الجمهورية الجزائرية الديمقراطية الشعبية
**République Algérienne Démocratique et Populaire**

وزارة التعليم العالي والبحث العلمي
**Ministère de l'Enseignement Supérieur et de la Recherche Scientifique**

جامعة الإخوة منتوري قسنطينة 1
Frères Mentouri Constantine 1 University
Université Frères Mentouri Constantine 1

**Faculté des Sciences de la Nature et de la Vie**
**Département de biologie appliquée**

**كلية علوم الطبيعة والحياة**
**قسم البيولوجيا التطبيقية**

**Mémoire présenté en vue de l'obtention du diplôme de Master**

**Domaine :** Sciences de la Nature et de la Vie
**Filière :** Science Biologique
**Spécialité :** *Bio-informatique*

**N° d'ordre :**
**N° de série :**

Intitulé :

---

# Deepredictor : un modèle de réseaux de neurones denses pour la prédiction de la mortalité des patients atteints d'un cancer du sein

---

**Présenté par :** BENOUAR Mohamed Salah Amine
CHERITI Abir

**Jury d'évaluation :**

**Président 1 :** HAMIDECHI Abdelhafid (Professeur - Université Frères Mentouri, Constantine 1).
**Encadreur :** CHEHILI Hamza (MCA - Université Frères Mentouri, Constantine 1).
**Examinateur :** GHERBOUDJ Amira (MCA - Université Frères Mentouri, Constantine 1).

**Année universitaire**
**2021 - 2022**

# *Remerciements*

*Nous remercions vivement le Dieu le tout puissant de son aide et de nous avoir donné la persistance, la volonté et la patience de mener ce travail à terme. Nous tenons aussi à remercier tous nos enseignants qui ont participé à notre formation et qui nous ont inculqué des bases solides dans ce domaine.*

*Nous dédions ce travail, notre reconnaissance et nos remerciements plus particulièrement à Mr CHEHILI Hamza qui nous a assisté afin de mener à bien ce travail. Ses précieux conseils et orientations nous ont poussés toujours le plus loin possible, rendant cette discipline une véritable passion.*

*Nous exprimons notre gratitude envers les membres du jury Mr HAMIDECHI Mohamed Abdelhafid et Mme GHERBOUDJ Amira qui auront à prendre la responsabilité d'examiner et évaluer notre modeste travail et dont nous souhaitons donner satisfaction et répondre à leurs exigences.*

*Un grand merci à nos parents, pour leur amour, leurs conseils ainsi que leur soutien inconditionnel, à la fois moral et économique, qui nous a permis de réaliser les études que nous voulions et par conséquent ce mémoire.*

*Résumé*

L'objectif de ce travail est de développer une architecture de réseau de neurones artificiels qui vise la prédiction binaire de la survie des patientes atteintes d'un cancer du sein, et de démontrer l'efficacité des DNN et leur capacité à apprendre et à déterminer les caractéristiques dominantes face à un jeu de données non seulement relativement peu volumineux par rapport aux standards des jeux de données médicales disponibles dans la littérature, qui est amputé d'une quantité considérable de ces données génétiques et cliniques, mais aussi extrêmement complexe du fait de leur diversité.

**Mots clés :** architecture, prédiction binaire, cancer du sein, DNN.

**Abstract**

The objective of this work is to develop an artificial neural network architecture that aims at the binary prediction of the survival of patients with breast cancer, and to demonstrate the efficiency of DNNs and their ability to learn and determine the dominant characteristics in the face of a dataset not only of relatively low volume compared to the standards of medical datasets available in the literature, which is amputated by a considerable amount of these genetic and clinical data, but also extremely complex due to the fact of their diversity.

Keywords : architecture, binary prediction, breast cancer, DNN.

**ملخص**

الهدف من هذا العمل هو تطوير بنية شبكة عصبية اصطناعية تهدف إلى التنبؤ الثنائي لبقاء مرضى سرطان الثدي على قيد الحياة، وإثبات كفاءة DNNs وقدرتهم على التعلم وتحديد الخصائص السائدة في مواجهة مجموعة البيانات ليس فقط ذات حجم منخفض نسبيًا مقارنة بمعايير مجموعات البيانات الطبية المتوفرة في الأدبيات، والتي تم بترها بواسطة قدر كبير من هذه البيانات الجينية والسريرية، ولكنها أيضًا معقدة للغاية بسبب حقيقة تنوعها.

الكلمات المفتاحية: بنية، التنبؤ الثنائي، سرطان الثدي، DNN.

## Liste des tableaux

## Liste des figures

**Table des matières**

Chapitre II : Matériel et méthode

Chapitre III : Résultats et discussion

# INTRODUCTION

## Introduction

Le cancer est une maladie causée par une transformation cellulaire anormale et hyperprolifération. Ces cellules dérégulées finissent par former des masses appelées tumeurs malignes. Les cellules cancéreuses ont tendance à envahir les tissus voisins et à se détacher de la tumeur. Ils migrent ensuite à travers les vaisseaux sanguins et lymphatiques pour former une autre tumeur (métastase), sachant que certaines tumeurs sont bénignes (ne deviennent pas envahissantes).

D'après l'organisation mondiale de la santé, À l'origine de près de 10 millions de décès en 2020, le cancer est l'une des principales causes de mortalité dans le monde. En 2020, les cancers les plus courants (en termes de nombre de cas recensés) étaient : le cancer du sein (2,26 millions de cas), le cancer du poumon (2,21 millions de cas), le cancer colorectal (1,93 million de cas), le cancer de la prostate (1,41 million de cas), le cancer de la peau (non-mélanome) (1,20 million de cas) et le cancer de l'estomac (1,09 million de cas).

En 2020, l'OMS a recensé 2,3 millions de femmes atteintes du cancer du sein et 685 000 décès par ce dernier dans le monde, ce qui fait du cancer du sein le cancer le plus courant à l'échelle mondiale. Ce type de cancer prend la plupart du temps la forme d'une masse ou d'un épaississement non douloureux dans le sein, un changement de taille et de forme ou d'apparence du sein.

Certains facteurs accroissent le risque de cancer du sein, à savoir, notamment, un âge grandissant, l'obésité, l'abus d'alcool, une exposition aux radiations, suivre un traitement hormonal prolongé et le tabagisme. Mais malheureusement, même si on contrôlait tous les facteurs de risque qui peuvent l'être, on ne parviendrait à réduire le risque de cancer du sein que de 30 % tout au plus.

Certaines mutations génétiques héréditaires « de haute pénétrance » accroissent fortement le risque de cancer du sein. Les plus importantes d'entre elles sont présentes dans les gènes BRCA1, BRCA2.

Dans les pays à revenu élevé, le taux de mortalité par cancer du sein comparatif par âge a chuté de 40 % entre les années 1980 et 2020. Les pays qui sont parvenus à réduire la mortalité par cancer du sein ont réussi à atteindre une réduction de la mortalité annuelle par

cancer du sein de 2 à 4 % par an. Si la mortalité annuelle baissait chaque année de 2,5 % dans le monde, on éviterait 2,5 millions de décès par cancer du sein entre 2020 et 2040.

Dans ce but, on propose un modèle qui prédit le taux de la survie du patient en prenant en compte une multitude d'attributs génétiques et cliniques.

Afin d'atteindre notre objectif, ce travail est organisé en trois chapitres :

Le premier chapitre constitue une synthèse bibliographique qui donne un aperçu général sur le cancer du sein et le traitement.

Le deuxième chapitre est consacré à la description des données cliniques et génétiques utilisées et la construction de l'architecture de réseau de neurones artificiel.

Le dernier chapitre, Résultats et discussion, présente et discute les différents résultats obtenus lors de la visualisation des graphes générés par notre étude statistique des données et visualisation des graphes d'estimation du modèle.

Enfin, nous présentons une conclusion générale et quelques perspectives.

# CHAPITRE I :

# Synthèse Bibliographique

## 1. Le cancer du sein

Le sein se compose de graisse, de lobules et des canaux. Les tissus mammaires sont influencés par des hormones (œstrogène et progestérone) produites tout au long de la vie.

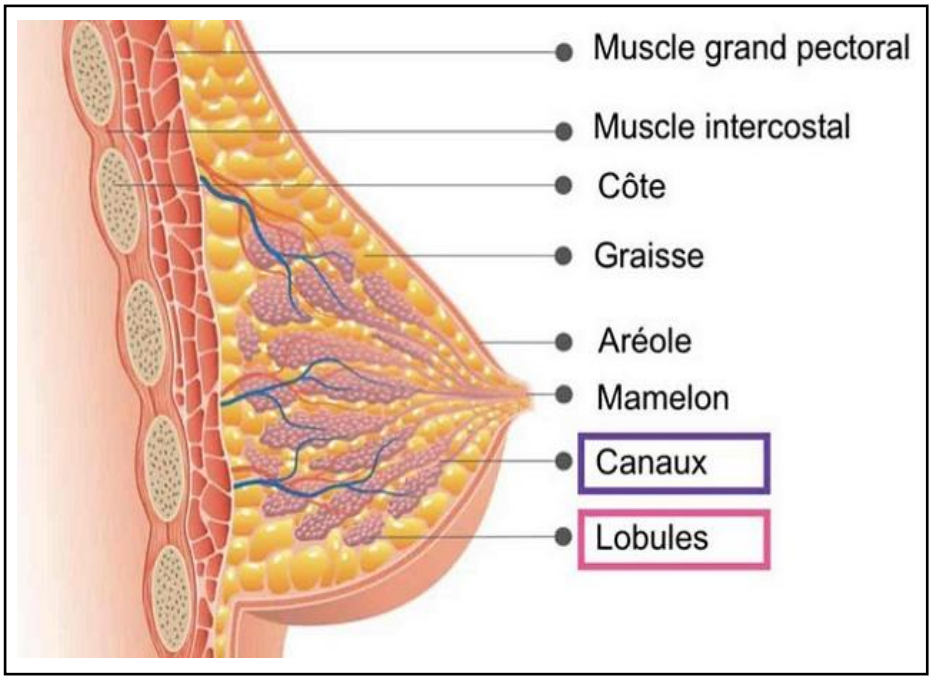



**Figure 1 :** Histologie du sein (Benchmark, 2022)

Le cancer du sein prend naissance dans les cellules mammaires. C'est le cancer le plus courant chez les femmes et provoque la destruction et l'invasion des tissus. Il se caractérise par la présence ou l'absence de trois types de molécules à sa surface : les récepteurs d'œstrogènes (ER), de la progestérone (PR) et un facteur de croissance appelé HER2. L'identification de ces signatures moléculaires peut diviser le cancer du sein en plusieurs catégories qui ne répondent pas aux mêmes traitements.

Ce type de cancer peut aussi toucher les hommes. Cependant, la proportion d'hommes touchés par ce type de cancer reste faible (environ 1 % des cas de cancer du sein touchant les hommes) (institut national du cancer, 2022).

## 2. La classification des différents cancers du sein

### 2.1. Le carcinome canalaire

Est le plus fréquent et se forme à l'intérieur des canaux de lactation, diagnostiqué assez tôt grâce à la mammographie. Il reste non invasif et on en guérit dans presque tous les cas avec un traitement adapté. En revanche, sans traitement, il peut poursuivre sa croissance et devient invasif en se propageant à l'extérieur des canaux.

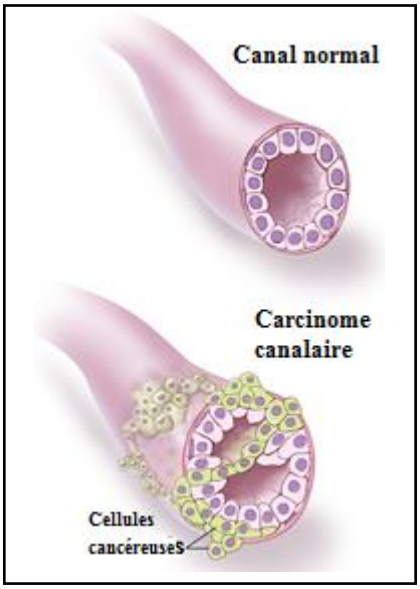



**Figure 2 :** le carcinome canalaire (Zemmouri, 2022)

### 2.2. Le carcinome lobulaire

Le carcinome lobulaire se forme dans les lobules. Les cellules cancéreuses traversent leur paroi et se développent dans les tissus environnant.

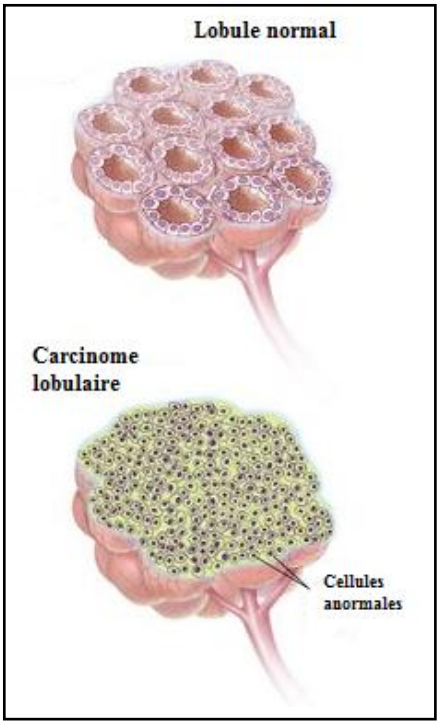



**Figure 3 :** Carcinome lobulaire (Zemmouri, 2022)

### 2.3. Le carcinome mucineux

Le seul critère pour dire que c'est un carcinome mucineux est l'absence des flaques de mucines situées au niveau stromale extracellulaire.

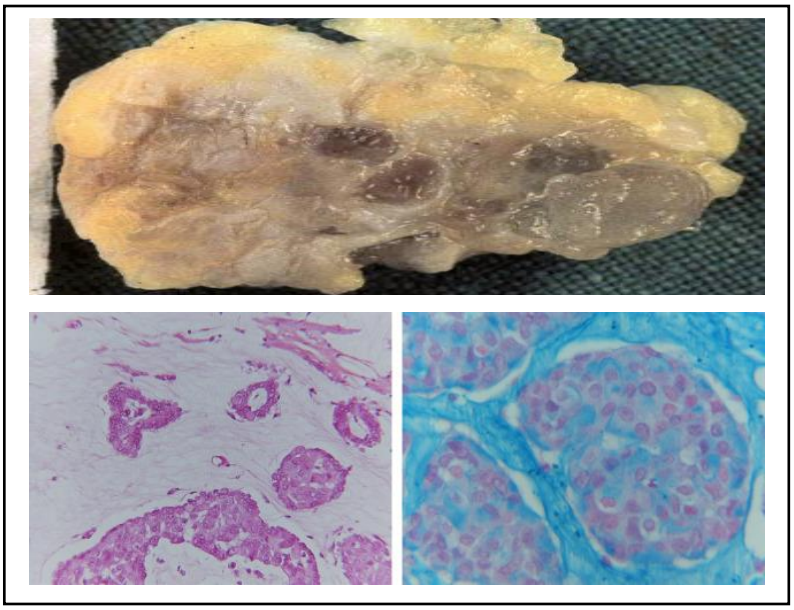


**Figure 4 :** Carcinome mucineux (Antoine *et al*., 2016)

### 2.4. Les sous-types

#### 2.4.1. Le cancer de type luminal

- luminal A : Il est fortement exprimé dans les récepteurs des œstrogènes et de la progestérone, mais pas dans HER2 (HER2). Ce type de tumeur est généralement moins agressif.
- Luminal B : L'expression des récepteurs hormonaux est faible, mais HER2 (HER2 +) peut être surexprimé. Ce type de tumeur est généralement prolifératif (Malhotra *et al*., 2010).

#### 2.4.2. Le cancer HER2

Il possède des récepteurs du facteur de croissance HER2, mais pas des récepteurs des œstrogènes et/ou de la progestérone (ER- et/o RP- et HER2+) (Schnitt, 2010).

#### 2.4.3. Le cancer triple-négatif

Le cancer du sein triple négatif est un sous-type de cancer du sein dépourvu de l'expression du récepteur d'œstrogène, du récepteur de la progestérone et du récepteur 2 du facteur de croissance épidermique humain (Wu *et al*., 2022).

## 3. Le traitement

### 3.1. La radiothérapie

La radiothérapie intervient en association avec la chirurgie car le chirurgien enlève la partie visible macroscopique de la tumeur alors que la radiothérapie s'adresse à la partie invisible. Cette technique traite l'ensemble du sein par des rayonnements ionisants qui traversent le sein et qui vont donc entrainer certain nombre de modifications au niveau de l'ADN des cellules anormales mais également des cellules normales (Cowen, 2017).

### 3.2. La chimiothérapie

La chimiothérapie consiste en une série de 4 à 8 cycles de traitement en général intraveineux. Elle se fait par perfusion intraveineuse courte de 10 à 60 minutes, sur un rythme soit hebdomadaire, soit toutes les 3 semaines. Cette perfusion intraveineuse est potentiellement toxique pour les veines. Ce type de traitement est utilisé afin de réduire le volume tumoral et permettre un traitement conservateur (Zemmouri, 2022).

### 3.3. Hormonothérapie

Son principe repose sur l'inactivation de l'action des œstrogènes au niveau des récepteurs nucléaires par suppression des œstrogènes eux même ou par action directe au niveau de leur récepteurs afin d'empêcher la prolifération.

## 4. Les techniques de mesure

### 4.1. L'immunohistochimie

Le principe de l'immunohistochimie repose sur la reconnaissance d'un antigène par un anticorps marqué dans des coupes de tissus. Cette technique utilise un seul anticorps et la procédure est courte et rapide. Cependant, elle est moins sensible en raison de la faible amplification du signal et donc rarement utilisée.

La notation basée sur l'immunohistochimie (IHC) du statut ER est utilisée pour classer les tumeurs ER positives (ER+) et ER négatives (ER-) (Pereira *et al*., 2016).

### 4.2. Microarray

La formation de tumeurs implique des changements simultanés dans des centaines de cellules et des variations dans les gènes. Les puces à ADN fournissent une plate-forme pour tester simultanément un grand nombre d'échantillons génétiques. Elle aide notamment à l'identification des polymorphismes mononucléotidiques (SNP) et des mutations, à

l'identification des gènes associés à la chimiorésistance et à la découverte de médicaments. Nous pouvons aussi comparer les différents modèles de niveaux d'expression génique entre un groupe de patients cancéreux et un groupe de patients normaux et identifier le gène associé à ce cancer particulier (Govindarajan *et al*., 2012).

### 4.3. Z-score

Fondamentalement, un z-score est le nombre d'écarts types par rapport à la moyenne d'un point d'information. Quoi qu'il en soit, il s'agit en fait d'une proportion du nombre d'écarts-types en dessous ou au-dessus de la population que représente un score brut. Un z-score est autrement appelé un score standard. Ce score est représenté par la formule générale : $z = (x - \mu)/\sigma$. Où z représente le z-score, x représente le score, μ représente la moyenne et σ représente l'écart type.

## 5. L'intelligence artificielle et le cancer

Le développement de nouvelles techniques d'imagerie médicale pour le cancer du sein ont provoqué l'augmentation exponentielle du volume des datasets disponibles pour les radiologistes. L'intelligence artificielle quand a elle rend possible l'identification et la stratification de pattern complexe dans les images médicales, la traduction clinique des phénotypes tumoral au génotype, ce qui a engendré l'apparition d'une multitude d'approche d'aide à la décision tel que CADe (computer-Aided detection) et CADx( computer-aided diagnosis) (Yang *et al*., 2022).

Des modèles de machine learning tels que les forets aléatoires/ l'arbre de décision et la régression logistique ont été aussi proposés pour identifier les facteurs de pronostics décisifs pour la survie des patients touchés par le cancer du sein (Ganggayah *et al*., 2019).

Des modèles de deep learning ont aussi été développés pour la détection du cancer métastatique du sein à partir d'imagerie de ganglions sentinelles (Dayong *et al*., 2016).

Des modèles DL ont aussi été développés pour la classification des sous types de cancer à partir des données OMICS de la base de données TCGA (The Cancer Genome Atlas) (Lin *et al*., 2020).

## 6. Qu'est-ce que l'intelligence artificielle (IA) ?

L'intelligence artificielle est l'étude et la conception d'algorithmes pouvant répliquer des actions humaines. Un système intelligent peut prendre plusieurs formes. Il peut être indifférenciable de l'humain. Ce que nous appelons l'intelligence artificielle générale. Il peut

prendre la forme d'assistant vocal tel que Siri, Alexa, Cortana ou Google Assistant. Il peut être une voiture autonome tel que les voitures Tesla. Les recommandations des boutiques online et les NPC ou non playable characters dans les jeux vidéo constituent eux aussi une IA. Les systèmes intelligents peuvent donc prendre des décisions visant à la résolution de problèmes sans l'intervention humaine, d'où vient l'aspect intelligent de ces derniers (Riedl, Mark, 2019).

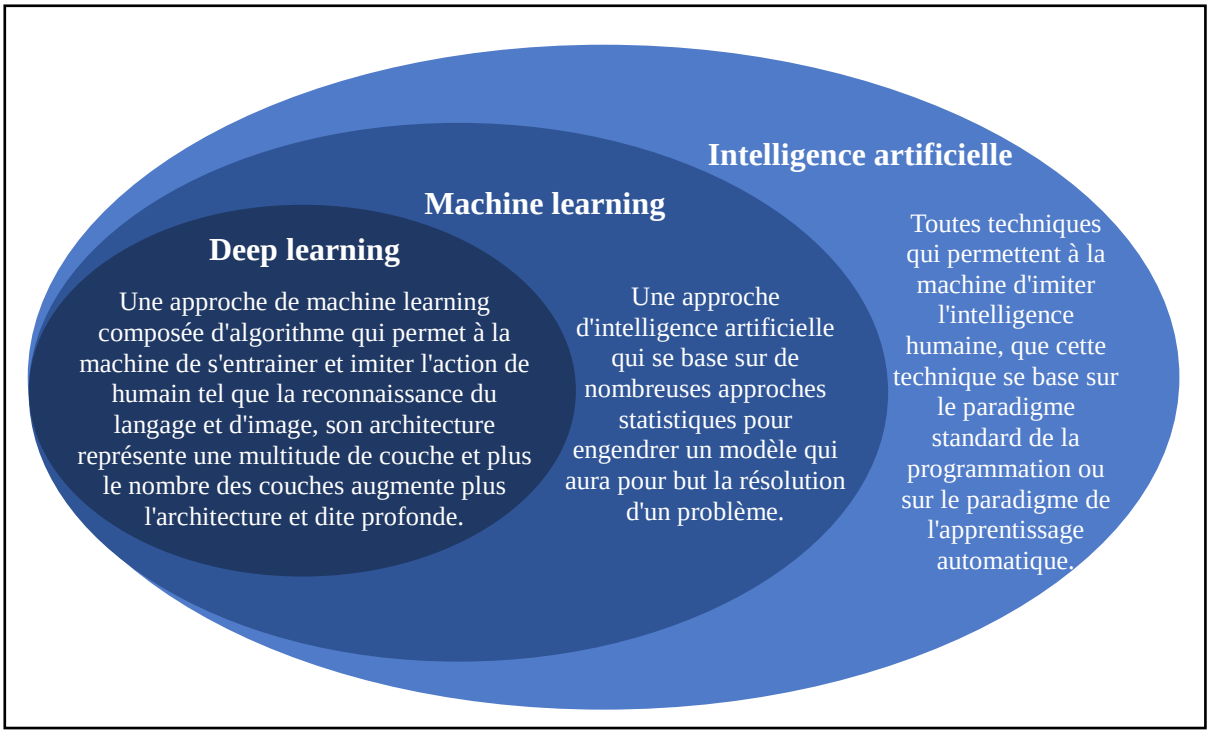



**Figure 5 :** Représentation de la différence entre l'intelligence artificielle, l'apprentissage automatique et l'apprentissage profond

## 7. L'apprentissage automatique (Machine Learning)

L'apprentissage automatique ou Machine Learning se base sur un paradigme diffèrent de celui de l'intelligence artificiel conventionnelle. Ce dernier est réalisé en appliquant des algorithmes qui apprennent de manière itérative à partir de données d'entrainement spécifiques à un problème. Ce qui permet aux ordinateurs de trouver des informations cachées, de soutirer les lois à partir de ces données et d'implémenter des modèles complexes sans être explicitement programmés. En particulier dans les tâches liées aux données de grande dimension telles que la classification, la régression et le clustering. L'apprentissage est théoriquement applicable sur quasiment tous genres de données et donc peut répondre à toute

sorte de question. En apprenant des calculs précédents et en extrayant les informations pertinentes de bases de données massives, cela peut aider à produire des décisions extrêmement fiables et reproductibles. Nous citons bien trois grandes catégories d'apprentissage automatiques.

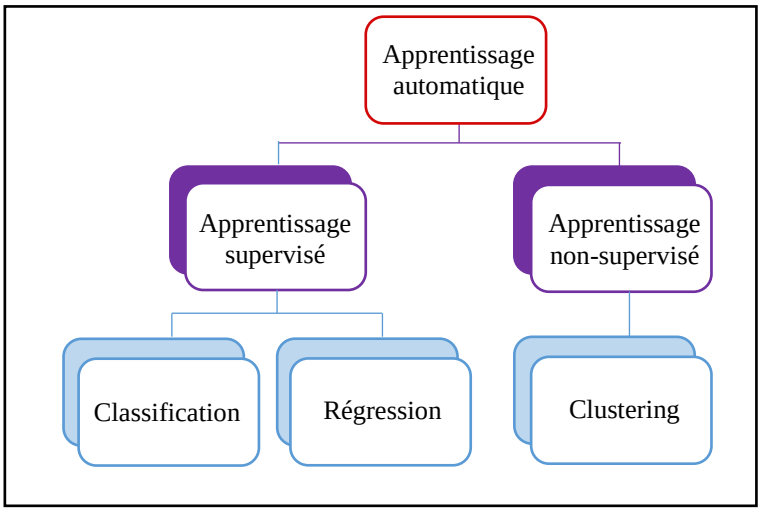



**Figure 6 :** Schéma de la classification des différentes approches de l'apprentissage automatique (Jaime, 2017)

### 7.1. Apprentissage automatique supervisé

L'apprentissage supervisé nécessite un ensemble de données d'entrainement constitué de données d'entrée et de sortie. L'algorithme compare les données de sortie réelle du dataset avec les données de sortie qu'il génère (prédiction, classification…etc.) et il modifie le model en se basant sur l'erreur calculée ou plus fréquemment appelée loss. L'amélioration du model reflètera la minimisation de la valeur loss.

### 7.2. Apprentissage automatique non supervisé

Comme son nom l'indique, cette approche de Machine Learning utilise des dataset ne contenant que les données d'entrée et aucune donnée de sortie et donc ne seront pas étiquetées. En effet, la machine ne peut pas voir la bonne réponse pour améliorer son apprentissage. L'objectif de cette approche est d'explorer les données et établir une structure d'attributs pour catégoriser les données et leur attribuer des données de sortie théoriques. L'approche d'apprentissage automatique non supervisée la plus connue et qui facilite le plus la compréhension de ce dernier est celle du Clustering.

### 7.3. Apprentissage automatique Semi-supervisé

Cette approche vise à utiliser un dataset constitué en même temps de données étiquetées et non étiquetées (étiqueté voulant dire contenant des données de sortie). La quantité de données étiquetées étant inferieur. Cette approche est utile quand il est coûteux d'étiqueter l'intégralité du dataset, que le coût soit financier ou humain. Cette approche est utile pour la classification, la régression et la prédiction (Ongsulee, 2017). Bien sûr nous pouvons citer d'autres mentions honorables et des approches extrêmement efficaces d'apprentissage automatique tel que l'apprentissage par renforcement.

### 7.4. Applications de l'apprentissage automatique

Il existe plusieurs applications de l'apprentissage automatique, la figure ci-dessous représente ces différentes applications :

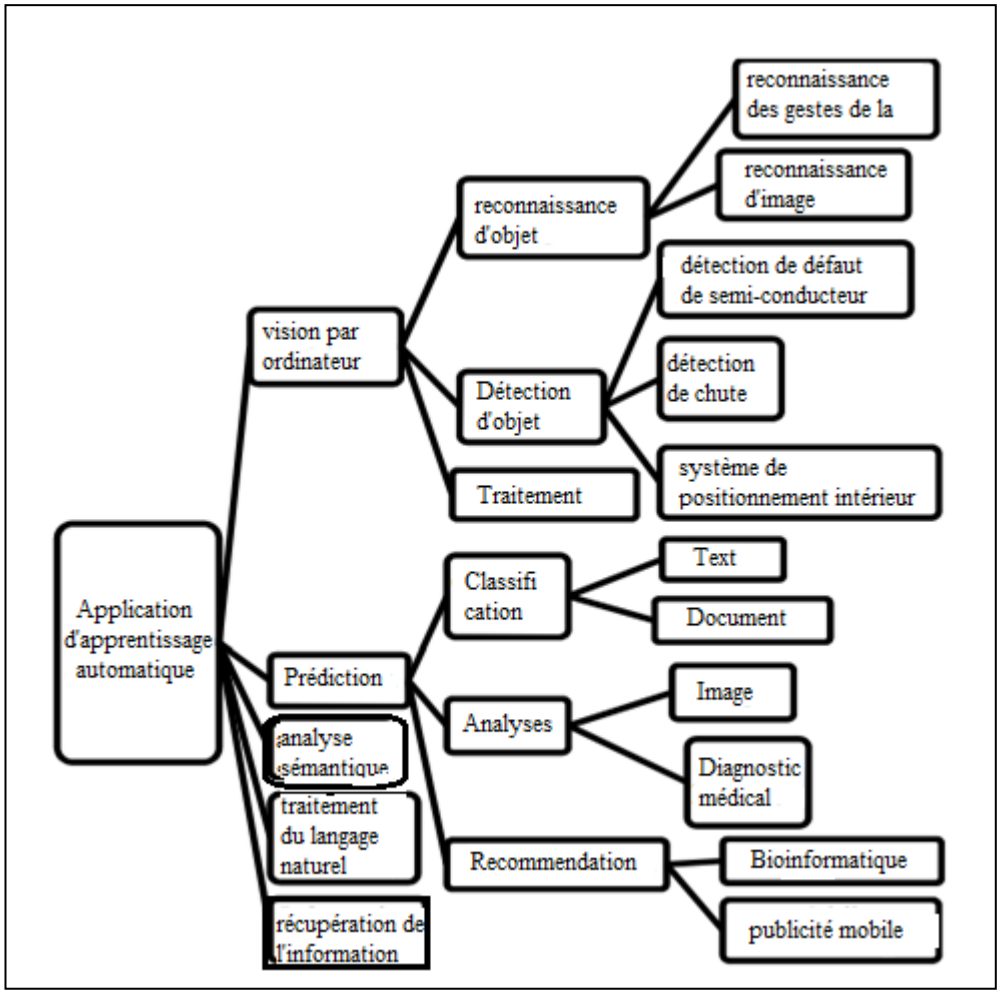



**Figure 7 :** Applications de l'apprentissage automatique (Shinde et Shah, 2018)

## 8. Apprentissage profond (Deep learning)

Les modèles d'apprentissage profond sont considérés comme les modèles prédictifs de pointe pour les grands ensembles de données uniquement au cours de la dernière décennie, même s'ils ont été théorisés pour la première fois dans les années 1980 (Rina ,1986).

Ces derniers ont été appliqués à de nombreux domaines tels que la reconnaissance vocale, le traitement du langage naturel, computer vision. Le terme "profond" dans l'apprentissage profond fait référence au nombre de couches à travers lesquelles les données sont transformées et ou l'information circule. En effet, il existe une multitude d'approches différentes d'apprentissage profond ou plutôt d'architecture différente. La différence entre ces derniers se traduit majoritairement par les différentes équations au niveau des neurones de chaque couche. Des types de computation, de fonctions utilisées, et la façon dont l'information circule au sein même de l'architecture. De toutes les approches disponibles dans la littérature, nous allons citer les trois approches les plus utilisées.

### 8.1. Réseaux de neurones artificiels

Les réseaux de neurones artificiels (ANN) ou réseaux de neurones denses (DNN) se sont inspirés des neurones et des réseaux qui constituent les cerveaux humains. L'ANN constitue un ensemble de modélisation de nœuds entièrement connectés (neurones). La propagation des stimuli des synapses cérébrales obéissent à la loi théorique de la neuroscience -neurones that wire together fire together- à travers le réseau neuronal. Ces architectures sont utilisées pour la sélection des fonctionnalités, la classification, la réduction de la dimensionnalité ou en tant que sous-module d'une architecture plus profonde comme les réseaux de neurones convolutifs (koumakis, 2020).

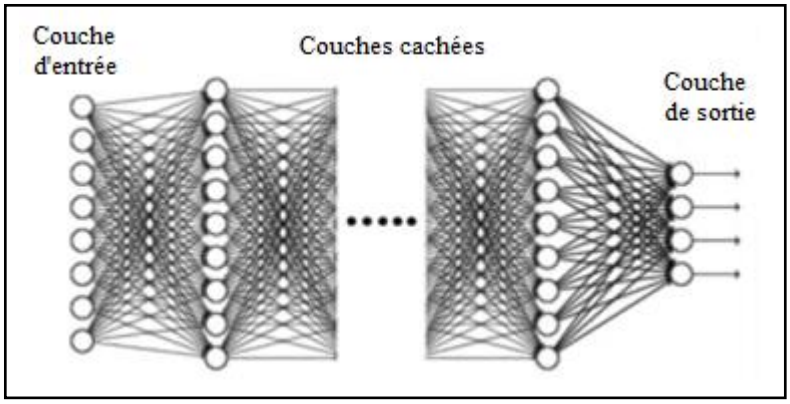



**Figure 8 :** Architecture générale d'un réseau neuronal profond avec plusieurs couches cachées (Sarker, 2021)

### 8.2. Réseaux de neurones convolutifs

Les réseaux de neurones convolutifs (CNN) sont basés sur les perceptrons multicouches et représentent des réseaux entièrement connectés où chaque nœud/neurone d'une couche est (entièrement) connecté à tous les nœuds de la couche suivante. Les ANNs sont des collections d'unités connectées et réglables qui peuvent transmettre un signal d'une unité à une autre. Au contraire, les CNN ont des couches d'unités de convolution matricielle qui reçoivent l'information des neurones de la couche précédente. Le principe fondamental de cette architecture est un calcul matriciel parallèle massif et de part ce calcul permettre une meilleure extraction de relations entre les attributs pertinents du dataset et la possibilité d'inférer des relations non linéaires entre le signal d'entrée et la donnée de sortie. Les CNN sont populaires pour l'extraction, la sélection et la réduction de caractéristiques, principalement pour la classification d'ensembles de données d'imagerie (Gu *et al.*, 2018).

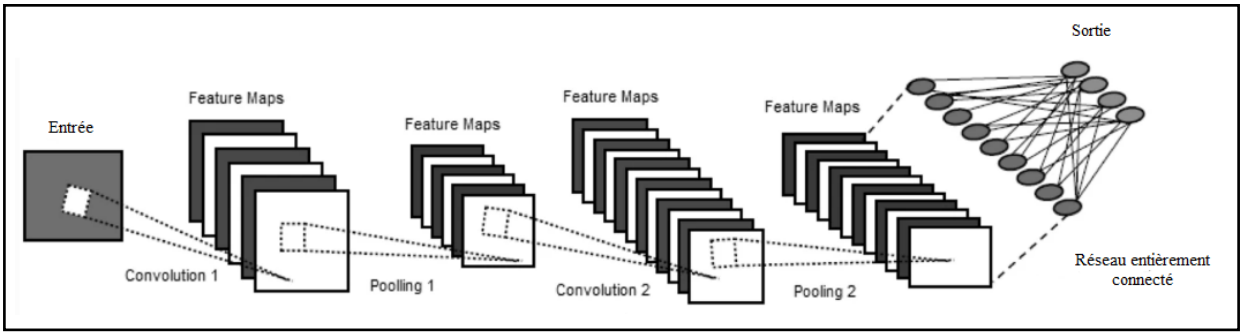



**Figure 9 :** Un exemple de réseau neuronal convolutif CNN comprenant plusieurs couches de convolution et de mise en commun (Sarker, 2021)

### 8.3. Les réseaux de neurones récurrents

Les réseaux de neurones récurrents (RNN) représentent des connexions entre les nœuds et forment un graphe orienté le long d'une séquence temporelle (Montana, Davis, 1989). Cela permet aux RNN de présenter un comportement dynamique temporel et leur permet d'intégrer une mémoire interne. Cette mémoire à court terme permet aux réseaux récurrents de littéralement se souvenir des informations des états précédemment analysés et choisir l'information pertinente. Par choix, nous voulons bien sur dire lui affecter une valeur et ne pas la réduire à 0. Une approche parfaite pour l'analyse séquentielle. Vulgairement décrite le point fort des RNN est sa capacité à utiliser l'information de la tâche précédente pour résoudre le problème présent (Williams, Zipser, 1989).

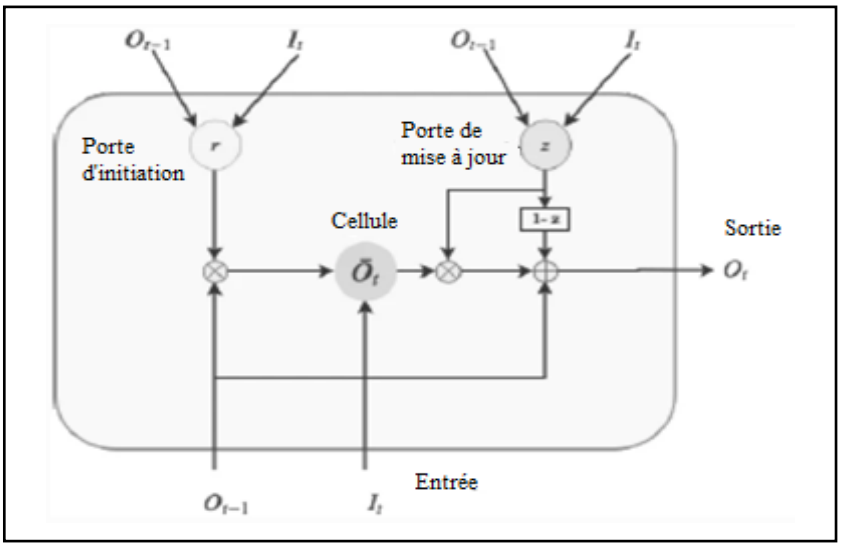



**Figure 10 :** Structure de base d'une cellule d'unité récurrente fermée (GRU) composée de portes de réinitialisation et de mise à jour (Sarker, 2021)

# CHAPITRE II :

# Matériel et Méthodes

## 1. Les données utilisées

La base de données Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) est un projet canado-britannique qui contient des données de séquençage ciblées de 1 980 échantillons primaires de cancer du sein. Les données cliniques et génomiques ont été téléchargées à partir de cBioPortal.

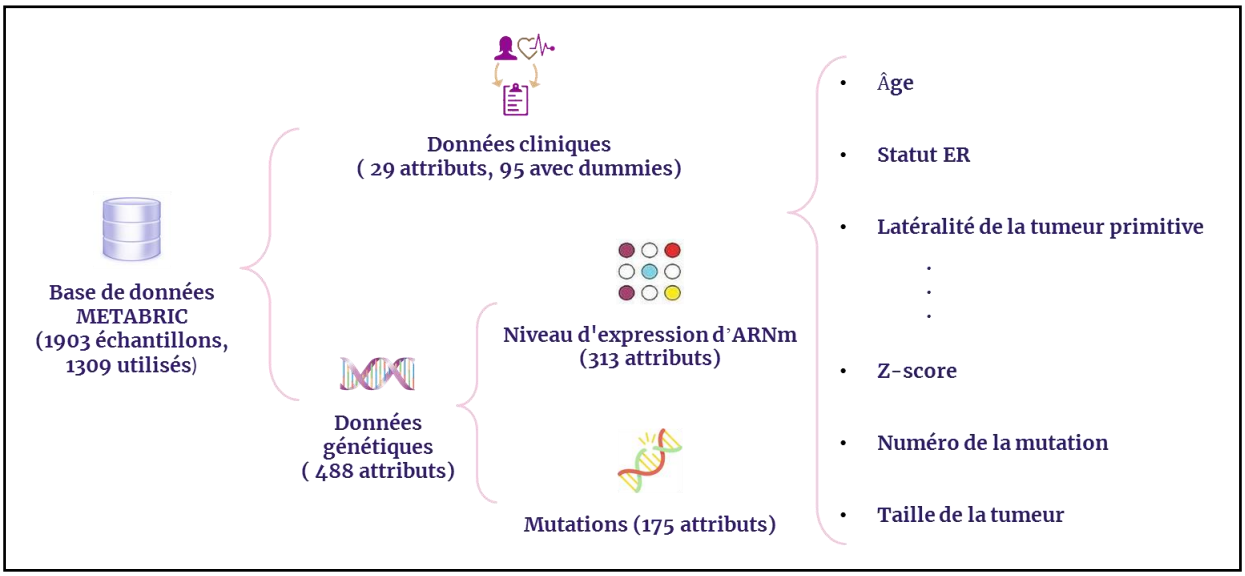



**Figure 11 :** Schéma représentant les données utilisées

L'ensemble des données a été recueilli par le professeur Carlos Caldas du Cambridge Research Institute et Professeur Sam Aparicio du British Columbia Cancer Center au Canada et publié sur Nature Communications (Pereira *et al*., 2016).

### 1.1. Les données cliniques

En plus de l'identifiant de l'étude, l'identifiant du patient et l'identifiant de l'échantillon, le tableau ci-dessous représente les 32 attributs des données cliniques utilisées.

**Tableau 1 :** les attributs cliniques utilisés

| Attribut | Description |
|---|---|
| age_at_diagnosis | L'âge du patient au moment du diagnostic |
| type_of_breast_surgery | Soit une mastectomie (qui fait référence à une intervention chirurgicale visant à retirer tout le tissu mammaire) ou la conservation du sein (qui fait |

| | |
|---|---|
| | référence à une urgence où seule la partie du sein qui a un cancer est enlevée) |
| cancer_type | Cancer du sein ou sarcome du sein (le plus dangereux) |
| cancer_type_detailed | Carcinome canalaire, carcinome lobulaire, carcinome mixte canalaire et lobulaire, carcinome mucineux |
| cellularity | La cellularité du cancer après la chimiothérapie (basse, élevée ou moyenne) |
| chemotherapy | Le patient a subit des séances de chimiothérapie ou non (oui/non) |
| pam50_+_Claudin-low_subtype | Un test de profilage tumoral qui aide à montrer si certains cancers du sein positifs aux récepteurs des œstrogènes (ER-positifs) et HER2-négatifs sont susceptibles de métastaser |
| Cohort | Un groupe de sujets qui partagent une caractéristique déterminante |
| er_status_measured_by_ihc | Pour évaluer si les récepteurs aux œstrogènes sont exprimés sur les cellules cancéreuses en utilisant l'immunohistochimie (positif/négatif) |
| er_status | Les cellules cancéreuses sont positives ou négatives pour les récepteurs aux œstrogènes |
| neoplasm_histologic_grade | Déterminé par la pathologie en regardant la nature des cellules, ont-elles l'air agressives ou non |
| her2_status_measured_by_snp6 | Évaluer si le cancer est positif pour HER2 ou non en utilisant des techniques moléculaires avancées |
| her2_status | Si le cancer est positif ou négatif pour HER2 |

| | |
|---|---|
| tumor_other_histologic_subtype | Type de cancer basé sur l'examen microscopique du tissu cancéreux |
| hormone_therapy | Que le patient ait eu ou non un traitement hormonal |
| inferred_menopausal_state | Que la patiente soit ménopausée ou non |
| integrative_cluster | Sous-type moléculaire du cancer basé sur l'expression de certains gènes |
| primary_tumor_laterality | Sein droit ou gauche |
| lymph_nodes_examined_positive | Prélever des échantillons du ganglion lymphatique pendant la chirurgie et voir s'il était impliqué par le cancer |
| mutation_count | Nombre de gènes qui ont des mutations pertinentes |
| nottingham_prognostic_index | Il est utilisé pour déterminer le pronostic après une chirurgie du cancer du sein. Sa valeur est calculée à partir de trois critères pathologiques : la taille de la tumeur ; le nombre de ganglions lymphatiques impliqués ; et le grade de la tumeur. |
| oncotree_code | OncoTree est une ontologie open source qui a été développée au Memorial Sloan Kettering Cancer Center (MSK) pour normaliser le diagnostic de type de cancer d'un point de vue clinique en attribuant à chaque diagnostic un code OncoTree unique. |
| overall_survival_months | Durée depuis le moment de l'intervention jusqu'au décès |
| overall_survival | Si le patient est vivant ou décédé. |
| pr_status | Les cellules cancéreuses sont positives ou négatives |

| | |
|---|---|
| | pour les récepteurs de la progestérone |
| radio_therapy | Si le patient a eu ou non une radiothérapie comme traitement |
| number_of_samples_per_patient | Nombre d'échantillon pris par patient |
| sample_type | Cancer primaire |
| 3-gene_classifier_subtype | Sous-types de classificateur à trois gènes, il prend une valeur parmi 'ER-/HER2-', 'ER+/HER2- haute prolifération', nan, 'ER+/HER2- faible prolifération','HER2+' |
| tumor_size | Taille de la tumeur mesurée par des techniques d'imagerie |
| tumor_stage | prend une valeur de 0 à 4 |
| death_from_cancer | Si le décès du patient était dû à un cancer ou non |

### 1.2. Les données génétiques

L'ensemble de données génétiques contient le nombre de mutations pour 175 gènes et le score z des niveaux d'ARNm pour 331 gènes qui est représenté par l'équation suivante :

$$z = \frac{l'expression\ dans\ l'échantillon\ tumoral - l'expression\ moyenne\ dans\ l'échantillon\ normal}{l'écarttype\ d'expressiondans\ l'échantillon\ de\ référence}$$

Cette mesure est utile pour déterminer si un gène est régulé à la hausse ou à la baisse par rapport aux échantillons normaux ou à tous les autres échantillons de tumeurs.

### 1.3. Préparation des données

Dans cette partie nous présentons les étapes de préparation des données dans la figure suivante :

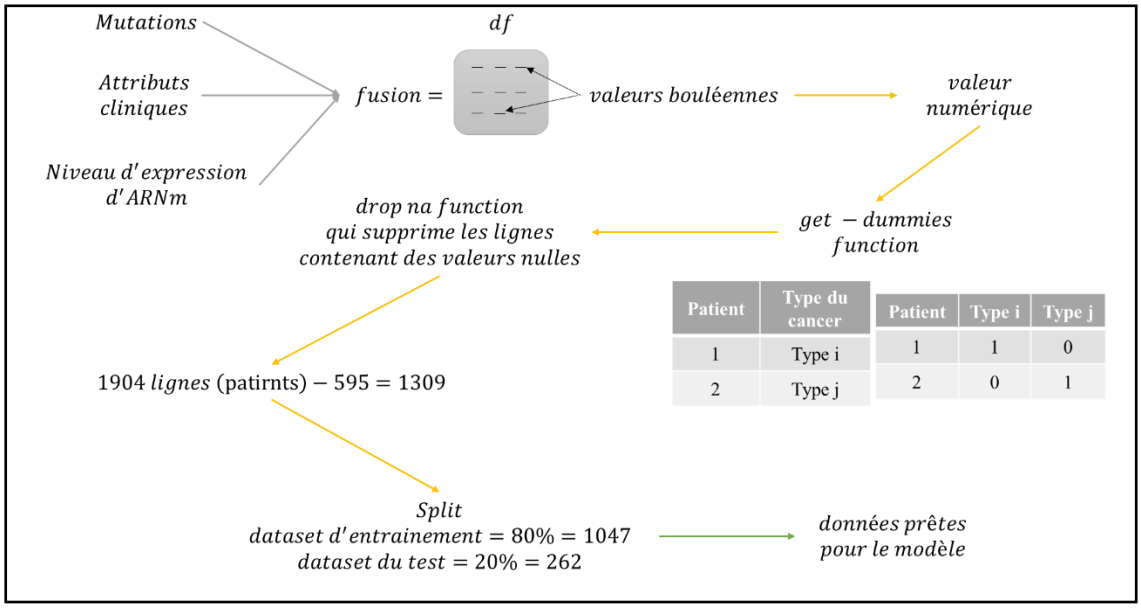



**Figure 12 :** Les étapes de préparation de données

## 2. Visualisation des données et statistiques

Dans cette partie, le code écrit nous informe sur nos données cliniques et génétiques, ce qui les rend plus clairs et nous permet de mieux les comprendre, comme :

- Visualisation des données cliniques
- Visualisation des données génétiques
- Distribution des deux classes de survie et de mort dans les colonnes d'attributs cliniques et génétiques
- Calcul de corrélation qui permet de savoir s'il existe un lien entre deux variables quantitatives ; si les valeurs des deux variables varient dans le même sens ou dans le sens contraire.
- Visualisation de l'expression génétique

## 3. Forward et Backward propagation

L'apprentissage des modèles DL se font en deux étapes : la première étant « Forward Propagation » ; sa finalité sera une prédiction quasiment aléatoire pour les premières itérations ou époch, en autre terme, le modèle fera des spéculations. La deuxième étape étant « Backward Propagation » ; sa finalité sera l'ajustement des paramètres et hyperparamètres du modèle pour minimiser l'erreur et rapprocher les valeurs prédites (les spéculations du modèle)

aux valeurs réelles. Nous présentons une figure simplifiée de l'ordre chronologique de ces deux étapes :

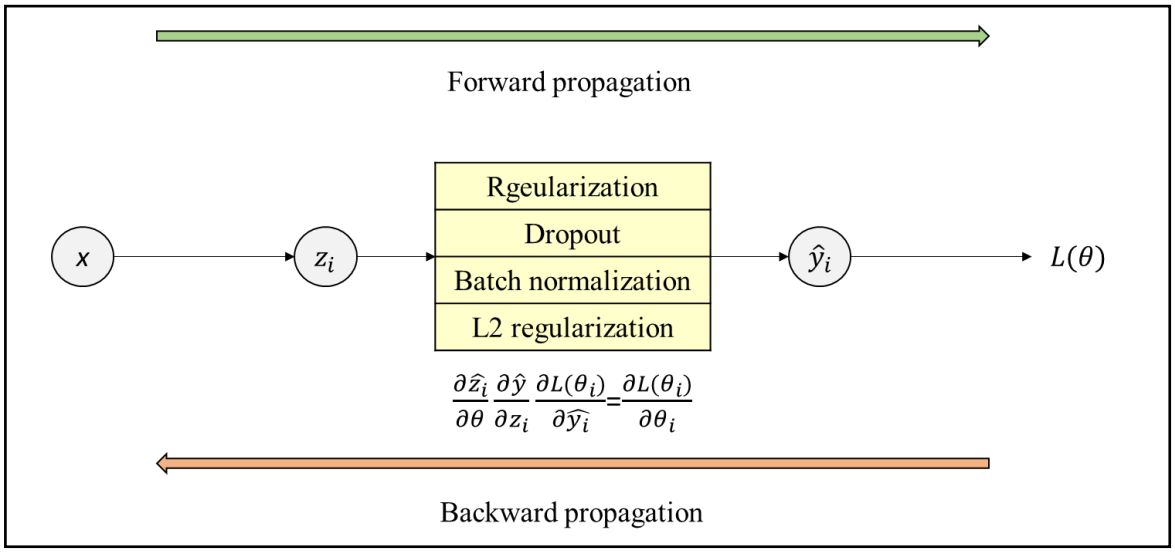



**Figure 13 :** Forward propagation et Backward propagation

## 3.1. Forward propagation

### 3.1.1. Architecture ANN choisie

Notre model représente un ANN contenant 6 couches. La première étant la couche de données d'entrée ayant 581 neurones qui reflète le nombre des attributs de notre dataset. La dernière étant les couches de sortie ayant un seul neurone celle de la classification finale. Et 4 couches cachées ayant chacune 1024, 512, 256 et 128 neurones respectivement.

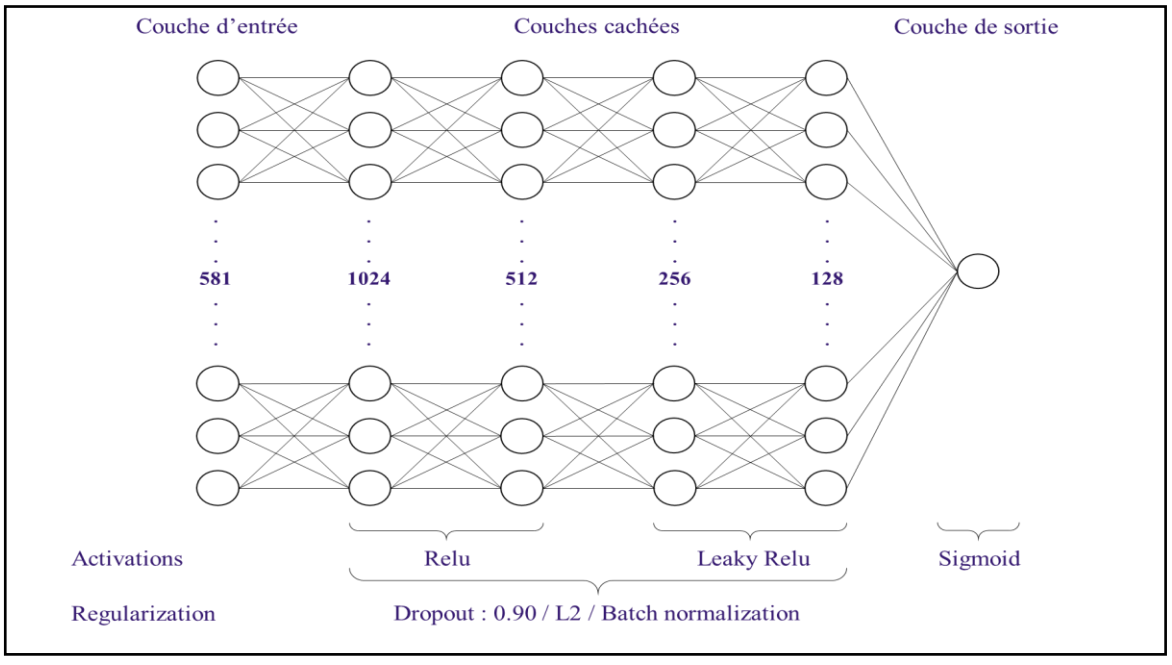



**Figure 14 :** Représentation de notre architecture

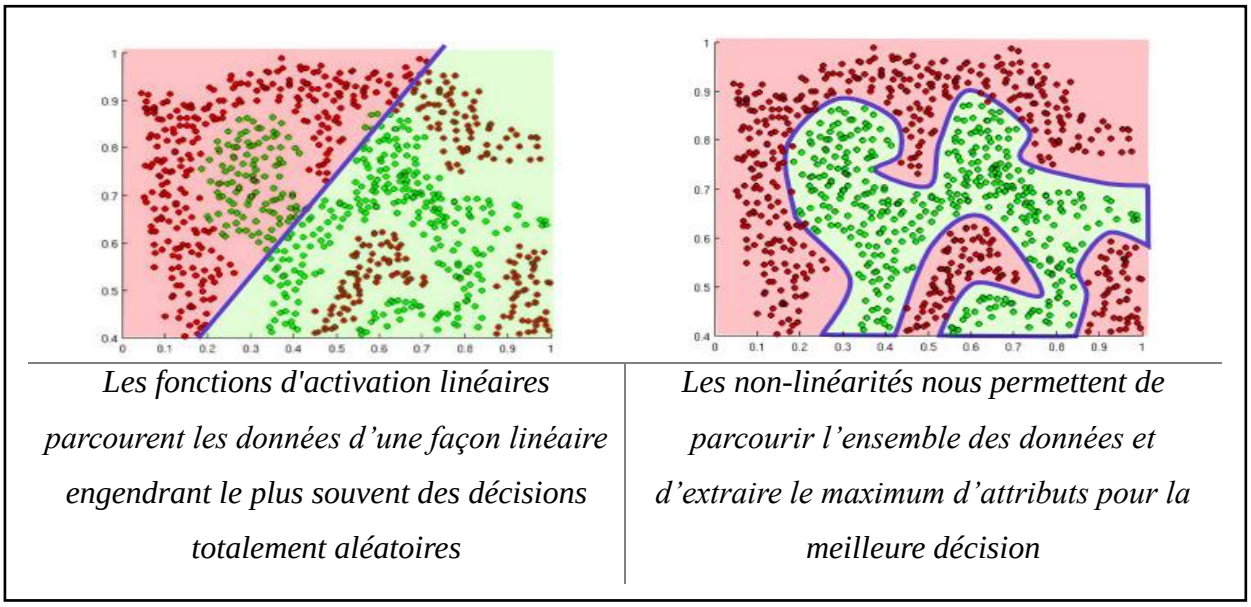
L'intégralité des neurones de notre architecture peuvent se traduire par la formule mathématique suivante :

$$Z^t = W^t\, a^{t-1} + B^t$$

Où $a^{t-1}$ est la valeur des neurones d'entrée dans le cas de la première couche. Pour les couches cachées, elle représente le résultat des couches précédentes après activation. $W^t$ Et $B^t$ représentent le poids qui sera calculé durant l'entrainement de notre modèle et de par sa valeur va définir les attributs dominants de notre dataset pour chaque patient, et le Bias est une valeur aussi calculée durant l'entrainement qui va déterminer le seuil de la somme du neurone pour qu'il soit activé, et $Z^t$ représente la somme du neurone en question qui va passer par une fonction d'activation non-linéaire.

### 3.1.2. Les fonctions d'activations

Chaque neurone de chaque couche constitue une équation linéaire, bien sur notre objectif et d'augmenter l'efficacité de notre model et pour se faire ce dernier doit généraliser ses prédictions sur l'intégralité des données d'entrainement et de test, nous devons donc rajouter une fonction non-linéaire à chaque neurone, leur utilité est mieux représentée dans le graphe suivant :



| *Les fonctions d'activation linéaires parcourent les données d'une façon linéaire engendrant le plus souvent des décisions totalement aléatoires* | *Les non-linéarités nous permettent de parcourir l'ensemble des données et d'extraire le maximum d'attributs pour la meilleure décision* |
|---|---|

**Figure 15 :** Impact des fonctions linéaires et non linéaires sur l'apprentissage de modèle

Les fonctions non linéaires utilisées dans notre model sont les suivantes :

Couche 1 et 2 : Relu ( rectified linear unit)

Couche 3 et 4 : LeakyReLU

Couche de sortie : Sigmoid

#### 3.1.2.1. Relu

ReLU est une fonction d'activation qui a de solides fondements biologiques et mathématiques. En 2011, il a été démontré qu'elle améliore grandement l'efficacité de l'entraînement et augmente la vitesse de ce dernier, elle fonctionne en mettons un seuil ( 0 ) au valeurs de $Z^t$.

Son équation est la suivante : f (x) = max(0,x)

Où x est représentatif de $Z^t$, si elle est inférieure à 0 elle recevra la valeur de 0, si elle est supérieure elle restera inchangée et f(x) est représentatif de $a^t$.

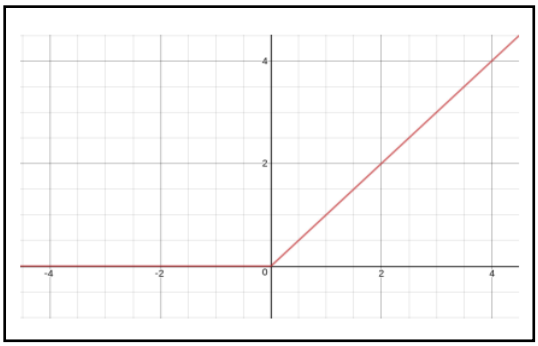



**Figure 16 :** représentation de la fonction d'activation de l'unité linéaire rectifiée (ReLU) qui produit 0 en sortie lorsque x < 0, puis produit une linéaire avec une pente de 1 lorsque x>0

#### 3.1.2.2. Leaky ReLU

Le $a_i$ est un paramètre fixe compris entre [1, +∞], la valeur la plus utilisé étant 0.01. LeakyReLU a presque le même impact que ReLU sur la propagation de l'information à l'exception que les valeurs négatives ne recevront pas un 0, mais une valeur minime extrêmement proche de ce dernier, la formule mathématique de cette dernière est la suivante :

$$f(x) = \begin{cases} x, & si\ x \geq 0 \\ \frac{x}{a_i}, & si\ x < 0 \end{cases}$$

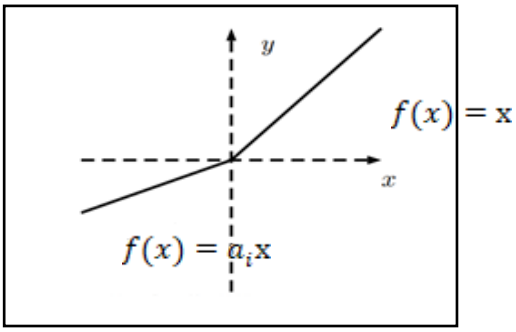



**Figure 17 :** représentation graphique de Leaky ReLU

#### 3.1.2.3. Sigmoid

La fonction Sigmoid représente la finalité de notre architecture, elle suit le dernier neurone et elle est utilisé pour les prédictions concernant la classification binaire, le résultat après cette fonction sera compris entre [0,1], plus la valeur de Z est élevée plus le résultat de la non linéarité sera proche de 1, dans le cas inverse elle sera proche de 0, la formule mathématique de cette dernière est la suivante :

$$f(x) = \left(\frac{1}{1 + exp^{-x}}\right)$$

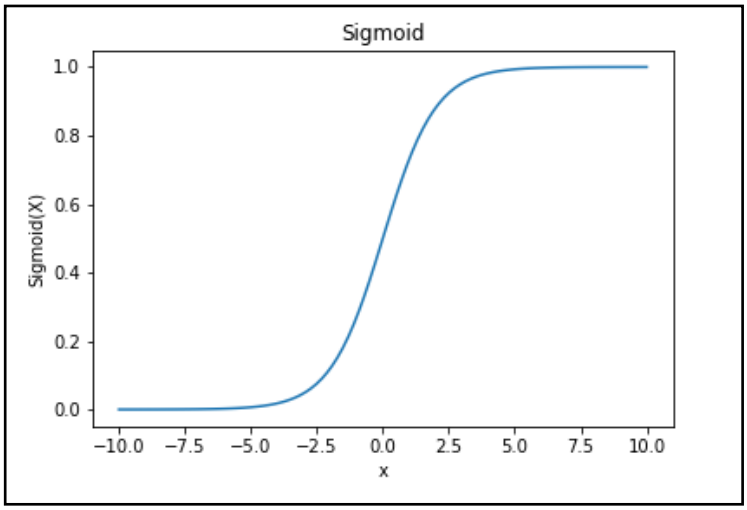



**Figure 18 :** représentation de la fonction sigmoid qui insère les valeurs reçues dans une intervalle de [0 ; 1]

### 3.1.3. Calcul de Loss

La valeur loss reflète la marge d'erreur de notre model et l'objectif de l'entrainement est de minimiser au maximum cette valeur dans les données d'entrainement et surtout dans les données de test ce qui va théoriquement maximiser la précision de notre model.

La fonction que nous avons choisi pour le calcul de de la valeur loss est la fonction « Binary cross entropy function », l'équation de cette dernière est la suivante :

$$Loss = -\frac{1}{nombre\ des\ prédictions} \sum_{i=1}^{nombres\ des\ prédictions} y_i\ .log\ \hat{y}_i + (1 - y_i)\ .\log(1 - \hat{y}_i)$$

Où $y_i$ est la valeur réelle représentant dans notre problème le décès ou la survie du patient donc sera soit 1 ou 0, et $\hat{y}_i$ est la valeur prédite par le modèle et sera comprise entre [0;1].

### 3.1.4. Régularisation

#### 3.1.4.1. L2 Régularisation

L2 est une pénalité qui minimise la valeur des poids de chaque neurone de notre modèle (plus lambda $\lambda$ est élevée plus le poids w est minimisé) ce qui va indirectement affecter la minimisation de la valeur de Loss et rendre le model moins complexe et donc relativement accéléré l'entrainement (Twan, 2017), la formule de cette dernière est la suivante :

$$L_\lambda(w) = L + \lambda \|w\|^{[t]2}$$

#### 3.1.4.2. Dropout

Le dropout étant la méthode de régularisation la plus simple disponible, car mathématiquement elle ne fait que multiplier le résultat des neurones choisie au hasard par 0, elle est néanmoins la méthode de régularisation la plus efficace qui soit car en éliminant plusieurs neurones à chaque entrainement, elle force les neurone restantes à être plus robustes, à apprendre par elles-mêmes, sans dépendre du résultat des neurones les précédents, dans ce contexte nous pouvant dérivé à partir de notre model plusieurs sous-modèles qui vont chacun s'entrainé sur un nombre limité de mini-batch spécifique.

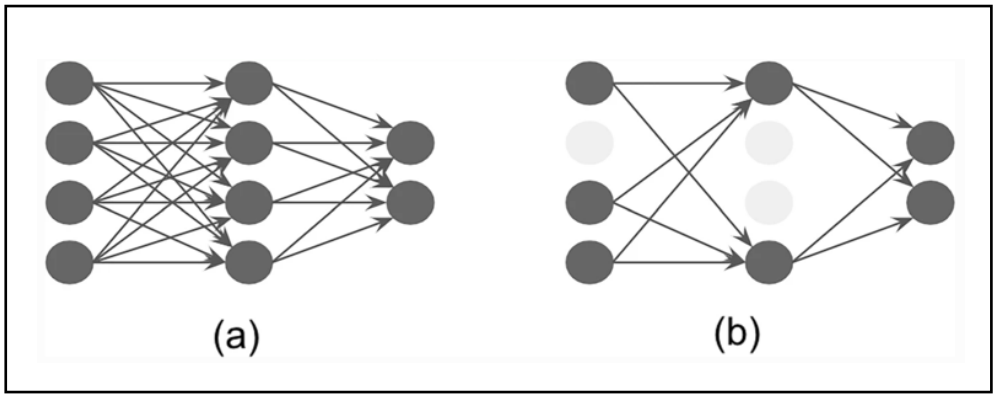



**Figure 19 :** Figure montrant un ANN standard (a) et un ANN avec une probabilité de dropout de 0.25 dans la première couche et 0.50 dans la deuxième (b)

#### 3.1.4.3. BATCH normalization

L'approche « batch normalization » nous permet d'insérer l'intégralité des résultats des couches concernées dans un même intervalle, ce qui va engendrer la minimisation du temps d'entrainement mais aussi réduire le bruit lors de ce dernier et agir entant que régulariseur pour réduire le sur apprentissage (Garbin *et al* ; 2020).

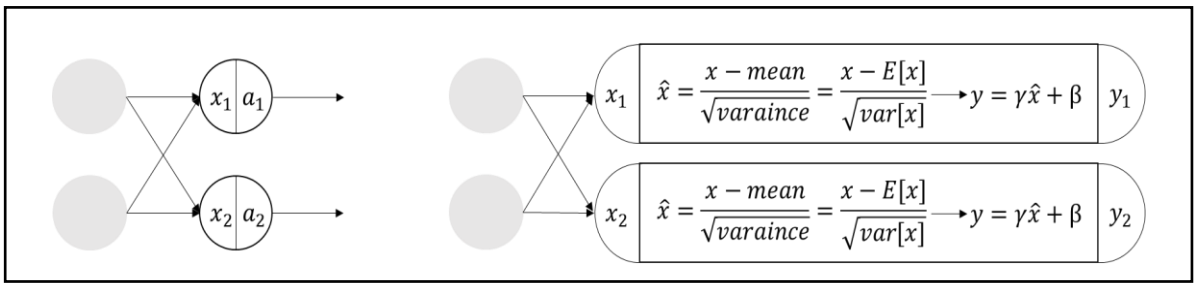



**Figure 20 :** Figure représentant la différence entre une couche standard et une couche avec la régularisation par « batch norm »

Où le mean et variance et BETA (β) et gamma (γ) sont tous de nouveaux paramètres entrainables qui vont s'ajuster après chaque epoch pour la normalisation la plus optimal (loffe *et al*., 2015).

#### 3.1.4.4. Autres méthodes de régularisation utilisées

A part les approches mentionnées auparavant et les nouveaux hyperparamètrex qu'on a introduit, nous pouvons considérer l'intégralité de la structure de notre architecture comme une régularisation, en effet, le nombre de couches, nombre de neurones par couche, la taille des mini-batch, nombre de epochs et optimiseur sont tous indirectement des hyperparamètres à choisir et tout a était choisis en connaissance de cause, la combinaison de tous ces

hyperparametres a rapporté les meilleurs résultats comparés aux autres combinaisons essayées.

**Tableau 2 :** le sommaire du modèle

| Couche (type) | Output shape | Paramètre |
|---|---|---|
| Dense (dense) | (None, 1024) | 595968 |
| Batch_normalization (BatchNormalization) | (None, 1024) | 4096 |
| Dropout (dropout) | (None, 1024) | 0 |
| Dense 1 (dense) | (None, 512) | 524800 |
| Batch_normalization_1 (BatchNormalization) | (None, 512) | 2048 |
| Dropout 1 (dropout) | (None, 512) | 0 |
| Dense 2 (dense) | (None, 256) | 131328 |
| Dropout 2 (dropout) | (None, 256) | 0 |
| Dense 3 (dense) | (None, 128) | 32896 |
| Batch_normalization_2 (BatchNormalization) | (None, 128) | 512 |
| Dropout 3 (dropout) | (None, 128) | 0 |
| Dense 4 (dense) | (None, 1) | 129 |
| Total des paramètres : 1, 291, 777<br>Paramètres entrainable : 1, 288, 449<br>Paramètres non-entrainables : 3, 328 | | |

### 3.2. Backward propagation

Après la forward propagation la machine devra ajuster l'intégralité de ses paramètres et hyper paramètres, cela se fera par la descente des gradients selon l'algorithme suivant :

ALGORITHME

1-initialiser le vecteur de paramètre θ au Hazard

2- tant que θ n'est pas convergé faire :

3- Calculer le gradient $\frac{\partial L(w)}{\partial w}$

4- mettre à jour le vecteur θ ou θ ← θ - $\alpha\frac{\partial L(\theta)}{\partial \theta}$

5-retourner θ

Où θ représente un des paramètres ou hyperparamètres ajustés, α représente le « learning rate » ou ‘’pas’’ qui va déterminer le rythme par lequel la descente des gradients sera effectuée.

### 3.2.1. Calcul du gradient $\frac{\partial L(\theta)}{\partial \theta}$

Selon la règle de la chaine :

$$\frac{\partial L(\theta)}{\partial \theta} = \frac{\partial L(w)}{\partial a}\frac{\partial a}{\partial z}\frac{\partial z}{\partial \theta}$$

Où le dérivé de la fonction « binary cross entropy » est représenté comme suit :

$$\frac{\partial L(\theta)}{\partial a} = -\frac{a - y}{a(1-a)}$$

Si nous prenons en compte la fonction d’activation Sigmoid, son dérivé est le suivant :

$$\frac{\partial a}{\partial z} = a(1-a)$$

Et pour notre équation de base $Z^t = W^t a^{t-1} + B^t$, son dérivé est le suivant :

$$\frac{\partial z}{\partial \theta} = \mathrm{a}$$

En plus des équations précédentes, nous avons le dérivé de la fonction ReLU g1 La dérivé de la fonction LeakyReLU g2:

$$\frac{\partial g1}{\partial z} = f(x) = \{0 \text{ si } x < 0 \text{ et } 1 \text{ si } x > 0$$

$$\frac{\partial g2}{\partial z} = f(x) = \{0.01 \text{ si } x < 0 \text{ et } 1 \text{ si } x > 0$$

Bien sûr l’équation du gradient est juste un exemple simplifié et explicatif, pour notre architecture lors de la backpropagation, chacun des 1.288.449 paramètres sera ajuster, une équation montrant le calcul de la descente des gradients en prenant en compte les dérivés de nos paramètres principaux ressemblera à ça :

$$\frac{\partial L(\theta)}{\partial \theta^t} = \frac{\partial L(\theta)}{\partial g_1^t} \frac{\partial g_1^t}{\partial Z^t} \frac{\partial Z^t}{\partial g_2^{t-1}} \frac{\partial g_2^{t-1}}{\partial Z^{t-1}} \frac{\partial Z^{t-1}}{\partial g_2^{t-2}} \frac{\partial g_2^{t-2}}{\partial Z^{t-2}} \frac{\partial Z_2^{t-2}}{\partial g_3^{t-3}} \frac{\partial g_3^{t-3}}{D\partial^{t-3}} \frac{\partial Z^{t-3}}{\partial g_3^{t-4}} \frac{\partial g_3^{t-4}}{\partial Z^{t-4}} \frac{\partial Z^{t-4}}{\partial \theta^t}$$

g1=Sigmoid /g2=ReLU /g3=LeakyReLU (Yash Garg, 2022).

### 3.2.2. Utilisation de ADAM

ADAM ou adaptive moment estimation est un algorithme extrêmement efficace en termes de calcul, nécessite peu de mémoire vive, et est très bien adapté pour les problèmes contenant une large quantité de données ou de paramètres. La méthode est également appropriée pour les objectifs non stationnaires et les problèmes de bruit au quels fait face la descente de gradients. Les hyper-paramètres ont des interprétations intuitives et nécessitent généralement aucun réglage, cet algorithme combine les avantages de RMSProp (Tieleman, Hinton, 2012), et Adagrad (Duchi *et al*., 2011)

L'algorithme :

1-Requière : $\alpha$: learning rate (rythme de l'apprentissage)

2-Requière: $\beta_1, \beta_2 \in [0, 1)$: hyper paramètre appelé exponential decay

3-Requière: $f(\theta)$: fonction stochastique avec le paramètre $\theta$

4-Requière: $\theta_0$: vecteur de paramètre initial

5- $m_0 \leftarrow 0$ (initialiser $1^{st}$ vecteur)

6- $v_0 \leftarrow 0$ (Initialiser $2^{nd}$ vecteur)

7- $t \leftarrow 0$ (Initialiser le moment t)

8- Tant que $\theta_t$ n'est pas convergé faire

9- $t \leftarrow t + 1$

10- $g_t \leftarrow \nabla_\theta f_t(\theta_t - 1)$ ( gradients de la fonction objective au moment t)

11- $m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$

12- $v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$

13- $\hat{m}_t \leftarrow m_t / (1 - \beta_1^t)$

14- $\hat{V}_t \leftarrow vt/(1 - \beta_2^t)$

15- $\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \hat{m}_t / (\sqrt{\hat{V}_t} + \varepsilon)$ (mettre à jour le nouveau paramètre)

16- Fin tant que

17-Retourner $\theta_t$. (Kingma *et al*., 2014).

Les paramètres d'ADAM :

Learning rate : 0.001 paramètre fixe

$\beta_1$ : 0.9 dans t=0/ paramètre entrainable dans t>0

$\beta_2$ : 0.999 dans t=0/ paramètre entrainable dans t>0

$\beta_1^t$ : $\beta_1$ puissance t

$\beta_2^t$ : $\beta_2$ puissance t

$\varepsilon$ : $10^{-8}$ paramètre fixe

### 3.3. Matrice d'estimation de la qualité du modèle :

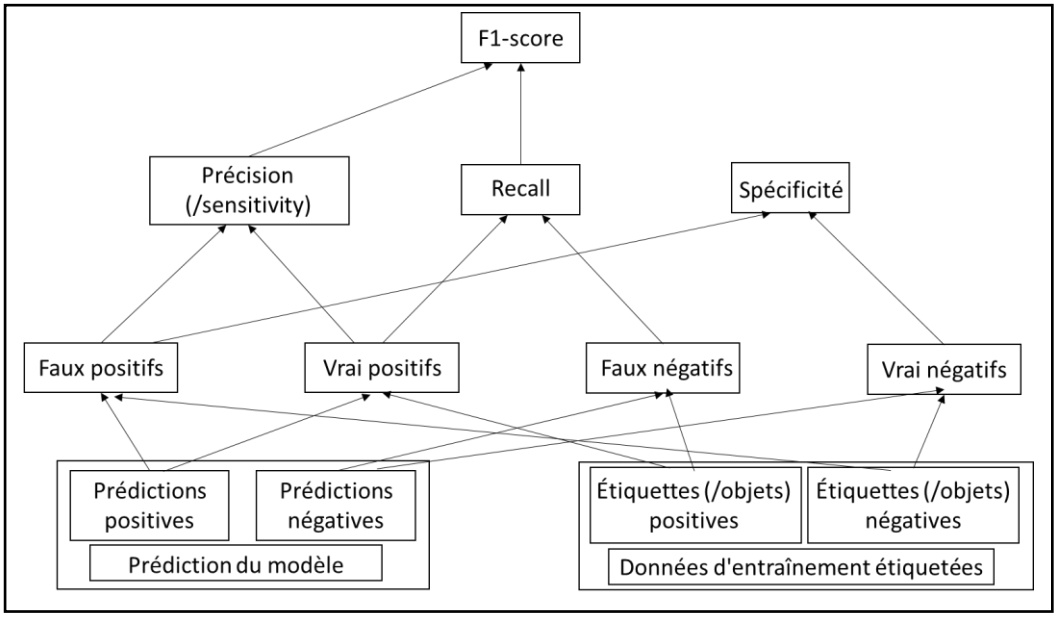



**Figure 21 :** représentation de la hiérarchie des matrices calculées

### 3.3.1. Accuracy

C'est la matrice de base lors de l'évaluation du model, elle représente le pourcentage des prédictions correctes par rapport au nombre total des cas, sa formule est la suivante :

$$\frac{VP + VN}{VP + VN + FP + FN} = \frac{predictions\ correctes}{toutes\ les\ predictions}$$

### 3.3.2. Précision

La précision vise à mesurer le nombre le pourcentage des prédictions positives correctes, la différence entre celle-ci et le Recall est que le Recall prend en compte les faux négatifs au lieu des faux positifs de la précision, sa formule est la suivante :

$$\frac{VP}{VP + FP} = \frac{predictions\ positives\ correctes}{total\ des\ predictions\ postive}$$

### 3.3.3. Spécificité :

Elle vise à mesurer le pourcentage total des prédictions négatives correctes prédites par notre modèle, sa formule est :

$$\frac{VN}{VN + VP} = \frac{predictions\ negatives\ correctes}{total\ des\ cas\ negatif\ du\ dataset}$$

### 3.3.4. Recall / Sensitivité

Le « Recall » vise à mesurer le nombre de cas positifs que notre model a correctement prédits, sur tous les cas positifs dans les données. Il est parfois aussi appelé Sensibilité. La formule pour cela est :

$$\frac{VP}{VP + FN} = \frac{predictions\ positives\ correctes}{total\ des\ cas\ positif\ du\ dataset}$$

### 3.3.5. F1-SCORE

Le F1-Score est un moyen de mesure combinant la précision et le recall, il calcul un ratio équilibré et harmonieux entre les deux, après l'accuracy cette approche de mesure est la mieux placée pour évaluer l'efficacité de notre model, sa formule est la suivante :

$$2 * \frac{Precision * recall}{Precision + recall}$$

## 4. Implémentation

L'étude statistique, le traitement de données, l'implémentation du modèle et l'intégralité des visualisations ont été implémentés en python (version 3.8), en utilisant l'environnement de développement « Spyder ».

Les modules utilisés sont les suivants :

- Numpy
- Scipy
- Matplotlib
- Seaborn
- Sklearn
- Tensorflow
- Keras

# CHAPITRE III :

# Résultats et Discussion

## 1. Résultats de la visualisation des données et statistiques

Nous présentons dans cette partie les résultats de visualisation de données et les statistiques.

### 1.1. Visualisation des données cliniques

Les deux classes cibles (la classe des survivants et la classe des morts) sont distribuées dans les colonnes des données cliniques numériques et représentées dans la figure ci-dessous, sous forme de représentations graphiques de la densité des survivants (en vert) par rapport à l'âge du patient au moment du diagnostic.

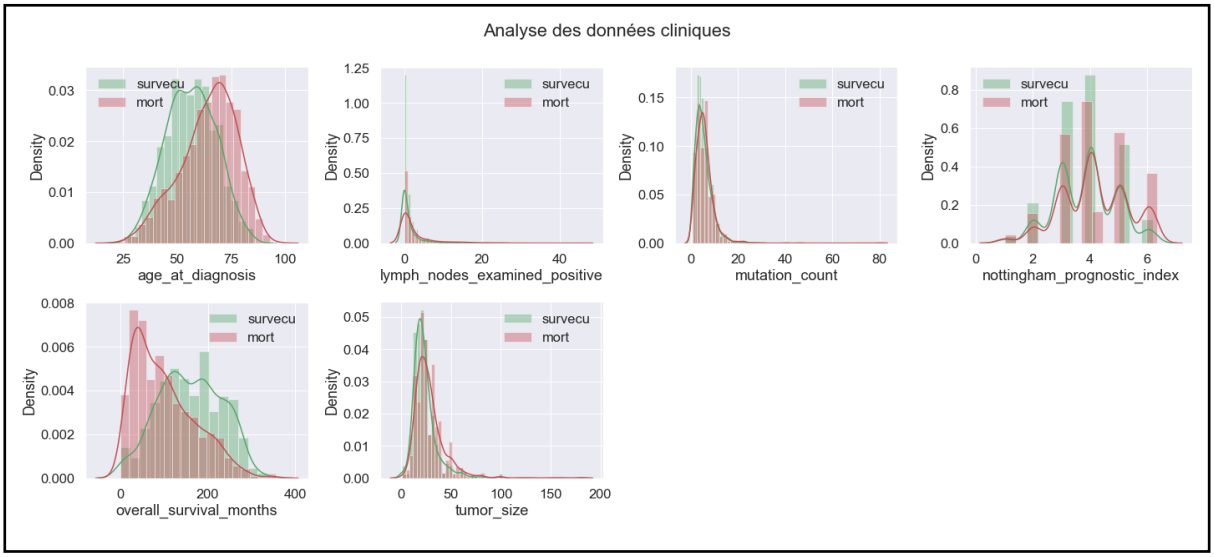



**Figure 22 :** La distribution des deux classes cibles (survivants et morts) dans les colonnes cliniques numériques de données

La figure ci-dessous représente la dispersion de la taille de la tumeur où la couleur rouge représente les morts et la couleur bleue représente les survivants.

On constate qu'à mesure que le stade de la tumeur augmente, la taille de la tumeur augmente également. De plus, si les stades tumoraux sont inférieurs, la probabilité de survie est plus élevée que lorsque le patient atteint le quatrième stade.

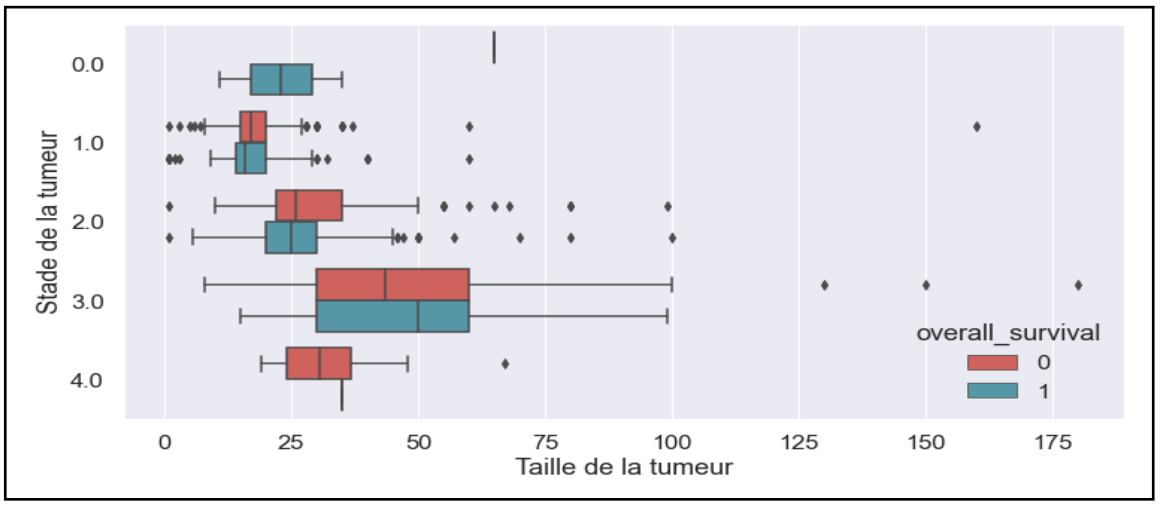



**Figure 23 :** Représentation graphique permettant de visualiser la dispersion de taille de la tumeur

Dans la figure suivante, on trouve que la médiane de la taille de la tumeur et du nombre de ganglions lymphatiques positifs est plus faible dans la classe des survivants que dans la classe des morts.

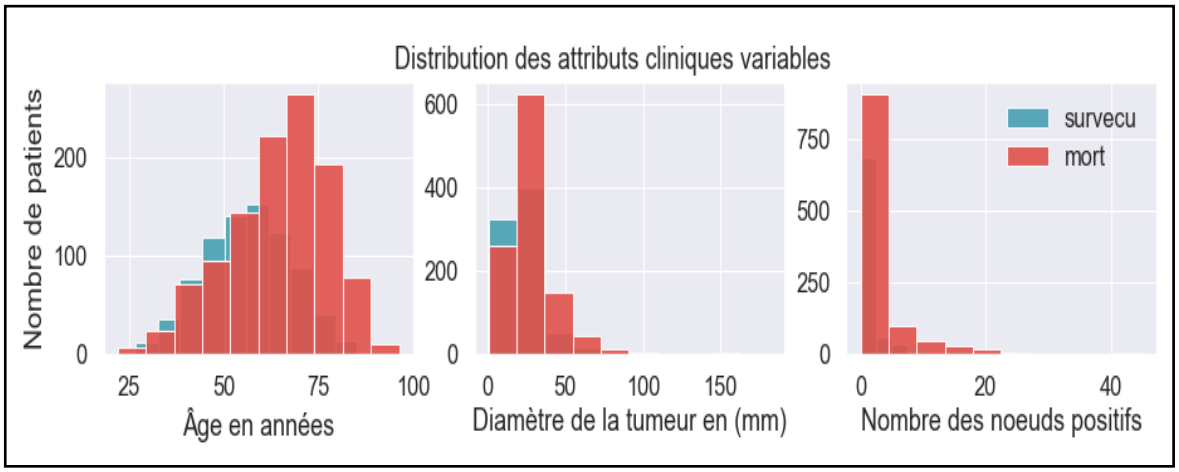



**Figure 24 :** Représentation graphique permettant de visualiser les deux classes dans les attributs cliniques variables (âge, taille de la tumeur et nombre de ganglions positifs)

L'histogramme ci-dessous représente la distribution des deux classes (survivants et morts) sur les trois colonnes du traitement (chimiothérapie, hormonothérapie et radiothérapie).

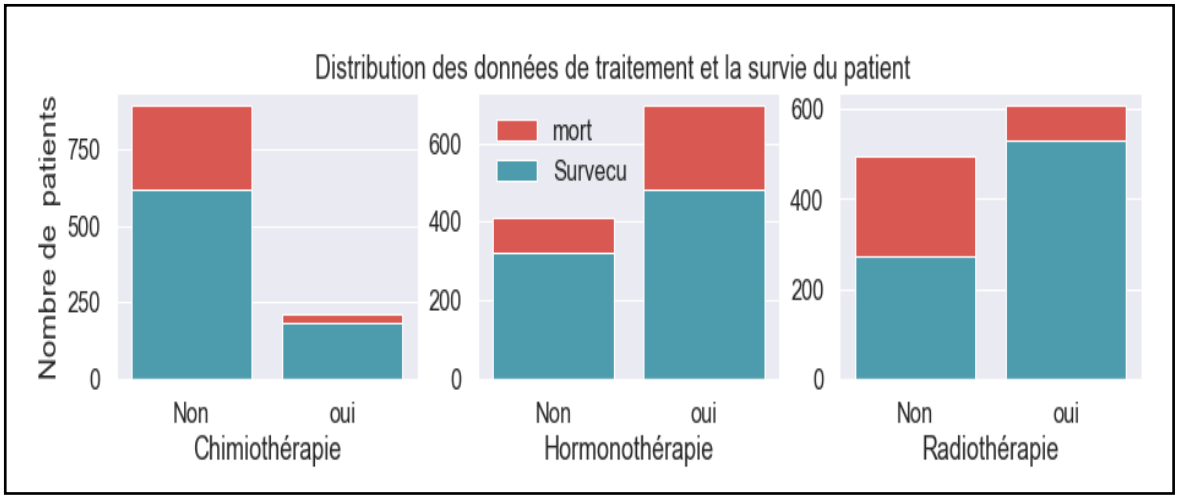



**Figure 25 :** Visualisation des données de traitement du cancer et la survie des patients

Dans la figure suivante, qui représente la corrélation entre les attributs cliniques, nous pouvons voir qu'il existe une forte corrélation entre certaines des colonnes.

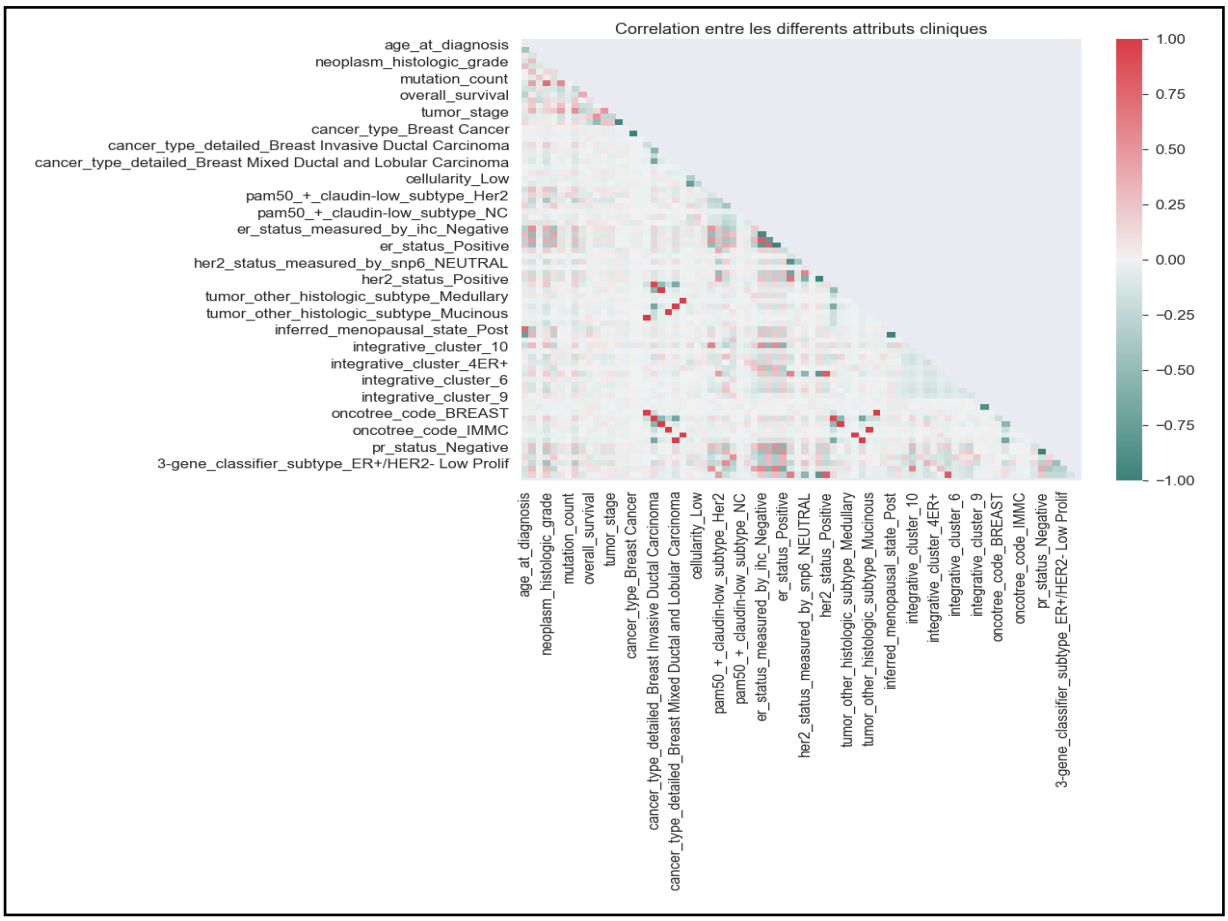



**Figure 26 :** Représentation graphique permettant de visualiser la corrélation entre les attributs cliniques

### 1.2. Visualisation des données génétiques

Dans la représentation ci-dessous, nous pouvons tirer les informations suivantes :

- Valeur maximal possible dans les données génétiques : 18.6351
- Valeur minimal possible dans les données génétiques : -6.4387

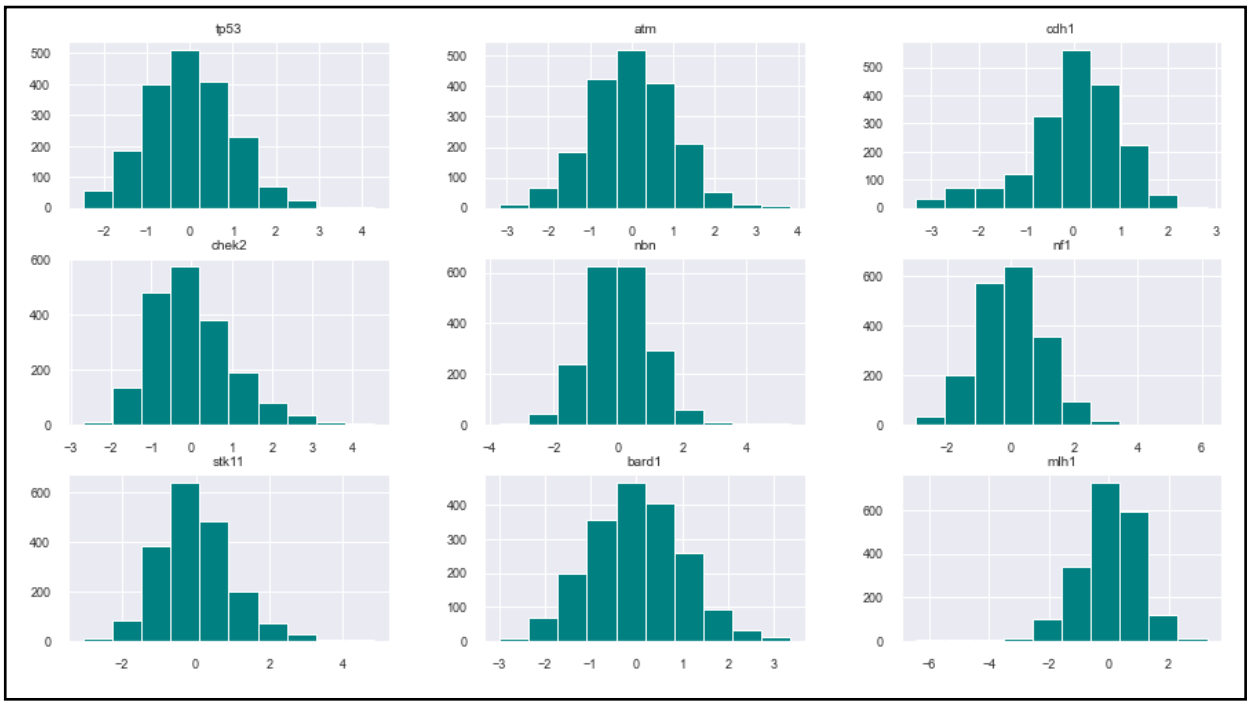



**Figure 21 :** Histogramme représentant l'expression génétique de quelques gènes

La figure représente la distribution des données dans les deux classes de survie est très similaire avec peu de valeurs aberrantes dans certains gènes.

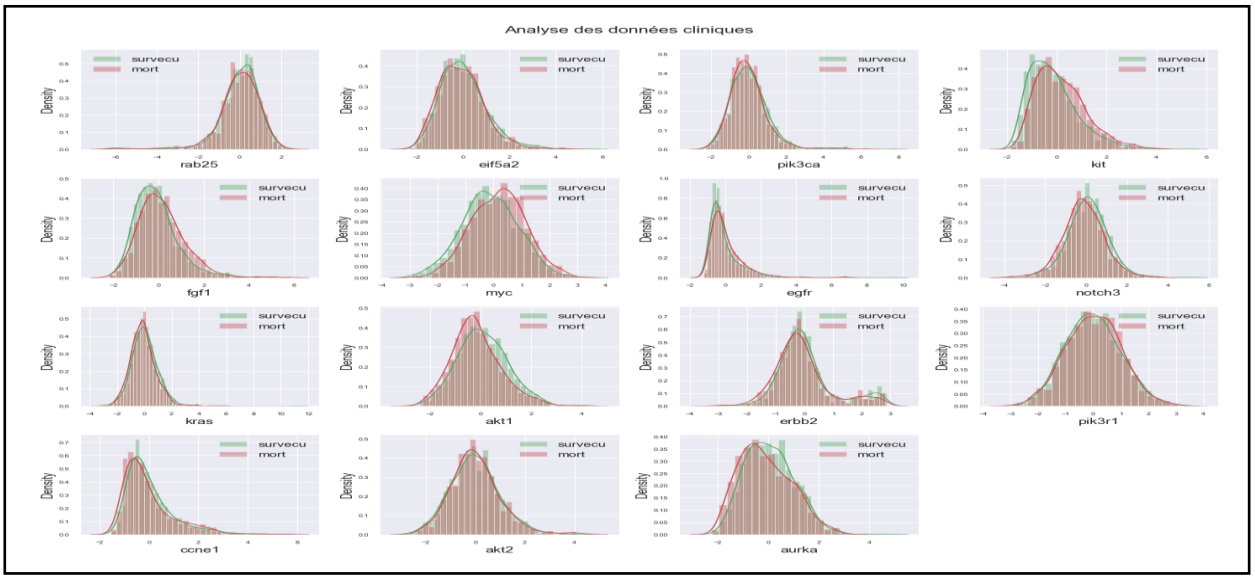



**Figure 28 :** Distribution des données des deux classes dans certains gènes

Dans la figure ci-dessous, dans l'histogramme 1, la corrélation entre notre cible et les caractéristiques génétiques montre que la plupart des caractéristiques ne sont pas réellement corrélées.

Dans l'histogramme 2, aucune corrélation entre la survie et les mutations, car nous avons changé la mutation en 0 et 1 au lieu de 0 s'il n'y a pas de mutation et le type de mutation s'il y a une mutation.

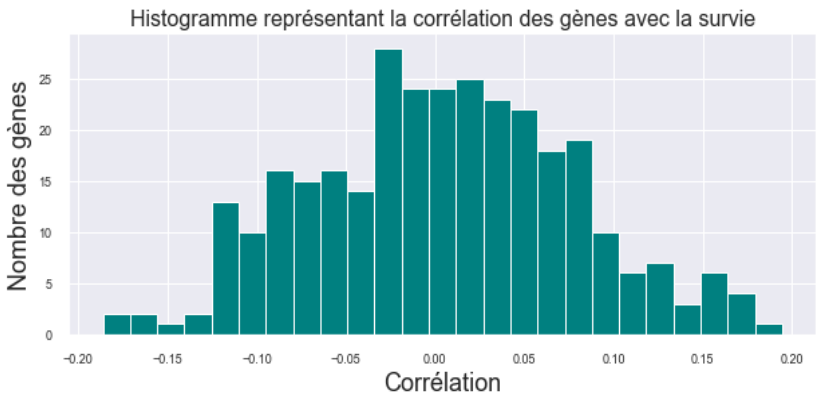
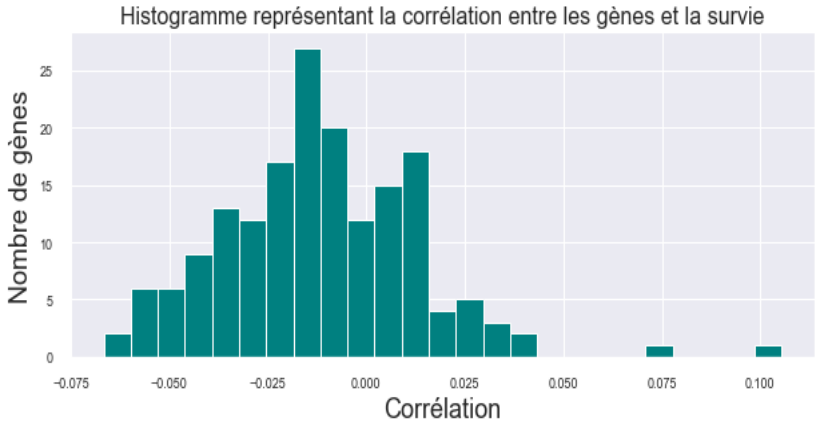
| | |
|---|---|
|  | Corrélation maximale :<br>0.194<br>Corrélation minimale :<br>-0.186<br>Corrélation moyenne :<br>0.004 |
|  | Corrélation maximale :<br>0.105<br>Corrélation minimale :<br>-0.067<br>Corrélation moyenne :<br>-0.012 |

**Figure 29 :** Histogrammes représentant la corrélation entre différents gènes et la survie

### 1.3. Résultats statistiques

D'après la lecture et la visualisation des différentes données, nous déduisons que :

- Nombre de patients qui n'ont eu aucun traitement : 289
- Proportion de survie dans ce groupe : 0.381
- Proportion minimale de survie : 0.421
- Âge moyen : 61.087

- Stade de la tumeur le plus fréquent : 2
- Type histopathologique le plus fréquent : 3
- Diamètre moyen de tumeur : 26.239
- Probabilité de survie : 0.421

## 2. Résultat du modèle

Dans cette partie nous présentons nos représentations graphiques et le tableau des résultats de nos matrices utilisées :

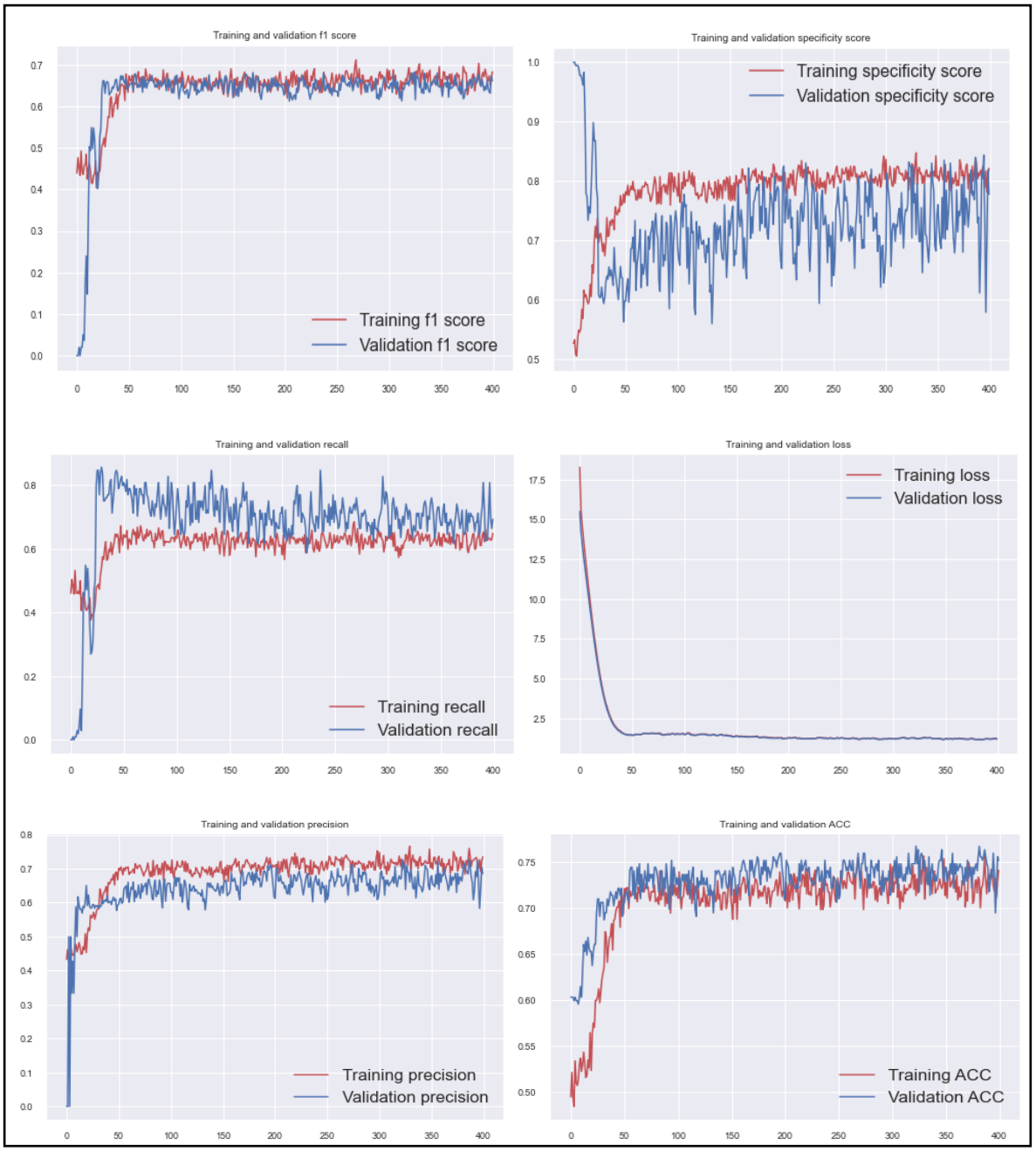



**Figure 30 :** Graphes représentant les 6 matrices utilisées

**Tableau 3 :** Les valeurs des paramètres utilisés

| Paramètres | Valeurs |
|---|---|
| Loss | 1.2123 |
| Accuracy | 0.7412 |
| Recall | 0.6479 |
| Precision | 0.7353 |
| Get_fel | 0.6827 |
| Specificity | 0.8212 |
| Val_loss | 1.1861 |
| Val_accuracy | 0.7519 |
| Val_recall | 0.6923 |
| Val_precision | 0.6857 |
| Val_get_fl | 0.6618 |
| Val_specificity | 0.7776 |

## 3. Discussion des résultats obtenus

Les graphes précédents montrent nos 6 matrices de mesures qui sont : loss, accuracy, precision, specificity, recall, F1-Scrore, après un entrainement de 400 epochs. Nous mettrons en lumière le fait que le model Deepredictor dans le dernier epoch montre une précision (accuracy) final de %74,12 et %75,19, une valeur Loss finale de 1.2123 et 1.1861, et un score f1 final de %68,27 et %66,18 sur les données d'entrainement et de test respectivement. Les valeurs de ces trois dernières matrices qui sont considérées dans la littérature comme les matrices de référence pour l'estimations sont très encourageantes et montre que notre architecture et les méthodes de régularisation employées, nous ont permis de faire face au sur-apprentissage qui est un réel challenge. Quand nous utilisons ces types de données avec leur nombre d'attributs relativement élevés par rapport au nombre réduit de patients aux données complètes, les graphes montrent que nous avons trouvé un parfait équilibre entre le sous

apprentissage qui reflète la capacité du modèle vis-à-vis des données d'entraînement, et le sur-apprentissage qui reflète son habitude à rester général sur des données inconnues ou autrement dit les données de test. Le bruit que l'on peut très facilement distinguer sur l'intégralité des graphes en est une autre preuve. Néanmoins ce sont là les meilleurs résultats qui reflètent la meilleure combinaison de paramètres que nous avons pu implémenter. Théoriquement, nous avons atteint le potentiel maximal des réseaux de neurones denses vis-à-vis du problème donné et qui est la prédiction de la mortalité d'un patient. Le seul hyper paramètre restant qui peut être ajusté pour l'amélioration du modèle est l'augmentation de la quantité de données et l'amélioration de ces derniers.

# Conclusion

# Et Perspectives

**Conclusion**

L'impact et le rôle que joue l'IA et plus précisément l'apprentissage profond dans les domaines de la médecine de précision et de l'ingénierie médicale ne cesse de prendre de l'importance. Cet impact bien sûr se traduit sur le terrain par une minimisation drastique des coûts de traitement des patients, l'amélioration de la longévité de ces derniers, les diagnostics précoces, la prédiction en termes d'efficacité des traitements...etc.

Nous avons présenté Deepredictor, un modèle de DNN capable de prédire avec une précision 75.19 % le taux de mortalité d'un patient en prenant en compte une multitude d'attributs qu'ils soient génétiques ou cliniques. Nous considérons que notre modèle peut dès à présent servir de fondement extrêmement fiable pour des travaux futurs concernant ce dataset et précisément cette problématique.

Pour nos perspectives, la première étape serait de développer une architecture DNN qui aura pour but la prédiction automatisée de l'intégralité des attributs manquants dans le dataset. En effet, de 1980 patients, nous n'avons pu utiliser les données que de 1309 par faute de manque d'attributs. Ce qui est extrêmement pénalisant vu le nombre déjà très faible des cas.

Une autre approche qui pourrait théoriquement rapporter de meilleurs résultats est l'utilisation de réseaux de neurones convolutifs à deux dimensions sur nos attributs sous forme matricielle vu leur nombre relativement élevé. Les CNN ont fait leur preuve mainte fois et pourrait grandement améliorer l'efficacité de notre model. Mais cette approche est connue pour être extrêmement gourmande quand il est question de volume de données. Donc l'évolution de la collecte de données par l'initiative METABRIC et sera toujours un facteur crucial pour tout travail à venir.

L'intégration des données génomiques brutes propres à chaque patient au dataset pourrait faire la différence. Une concaténation d'une architecture de RNN traitant ces données avec une architecture CNN ou DNN traitant les données cliniques ainsi que les profils d'expression pourraient aboutir à un résultat extrêmement satisfaisant.

Bien sûr jusqu'à maintenant nous parlons de la problématique de la prédiction de la mortalité du patient. Nous pouvons réorienter nos modèles vers :

- la prédiction du temps de survie d'un patient.
- l'efficacité du traitement clinique de celui-ci.
- la prédiction du sous-type de cancer du sein auquel nous sommes confrontés.

# Références Bibliographiques

1. L'organisation mondiale de la santé, https://www.who.int/fr/news-room/fact-sheets/detail/breast-cancer , visité le (19 mai 2022).
2. CCM Benchmark, https://sante.journaldesfemmes.fr/fiches-anatomie-et-examens/2571039-sein-anatomie-examens-et-maladies/ , visité le (03 juin 2022).
3. Institut national du cancer, www.e-cancer.fr , visité le (20 mai 2022).
4. Zemmouri Nadjib, http://www.docteurzemmourinajib.ma/Mon_site/Cancer_du_sein.html , visité le (03 juin 2022).
5. Pereira, B., Chin, SF., Rueda, O. et al. The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. Nat Commun 7, 11479 (2016). https://doi.org/10.1038/ncomms11479
6. Cours de deep learning de MIT (Massachusetts Institute of technologie) sous le nom de « Introduction to deep learning MIT 6.S191) page 42.
7. Twan van Laarhoven. L2 Regularization versus Batch and Weight Normalization. https://doi.org/10.48550/arXiv.1706.05350
8. Antoine Georges, Joëlle Lacroix & Véronique Bouté. Mucinous carcinoma: A rare malignant breast tumour (2016), Pages 8-20, https://doi.org/10.1016/j.femme.2016.03.001.
9. Gautam K. Malhotra, Xiangshan Zhao, Hamid Band & Vimla Band Histological, molecular and functional subtypes of breast cancers, Cancer Biology & Therapy (2010), 10:10, 955-960, DOI: 10.4161/cbt.10.10.13879
10. Schnitt SJ. Classification and prognosis of invasive breast cancer: from morphology to molecular taxonomy. Mod Pathol. (2010) May;23 Suppl 2:S60-4. doi: 10.1038/modpathol.2010.33. PMID: 20436504.
11. Wu, Q., Nie, D.Y., Ba-allawi, W. *et al.* PRMT inhibition induces a viral mimicry response in triple-negative breast cancer. *Nat Chem Biol* (2022).
12. Pr Didier COWEN chef de service de radiothérapie Hôpital de la Timone, Hôpital Nord – Marseille (octobre 2017).
13. Govindarajan, R., Duraiyan, J., Kaliyappan, K., & Palanisamy, M. Microarray and its applications. Journal of pharmacy & bioallied sciences (2012), *4*(Suppl 2), S310–S312. https://doi.org/10.4103/0975-7406.100283
14. Yang Zhang, Siwa Chan, Vivian Youngjean Park, Kai-Ting Chang, Siddharth Mehta, Min Jung Kim, Freddie J. Combs, Peter Chang, Daniel Chow, Ritesh Parajuli, Rita S. Mehta, Chin-Yao Lin, Sou-Hsin Chien, Jeon-Hor Chen, Min-Ying

Su, Automatic Detection and Segmentation of Breast Cancer on MRI Using Mask R-CNN Trained on Non–Fat-Sat Images and Tested on Fat-Sat Images, Academic Radiology, 10.1016/j.acra.2020.12.001, **29**, (S135-S144), (2022).

15. Ganggayah, M.D., Taib, N.A., Har, Y.C. *et al.* Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC Med Inform Decis Mak* **19,** 48 (2019). https://doi.org/10.1186/s12911-019-0801-4
16. Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, Andrew H Beck Beth. Deep Learning for Identifying Metastatic Breast Cancer, Israel Deaconess Medical Center, Harvard Medical School, CSAIL, Massachusetts Institute of Technology (2016)
17. Lin Y, Zhang W, Cao H, Li G, Du W. Classifying Breast Cancer Subtypes Using Deep Neural Networks Based on Multi-Omics Data. Genes (Basel). (2020 Aug 4);11(8):888. doi: 10.3390/genes11080888. PMID: 32759821; PMCID: PMC7464481
18. Siri : https://www.apple.com/fr/siri/, visité le (15 juin 2022).
19. Alexa : https://developer.amazon.com/en-US/alexa, visité le (15 juin 2022).
20. cortana : https://www.microsoft.com/en-us/cortana, visité le (15 juin 2022).
21. google assistant : https://assistant.google.com/, visité le (15 juin 2022).
22. Riedl, Mark O. Human-centered artificial intelligence and machine learning. Human Behavior and Emerging Technologies (2019), e117–. doi:10.1002/hbe2.117
23. Jaime Vitola, Francesc Pozo, Diego A. Tibaduiza and Maribel Anaya. Distributed Piezoelectric Sensor System for DamageIdentification in Structures Subjected toTemperature Changes. Sensors (2017), 17, 1252; doi:10.3390/s17061252
24. P. Ongsulee, "Artificial intelligence, machine learning and deep learning," 2017 15th International Conference on ICT and Knowledge Engineering (ICT&KE), (2017), pp. 1-6, doi: 10.1109/ICTKE.2017.8259629.
25. P. P. Shinde and S. Shah, "A Review of Machine Learning and Deep Learning Applications," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), (2018), pp. 1-6, doi: 10.1109/ICCUBEA.2018.8697857.
26. Rina Dechter. Learning While Searching In Constraint-Satisfaction-Problems. Ann Math (1986).

27. Sarker, I.H. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. SN COMPUT. SCI. 2, 420 (2021). https://doi.org/10.1007/s42979-021-00815-1
28. Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, Tsuhan Chen,Recent advances in convolutional neural networks, Pattern Recognition, Volume 77,(2018), Pages 354-377, ISSN 0031-3203,
29. Montana DJ, Davis L. Training Feedforward Neural Networks Using Genetic Algorithms. Proc 11th Int Jt Conf Artif Intell - Vol 1 (1989).
30. Williams RJ, Zipser D. A Learning Algorithm for Continually Running Fully Recurrent Neural Networks. Neural Comput (1989). https://doi.org/10.1162/neco.1989.1.2.270.
31. Garbin, C., Zhu, X. & Marques, O. Dropout vs. batch normalization: an empirical study of their impact to deep learning. Multimed Tools Appl 79, 12777–12815 (2020). https://doi.org/10.1007/s11042-019-08453-9
32. Tieleman, T. and Hinton, G. Lecture 6.5-rmsprop: Divide the Gradient by a Running Average of Its Recent Magnitude. COURSERA: Neural Networks for Machine Learning, 4, (2012), 26-31.
33. Ioffe, Sergey & Szegedy, Christian. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift (2015).
34. Yash Garg, https://yashgarg1232.medium.com/derivative-of-neural-activation-function-64e9e825b67 , visité en (avril 2022).
35. Duchi, John & Hazan, Elad & Singer, Yoram. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. Journal of Machine Learning Research. (2011). 12. 2121-2159.
36. Kingma, Diederik & Ba, Jimmy. Adam: A Method for Stochastic Optimization. International Conference on Learning Representations (2014).
37. Van Rossum, G., & Drake, F. L. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace (2009).
38. Raybaut, P. Spyder-documentation. Available Online at: Pythonhosted. Org, (2009).
39. Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy. Nature 585, 357–362 (2020). DOI: 10.1038/s41586-020-2649-2.

40. Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E.A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods, 17(3), (2020), 261-272.

41. J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, (2007).

42. Waskom, M. et al., 2017. mwaskom/seaborn: v0.8.1 (September 2017), Zenodo. Available at: https://doi.org/10.5281/zenodo.883859.

43. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, (2011).

44. Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Rafal Jozefowicz, Yangqing Jia, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Mike Schuster, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, (2015). Software available from tensorflow.org.

*Présenté par :*

- **BENOUAR Mohamed Salah Amine**
- **CHERITI Abir**

*Année académique*
*2021 / 2022*

**Mémoire présenté en vue de l'obtention du diplôme de Master**

**Domaine : Sciences de la Nature et de la Vie**
**Filière : Science Biologique**
**Spécialité : *Bio-informatique***

# Deepredictor : un modèle de réseaux de neurones denses pour la prédiction de la mortalité des patients atteints d'un cancer du sein

Résumé

**L'objectif de ce travail est de développer une architecture de réseau de neurones artificiels qui vise la prédiction binaire de la survie des patientes atteintes d'un cancer du sein, et de démontrer l'efficacité des DNN et leur capacité à apprendre et à déterminer les caractéristiques dominantes face à un jeu de données non seulement relativement peu volumineux par rapport aux standards des jeux de données médicales disponibles dans la littérature, qui est amputé d'une quantité considérable de ces données génétiques et cliniques, mais aussi extrêmement complexe du fait de leur diversité**

**Mots clés : architecture, prédiction binaire, cancer du sein, données génétiques et cliniques.**

**Jury d'évaluation :**

| | | |
|---|---|---|
| **Président 1 :** | HAMIDECHI Abdelhafid | (Professeur - Université Frères Mentouri, Constantine 1). |
| **Encadreur :** | CHEHILI Hamza | (MCA - Université Frères Mentouri, Constantine 1). |
| **Examinateur :** | GHERBOUDJ Amira | (MCA - Université Frères Mentouri, Constantine 1). |

Déposé le : 19/06/2022