâce à l'apprentissage automatique (une branche de l'intelligence artificielle), ont permis de révolutionner l'exploitation des données, particulièrement dans des domaines critiques comme la médecine. Ces technologies permettent désormais de traiter des volumes massifs d'informations médicales, ouvrant ainsi des perspectives inédites pour des approches de plus en plus personnalisées et précises, comme en témoigne l'émergence de la médecine de précision [27].

L'analyse des données de survie, une sous-discipline spécifique de l'analyse statistique, se concentre sur l'étude du temps jusqu'à la survenue d'un événement particulier, que ce soit la mort d'un individu ou l'apparition d'une maladie. En médecine, cette approche est d'une importance capitale, car elle permet de mieux comprendre les mécanismes des maladies, d'optimiser les stratégies thérapeutiques et d'améliorer les décisions cliniques. L'analyse de survie est particulièrement utilisée en recherche biomédicale, surtout dans les cas de maladies graves où la durée de survie est un facteur clé pour évaluer l'efficacité des traitements et la gestion des patients.

Dans le cadre de cette étude, nous nous intéressons spécifiquement à l'analyse

de survie pour le cancer de l'estomac, une pathologie qui reste un défi majeur dans le domaine médical en raison de sa complexité et de ses nombreux facteurs de risque. Le cancer gastrique se développe lorsque des cellules de l'estomac subissent des mutations et se multiplient de manière incontrôlée, compromettant ainsi le bon fonctionnement de l'organe.

Malheureusement, ce cancer est souvent diagnostiqué à un stade avancé, ce qui limite les options thérapeutiques et compromet le pronostic des patients [16]. Plusieurs facteurs, tels que l'infection chronique par *Helicobacter pylori*, les antécédents familiaux, le régime alimentaire, le tabagisme et l'alcoolisme, jouent un rôle important dans le développement de la maladie [28]. Ces éléments, ainsi que les métastases pouvant se propager à d'autres parties du corps, rendent l'étude de cette maladie particulièrement complexe.

Bien que des progrès aient été réalisés dans la prise en charge du cancer gastrique, le pronostic reste préoccupant, avec des taux de survie à cinq ans toujours faibles, surtout dans les formes avancées. Cette situation souligne la nécessité de mieux comprendre les facteurs influençant la survie des patients et de développer des outils capables de prédire avec plus de précision les risques associés à la maladie.

Les modèles statistiques classiques, tels que l'estimateur de Kaplan-Meier ou les modèles de régression de Cox, ont permis des avancées importantes dans l'analyse des données de survie. Toutefois, ces modèles nécessitent aujourd'hui d'être complétés par des approches plus modernes, comme les algorithmes d'apprentissage automatique, afin d'exploiter pleinement le potentiel des données disponibles [34]. L'intégration des techniques classiques et de l'apprentissage automatique ouvre la voie à des solutions d'aide à la décision plus efficaces et adaptées aux besoins actuels du domaine médical.

Problématique

Le cancer gastrique représente une lourde charge pour les systèmes de santé à l'échelle mondiale. En 2022, il figurait parmi les cancers les plus fréquents, occupant la cinquième position pour l'incidence et la quatrième pour la mortalité, avec près d'un million de nouveaux cas et plus de 660 000 décès [7]. En Afrique, la situation est particulièrement inquiétante, avec 33 352 nouveaux cas et un taux de mortalité de 86,14% [7]. Le Sénégal, pour sa part, connaît un taux de mortalité élevé pour ce type de cancer, avec 88% des patients décédant dans les deux ans suivant le diagnostic [7].

Malgré de nombreuses études pronostiques, la majorité se concentrent sur la classification binaire de l'événement (décès ou non), négligeant la dimension temporelle, qui est essentielle pour comprendre l'évolution de la maladie [31, 30, 29, 33]. Cette lacune met en évidence le besoin d'approches analytiques plus fines, capables de modéliser non seulement la survenue de l'événement, mais aussi la durée de survie des patients, afin d'offrir un pronostic plus précis et utile dans la pratique clinique.

Objectif principal de l'étude

L'objectif principal de cette étude est d'analyser la survie des patients atteints de cancer gastrique, en se concentrant particulièrement sur la durée jusqu'au décès. Nous nous proposons de comparer plusieurs méthodes statistiques et d'apprentissage automatique pour évaluer l'impact des facteurs de risque sur la survie des patients.

Les objectifs spécifiques sont les suivants :

- Utiliser l'estimateur de Kaplan-Meier pour comparer la distribution de survie entre différents groupes de patients.
- Appliquer le modèle de régression de Cox pour évaluer l'influence de chaque variable sur la survie des patients.
- Mettre en œuvre des techniques avancées d'apprentissage automatique, telles que le Random Survival Forest, le Gradient Boosting Survival Tree et les modèles de survie profonds, afin de maximiser la prédiction de la durée de survie.

Chapitre 1

Méthodes

1.1 Type, durée et environnement de l'étude

Nous avons réalisé une étude rétrospective qui couvre une période de 13 ans (2007-2020) au service de cancérologie et au service de chirurgie générale de l'hôpital Aristide Le Dantec (HALD) à Dakar. La période de suivi était de 5 ans.

1.2 Population étudiée

Critère d'inclusion : Ont été inclus dans notre étude les patient ayant bénéficié d'un traitement complet composé de chirurgie exclusive ou une chirurgie associée à la thérapie et suivi en post-opératoire. Les décès survenu dans les 30 jours après la chirurgie étaient exclus. Car ce décès était considère liée à l'intervention chirurgicale non à la maladie. Au total 337 patients ont été colligés, respectant les critères d'inclusion.

1.3 Présentation des données

Les données collectées incluent :

— Paramètres cliniques : âge, sexe, antécédents médicaux, motifs de consultation.
— Signes biologiques et cliniques : analyses hématologiques, aspects tumoraux, métastases.
— Traitements et survie : type de traitement, survenu de décès, censure.

Dans le cadre de notre étude, nous avons rassemblé des informations sur 337 patients avec 18 critères différents. Ces critères, que vous pourrez retrouver dans le tableau ci-dessous, ont été recueillis pour mener notre analyse (Tab. 1.1).

TABLE 1.1 – Liste des variables étudiées

	Variables	**Paramètres**	**Mesure**
Identité	AGE	État civil	Continue
	SEXE	État civil	Nominale
Antécédent	Cardiopathie		Nominale
	Ulceregastrique		Nominale
	Douleurepigastrique		Nominale
	Ulcero-bourgeonnant		Nominale
	Constipation		Nominale
	Dénutrition		Nominale
	Tabac		Nominale
Type Adénocarcinome	Mucineux		Nominale
	Tubuleux		Nominale
Aspect de la tumeur	Infiltrant		Nominale
	Sténosant		Nominale
Envahissement	Métastases		Nominale
	Adénopathie		Nominale
	Traitement		Nominale
Évolution	Temps de suivi (mois)		Continue
	Décès		Nominale

1.4 Analyses exploratoires des données

Dans cette étude, 223 patients, soit 66,2% de l'effectif total, avaient présenté l'événement d'intérêt (le décès), tandis que 114 patients (33,8%) étaient censurés, c'est-à-dire qu'ils n'avaient pas connu l'événement au moment de la fin du suivi.

La base de données a été divisée en deux sous-ensembles : 80% des patients ont été affectés à l'échantillon d'entraînement, et les 20% restants à l'échantillon de test, en vue de la modélisation.

1.4.1 Caractéristiques démographiques

Notre étude démontre que l'âge des patients variait entre 31 et 81 ans, avec une moyenne de 52,14 ans et un écart type de 9,85.
On a observé une légère prépondérance masculine, avec 181 hommes contre 156 femmes, ce qui donne un ratio hommes/femmes de 1,16.

Les données ont révélé que la tranche d'âge la plus dominée était celle des 41–50 ans, regroupant 123 personnes, soit 36,50% du total. Les groupes 51–60 ans et 31–40 ans venaient ensuite avec des proportions de 30,27% et 12,46%, respectivement. À l'inverse, le groupe des 71–81 ans était le moins nombreux, comptant seulement 17 individus, ce qui représente 5,04%. Globalement, les

femmes étaient plus présentes dans presque toutes les catégories d'âge, sauf pour la tranche des 41–50 ans, où les hommes étaient prédominants (Tab. 1.2).

TABLE 1.2 – Répartition des patients par classe d'âge et sexe

Classe d'âge	**Femmes**	**Hommes**	**Total**	**Pourcentages (%)**
31–40	21	21	42	12.46%
41–50	34	89	123	36.50%
51–60	60	42	102	30.27%
61–70	31	22	53	15.73%
71–81	10	7	17	5.04%

1.4.2 Profil clinico-pathologique et modalités thérapeutiques

Les patients inclus dans l'étude présentaient diverses conditions pathologiques et cliniques.
Parmi eux, 32,5% souffraient de cardiopathies, 38,6% d'ulcères gastriques et 86,7% ressentaient des douleurs épigastriques. Un ulcère bourgeonnant était identifié chez 33,4% des cas, la constipation chez 51,8% et la dénutrition chez 34,3%. Concernant les habitudes de vie, 53,9% des patients étaient fumeurs.

Sur le plan histologique, les lésions étaient principalement de type mucineux (64,8%) et tubulaire (72,6%).
De plus, 51,9% des cas montraient un caractère infiltrant, et 45,5% avaient des lésions sténosantes. Les métastases étaient présentes dans 42,5% des cas, et des adénopathies étaient détectées dans 40,1% des patients.

En ce qui concerne le traitement, 44,9% des patients ont subi une chirurgie associée à une chimiothérapie, tandis que 55,1% ont été traités uniquement par chirurgie. Le taux de mortalité au cours de l'étude était de 66%.

1.4.3 Évolution du taux de survie

La survie globale à 6 mois, 1 an et 5 ans après le traitement était respectivement de 78%, 64% et 28,2%.
Ces chiffres mettent en avant le caractère rapide et sévère de la maladie, soulignant l'importance d'une prise en charge précoce, bien que certains patients connaissent une progression plus lente sur le long terme.

Le taux de survie des patients a progressivement diminué au fil des mois. Entre 2 et 15 mois, 60,53% des patients étaient encore en vie. Ce taux a chuté

à 48,66% entre 16 et 30 mois, puis à 43,91% entre 31 et 45 mois. Enfin, entre 46 et 60 mois, seulement 33,82% des patients étaient toujours vivants (Tab. 1.3).

TABLE 1.3 – Évolution du taux de survie des patients

	2-15 MOIS	16-30 MOIS	31-45 MOIS	46-60 MOIS
Décès	133	40	16	34
Décès Cumule	133	173	189	223
Vivant	204	164	148	114
Taux de Survie	60,53%	48,66%	43,91%	33,82%

Note :

$$\text{Taux de survie} = \frac{\text{Vivant}}{\text{Population total}} * 100$$

$$\text{Vivant} = \text{Population total} - \text{Décès Cumulé}$$

1.5 Hypothèses de recherche

Les variables dont la distribution de survie est significative par rapport à la survenue instantanée du décès. L'estimation des temps de survie en fonction de l'impact des variables. Montrer que les méthodes IA permettent de mieux explorer les relations entre les variables étudiées par rapport aux modèles classiques.

1.6 Métrique d'évaluation

1.6.1 Indice de concordance

L'indice de concordance, introduit par Harell [10], est la mesure de performance la plus utilisée pour l'analyse du temps jusqu'à l'événement. Il mesure la fraction de paires de sujets qui sont correctement ordonnées au sein des paires qui peuvent être ordonnées. La valeur la plus élevée (et la meilleure) que l'on peut obtenir est 1, ce qui signifie qu'il y a une concordance complète entre l'ordre des temps observés et des temps prédits.

La valeur la plus basse que l'on peut obtenir est 0, ce qui signifie qu'il s'agit d'un modèle parfaitement erroné, tandis qu'une valeur de 0,5 signifie qu'il s'agit d'un modèle aléatoire. Pour calculer l'indice de concordance, nous prenons d'abord chaque paire de l'ensemble de test de sorte que le temps observé précédemment ne soit pas censuré. Ensuite, nous ne considérons que les paires (i, j) telles que$i < j$ et nous éliminons également les paires pour lesquelles les

temps sont égaux, sauf si au moins l'une d'entre elles a une valeur d'indicateur d'événement de 1.

Ensuite, nous calculons pour chaque paire (i, j) un score $C_(i, j)$ qui pour $Y_i = Y_j$ est 1 si le sujet avec un temps antérieur (entre i et j) a un risque prédit plus élevé (entre i et j), est de 0,5 si les risques sont égaux et 0 sinon. Pour $Y_i = Y_j$et $\delta_i = \delta_j = 1$, nous définissons $C_(i, j) = 1$ si les risques sont liés et 0,5 dans le cas contraire. Si un seul des éléments $\delta_i ou \delta_j$ est égal à 1, nous définissons $C_(i, j) = 1$ si le risque prédit est plus élevé pour le sujet avec $\delta = 1$ et $0, 5$ dans le cas contraire.

Enfin, nous calculons l'indice de concordance comme suit :

$$\frac{1}{|P|} \sum_{(i,j)\in P} C_(i; j), \tag{1.1}$$

Oú P représente l'ensemble des paires éligibles (i, j).

1.6.2 Score de Brie

Le score de Brier - Brier Score - (BS) permet d'évaluer la précision de l'estimation de la fonction de survie à un instant t [6]. Il mesure la distance moyenne entre le statut et la fonction de survie prédite.

Le score de Brie à un instant donné t est défini comme :

$$B(t) = \frac{1}{n} \sum_{i=1}^{n} \delta_i (S(t|x_i)) - y_i)^2 \omega_i, \tag{1.2}$$

Où :

- $S(t|x_i)$ est la probabilité de survie prédite pour l'individu i à l'instant t.
- y_i est la survie observée (égale à 1 si l'événement (Décès) survient avant ou à t , et 0 sinon).
- δ_i est l'indicateur de non-censure (1 si l'individu i est observé non censuré, 0 sinon).
- ω_i est une pondération pour tenir compte de la censure des données

Le score de Brier intégré - Integrated Brier Score (IBS)

Le IBS est ensuite calculé comme l'intégrale du score de Brier à travers le temps, pondérée par une fonction de risque de censure.

$$IBS = \int_0^x B(t)\omega(t)dt, \tag{1.3}$$

où :

— T représente la durée maximale de suivi dans les données.
— $\omega(t)$ est une fonction de pondération basée sur la survie estimée.

L'IBS résume la précision des prédictions du modèle sur l'ensemble de la période de suivi. Un IBS faible indique un modèle performant, capable de produire des estimations précises des probabilités de survie à travers le temps. Cette métrique est couramment utilisée pour comparer des modèles, particulièrement dans des contextes où les données sont censurées ou hétérogènes [32].

Chapitre 2

Les Méthodes Classiques d'analyses de Survie

2.1 Introduction

Dans le cadre du cancer de l'estomac, l'objectif de l'analyse de survie est d'étudier le lien entre diverses variables cliniques et le moment où survient un événement, tel que le décès du patient.
Cette analyse, couramment utilisée dans la recherche clinique, se distingue des méthodes traditionnelles de régression en raison de la nature particulière de ses données [26].
En effet, l'une des spécificités de l'analyse de survie réside dans le fait que certaines observations peuvent être censurées, c'est-à-dire que nous ne disposons pas d'une information complète sur le temps jusqu'à l'événement pour tous les individus suivis. Cela peut survenir si l'événement, décès dans notre cas, n'a pas eu lieu pendant la période d'observation ou si le patient quitte l'étude avant la survenue de l'événement.

On distingue trois types de censure :

- La censure à droite, la forme la plus courante, se produit lorsque le patient n'a pas expérimenté l'événement avant la fin de l'étude, ou lorsque celui-ci quitte l'étude prématurément.
- En revanche, la censure à gauche survient lorsque l'on sait uniquement que l'événement a eu lieu avant une certaine date.
- La censure par intervalle, quant à elle, indique que l'on connaît une plage temporelle durant laquelle l'événement s'est produit.

Ces différentes formes de censure introduisent des défis dans l'analyse des données, car il devient essentiel de prendre en compte ces informations partielles pour estimer correctement les risques et le temps de survie des patients. Les modèles de survie, tels que l'estimateur de Kaplan-Meier et le modèle de ré-

gression de Cox, sont des outils cruciaux pour analyser ces données censurées [29].
Ces méthodes permettent de prédire non seulement la probabilité de survenue de l'événement (décès) mais aussi d'estimer le temps de survie des patients en tenant compte des informations censurées et non censurées.
Dans cette étude, nous avons explorer ces approches pour mieux comprendre la survie des patients atteints de cancer de l'estomac au Sénégal.

Fonctions associées aux distributions de survie

Soit X une variable aléatoire de durée de vie. Plusieurs fonctions peuvent caractériser la loi de probabilité de X :

La fonction de probabilité de X

$$f(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq X \leq \Delta t)}{\Delta t}.$$

$f(t)$ représente la probabilité que l'évènement se produit juste après t.

Fonction de répartition F(x) :

$$F(t) = \mathbb{P}(X \leq t) = \int_0^t f(u)du.$$

Cette fonction représente la probabilité de présenter l'évènement entre 0 et t, elle est liée à la fonction de densité de probabilité.

La fonction de Survie S(x) :

$$S(t) = \mathbb{P}(X > t) = 1 - F(t) = \int_0^t f(u)du.$$

La fonction de survie représente la probabilité de survivre au-delà de t.

La fonction de risque instantané $h(t)$

Cette fonction de risque représente la probabilité qu'un décès dû au cancer de l'estomac survient à l'instant t, sachant qu'il n'a pas encore eu lieu. Elle est définie à chaque instant t par :

$$h(t) = \frac{\mathbb{P}(t \leq T \leq t + h/T \geq t)}{h}.$$

La fonction de risque cumulé $\Lambda(x)$:

$$\Lambda(t) = \int_0^t h(t)dt.$$

C'est l'intégrale de la fonction de risque [24].

2.2 Le modèle de Kaplan-Meier

2.2.1 Introduction

Le modèle de Kaplan-Meier est une méthode statistique non paramétrique utilisée pour estimer la fonction de survie à partir de données sur la durée de vie.
Introduite par Edward Kaplan et Paul Meier en 1958 [15], cette méthode permet d'analyser la probabilité de survie à différents intervalles de temps tout en prenant en compte les données censurées (observations pour lesquelles la durée de survie exacte n'est pas connue). Ce modèle est largement employé dans le domaine médical pour comparer les distributions de survie entre différents groupes de patients [9].

2.2.2 Équation mathématique de Kaplan Meier

L'éstimateur $\hat{S}(t)$ est également appelé Produit Limite car il s'obtient comme limite d'un produit. $\hat{S}(t)$ est une fonction en escalier décroissante, continue à droite [8].
Soit t le temps qui s'écoule entre le début de l'étude et un événement donné de la survenue du décès. $\hat{S}(t)$ est la probabilité qu'un individu survive au-delà du temps t.
Estimation par Kaplan-Meier :

$$\hat{S}(t) = \prod_{t_1 \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

où
— t_i : le temps qui s'écoule pour chaque événement observé (décès).
— d_i : le nombre d'événements (décès) survenus au temps t_i
— n_i : nombre d'individus (en risque) juste avant t_i. C'est-à-dire le nombre d'individus encore en observation et n'ayant pas encore eu l'événement ou été censurés.

2.3 Résultats : Kaplan-Meier

2.3.1 Fonction de survie

La représentation de la Fig. 2.1 montre comment la probabilité de survie diminue au fil du temps (axe X en mois). Chaque descente correspond à un événement (un décès). Les points représentent les individus ayant quitté le suivi avant l'événement (décès). Ils ne modifient pas directement la courbe mais réduisent le nombre de participants.
L'intervalle de confiance est plus étroit au début (à t=0) et s'élargit au fil du temps, devenant plus large à t=60 (Fig. 2.1).

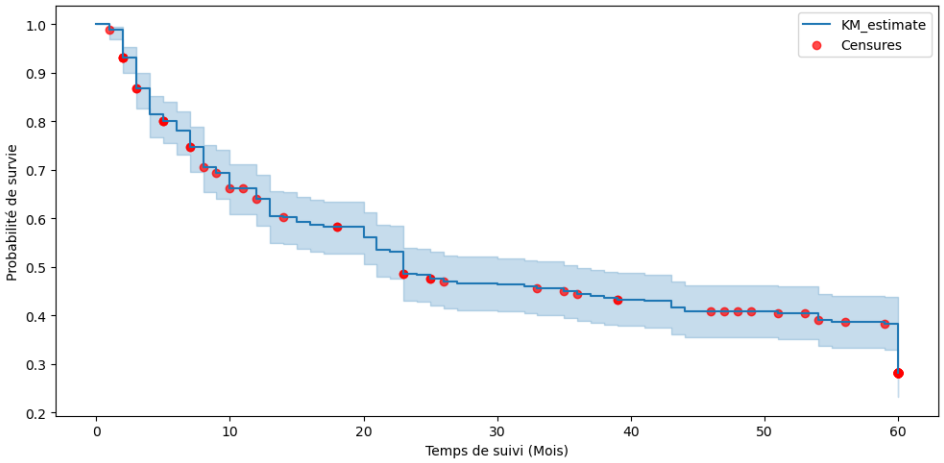


FIGURE 2.1 – Fonction de survie

Les résultats de cette courbe montrent qu'au début (0-20 mois), la probabilité de survie diminue avec le temps, indiquant un risque élevé de décès. Entre le 20ème et le 45ème mois, on observe une légère diminution de la probabilité de survie.
Après 45 mois, la probabilité devient plus stable, indiquant un risque de décès plus faible pour les personnes restantes. La courbe offre une vision claire de la survie globale, et les censures sont bien visualisées, permettant une interprétation équilibrée des données.

2.3.2 Distributions de survie en fonction du type de traitement

Les traitements proposés aux patients de l'étude étaient la chirurgie exclusive et la chirurgie associée à une chimiothérapie. La représentation graphique des distributions a montré qu'il n'y avait pas de différence significative entre les effets des deux traitements (Fig. 2.2). La p-value du test de log-rank était de 0,238, supérieure au seuil de significativité de 5%.

Cependant, la moyenne de survie pour les patients ayant bénéficié de la chirurgie avec chimiothérapie (34,73 mois) était supérieure à celle des patients traités uniquement par chirurgie (29,74 mois).
Il en allait de même pour les médianes de survie (41 mois pour la chirurgie avec chimiothérapie contre 21 mois pour la chirurgie seule) (Tab. 3.2).

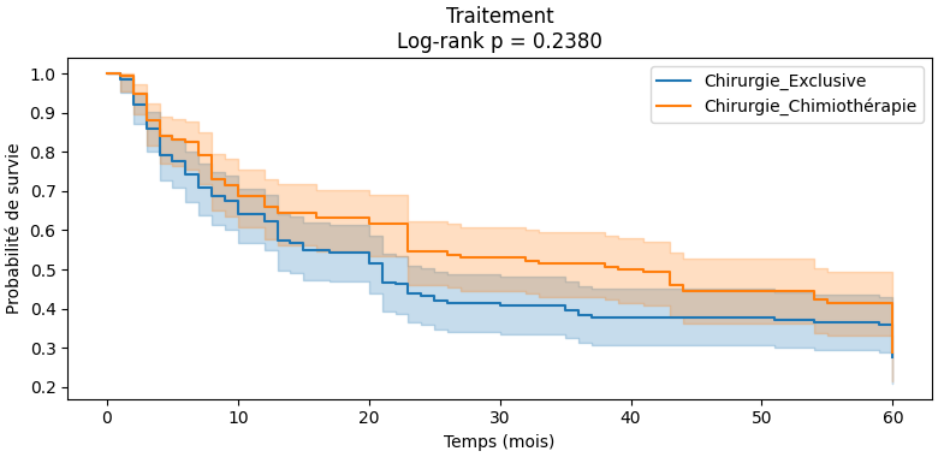


FIGURE 2.2 – Distributions de Survie selon le Traitement Administré

2.3.3 Distributions de survie

Variables avec des différences statistiquement significatives

Nous avons mis en évidence plusieurs variables pour lesquelles la différence de survie entre les patients présentant ces caractéristiques et ceux ne les présentant pas était statistiquement significative.
Ces variables incluent la Cardiopathie, la présence d'une Ulcero-bourgeonnant, le Tabagisme, un caractère Infiltrant, des Métastases et des Adénopathies, avec des p-values égales à 0,0001 (Fig. 2.3).

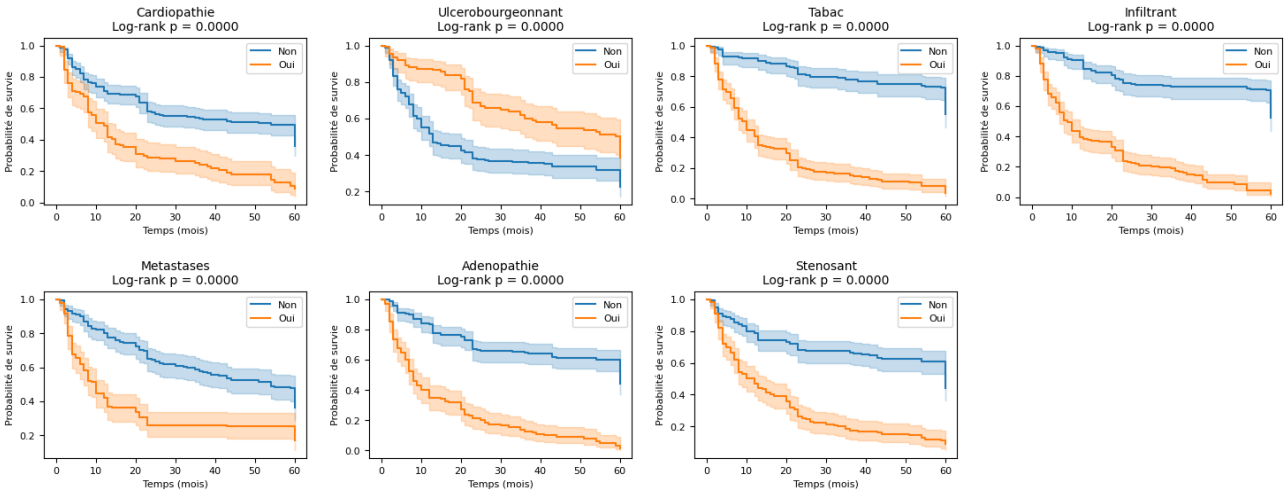


FIGURE 2.3 – Les distributions de survie en fonction des variables significatives

Variables avec des différences non significatives ou avec des courbes qui se croisent

Pour certaines variables, les distributions de survie ne montrent pas de différences significatives, avec des p-values supérieures au seuil de 5%. Ces variables comprennent la Douleur épigastrique (log-rank p = 0,151), le type Mucineux de l'adénocarcinome (log-rank p = 0,771), le Traitement (log-rank p = 0,238) et le Sexe (log-rank p = 0,741).

Cependant, d'autres variables, bien que présentant des p-values inférieures à 5%, ont montré des distributions de survie qui se croisaient, ne respectant pas ainsi les conditions d'application du test de log-rank. Ces variables sont le type Tubulaire (log-rank p = 0,000), les Ulcères gastriques (log-rank p = 0,000), la Constipation (log-rank p = 0,013) et la Dénutrition (log-rank p = 0,003) (Fig. 2.4).

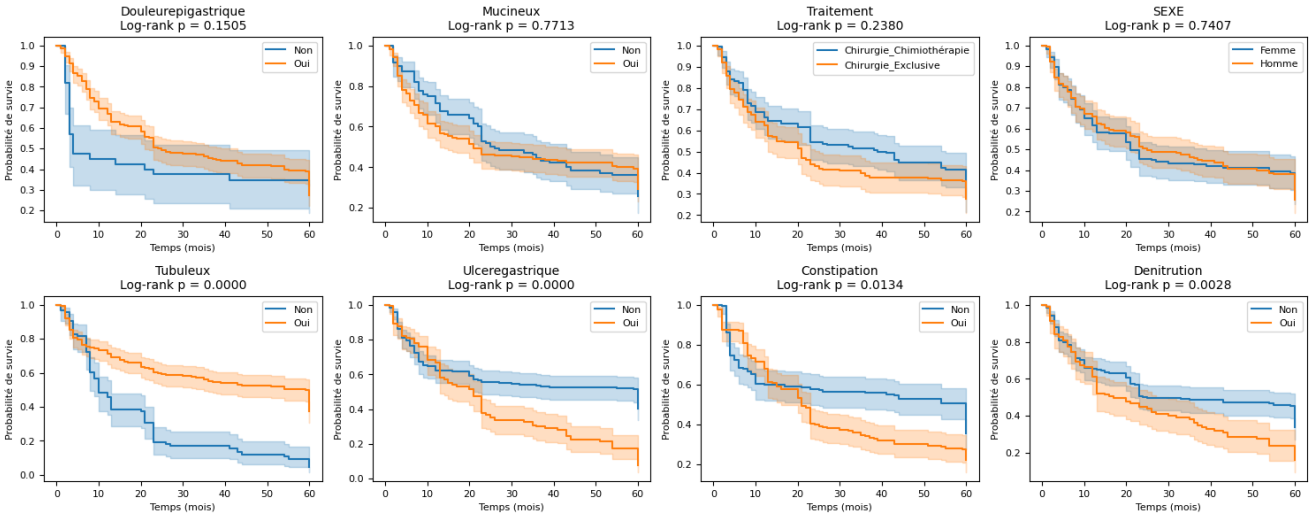


FIGURE 2.4 – Les distributions de survie en fonction des variables non significatives

2.4 Test statistique pour évaluer la performance du modèle

Test de Log-Rank

Le test de Log-Rank est utilisé pour comparer deux ou plusieurs fonctions de survie, à condition que les courbes de survie ne présentent pas d'intersections. Hypothèse à vérifier est l'existence d'une différence significative entre les distributions de survie.

$$\begin{cases} H_0 : \text{la différence est significative} \\ H_1 : \text{la différence n'est pas significative} \end{cases}$$

Son principe est que : à chaque occurrence de décès, la proportion de sujets à risque dans chaque groupe est calculée, et la différence entre les décès observés et les décès attendus est cumulée jusqu'à la date du dernier décès enregistré. La statistique de test procède ainsi à une sommation de ces écarts, et sa significativité est évaluée en divisant la différence cumulée par sa variance.

Soit A et B les groupes de la distribution.
Soit $n_{A,i}$ et $n_{B,i}$ le nombre de sujets à risque au temps t_i dans les groupes A et B, respectivement. Le nombre total de sujets à risque au temps t_i est donné par : $n_i = n_{A,i} + n_{B,i}$

Soit $d_{A,i}$ et $d_{B,i}$ le nombre d'événements observés au temps t_i dans les groupes A et B, respectivement. Le nombre total d'événements observés au temps t_i est donné par : $d_i = d_{A,i} + d_{B,i}$

La statistique de test est définie par :

$$U = \sum_{t_i} d_{B,i} - \frac{d_i n_{B,i}}{n_i}$$

Sa variance est donnée par :

$$\mathrm{Var}(U) = \sum_{t_i} d_i \left(\frac{n_i - d_i}{n_{i-1}}\right) \frac{n_{A,i} n_{B,i}}{n_i^2}$$

Enfin, la statistique normalisée suit une loi du chi-deux à un degré de liberté :

$$\frac{U^2}{\mathrm{Var}(U)} \xrightarrow{\mathcal{D}} \chi_1^2$$

2.4.1 Résultats : Test de Log-Rank

Le tableau récapitulatif du test de log-rank présente les résultats des comparaisons de survie entre les différentes catégories des variables étudiées.
Les p-values indiquent si les différences observées sont statistiquement significatives, avec un seuil fixé à 5%. Les variables dont les courbes de survie diffèrent de manière significative (p-values $< 0{,}05$) sont mises en évidence, tandis que celles présentant des p-values au-delà de ce seuil ne montrent pas de différence notable.

Ce tableau permet d'identifier les différences marquantes sur la survie des patients, tout en soulignant les limites du test, notamment en ce qui concerne les variables dont les courbes se croisent, défiant ainsi les hypothèses du test (Tab. 3.2).

2.5 Conclusion

La méthode de Kaplan-Meier nous a permis d'identifier que les variables telles que la cardiopathie, les aspects ulcéro-bourgeonnant, infiltrant et sténosant de la tumeur, le tabagisme, ainsi que la présence de métastases et d'adénopathies, présentaient des différences significatives au niveau de leurs distributions de survie respectives. Ces différences ont été attestées par le test de log-rank au seuil de 5En revanche, pour les autres variables (le sexe, le traitement, les douleurs épigastriques et le caractère mucineux), aucune différence significative n'a été observée. Bien que la méthode de Kaplan-Meier soit utile pour visualiser et comparer les distributions de survie, elle présente certaines limitations. Pour les variables telles que le type tubulaire de la tumeur, l'ulcère gastrique, la constipation et la dénutrition, le test de log-rank ne pouvait pas s'appliquer car les distributions de survie se croisaient. Cette situation constitue une limite majeure à l'application de ce test pour évaluer les courbes de Kaplan-Meier. Un autre inconvénient de l'utilisation du modèle de Kaplan-Meier est qu'il considère que la survenue du décès est uniquement liée au temps, ce qui ne permet pas d'intégrer les covariables étudiées. En réalité, les covariables étudiées peuvent impacter significativement la survenue de décès. Pour estimer leurs contributions, nous avons appliqué, dans la partie suivante, le modèle de régression à risques proportionnels de Cox.

2.6 Le Modèle de risques proportionnels de Cox

2.6.1 Introduction

Le modèle de Cox, aussi appelé modèle de risques proportionnels, est un modèle semi-paramétrique qui permet de modéliser la relation entre le temps de survie et des covariables, utilisé pour analyser les données de survie. Il permet d'étudier l'influence de plusieurs variables explicatives (comme l'âge, le sexe ou d'autres caractéristiques) sur le temps avant qu'un événement d'intérêt (comme décès, rechute ou guérison) ne survienne. La spécificité du modèle de Cox est qu'il ne fait pas d'hypothèse sur la forme précise de la fonction de survie dans le temps

2.6.2 Conditions d'application du modèle de Cox

Pour que le modèle de Cox soit valide et que ses résultats soient interprétables, deux conditions principale doivent être respectées :

— **Risques Proportionnels :** L'hypothèse des risques proportionnels est

centrale dans le modèle de Cox. Elle stipule que le rapport des risques (Hasard Ratios) entre les groupes comparés est constant au cours du temps.

— **Linéarité des Covariables** : Les covariables doivent avoir un effet linéaire sur le logarithme du risque. Cela signifie que chaque unité de changement dans une covariable est associée à une augmentation ou une diminution proportionnelle du risque de l'événement.

Le modèle de Cox repose sur la fonction de risque $h(t)$, définie à chaque instant t, permettant d'étudier l'association entre les covariables et le risques de survenu instantané d'un évènement final, depuis un évènement initial [24].

Cette fonction de risque représente une probabilité d'observer l'événement à l'instant t sachant qu'il n'a pas été observé avant t (risque instantanée), et se modélise comme suit dans le modèle de Cox :

$$h(t; X_1, \ldots, X_p) = h_0(t)exp(\beta_1 X_1 + ... + \beta_p X_p = h_0(t)exp(\beta^T X)$$

Avec
— h_0 : La fonction de risque de base,
— $X = (X1, ..., Xp)^T$: Le vecteur contenant les covariables,
— $\beta = (\beta_1, ..., \beta_p)^T$: Le vecteur de coefficients associés aux covariables.

2.6.3 Risques proportionnels et hypothèse de proportionnalité

La fonction h_0 est la fonction de risque lorsqu'aucune covariable n'est introduite dans l'équation. Et cette fonction n'est pas calculer dans l'eqaution de Cox. Elle pressente une fonction en paramètre de nuisance.
Les covariables forment ainsi une combinaison linéaire dans l'exponentielle, mais sont reliées à la fonction de risque par une fonction non linéaire. On parle dans ce cas de modèle linéaire généralisé. Selon le modèle de Cox le rapport des fonctions de risque de deux patients j et k qui ont pour covariables $X^{(j)}$ et $X^{(k)}$ est constant au cours du temps :

$$\frac{h(t, X^{(j)}))}{h(t, X^{(k)}))} = \frac{exp(\beta^T X^{(j)})}{exp(\beta^T X^{(k)})} = c$$

Avec t le temps.
Pour un patient i avec des données de survie (t_i, δ_i) tel que $\delta_i = 1$ (i.e. l'événement est observé, donnée non-censurée), la quantité $\frac{exp(\beta^T X^i)}{exp(\beta^T X^t)}$ correspond à la probabilité que l'individu i subisse effectivement l'événement en ti sachant qu'un événement a eu lieu en t_i. Le vecteur de coefficients β modèle de Cox peut

ainsi être estimé en maximisant la pseudo-vraisemblance. Ainsi, l'hypothèse de proportionnalité, en assurant la constance des ratios de risque dans le temps, constitue un pré-requis essentiel pour garantir la validité et l'interprétation des résultats obtenus [24].

2.6.4 Fonction de vraisemblance partielle

La vraisemblance partielle, également appelée vraisemblance de Cox, correspond au produit des probabilités conditionnelles d'observer l'événement A_i à l'instant t_i, étant donné le groupe A_i de sujets à risque juste avant t_i. L'estimation du vecteur des paramètres β nécessite de maximiser cette vraisemblance. Soit $L(\beta)$ la vraisemblance partielle du modèle de Cox, définie par la fonction suivante :

$$L(\beta) = \prod_{i=1}^{n} \frac{exp(X_i\beta)}{\sum_{j \in A_i} exp(X_j\beta)}$$

Cette fonction est appelée « pseudo-vraisemblance » car elle est définie comme un produit de probabilités conditionnelles, et non comme un produit de fonctions de densité. Pour faciliter l'optimisation et le calcul numérique, la log-pseudo-vraisemblance l (et non la pseudo-vraisemblance) est utilisée pour estimer β :

$$\hat{\beta} = \arg\left(l(\beta)\right)$$

Avec $l(\beta) = \log(L(\beta))$ la log-pseudo-vraisemblance (vraisemblance partielle). Il est important de souligner que le modèle de Cox peut être difficile à appréhender, car la relation entre les variables explicatives et le temps de survie (variable à expliquer) n'est pas définie de manière directe, mais plutôt par l'intermédiaire de la fonction de risque h. Cependant, la pseudo-vraisemblance ainsi définie permet de prendre efficacement en compte les censures, tout en s'affranchissant de modéliser, d'estimer, et d'émettre des hypothèses sur la fonction de risque de base h_0.

2.6.5 Les coefficients de régression

Les coefficients estimés du modèle de Cox, notés β_i, mesurent l'effet des covariables sur le risque de survenue de l'événement. Ces coefficients correspondent aux logarithmes des Hazard Ratio ($logHR$). Le hazard Ratio est le ratio de risque instantané de survenue de l'événement entre deux groupes. Pour le comprendre, envisageons un modèle de Cox simple, avec une seule variable explicative X. Supposons que nous observons deux individus avec des modali-

tés X_1 pour l'un (par exemple « Hommes ») et X_2 pour l'autre (par exemple « Femmes »). Les risques instantanés de ces deux individus sont donnés par :

$$h(t|X_1) = h_0(t)exp(\beta_1 X_1)$$
$$h(t|X_2) = h_0(t)exp(\beta_1 X_2)$$

Le rapport de risque (Hazard Ratio) entre ces deux individus est donné par :

$$\frac{h(t|X_1)}{h(t|X_2)} = \frac{h_0(t)exp(\beta_1 X_1)}{h_0(t)exp(\beta_1 X_2)} = exp(\beta_1(X_1 - X_2))$$

En utilisant la transformation logarithmique, nous obtenons :

$$\log(\frac{h(t|X_1)}{h(t|X_2)}) = \log(\frac{h_0(t)exp(\beta_1 X_1)}{h_0(t)exp(\beta_1 X_2)}) = \beta_1(X_1 - X_2)$$

Il est important de noter que le hazard ratio est indépendant du temps. Cela signifie que l'effet des covariables sur le risque est multiplicatif et constant au cours du temps. Cela illustre le principe des « risques proportionnels ».
Par exemple, si le Hazard Ratio associé à une covariable est de 2, alors le risque de l'événement pour une personne avec cette covariable sera toujours deux fois plus élevé que pour une personne sans cette covariable, et ce à n'importe quel moment de l'étude [24].

2.6.6 Les tests associés au coefficient

Pour la validité des résultats de notre modèle de Cox nous utilise :
Le test de Wald est utilisé pour tester l'hypothèse nulle selon laquelle un coefficient ($logHR$) est égal à zéro (c'est-à-dire qu'il n'a pas d'effet significatif). La statistique z est calculé comme le rapport du coefficient estimé $\hat{\beta}$ à son erreur standard (ES) [24].

$$Z = \frac{\hat{\beta}}{ES}$$

Les valeurs de Z suivent une distribution normale standard sous l'hypothèse nulle. Par conséquent, une valeur absolue élevée de Z indique que le coefficient est significativement différent de zéro. Si nous raisonnons en termes de Hazard Ratio (HR) plutôt qu'en $logHR$, alors le test de Wald permet de tester l'hypothèse nulle selon laquelle le hazard ratio associé à une covariable est égal à 1. Une p-value inférieure à un seuil alpha (généralement 0,05) suggère que le Hazard ratio significativement différent de 1, ce qui implique que la covariable a un effet significatif sur le risque de survenue de l"événement [24].

2.6.7 Interprétation des Hazard Ratios

Le Hazard Ratio (HR) est une mesure clé dans l'analyse de survie avec le modèle de Cox. Il permet d'évaluer l'effet des covariables sur le risque de survenue d'un événement.

- **HR = 0 :** indique une absence totale de risque associé à la variable étudiée. Cela signifie que l'événement ne peut pas survenir chez les individus concernés par cette variable.
- **HR = 1 :** Lorsque le Hazard Ratio est égal à 1, cela signifie que la covariable n'a aucun effet sur le risque de survenue de l'événement. Autrement dit, la présence ou l'absence de cette covariable n'affecte pas le taux auquel l'événement se produit.
- **HR < 1 :** Lorsque le Hazard Ratio est significativement inférieur à 1, cela signifie que la covariable est associée à une réduction du risque de survenue de l'événement
- **HR > 1 :** Lorsque le Hazard Ratio est significativement supérieur à 1, cela indique que la covariable est associée à une augmentation du risque de survenue de l'événement.

L'interprétation des Hazard Ratios doit toujours être faite dans le contexte de l'étude et en tenant compte de la signification statistique des résultats. Un Hazard Ratio proche de 1 peut ne pas être cliniquement significatif même s'il est statistiquement significatif, et vice versa [24].

2.7 Résultats : modèle de Cox

2.7.1 Test de proportionnalité des risques

Le tableau (Tab : 2.1) offre une vue d'ensemble du modèle de Cox ajusté, incluant les détails généraux du modèle ainsi que ses statistiques de performance. Ce modèle a été appliqué à 337 observations, affichant un taux de concordance de 0,85 et un AIC partiel de 2024,20. Le test sur la proportionnalité des risques a révélé un bon ajustement, avec une distribution selon χ^2 et un degré de liberté de 1.

TABLE 2.1 – Résumé du modèle de Cox proportionnel ajusté

Informations générales	
Modèle	`lifelines.CoxPHFitter`
Colonne durée	`Temps de suivi (Mois)`
Colonne événement	`Décès`
Méthode d'estimation de la baseline	`breslow`
Nombre d'observations	337
Nombre d'événements observés	223
Performances et statistiques de test	
Concordance index	0.85
Partial AIC	2024.20
Vraisemblance partielle (log-partial likelihood)	-996.10
Test du rapport de vraisemblance	356.27 sur 16 ddl
$-\log_2(p)$ du test de vraisemblance	216.90
Test de proportionnalité des risques	
Nom du test	`proportional_hazard_test`
Distribution nulle	χ^2
Degrés de liberté	1

Le tableau (Tab : 2.2) affiche les résultats de l'évaluation de la proportionnalité des risques pour chaque variable. Les termes **km** et **rank** font référence aux différentes approches utilisées dans ce test :**km** désigne l'estimation utilisant la méthode de Kaplan-Meier, tandis que **rank** correspond à celle basée sur les rangs de Cox. Certaines variables possèdent des p-valeurs inférieures à 0.05, suggérant un éventuel manquement à l'hypothèse de proportionnalité des risques.

TABLE 2.2 – Test de proportionnalité des risques pour chaque variable

Variable	**Méthode**	**test_statistic**	**p**	$-\log_2(p)$
AGE	km	3.06	0.08	3.64
	rank	3.18	0.07	3.75
Adénopathie	km	0.76	0.38	1.38
	rank	0.66	0.41	1.27
Cardiopathie	km	0.02	0.88	0.19
	rank	0.00	0.96	0.07
Constipation	km	4.64	0.03	5.00
	rank	6.27	0.01	6.35
Dénutrition	km	6.11	0.01	6.22
	rank	6.08	0.01	6.19
Douleur épigastrique	km	19.82	<0.005	16.85
	rank	21.20	<0.005	17.89
Infiltrant	km	0.19	0.67	0.59
	rank	0.20	0.66	0.61
Métastases	km	3.35	0.07	3.89
	rank	4.10	0.04	4.55
Mucineux	km	3.89	0.05	4.36
	rank	4.71	0.03	5.06
SEXE	km	0.03	0.87	0.20
	rank	0.00	0.95	0.08
Stenosant	km	0.99	0.32	1.65
	rank	1.63	0.20	2.31
Tabac	km	5.06	0.02	5.35
	rank	4.79	0.03	5.13
Traitement	km	0.91	0.34	1.55
	rank	0.98	0.32	1.63
Tubuleux	km	3.78	0.05	4.27
	rank	4.21	0.04	4.64
Ulcere gastrique	km	4.76	0.03	5.11
	rank	4.02	0.04	4.48
Ulcero-bourgeonnant	km	1.37	0.24	2.05
	rank	1.47	0.22	2.15

2.7.2 Résultats du Test de Wald

À l'aide du test de Wald, nous avons identifié des variables significatives et non significatives parmi toutes les variables incluses.
Parmi les plus importantes, on trouve les métastases, le sexe, l'adénopathie, la douleur épigastrique et le caractère mucineux, tandis que parmi les moins

importantes figurent la constipation, l'âge et le type de traitement.

TABLE 2.3 – Le test de Wald au seuil de 5%

Variables	Coefficient	Rapport de risque (HR)	Stat Wald	P-Value
Métastases	1.23	3.42	5.93	<0.005
Tabac	1.18	3.25	5.31	<0.005
Infiltrant	1.05	2.86	5.14	<0.005
Adénopathie	0.99	2.68	4.98	<0.005
Cardiopathie	0.49	1.64	3.06	<0.005
Sténosant	0.49	1.63	2.91	<0.005
Ulcère Gastrique	0.49	1.63	3.05	<0.005
Traitement	0.05	1.06	0.39	0.70
Constipation	-0.01	0.99	-0.07	0.94
AGE	-0.02	0.98	-3.04	<0.005
SEXE	-0.17	0.84	-1.14	0.25
Tubuleux	-0.27	0.76	-1.71	0.09
Ulcèro-bourgeonnant	-0.47	0.62	-2.74	0.01
Dénutrition	-0.59	0.55	-3.52	<0.005
Douleur Épigastrique	-0.71	0.49	-3.04	<0.005
Mucineux	-0.71	0.49	-3.62	<0.005

Pour identifier les paramètres indépendants associés à la survie, nous avons utilisé le modèle de Cox à risques proportionnels. Les coefficients β (HR) obtenus permettent de quantifier l'effet relatif des covariables.

La sélection des variables dans ce modèle a été réalisée à l'aide d'une régression logistique automatisée (Stepwise). Cette approche est particulièrement utile dans des contextes exploratoires, notamment lorsque les connaissances sur les prédicteurs potentiels sont limitées ou que l'on souhaite réduire un grand nombre de variables en identifiant celles qui ont la plus grande contribution au modèle.
Elle permet de construire des hypothèses solides en éliminant automatiquement les variables non significatives, tout en prenant en compte les interactions potentielles entre les covariables.

Les résultats de cette analyse ont mis en évidence plusieurs variables significatives associées au risque de décès. Parmi les facteurs de risque majeurs, on retrouve les métastases (HR = 3,70, p < 0,0001), le tabagisme (HR = 3,43, p < 0,0001) et l'aspect infiltrant (HR = 2,87, p < 0,0001), qui montrent une forte augmentation du risque instantané de décès.
D'autres variables, comme la cardiopathie (HR = 1,60, p = 0,001), l'ulcère gastrique (HR = 1,67, p = 0,004) et le caractère sténosant (HR = 1,56, p = 0,01), contribuent également de manière significative.
Par ailleurs, certaines covariables apparaissent comme protectrices, notamment

l'âge (HR = 0,98, p = 0,01), le caractère mucineux (HR = 0,48, p = 0,007), le caractère ulcéro-bourgeonnant (HR = 0,61, p = 0,009), la douleur épigastrique (HR = 0,56, p = 0,01), la dénutrition (HR = 0,53, p < 0,0001) et le caractère mucineux (HR = 0,48, p = 0,0001).

À l'inverse, certaines variables, comme le sexe, la constipation, le caractère tubulaire et les différents traitements, n'ont pas été retenues dans le modèle. Cela peut s'expliquer par leur faible contribution significative ou par leur redondance avec d'autres covariables incluses dans le modèle.
Ces exclusions soulignent l'utilité de la méthode "stepwise" pour éviter la surcharge du modèle et garantir une interprétation plus claire des résultats. Ses principaux avantages incluent sa capacité à gérer des données censurées, à interpréter directement l'effet des covariables grâce aux Hazard Ratios, et à fournir des estimations robustes des risques (Tab. 2.4).

TABLE 2.4 – Le test de Wald avec les variables significatives au seuil de 5%

Variables	Coefficient	HR	Erreur Standard	Z (Test Wald)	P-value
Métastases	1.31	3.70	0.20	6.44	<0.005
Tabac	1.23	3.43	0.22	5.65	<0.005
Infiltrant	1.06	2.87	0.21	5.13	<0.005
Adénopathie	0.98	2.67	0.20	4.96	<0.005
Ulcère gastrique	0.51	1.67	0.16	3.29	<0.005
Cardiopathie	0.47	1.60	0.16	2.95	<0.005
Sténosant	0.44	1.56	0.16	2.79	0.01
AGE	-0.02	0.98	0.01	-2.73	0.01
Ulcero-bourgeonnant	-0.49	0.61	0.17	-2.85	<0.005
Douleur épigastrique	-0.58	0.56	0.22	-2.61	0.01
Dénutrition	-0.63	0.53	0.17	-3.76	<0.005
Mucineux	-0.73	0.48	0.20	-3.73	<0.005

L'équation de la fonction de risque estimée par le modèle de Cox, en fonction des variables significatives identifiées, s'écrit comme suit :

$$\begin{aligned} h(t) = h_0(t) \cdot \exp(&1.31 \cdot \text{Métastases} + 1.23 \cdot \text{Tabac} + 1.06 \cdot \text{Infiltrant} \\ &+ 0.98 \cdot \text{Adénopathie} + 0.51 \cdot \text{Ulcère gastrique} \\ &+ 0.47 \cdot \text{Cardiopathie} + 0.44 \cdot \text{Sténosant} - 0.02 \cdot \text{Âge} \\ &- 0.49 \cdot \text{Ulcero-bourgeonnant} - 0.58 \cdot \text{Douleur épigastrique} \\ &- 0.63 \cdot \text{Dénutrition} - 0.73 \cdot \text{Mucineux}) \end{aligned}$$

Le modèle de Cox ainsi obtenu avait montré de bonnes performances prédictives, avec un indice de concordance (C-index) de 0,85 et un score d'erreur intégrée (IBS) de 0,080.

Remarque

Le fait de choisir entre une chirurgie seule et une chirurgie combinée avec chimiothérapie ne change pas le résultat en lui-même. Cependant, cela ne signifie pas que le traitement en question n'a aucun effet. De même, le choix entre une femme et un homme n'affecte pas le résultat directement. Cependant, cela ne signifie pas que le sexe n'ait aucun effet.

2.7.3 Valeurs SHAP et importance de permutation

Les valeurs SHAP et l'importance des permutations des variables, selon le modèle de Cox, ont révélé l'impact de chacune d'elles sur la prédiction de la durée de survie des patients atteints d'un cancer de l'estomac après traitement.

Les résultats ont mis en lumière que des variables telles que les métastases, l'adénopathie, les infiltrants, le tabagisme et le caractère mucineux jouaient un rôle essentiel dans cette prédiction. Ces observations soulignent l'importance d'intégrer ces paramètres pour améliorer la précision des modèles de survie dans ce cadre clinique (Fig. 2.5).

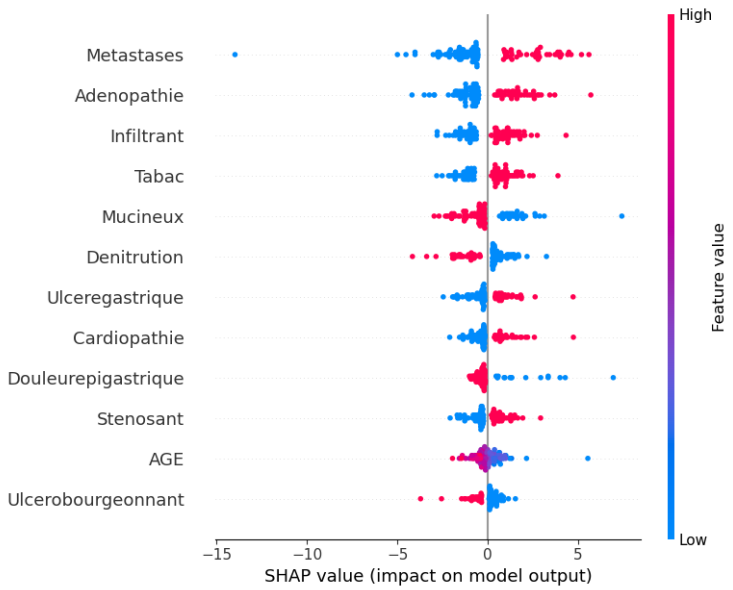


FIGURE 2.5 – Pertinence des variables selon le modèle de Cox

2.8 Évaluation des effets des covariables

Le modèle de Cox nous a aidés à déterminer que des facteurs comme la cardiopathie, les aspects mucineux, ulcéro-bourgeonnants, ulcères gastriques, tumeurs infiltrantes et sténosantes, ainsi que le tabagisme, la présence de métastases, l'âge des patients, les douleurs épigastriques, la dénutrition et les adé-

nopathies exercent un effet significatif sur la probabilité immédiate de décès. Ces distinctions ont été corroborées par le test de Wald à un seuil de 5%.

En revanche, le sexe, le choix du traitement et la constipation n'ont montré aucune différence significative. Cependant, bien que le modèle de Cox soit utile pour vérifier l'impact significatif de diverses variables, il présente certaines limites. Ces limites incluent : l'exigence des risques proportionnels, la sensibilité à l'information censurée, l'absence de colinéarité entre les variables explicatives, la nécessité d'une relation linéaire entre les covariables, et une interprétation indirecte via la fonction de risque.

2.9 Conclusion

Dans ce chapitre, nous avons examiné les méthodes classiques d'analyse de survie pour étudier l'impact des différentes variables sur la survie des patients.

Nous avons d'abord utilisé la méthode de Kaplan-Meier, ce qui nous a permis de visualiser les courbes de survie et d'identifier les variables présentant des différences significatives grâce au test du log-rank. Nous avons constaté que certains facteurs, tels que la cardiopathie, les aspects ulcéro-bourgeonnants, infiltrants, et sténosants, le tabagisme ainsi que la présence de métastases et d'adénopathies, avaient une influence notable sur la survie.

Ensuite, nous avons appliqué le modèle de Cox pour quantifier l'effet des covariables sur le risque de décès. Ce modèle nous a permis de confirmer l'impact significatif de plusieurs facteurs et d'identifier ceux qui augmentaient ou diminuaient le risque. Toutefois, nous avons également relevé certaines limites de ces méthodes classiques, notamment l'incapacité à intégrer toutes les variables et la difficulté à appliquer certains tests lorsque les courbes se croisaient.

Pour surmonter ces limitations, nous avons recours aux méthodes d'apprentissage automatique.

Chapitre 3

Modèles d'apprentissage Automatique sur l'analyse de Survie

3.1 Introduction

L'apprentissage automatique (ou apprentissage artificiel) est, suivant la définition de Tom Mitchell [23], l'étude des algorithmes qui permettent aux programmes de s'améliorer automatiquement par expérience.

L'intelligence artificielle (IA) a révolutionné de nombreux domaines, notamment l'analyse des données de survie. Cette discipline statistique, essentielle en médecine et en biologie, vise à estimer la durée jusqu'à la survenue d'un événement spécifique, tel que le décès ou la rémission d'une maladie [2].
Avec l'émergence de l'IA, en particulier d'apprentissage automatique et d'apprentissage profond, plusieurs modèles ont été développé pour l'exploitation des données censures [14].

Dans le contexte de l'analyse de survie, des techniques telles que les forêts aléatoires de survie (Random Survival Forest) et les modèles de gradient Boosting ont été appliquées pour améliorer la précision des prédictions de survie [11].

Le Deep learning, une sous-catégorie de la machine learning, utilise des réseaux de neurones profonds pour modéliser des relations complexes dans les données. Des modèles tels que Deep Surv ont démontré une capacité supérieure à prédire les issues de survie par rapport aux méthodes traditionnelles. Ces approches permettent d'intégrer des données non structurées, comme des images médicales ou des séquences génétiques, offrant ainsi une vision plus complète du patient [11].

L'application de l'IA aux données de survie permet de développer des application dynamiques pour une aide à la prise de décision base sue les données.

3.1.1 Problématique

La méthode classique de régression de Cox, bien que largement utilisée pour l'analyse des données de survie, présente plusieurs limitations qui peuvent affecter sa performance dans des situations complexes.
Tout d'abord, le modèle de Cox repose sur deux hypothèses fondamentales :
L'hypothèse de proportionnalité des risques et l'hypothèse de log-linéarité.

La première suppose que les rapports des risques entre les groupes restent constants au fil du temps, ce qui n'est pas toujours le cas dans des contextes cliniques.

La seconde suppose que l'effet des variables explicatives sur le risque de l'événement suit une relation linéaire, ce qui peut être trop restrictif pour certaines données non linéaires ou avec des interactions complexes entre les variables. Lorsqu'une ou les deux hypothèses ne sont pas vérifiées le modèle de Cox n'était pas applicable.

En outre, le modèle de Cox a du mal à prendre en compte les effets hétérogènes des traitements ou des interventions variables dans le temps, et des interactions entre les covariables. Par exemple, dans le cadre d'une étude sur le cancer, différents traitement peut être proposé au patient par intervalle de temps. Le modèle de Cox ne permet pas d'identifier les effets de chacune de traitement sur le patient.

Ces limites peuvent réduire la précision des prédictions, particulièrement dans des domaines où les interactions entre les facteurs sont multiples et complexes. Face à ces défis, les approches d'AI pour l'analyse de survie offrent des solutions potentielles pour l'exploitation des données censures.

3.1.2 Objectifs

L'objectif pour cette chapitre était d'utiliser les méthodes d'apprentissage automatique pour l'analyse de survie.
Il s'agissait spécifiquement de comparer les performances des différents algorithmes d'apprentissage automatique par rapport aux méthodes classiques.
Nous avons développé trois algorithmes d'apprentissage automatique à savoir Random Survival Forest, Gradient Boosting Survival Analytics et Deep Survival Model.

3.2 Méthode

3.2.1 Données employées :

Les variables cliniques et tumorales retenues pour l'analyse du risque de décès chez les patients atteints d'un cancer de l'estomac ont été sélectionnées selon leur pertinence statistique dans le modèle de Cox et leur contribution explicative mesurée par les valeurs "shap et l'importance de permutation".

Parmi ces variables figuraient : l'âge des patients, la présence de comorbidités (cardiopathies), des symptômes cliniques (douleur épigastrique, ulcère gastrique, lésion ulcéro-bourgeonnante), des facteurs liés au mode de vie (tabagisme, dénutrition), ainsi que des caractéristiques tumorales spécifiques (forme mucineuse, infiltration, sténose, métastases et adénopathies).

Ces critères ont été retenus car ils ont démontré, à travers le modèle de Cox une influence significative sur la réduction de la survie.
La présence significative de ces variables, validée à la fois par une approche statistique classique et par une analyse d'interprétabilité moderne (SHAP), fournit une base solide pour les explorations ultérieures.

3.2.2 Optimisation des hyperparamètres

L'optimisation des hyperparamètres sur le model random survival et le gadient boostin survival tree a été réalisée à l'aide de la méthode GridSearchCV, qui teste systématiquement toutes les combinaisons prédéfinies d'hyperparamètres pour identifier la configuration optimale du modèle. Cette approche exhaustive, bien que coûteuse en calcul, a permis d'explorer rigoureusement l'espace des hyperparamètres, comme recommandé dans les bonnes pratiques d'apprentissage automatique [19]. L'algorithme effectue toutes les combinaisons possibles et fournit les meilleures valeurs d'hyperparamètres ayant obtenu les meilleurs scores lorsqu'elles sont appliquées aux données disponibles par validation croisée.

Pour le Deep Survival, l'optimisation s'était appuyée sur **ManualGridSearch** [1], une approche plus rapide explorant un sous-ensemble aléatoire d'hyperparamètres. Ces stratégies avaient permis d'assurer une évaluation rigoureuse et reproductible des modèles.

3.2.3 Importance des variables selon chaque modèle

La sélection d'importance des variables selon chaque modèle a été effectuée à l'aide de deux techniques complémentaires : Les valeurs SHAP et l'importance de permutation.
Les valeurs SHAP, issues de la théorie des jeux coopératifs, ont permis de quantifier la contribution individuelle de chaque variable aux prédictions du modèle, en identifiant celles qui influençaient le plus les résultats [22] . Cette approche, proposée par [18], a offert une interprétation transparente des décisions du modèle, même pour des données complexes.
Parallèlement, l'importance de permutation a mesuré l'impact de chaque variable sur la performance globale du modèle en calculant la baisse de précision lorsque ses valeurs étaient aléatoirement mélangées, comme décrit dans les travaux de Breiman sur les forêts aléatoires [4].
En combinant ces méthodes, une analyse robuste a été réalisée : les valeurs SHAP ont mis en lumière les relations locales entre variables et prédictions, tandis que l'importance de permutation a validé leur rôle global.

3.2.4 Métriques d'évaluation

L'évaluation des performances des modèles a reposé sur deux métriques clés adaptées aux données de survie.
L'indice de concordance de Harell a été utilisé pour mesurer la capacité du modèle à ordonner correctement les temps de survie des individus, en comparant les prédictions aux événements (décès) observés.
Cette métrique, inspirée des travaux de [10], sur les modèles de régression en survie, a tenu compte des données censurées pour éviter les biais d'interprétation.

En complément, le score de IBS a permis d'évaluer la précision des prédictions à différents horizons temporels. Le score de Brier intégre a quantifié l'erreur moyenne entre les probabilités prédites et les événements réels, suivant les recommandations de Blanche et al. pour les analyses dynamiques [3].

3.3 Modèles de Machine Learning en Survie

3.3.1 Random Survival Forest

Le modèle Random Survival Forest (RSF) est une variante des forêts aléatoires particulièrement adaptée à l'analyse des données de survie avec censure à droite. En utilisant l'apprentissage par ensemble, le RSF crée de nombreux arbres de décision à partir de sous-échantillons bootstrap, ce qui lui permet

de capturer et de modéliser des relations complexes et non linéaires entre les variables explicatives.

Chaque arbre évalue la fonction de survie, et la combinaison de ces arbres fournit une estimation robuste et globale du risque, souvent supérieure aux modèles traditionnels comme la régression de Cox. Plusieurs études empiriques ont confirmé l'efficacité de ce modèle.

Par exemple, Jiaxi Lin et ses collègues ont montré la supériorité du RSF dans la prédiction de la survie des patients atteints de cancer du pancréas, avec un indice C de 0,723 contre 0,670 pour la régression de Cox [21]. De même, Roya Najafi-Vosough et son équipe ont utilisé le RSF pour évaluer la survie des patientes atteintes de cancer du sein en tenant compte des risques concurrents, mettant en évidence le rôle crucial de la variable HER2 dans le pronostic [25].

Plus récemment, Lin Wei Li et ses collègues ont appliqué ce modèle à des jeunes patientes atteintes de cancer du sein à un stade précoce et ont obtenu un indice C impressionnant de 0,920 sur leur ensemble d'entraînement, confirmant ainsi la robustesse du RSF par rapport aux méthodes traditionnelles telles que la régression de Cox [20].

Ces exemples illustrent à la fois la souplesse et la puissance prédictive du RSF dans divers contextes cliniques.

3.3.2 Gradient Boosting Survival Analytics

Le Gradient Boosting Survival Analytics (GBSA) est une technique d'apprentissage automatique qui applique les principes du boosting aux problèmes de survie [33]. Cette méthode consiste à construire de manière séquentielle des arbres de décision simples qui, en se concentrant sur les erreurs résiduelles des étapes précédentes, améliorent progressivement la prédiction du risque ou de la probabilité de survie.

Pour atteindre cet objectif, le modèle intègre des fonctions de perte spécifiques aux données de survie, telles que la fonction de perte de Cox ou la perte de rang, qui permettent de prendre en compte la censure des observations, un aspect essentiel dans l'analyse de survie.

Bien que la littérature sur le GBSA ne détaille pas systématiquement son application aux cancers, ce modèle reste l'un des outils les plus couramment utilisés en analyse de survie, grâce à sa capacité à identifier des relations com-

plexes entre les variables tout en minimisant le risque de sur-ajustement [13].

Ses performances élevées et son interprétabilité partielle, notamment à travers l'évaluation de l'importance des variables à l'aide de métriques telles que les valeurs SHAP, en font une méthode robuste et flexible pour prédire des événements temporels dans divers contextes médicaux et autres domaines d'application.

3.3.3 Deep Survival Modèle

DeepSurv est un modèle d'apprentissage profond qui s'adresse à l'analyse de survie et vise à prévoir le temps jusqu'à ce qu'un événement crucial, comme la mort, se produise. Contrairement aux méthodes traditionnelles telles que le modèle de Cox à risques proportionnels, qui nécessite une sélection préalable des variables, DeepSurv utilise un réseau de neurones pour apprendre de façon dynamique et capturer les interactions non linéaires entre les covariables. Le concept principal du modèle repose sur l'intégration d'un réseau de neurones profond avec le cadre du modèle de Cox, permettant ainsi d'estimer le risque relatif associé aux caractéristiques des patients tout en évitant de faire des hypothèses contraignantes sur le lien entre ces caractéristiques et le risque d'événement.

Ce mécanisme permet à DeepSurv de saisir des interactions complexes dans les données cliniques, ce qui est un atout majeur dans le domaine médical, où les interactions entre variables sont souvent difficiles à comprendre avec les méthodes classiques. Selon la littérature, plusieurs recherches ont montré l'efficacité de DeepSurv. Cheng et al. l'ont utilisé pour prédire la survie spécifique au cancer chez des patients atteints de chordome spinal, obtenant un indice C de 0,830, ce qui représente une amélioration par rapport aux méthodes traditionnelles [5].

De même, Kim Dong Wook et ses collègues ont appliqué DeepSurv pour estimer la survie des patients atteints de carcinome épidermoïde buccal, obtenant des indices C de 0,810 et 0,781 pour les ensembles d'entraînement et de test, surpassant ainsi les performances de la forêt de survie aléatoire et de la régression de Cox [17].

Enfin, Zhi Huang et al. ont examiné l'utilisation de DeepSurv dans le contexte des données transcriptomiques RNA-seq, renforçant ainsi la solidité et la diversité de cette approche dans divers domaines d'application [12].

3.4 Résultats

3.4.1 Optimisation des hyperparametre

L'optimisation des hyperparamètres a été effectuée à l'aide de **GridSearchCV**. Cette méthode a été appliquée aux modèles Random Survival Forest et Gradient Boosting Survival Analysis. Pour le modèle de Deep Survival, nous avons opté pour **ManualGridSearch**. Les meilleurs hyperparamètres obtenus sont répertoriés dans le tableau suivant (Tab. 3.4) :

3.4.2 Entrainement du modèle

Les résultats obtenus indiquent des performances globalement solides des modèles formés pour prédire la durée de survie.

Pour le modèle de Forêt de Survie Aléatoire (RSF), le C-index sur les données d'entraînement était de 0.95, tandis qu'il atteignait 0.87 sur les données de test, avec un intervalle de confiance de [0.82, 0.91]. L'IBS était de 0.011 pour l'entraînement et de 0.078 pour le test, ce qui suggère une bonne calibration du modèle, même si l'erreur a légèrement augmenté pour les données non observées.

Le Gradient Boosting Survival Tree a présenté un C-index de 0.90 sur l'ensemble d'entraînement et de 0.87 sur le jeu de test. L'intervalle de confiance du C-index sur les données de test était de [0.82, 0.91], avec une moyenne bootstrap de 0.87. L'IBS était de 0.058 pour l'entraînement et de 0.086 pour le test, indiquant une légère augmentation de l'erreur tout en montrant une cohérence globale du modèle.

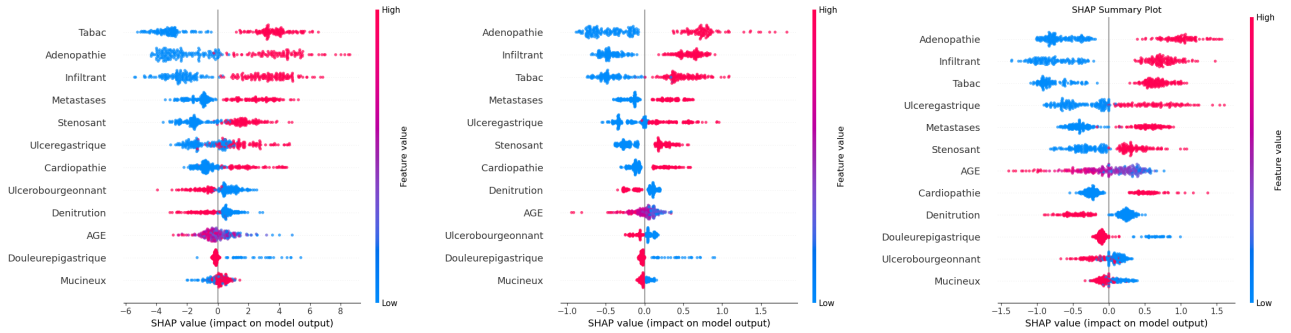
Pour finir, les modèles de Deep Survival ont montré un C-index de 0.91 sur l'entraînement. Pour les données de test, le C-index atteignait 0.89, avec une moyenne bootstrap similaire et un intervalle de confiance de [0.81, 0.90]. L'IBS est resté stable entre l'entraînement est de 0,046 et 0,0755 pour la base test, ce qui reflète une bonne capacité de généralisation du modèle.

En résumé, ces résultats montrent que les modèles ont bien compris les relations entre les différentes variables et les durées de survie, tout en maintenant de bonnes performances sur les données de test.

3.4.3 Valeurs SHAP selon les modèles d'apprentissage automatique

L'analyse des valeurs SHAP visait à déterminer l'influence de chaque variable sur les prédictions du modèle. Les valeurs SHAP montraient de quelle manière

chaque variable affectait le résultat, en prenant en compte son effet direct ainsi que les interactions possibles avec d'autres variables (Tab. 3.3). En résumé, elles rendaient le modèle plus compréhensible en identifiant les variables les plus influentes, notamment **tabac, Adénopathie, Métastases et Infiltrant** qui avaient les impacts les plus significatifs sur les prédictions des différents modèles (Fig. 3.1).

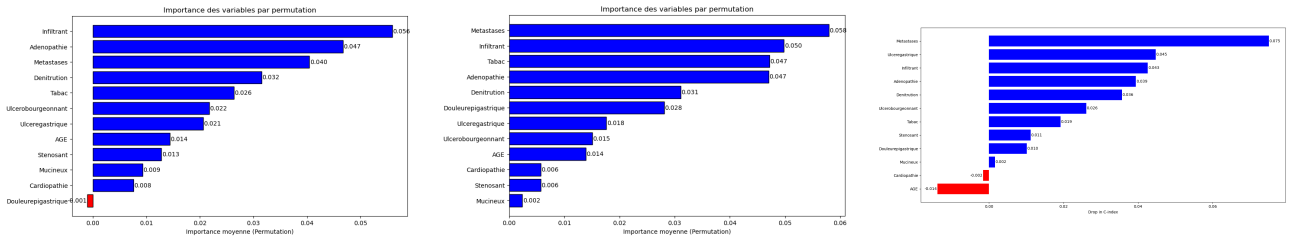


RSF GBST Deep Survival

FIGURE 3.1 – Comparaison des valeurs SHAP selon les modèles de survie

3.4.4 Importance par permutation

L'analyse d'importance par permutation a permis de déterminer l'impact de chaque variable sur la précision du modèle prédictif. Cette technique évalue comment l'erreur du modèle varie lorsqu'on échange une variable, ce qui révèle son influence globale. Elle aide à repérer les variables les plus cruciales, en particulier **Adénopathie, Métastases, Tabac, Infiltrant et Mucineux**, qui se sont avérées être les plus influentes dans différents modèles (Fig. 3.2).



RSF GBST Deep Survival

FIGURE 3.2 – Importance de permutation des variables selon les modèles

3.5 Comparaison des modèles

L'ensemble des modèles évalués avait montré de bonnes performances en termes de discrimination et de calibration. Le modèle de Forêt de Survie Aléatoire avait obtenu les meilleures performances sur les données d'entraînement,

avec un C-index de 0,95 et un IBS très faible (0,011), traduisant une excellente capacité à prédire les événements. Toutefois, une légère diminution des performances avait été observée sur les données de test (C-index = 0,87 ; IBS = 0,078).

Le modèle Gradient Boosting Survival Tree avait également affiché une bonne robustesse, avec des scores constants entre l'entraînement et le test (C-index = 0,90 et 0,87 respectivement), bien que l'IBS ait légèrement augmenté sur le jeu de test (0,058 à 0,086).

Le modèle Deep Survival, quant à lui, avait présenté une excellente stabilité entre les ensembles d'entraînement et de test, avec un C-index de 0,91 et 0,89, ainsi qu'un IBS contenu (0,046 en entraînement et 0,0755 en test), ce qui reflétait une bonne capacité de généralisation (Tab. 3.1).

TABLE 3.1 – Comparaison des performances des modèles de survie selon le C-index et l'IBS

Modèle	C-index (train)	C-index (test)	IC C-index (test)	IBS (train)	IBS (test)
CoxPH	0.85	0.85	–	–	0.080
RSF	0.95	0.87	[0.82, 0.91]	0.011	0.078
GBST	0.90	0.87	[0.82, 0.91]	0.058	0.086
DeepSurv	0.91	0.89	[0.81, 0.90]	0.046	0.0755

3.5.1 Visualisation des Brier Score Intégré

Le graphique illustrait comment les Brier Scores évoluaient au fil du temps pour divers modèles de survie, montrant globalement que l'erreur de prédiction augmentait avec le temps. Le modèle de survie profonde affichait un score IBS très bas comparé aux autres modèles, même avec une période de suivi plus longue, ce qui indique une prédiction plus fiable et mieux calibrée (Fig. 3.3).

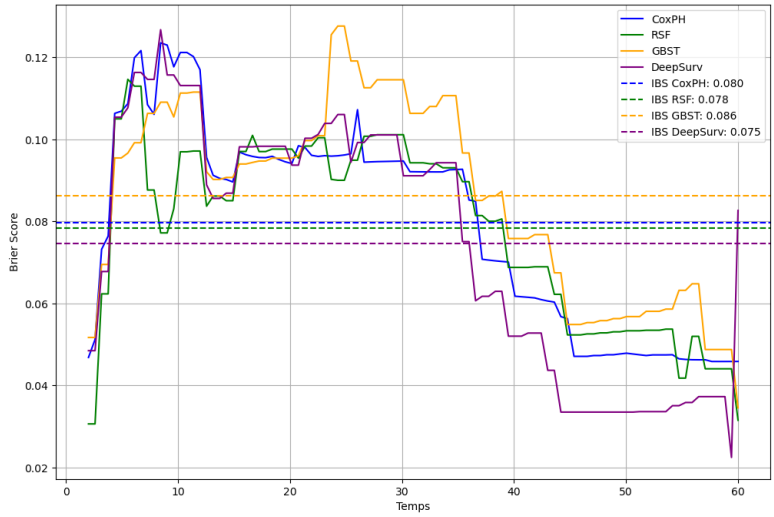


FIGURE 3.3 – Comparaison des Brice Score Intégré des Modèles

3.6 Conclusion

Les résultats ont montré que les modèles de machine learning avaient réussi à estimer avec précision le temps de survie, chaque modèle affichant des performances solides, une bonne calibration (faibles IBS) et des scores de concordance satisfaisants sur l'ensemble de test. En outre, l'examen des valeurs SHAP et de l'importance par permutation a permis d'identifier des facteurs clés (notamment tabac, Adénopathie, Métastases, et Infiltrant) ayant une influence significative sur les prédictions, ce qui a accru la compréhension et la transparence des modèles. Ces analyses ont confirmé que les approches les plus performantes, en particulier le Deep Survival, étaient supérieures pour prévoir l'évolution des patients à long terme. Globalement, les résultats ont démontré que les modèles de machine learning pouvaient être des outils très précieux pour prédire les temps de survie, combinant précision et facilité d'interprétation.

Conclusion Générale

Cette recherche avait pour but d'examiner la survie des patients atteints de cancer de l'estomac au Sénégal en utilisant à la fois des méthodes statistiques traditionnelles et des techniques d'apprentissage automatique.
Elle a permis de mettre en lumière que divers facteurs, comme la présence de métastases, le tabagisme, les caractéristiques des tumeurs et certaines comorbidités, influençaient de manière significative la survie des patients.

Les modèles Kaplan-Meier et Cox avaient révélé des différences importantes dans la durée de survie selon les variables analysées, malgré certaines contraintes tels que l'hypothèse des risques proportionnels et la gestion de la censure.

En outre, l'application des techniques de machine learning, en particulier les forêts aléatoires de survie, le Gradient Boosting Survival Tree et Deep Survival, avait apporté des perspectives nouvelles en augmentant la robustesse et la précision des prédictions.

Cette approche combinée a ainsi enrichi la compréhension des interactions complexes entre les variables cliniques et tumorales, offrant un regard inédit sur les mécanismes qui influencent la survie des patients atteints de cancer de l'estomac.

Perspectives

Pour aller plus loin dans cette recherche, nous avions pensé à intégrer des données supplémentaires, comme des marqueurs génétiques, des informations provenant de l'imagerie médicale et la phase du cancer.

L'objectif serait de développer un modèle multi-tâches capable de prédire le traitement le plus adapté (chirurgie seule ou chirurgie associée à la chimiothérapie,....), ainsi que l'événement (décès) et la durée de survie.

Annexe

TABLE 3.2 – Tableau Récapitulatif des Résultats du Test de Log-Rank

Variable	Modalités	Décès		Moyenne	Mediane	Test Statistic	P-value
		Non	Oui				
Cardiopathie	Non	92	136	37,19	54	42,3	7,83E-11
	Oui	22	87	20,4	12		
Ulcère Gastrique	Non	94	114	35,92	60	24,108	9,11E-07
	Oui	20	109	25,78	21		
Douleur épigastrique	Non	15	29	25,11	4	2,067	1,51E-01
	Oui	99	194	33,03	24		
Ulcère Bourgeonnant	Non	63	161	26,87	13	19,282	1,13E-05
	Oui	51	62	41,82	60		
Constipation	Non	65	96	35,86	60	6,112	1,34E-02
	Oui	49	127	28,49	21		
Denitrution	Non	85	136	34,35	25	8,92	2,82E-03
	Oui	29	87	27,45	17		
Tabac	Non	93	62	49,4	10	166,689	3,91E-38
	Oui	21	161	16,75			
Mucineux	Non	39	79	3,213	26	0,084	7,71E-01
	Oui	75	144	31,26	21		
Tubuleux	Non	9	85	18,63	10	45,812	1,30E-11
	Oui	105	138	37,33	59		
Infiltrant	Non	90	71	47,45		153,764	2,61E-35
	Oui	24	152	17,07	9		
Stenosant	Non	91	92	42,68	60	80,205	3,37E-19
	Oui	23	131	19,28	12		
Métastases	Non	82	111	39,57	54	37,322	1,00E-09
	Oui	32	112	21,57	10		
Adénopathie	Non	98	102	42,72	60	138,128	6,83E-32
	Oui	16	121	15,44	8		
Traitement	Exclusive	61	125	29,74	21	1,392	2,38E-01
	Chimio	53	98	34,73	41		
SEXE	Non	57	99	31,43	21	0,11	7,41E-01
	Oui	57	124	32,43	25		

TABLE 3.3 – Tableau récapitulatif des valeurs Shap selon chaque modèle

Variable	Shap_Cox	Shap_rsf	Shap_gbst	Shap_DeepSurv
Metastases	2.11	3.22	0.25	0.50
Adenopathie	1.45	5.2	0.59	0.34
Infiltrant	1.17	5.22	0.51	0.68
Tabac	1.13	4.97	0.47	0.54
Mucineux	1.12	0.33	0.04	0.23
Denitrution	0.96	1.28	0.14	0.35
Ulceregastrique	0.83	1.69	0.25	0.10
Cardiopathie	0.74	1.62	0.17	0.27
Douleurepigastrique	0.70	0.29	0.08	0.11
Stenosant	0.69	2.85	0.25	0.23
AGE	0.52	0.72	0.12	0.13
Ulcerobourgeonnant	0.51	1.43	0.09	0.07

TABLE 3.4 – Hyperparamètres optimisés pour différents modèles de survie

Modèle	**Paramètre**	**Valeur**
Forêt de Survie Aléatoire	min_samples_leaf	1
	min_samples_split	2
	n_estimators	100
Gradient Boosting Survival Tree	learning_rate	0.2
	min_samples_leaf	1
	min_samples_split	10
	n_estimators	50
Deep Survival	num_nodes	[64, 64, 64]
	dropout	0.1
	learning_rate	0.001
	batch_size	56
	epochs	100

Script python utilise pour l'analyse et la modélisation

[ ]:
```
import pandas as pd
import numpy as np
import seaborn as sns
%matplotlib inline
import scipy.stats as stats
import statsmodels.api as sm
import matplotlib.pyplot as plt
from itertools import product
import matplotlib.pyplot as plt
from lifelines import CoxPHFitter
from lifelines import KaplanMeierFitter
from sklearn.model_selection import train_test_split
```

[ ]:
```
from google.colab import drive
drive.mount('/content/drive')
# Lire le fichier en sautant la première ligne
df = pd.read_excel("/content/.../data.xlsx", skiprows=1)
# Identifier les colonnes catégoriques
CatCols = df.select_dtypes(include=['object']).columns
# Encodage des variables catégoriques
#Label Encoding
label_encoder = LabelEncoder()
for col in CatCols:
    df[col] = label_encoder.fit_transform(df[col].astype(str))

# Afficher les premières lignes du dataframe
df.head()
```

Approche non parametrique de Kaplan - Meier

[ ]:
```
# Initialisation du Kaplan-Meier
kmf = KaplanMeierFitter()

# Ajustement du modèle avec les données
kmf.fit(df['Tempsdesuivi (Mois)'], event_observed=df['Deces'])

# Tracer la fonction de survie avec les IC
ax = kmf.plot_survival_function(ci_show=True)

# Ajouter les points de censure
censored_times = df.loc[df['Deces'] == 0, 'Tempsdesuivi (Mois)']
survival_probabilities = [float(kmf.
 ↪survival_function_at_times(time).iloc[0]) for time in␣
 ↪censored_times]

plt.scatter(censored_times,
            survival_probabilities,
            color='red',
            label='Censures',
            alpha=0.7)

# Ajout des titres et légendes
plt.title('Fonction de survie avec IC et censures')
plt.xlabel('Mois')
plt.ylabel('Probabilité de survie')
plt.legend()
plt.show()
```

Le code complet associé à ce mémoire est disponible sur GitHub à l'adresse suivante : https://github.com/sefdineahmed/shahidi.
Ce dépôt contient l'ensemble des scripts nécessaires pour la reproduction des expériences et l'analyse des résultats présentés dans ce mémoire.

Bibliographie

[1] Sigrún Andradóttir. A review of random search methods. *Handbook of simulation optimization*, pages 277–292, 2014.

[2] Ghalib A. Bello, Timothy J. W. Dawes, Jinming Duan, Carlo Biffi, Antonio de Marvao, Luke S. G. E. Howard, J. Simon R. Gibbs, Martin R. Wilkins, Stuart A. Cook, Daniel Rueckert, and Declan P. O'Regan. Deep learning cardiac motion analysis for human survival prediction. *CoRR*, abs/1810.03382, 2018. URL http://arxiv.org/abs/1810.03382.

[3] P. et al. Blanche. Time-dependent auc in survival analysis : A tutorial. *Statistics in Medicine*, 2015. URL https://onlinelibrary.wiley.com/doi/10.1002/sim.6423.

[4] L. Breiman. Random forests. *Machine Learning*, 45(1) :5–32, 2001. URL https://link.springer.com/article/10.1023/A:1010933404324.

[5] Debin Cheng, Dong Liu, Xian Li, Zhao Zhang, Zhenzhou Mi, Weidong Tao, Jun Fu, and Hongbin Fan. Deep-learning–based model for the prediction of cancer-specific survival in patients with spinal chordoma. *World Neurosurgery*, 178 :e835–e845, 2023. ISSN 1878-8750. doi : https://doi.org/10.1016/j.wneu.2023.08.032. URL https://www.sciencedirect.com/science/article/pii/S1878875023011373.

[6] Thomas A Gerds and Martin Schumacher. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6) :1029–1040, 2006.

[7] GLOBOCAN. Global cancer observatory. https://gco.iarc.fr, 2022.

[8] Ted A Gooley, Wendy Leisenring, John Crowley, and Barry E Storer. Estimation of failure probabilities in the presence of competing risks : new representations of old estimators. *Statistics in medicine*, 18(6) :695–706, 1999.

[9] D. G. Haller, J. Tabernero, J. Maroun, et al. Capecitabine plus oxaliplatin compared with fluorouracil and oxaliplatin as adjuvant therapy for stage iii

colon cancer. *Journal of Clinical Oncology*, 29(11) :1465–1471, 2011. doi : 10.1200/JCO.2010.33.6297.

[10] F. et al. Harrell. Regression modeling strategies for prognostic prediction. *Lifetime Data Analysis*, 1996. URL https://link.springer.com/article/10.1007/s10985-006-9022-0.

[11] Y. Huang, J. Li, M. Li, and R. R. Aparasu. Application de l'apprentissage automatique à la prédiction des résultats de survie à partir de données réelles : une étude de portée. *BMC Medical Research Methodology*, 23, 2023. doi : 10.1186/s12874-023-02078-1.

[12] Y. et al Huang. Application of deep learning in medical image processing for survival prediction. *Journal of Medical Systems*, 45(9) :34–56, 2021. doi : 10.1007/s10916-021-01757-7.

[13] Yinan Huang, Jieni Li, Mai Li, and Rajender R. Aparasu. Application of machine learning in predicting survival outcomes involving real-world data : a scoping review. *BMC Medical Research Methodology*, 23(1) :268, 2023. ISSN 1471-2288. doi : 10.1186/s12874-023-02078-1. URL https://doi.org/10.1186/s12874-023-02078-1.

[14] Yinan Huang, Jieni Li, Mai Li, and Rajender R Aparasu. Application of machine learning in predicting survival outcomes involving real-world data : a scoping review. *BMC medical research methodology*, 23(1) :268, 2023.

[15] E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282) : 457–481, 1958.

[16] R. Katz and D. Velasco. Predicting patient survival with machine learning : Applications in health systems and hospitals. *IEEE Access*, 6 :7779–7785, 2018. doi : 10.1109/ACCESS.2018.2800864.

[17] D. W. Kim, S. Lee, S. Kwon, W. Nam, I. H. Cha, and H. J. Kim. Deep learning-based survival prediction of oral cancer patients. *Scientific Reports*, 9(1), 2019. doi : 10.1038/s41598-019-43372-7.

[18] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 1188–1196, 2017.

[19] F. Li, J. Li, L. Liu, L. Huang, L. Zhou, and H. He. Modèle calibré basé sur l'apprentissage automatique pour la fonction de cartographie de prévision

de vienne 3 délai humide zénithal. *Remote Sensing*, 15, 2023. doi : 10.3390/rs15194824.

[20] Shen ML Zhao MJ Liu H Li LW, Liu X. Development and validation of a random survival forest model for predicting long-term survival of early-stage young breast cancer patients based on the seer database and an external validation cohort. *Oncology Journal of Tianjin*, 40(8), 2024. doi : 10.62347/OJTY4008.

[21] Jiaxi Lin. Développement d'un modèle de prédiction basé sur la forêt de survie aléatoire pour le pronostic postopératoire du cancer du pancréas : une étude basée sur seer. *Cancers*, 14, 2022. doi : 10.3390/cancers14194667.

[22] S. M. Lundberg and S. I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, page 4765–4774, 2017.

[23] Tom M Mitchell. *Machine learning*, volume 1. McGraw-hill New York, 1997.

[24] Dirk F Moore et al. *Applied survival analysis using R*, volume 473. Springer, 2016.

[25] Roya Najafi-Vosough, Javad Faradmal, Leili Tapak, Behnaz Alafchi, Khadijeh Najafi-Ghobadi, and Tayeb Mohammadi. Prediction the survival of patients with breast cancer using random survival forests for competing risks. *Journal of Preventive Medicine and Hygiene*, 63(2) :E298, 2022.

[26] J-P Nakache, A Gueguen, M Zins, and M Goldberg. Analyse de données de survie groupées avec covariables dépendant du temps : application à l'étude de l'effet prédictif de l'état de santé perçu sur le décès, chez les hommes de la cohorte gazel observés dans la période 1989-1999. *Revue de statistique appliquée*, 52(2) :27–49, 2004.

[27] Ronald Wihal Oei, Yingchen Lyu, Lulu Ye, Fangfang Kong, Chengrun Du, Ruiping Zhai, Tingting Xu, Chunying Shen, Xiayun He, Lin Kong, et al. Progression-free survival prediction in patients with nasopharyngeal carcinoma after intensity-modulated radiotherapy : machine learning vs. traditional statistics. *Journal of Personalized Medicine*, 11(8) :787, 2021.

[28] P. Rawla, A. Barsouk, and A. Barsouk. Epidemiology of gastric cancer : Global trends, risk factors, and prevention. *Przegląd Gastroenterologiczny*, 14(1) :26–38, 2019. doi : 10.5114/pg.2018.80001.

[29] Patrick Royston and Douglas G Altman. External validation of a cox prognostic model : principles and methods. *BMC medical research methodology*, 13 :1–15, 2013.

[30] I. Sy, M. Bousso, A. Correa, and M. Dieng. State of the art of artificial intelligence applications in oncology. *Open Journal of Applied Sciences*, 13 :2245–2262, 2023. doi : 10.4236/ojapps.2023.1312175.

[31] Khadidiatou Toure. Étude de survie et prédiction du risque de décès pour les patients atteints du cancer de l'estomac. Master's thesis, Université Alioune Diop, May 2021. Mémoire soutenu le 22 mai 2021.

[32] Hajime Uno, Tianxi Cai, Michael J Pencina, Ralph B D'Agostino, and Lee-Jen Wei. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10) :1105–1117, 2011.

[33] Yucan Xu, Lingsha Ju, Jianhua Tong, Chengmao Zhou, and Jianjun Yang. Supervised machine learning predictive analytics for triple-negative breast cancer death outcomes. *OncoTargets and therapy*, pages 9059–9067, 2019.

[34] X. Zhao, J. Du, T. Ge, Y. Wang, and J. He. Machine learning in survival analysis : Methods and applications. *Computational and Structural Biotechnology Journal*, 20 :600–612, 2022. doi : 10.1016/j.csbj.2021.12.005.