الجمهورية الجزائرية الديمقراطية الشعبية

République Algérienne Démocratique et Populaire

وزارة التعليم العالي والبحث العلمي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

**Université Constantine 1 Frères Mentouri**
**Faculté des Sciences de la Nature et de la Vie**

**الإخوة منتوري1جامعة قسنطينة**
**كلية علوم الطبيعة والحياة**

**Département : Biologie Apliquée**

**قسم : قسم البيولوجيا التطبيقية**

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

**Domaine :** Sciences de la Nature et de la Vie

**Filière :** Sciences Biologiques

**Spécialité :** Bio-Informatique

**N° d'ordre :**
**N° de série :**

Intitulé :

# Prédiction de profil de cancer du sang à partir des données d'expression des gènes basée sur le Deep learning

**Présenté par :** BENCHAOUI Fella Mounira **Le : 10/06/2024**

**Jury d'évaluation :**

**Président :** HAMIDECHI .M (PROFESSEUR - U Constantine 1 Frères Mentouri).

**Encadrant :** DAAS.S (MAITRE DE CONFÉRENCES A - U Constantine 1 Frères Mentouri)

**Examinateur(s) :** BELLIL.I (PROFESSEUR - U Constantine 1 Frères Mentouri).

**Année universitaire 2023 - 2024**

# REMERCIEMENTS

Je tiens à exprimer mes profonds remerciements à toutes les personnes qui ont contribué à la réalisation de ce mémoire de fin d'étude.

Je souhaite tout d'abord adresser mes plus sincères remerciements :

- à Mr le professeur Hamidechi M.A, qui a accepté de présider le jury de mon mémoire. Son expertise et son savoir sont un exemple pour tous,

- à Monsieur Skander Daas qui a accepté d'encadrer ce travail. Je suis honorée d'avoir pu travailler sous votre direction. Vos conseils avisés et votre expertise m'ont été d'une aide inestimable,

- à Mme BELLIL Ines, examinateur, pour votre gentillesse, votre disponibilité, votre soutien continu et votre précieuse guidance tout au long de ce projet,

- à Mr ALIOUANE Salah Eddine, pour sa disponibilité, son aide, ses remarques constructives à chaque étape de ce travail.

# *Dédicaces*

**À ma mère adorée « Aicha » et à mon père bien-aimé « Khaled »,**

Vous êtes les phares qui éclairent ma route, les sources inépuisables d'amour et de soutien qui m'ont permis de grandir et de m'épanouir. Je vous dédie ce travail en signe de gratitude infinie pour votre présence inconditionnelle et vos sacrifices quotidiens. Merci de m'avoir transmis les valeurs qui me guident et de m'avoir donné la force de réaliser mes rêves.

**À ma tante « Nanou » bienveillante et à ma grand-mère paternelle aimée « Mémé »,**

Vos sages conseils et votre tendresse infinie ont toujours été une source de réconfort et d'inspiration pour moi. Je vous dédie ce travail en témoignage de mon respect profond et de mon affection sincère. Merci de m'avoir transmis votre sagesse et votre amour de la vie.

**À mes sœurs chéries, Meriem et Malak,**

Vous êtes mes confidentes, mes complices et mes plus grandes admiratrices. Nos liens indéfectibles et nos moments de partage comptent parmi mes plus précieux trésors. Je vous dédie ce travail en signe d'amitié indéfectible et d'amour fraternel. Merci de votre soutien indéfectible et de votre joie de vivre contagieuse.

**À mon petit frère Sidou,**

Tu es mon rayon de soleil, celui qui me fait rire et qui illumine mes journées. Je te dédie ce travail en signe d'affection profonde et de tendresse infinie. Merci d'être toujours là pour moi, de me faire sourire et de me donner du courage.

**À mes amies précieuses, Assala, Ahlem et Yesmine,**

Vous êtes des étoiles dans ma vie, des confidentes précieuses et des sources de soutien inestimable. Nos conversations et nos moments de complicité m'apportent tant de joie et de réconfort. Je vous dédie ce travail en signe d'amitié sincère et de gratitude infinie. Merci d'être présentes dans ma vie et de m'apporter votre amitié indéfectible.

# RÉSUMÉ

Ce mémoire présente le développement d'un modèle d'apprentissage profond basé sur un réseau de neurones convolutif unidimensionnel (Conv1D) pour prédire le cancer du sang à partir de données d'expression génique. Le dataset comprend des données d'expression génique, annotées par plusieurs types de cancer du sang et des cas normaux.

Le travail a suivi plusieurs étapes clés : préparation des données, conception et entraînement du modèle Conv1D, évaluation des performances et interprétation des résultats. Les données ont été divisées en ensembles d'entraînement (80%) de test (10%). Le modèle a atteint une précision de 99.45% sur l'ensemble d'entraînement, 98.8% sur l'ensemble de test, montrant une bonne capacité de généralisation.

La matrice de confusion a révélé une forte proportion de prédictions correctes avec peu de faux positifs et de faux négatifs. Les résultats ont été analysés pour identifier les gènes les plus significatifs dans la classification des différents types de cancer du sang.

Cette étude montre que les réseaux de neurones convolutifs unidimensionnels peuvent efficacement classer les données d'expression génique, offrant un outil puissant pour le diagnostic et la recherche sur les cancers hématologiques.

**Mots-clés** : intelligence artificielle, apprentissage profond, réseau de neurones convolutif unidimensionnel, expression génique, classification, cancer du sang, biomédecine.

# ABSTRACT

This thesis presents the development of a deep learning model based on a one-dimensional convolutional neural network (Conv1D) to predict blood cancer from gene expression data. The dataset includes gene expression data annotated by several types of blood cancer and normal cases

The work followed several key steps: data preparation, design and training of the Conv1D model, performance evaluation, and results interpretation. The data was divided into training (80%) and test (20%) sets. The model achieved an accuracy of 99.45% on the training set, 98.8% on the test set, demonstrating good generalization ability.

The confusion matrix revealed a high proportion of correct predictions with few false positives and false negatives. The results were analyzed to identify the most significant genes in the classification of different types of blood cancer.

This study shows that one-dimensional convolutional neural networks can effectively classify gene expression data, offering a powerful tool for the diagnosis and research of hematological cancers.

**Keywords**: artificial intelligence, deep learning, one-dimensional convolutional neural network, gene expression, classification, blood cancer, biomedicine.

# الملخص

يقدم هذا البحث تطوير نموذج تعلم عميق قائم على شبكة عصبية تلافيفية أحادية الأبعاد (Conv1D) للتنبؤ بسرطان الدم من بيانات التعبير الجيني. يتضمن مجموعة البيانات قيم التعبير الجيني المشروحة بأنواع مختلفة من سرطان الدم والحالات العادية.

اتبع العمل عدة خطوات رئيسية: إعداد البيانات، تصميم وتدريب نموذجConv1D ، تقييم الأداء وتفسير النتائج. تم تقسيم البيانات إلى مجموعات تدريب (80%)، واختبار (20%). حقق النموذج دقة بنسبة 99.45% على مجموعة التدريب، و98.8% على مجموعة الاختبار، مما يدل على قدرة جيدة على التعميم.

كشفت مصفوفة الالتباس عن نسبة عالية من التنبؤات الصحيحة مع عدد قليل من الإيجابيات والسلبيات الكاذبة. تم تحليل النتائج لتحديد الجينات الأكثر أهمية في تصنيف الأنواع المختلفة من سرطان الدم.

تظهر هذه الدراسة أن الشبكات العصبية التلافيفية أحادية الأبعاد يمكنها تصنيف بيانات التعبير الجيني بفعالية، مما يوفر أداة قوية لتشخيص وبحث سرطانات الدم.

**الكلمات الرئيسية**: الذكاء الاصطناعي، التعلم العميق، الشبكة العصبية التلافيفية أحادية الأبعاد، التعبير الجيني، التصنيف، سرطان الدم، الطب الحيوي..

# LISTES DES FIGURES

# LISTE DES TABLEAUX

# ACRONYMES

- ✓ ADN : Acide Désoxyribonucléique
- ✓ AI : Intelligence Artificielle
- ✓ APA : Apprentissage Automatique
- ✓ CNN : Convolutional Neural Network (Réseau Neuronal Convolutif)
- ✓ RNA : Acide Ribonucléique*
- ✓ CSV : Comma-Separated Values
- ✓ DL : Deep Learning (Apprentissage profond)
- ✓ ML : Machine Learning (Apprentissage Automatique)
- ✓ PCR : Polymérase Chain Reaction (réaction de polymérisation en chaîne)
- ✓ RNA-seq) : Séquençage d'ARN
- ✓ NLP : Natural Language Processing (Traitement automatique du langage naturel)

# TABLE DE MATIÈRE

# TABLE DES MATIÈRES

## Table des matières

# Introduction

# INTRODUCTION

Il est essentiel de prendre des mesures précoces et précises pour détecter les cancers. Dans mon mémoire j'ai travaillé sur le cancer du sang, étant donné la fréquence élevée de cette maladie et son impact sur la mortalité des gens à travers le monde. Grâce aux progrès technologiques en génomique et en apprentissage automatique, de nouvelles opportunités sont ouvertes pour améliorer les techniques de diagnostic et de pronostic du cancer. Ce mémoire fait partie de cette nouvelle dynamique, qui vise à étudier l'utilisation des réseaux de neurones convolutifs (CNN) pour analyser l'expression génique et classifier le cancer du sang.

Le profilage de l'expression génique permet la quantification des niveaux d'ARN messager dans les cellules, ce qui permet d'obtenir une vision approfondie de l'activité génétique. Des signatures génétiques distinctives ont été découvertes entre les cellules cancéreuses et normales grâce à cette technique, ce qui ouvre la voie à des méthodes diagnostiques basées sur les données. Toutefois, l'étude et l'interprétation de ces grandes quantités de données génomiques présentent des difficultés importantes, demandant l'utilisation d'outils analytiques avancés.

Les algorithmes d'apprentissage profond, tels que les réseaux de neurones convolutifs, ont montré une grande efficacité dans différents domaines tels que la reconnaissance de motifs et la classification d'images. Ils sont des candidats parfaits pour l'analyse des données d'expression génique en raison de leur aptitude à capturer des caractéristiques complexes à travers des couches hiérarchiques. Dans cette situation, notre étude propose d'employer cette méthode des CNN pour distinguer les profils d'expression génique associés aux tissus cancéreux et normaux du sang.

Ce mémoire est structuré en quatre chapitres :

- Le premier chapitre introduit les concepts d'expression génétique, les méthodes traditionnelles et modernes comme l'analyse de microarray par ADN et la puce à ADN la plus fréquentes GeneChip Humain Genome U133 Array.

- Le deuxième chapitre explore les fondements de l'intelligence artificielle, en particulier l'apprentissage automatique, l’apprentissage profond avec son architecture et leur application dans la classification des cancers.

- Le troisième chapitre détaille la méthodologie utilisée dans mon étude, y compris la collecte des données, la préparation des données, et la conception du modèle de réseau de neurones convolutifs.

- Le quatrième chapitre présente les résultats obtenus, discute leur pertinence et propose des pistes d'amélioration pour les travaux futurs.

# PARTIE 1 :

# RECHERCHE BIBLIOGRAPHIQUE

# CHAPITRE 1 : COMPREHENSION DU PROFILAGE DE L'EXPRESSION GENIQUE

# INTRODUCTION

Le profilage de l'expression génique est une technique fondamentale en biologie moléculaire et en génomique, permet de quantifier l'activité des gènes dans différentes conditions biologiques et pathologiques. Cette approche est particulièrement cruciale pour comprendre les mécanismes moléculaires sous-jacents à diverses maladies, notamment le cancer, où les altérations de l'expression génique jouent un rôle central dans la progression de la maladie.

Ce premier chapitre vise à fournir une compréhension approfondie des concepts et des méthodes associés au profilage de l'expression génique. Nous commencerons par explorer les fondamentaux de la biologie moléculaire, en examinant la nature de l'information génétique et les processus par lesquels elle est transcrite et traduite en protéines fonctionnelles. Ces protéines, essentielles à la vie, régulent de nombreux aspects des processus biologiques et déterminent les caractéristiques phénotypiques des organismes.

Ensuite, nous introduirons le concept d'expression génique et son importance dans le contexte biologique et médical. Nous discuterons de la régulation de l'expression génique, un processus complexe et finement contrôlé qui assure que les gènes sont exprimés au bon moment, au bon endroit et à des niveaux appropriés. Cette régulation est essentielle pour le développement normal, le fonctionnement des cellules et la réponse aux stimuli environnementaux.

Le chapitre se poursuit avec une présentation des différentes méthodes utilisées pour étudier l'expression génique. Nous examinerons les techniques traditionnelles comme le Northern blot, ainsi que les approches modernes à haut débit telles que les microarrays et les séquençages de nouvelle génération.

Enfin, nous mettrons en lumière le rôle du profilage de l'expression génique dans la recherche sur le cancer.

Ces informations sont cruciales pour le diagnostic, le pronostic et le développement de thérapies ciblées, ouvrant la voie à une médecine de précision plus efficace et personnalisée.

Ainsi, ce chapitre établira les bases nécessaires pour comprendre les données d'expression génique et leur analyse, ouvrant la voie à l'application des techniques de Deep Learning pour prédire les profils de cancer à partir de ces données dans les chapitres ultérieurs de ce mémoire.

# 1. Fondamentaux de la biologie moléculaire

## 1.1. L'information génétique :

L'information génétique représente fondamentalement le code génétique qui définit les caractéristiques biologiques d'un être vivant. Elle est principalement présente dans l'ADN chez la plupart des êtres vivants, même si certains virus utilisent l'ARN pour stocker leur information génétique. (Dillenburg, 2017)

Cette information est structurée en gènes, des segments spécifiques de l'ADN, renfermant les instructions nécessaires à la production de protéines. Ces dernières sont cruciales dans de multiples processus biologiques tels que la croissance, le métabolisme et la réparation des tissus. (*L'information génétique et la molécule d'ADN - 2nde - Cours SVT - Kartable*, s. d.)

L'information génétique détermine non seulement les traits physiques d'un organisme, mais également de nombreux aspects de son comportement et de son métabolisme. Transmise de génération en génération pendant le processus de reproduction, elle combine une partie de l'information génétique de chaque parent pour former un nouvel être. Les variations présentes dans l'information génétique, crées par des mutations ou par la recombinaison génétique, sont à l'origine de la diversité des espèces vivantes. (Choudhuri, 2014)

Le processus par lequel l'information génétique est utilisée pour créer des protéines fonctionnelles dans une cellule implique deux étapes principales : la transcription et la traduction (Voir la figure 1). (*Transcription ADN : cours et explications │ StudySmarter*, s. d.)

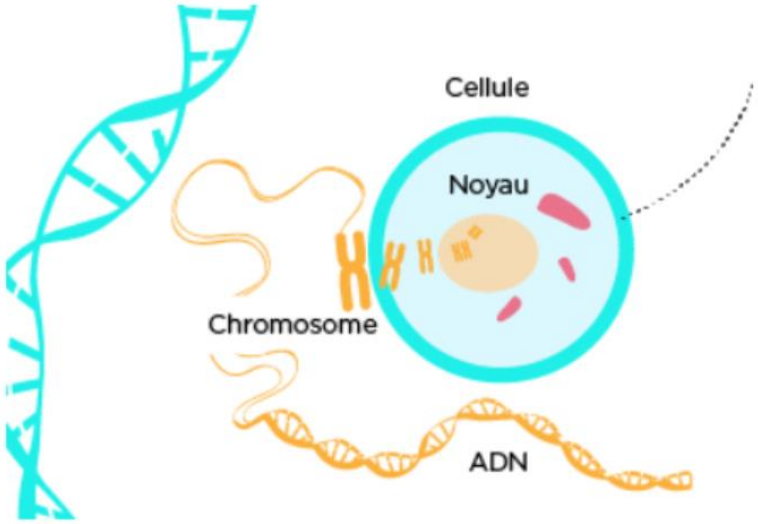


***Figure 1 :*** *Représente le cycle de vie de l'information génétique*

## 1.2. La transcription :

La transcription est un processus essentiel qui implique la synthèse d'une molécule d'ARN à partir d'une portion d'ADN. Ce processus se déroule en trois étapes clés : l'initiation, l'élongation et la terminaison. (*Transcription ADN : cours et explications | StudySmarter*, s. d.)

Tout d'abord, l'ARN polymérase se lie au promoteur de l'ADN, ce qui marque le début de la synthèse de l'ARN. Ensuite, l'ADN se déroule, permettant à l'ARN de commencer sa formation. (*Les étapes de la transcription (leçon) | Khan Academy*, s. d.)

Pendant l'élongation, le complexe de transcription (composé de l'ARN polymérase et des facteurs de transcription, qui sont des protéines qui se lient à des séquences spécifiques sur l'ADN et qui interagissent avec l'ARN polymérase pour initier et réguler la transcription) se déplace le long de l'ADN matrice pour produire un ARN. Un mécanisme de correction d'erreurs est activé pour remplacer tout nucléotide incorrectement ajouté. La transcription se termine lorsque le complexe rencontre un codon stop sur l'ADN matrice. Cette phase est caractérisée par la formation d'une boucle d'épingle secondaire, ce qui permet au complexe de transcription de se détacher de l'ADN. Le transcrit primaire subit ensuite des modifications pour devenir mature, notamment l'ajout d'une structure de cap à l'extrémité 5' et d'une queue poly A à l'extrémité 3', stabilisant ainsi la molécule. (*Aperçu de la transcription (leçon) | Khan Academy*, s. d.)

Les introns sont éliminés du transcrit primaire et les exons restants sont réunis pour former l'ARN mature, qui est ensuite transporté dans le cytoplasme pour la synthèse des protéines. (Dillenburg, 2017)

## 1.3. Traduction :

Le mécanisme de la traduction génétique consiste à transcrire les informations de l'ARNm en séquences d'acides aminés, permettant ainsi la formation des protéines essentielles au bon fonctionnement des cellules. (*Cours : Théorie de la traduction*, s. d.)

L'ARNm transporte ces informations du noyau cellulaire vers le cytoplasme, où la traduction a lieu sur les ribosomes. À chaque étape, un ARNm est lié à un ribosome, et les acides aminés sont apportés par des ARNt correspondants. Ces acides aminés sont assemblés selon l'information portée par l'ARNm pour former une chaîne polypeptidique. (*Les bases de la biologie moléculaire - Principe de la traduction*, s. d.)

Ce processus se répète jusqu'à ce qu'un codon d'arrêt spécifique (UAA, UAG ou UGA) soit rencontré, indiquant la fin de la chaîne polypeptidique. (*Les bases de la biologie moléculaire - Principe de la traduction*, s. d.)

La traduction est un processus fondamental car les protéines synthétisées lors de ce processus sont essentielles pour déterminer les caractéristiques physiques et fonctionnelles d'un organisme. (*Traduction ADN*, s. d.)

Ces protéines influent directement sur le phénotype, c'est-à-dire l'ensemble des traits observables d'un organisme, tels que la couleur des yeux, la taille ou la capacité à digérer certains aliments. Ainsi, la transcription et la traduction jouent un rôle crucial dans la création et la diversité des phénotypes au sang des espèces vivantes (Voir la figure 2). (Dillenburg, 2017)

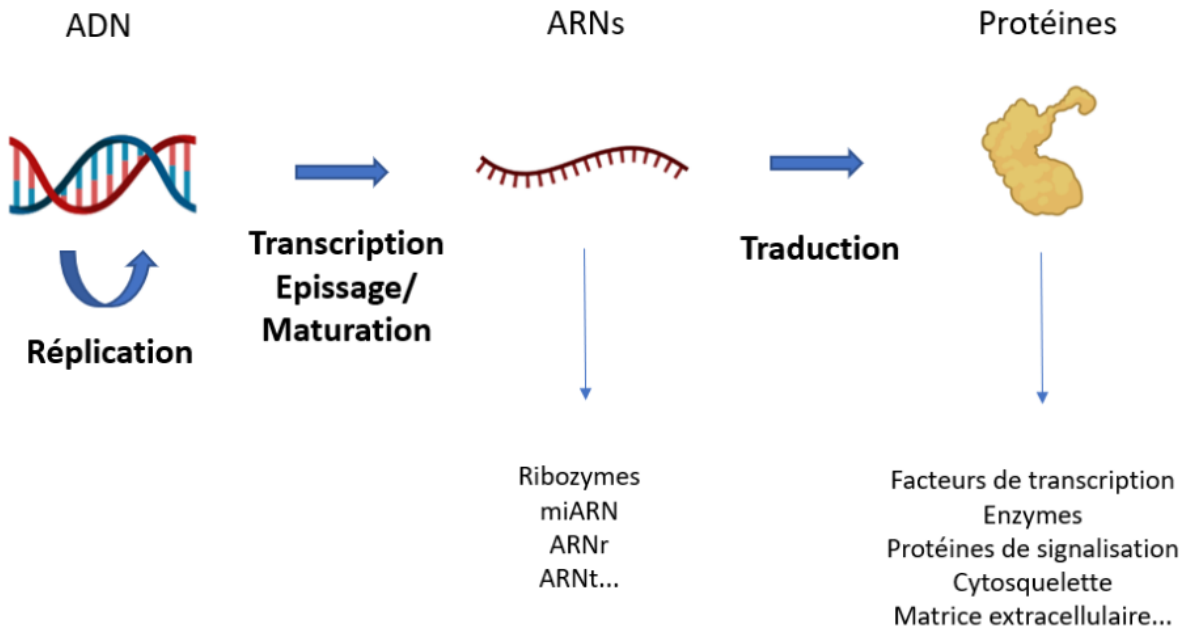


***Figure 2 :*** *Représente Le processus de transcription et de traduction*
*(« Contrôle de la transcription », 2021)*

## 2. L'expression génique et son importance :

L'expression génique désigne le processus par lequel les informations encodées dans un gène sont transformées en une fonction spécifique (*Expression génétique : définition & étapes | StudySmarter*, s. d.)

Ce processus se déroule en différentes étapes, telles que la transcription de l'ADN en ARN et la traduction de l'ARN en protéines. L'expression génique peut être envisagée comme un système de régulation complexe, agissant à la fois comme un "interrupteur" pour contrôler le moment et le lieu de production des ARN et des protéines, ainsi qu'un "réglage de volume" pour déterminer leur quantité. Ce processus est minutieusement régulé et varie selon les conditions et les types cellulaires. (*Gene Expression*, s. d.)

Les éléments produits par la transcription de nombreux gènes jouent un rôle primordial dans la régulation de l'expression d'autres gènes.

L'analyse de l'expression génique permet de déterminer quels gènes sont exprimés au niveau de la transcription dans des cellules spécifiques *(Marti et al., 2002)* .

Cette analyse est nécessaire pour étudier les gènes régulateurs impliqués dans certaines maladies telles que les différents types de cancer, ainsi que pour comprendre les réponses cellulaires à l'environnement. Elle permet également d'évaluer l'activité fonctionnelle des

produits géniques, d'observer les phénotypes associés à ces gènes et de contrôler les caractéristiques physiques et fonctionnelles des organismes vivants. En effet, elle détermine quels gènes sont activés et à quel niveau dans une cellule ou un tissu donné, en réponse à des signaux internes et externes (Voir la figure 3). (*Gene Expression*, s. d.)

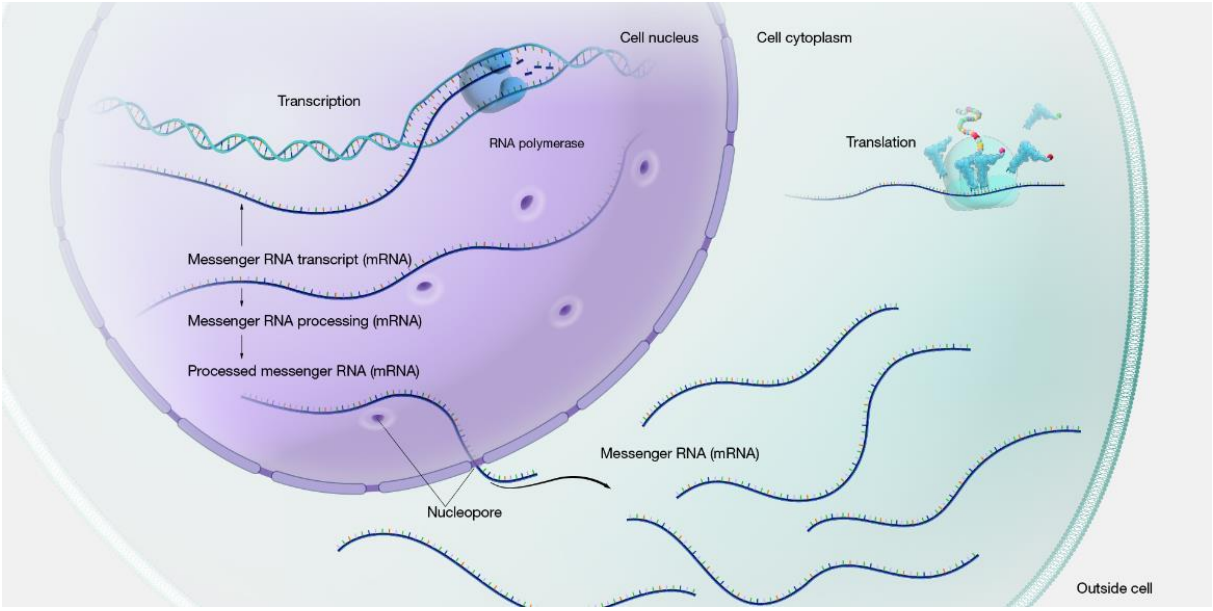


***Figure 3 :*** *Représente les étapes de l'expression génétique*

(*Gene Expression*, s. d.)

## 3. Méthodes d'étude de l'expression génique

Pour mesurer l'expression génique, plusieurs méthodes sont utilisées. Parmi les techniques courantes, on trouve :

- Analyse par puce à ADN (microarray ) : est une méthode largement utilisée pour analyser l'expression génique, permettant de mesurer simultanément l'expression de milliers de gènes. (*Puce à ADN : pourquoi et pour qui ?*, s. d.)
- Sequençage d'ARN (RNA-seq) : permettant une analyse exhaustive et précise de l'expression génique à l'échelle du génome entier. (*Séquençage d'ARN - Séquençage de nouvelle génération - GENEWIZ*, s. d.)

La PCR quantitative (qPCR) : est une technique de biologie moléculaire extrêmement sensible et précise utilisée pour quantifier la quantité d'ARN messager (ARNm) transcrit à partir d'un gène spécifique. (*Quantification et expression d'un gène*, s. d.)

Ces méthodes offrent des informations précieuses sur les schémas d'expression génique dans différents contextes biologiques et pathologiques.

## 3.1. Méthodes traditionnelles

Bien que des méthodes traditionnelles d'analyse de l'expression génique, par exemple le Northern Blot, existent toujours, elles sont aujourd'hui remplacées par des techniques plus modernes.

En effet, ces dernières offrent des avantages considérables en termes de sensibilité, de spécificité, de précision et de rapidité. (*Gene Expression*, s. d.)

## 3.2. Techniques à haut débit (Analyse par puce à ADN : microarray)

Les technologies à puces offrent une méthode puissante pour observer les interactions entre différentes molécules telles que des fragments d'ADN ou des protéines, et une sélection prédéterminée de sondes moléculaires. Actuellement, parmi les avancées les plus significatives dans ce domaine, figure l'utilisation de puces à ADN, aussi appelées puces ADN ou microarrays. Ces dispositifs permettent de quantifier simultanément les niveaux d'expression génique en ARNm dans une cellule vivante, fournissant ainsi des données précieuses sur l'activité génétique à grande échelle. (*DNA Microarrays and Gene Expression*, s. d.)

Le microarray, est une technique à haut débit puissante qui offre une vision globale des schémas d'expression génique dans les échantillons biologiques. Grâce à une seule puce de hybridation microarray, des milliers de gènes peuvent être analysés simultanément (STEKEL, 2003). Cette méthode joue un rôle important dans le développement des connaissances sur les origines et les mécanismes sous-jacents à divers troubles complexes. Elle permet aux chercheurs de comparer le comportement moléculaire de différentes lignées cellulaires ou de tissus spécifiques exposés à des conditions pathologiques ou expérimentales, (Dillenburg, 2017) offrant ainsi des informations sur les processus physiologiques et facilitant l'identification de nouveaux marqueurs biologiques exploitables dans le diagnostic, le pronostic et le traitement pharmacologique d'une grande variété de maladies. (Dillenburg, 2017)

Le processus standard de traitement des données de microarray et leur analyse est décrit dans la Figure 1 Il implique trois étapes majeurs (Voir la figure 4) :

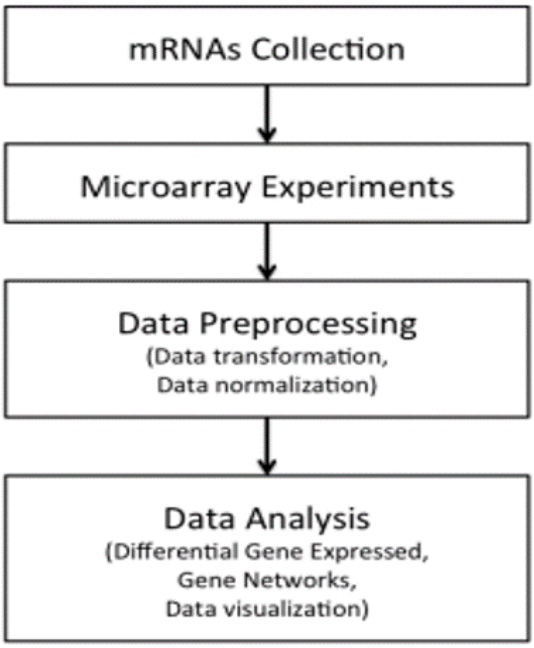


***Figure 4 :*** *Les étapes d'analyse et de traitement de données microarray*

(Dillenburg, 2017)

1) La préparation de l'échantillon : les échantillons d'ARN sont collectés à partir de tissus provenant d'organismes biologique (de patients atteints de maladies ou en bonne santé, ou de cultures de cellules homogènes) en fonction de la nature spécifique du problème étudié. Puis, l'ARN est extrait de ces cellules. (Dillenburg, 2017)
2) La fabrication/ préparation de la puce : une expérience de microarray est réalisée. Il existe deux types principaux selon la manière dont les sondes sont disposées sur la lame :

- les microarrays spotted ou cDNA, où les sondes sont imprimées mécaniquement sur la puce après avoir été synthétisées séparément . (Dillenburg, 2017)
- les puces à oligonucléotides , comme les Affymetrix GeneChip (les représentants principaux), où les sondes sont directement synthétisées sur la surface. Dans ce dernier cas, un gène est représenté par un ensemble de sondes appelé un jeu de sondes. (Dillenburg, 2017) (Voir la figure 5)

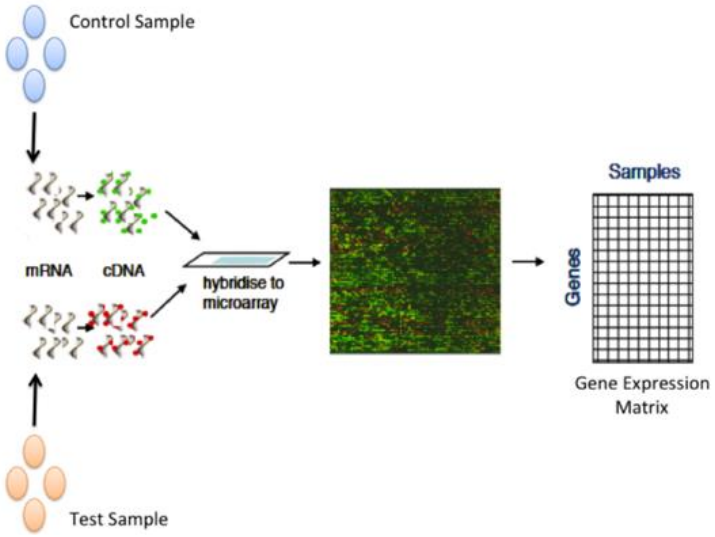


***Figure 5 :*** *Les étapes de détection des gènes exprimés par analyse du microarray*

(Dillenburg, 2017)

***Tableau 1 :*** *La description des étapes de détection des gènes éxprimés par analyse microarray*

(« Puce à ADN », 2024)

| Fabrication de la puce | Préparation, marquage et hybridation de l'ARNm | Balayage de la puce |
|---|---|---|
| • préparation d'une petite puce avec du verre ou du nylon, sur laquelle on fixe des milliers de petites sondes. Chaque sonde correspond à une séquence d'ADN. | • La préparation des échantillons d'ARNm (un témoin et un test) implique leur conversion en ADN complémentaire (cDNA), qui est ensuite marqué avec des colorants fluorescents ou des isotopes radioactifs. Ensuite, ces échantillons marqués sont hybridés avec les séquences clonées sur la surface de la puce. Cette interaction permet de générer un signal détectable, facilitant ainsi l'analyse des données. | • Une fois les échantillons hybridés, la puce est scannée pour mesurer l'intensité du signal émis par les cibles marquées.Chaque point lumineux sur l'image de la puce correspond à une sonde. cette image est traitée et convertie en données numériques, qui servent de base à l'analyse de l'expression génique. |

3) Prétraitement et analyse des données : avant toute analyse, il est souvent nécessaire de prétraiter les données, ce qui inclut la transformation et la normalisation des données.Ensuite, les données du microarray peuvent être représentées par une matrice bidimensionnelle X = xij, où chaque ligne représente un gène, chaque colonne représente un échantillon biologique (une maladie ou un tissu normal, ou différents moments dans le temps), et chaque valeur xij enregistre le niveau d'expression du gène i dans l'échantillon j (ou dans la condition j). (Dillenburg, 2017)
Les données issues des microarrays présentent certains défis, notamment celui de la dimensionnalité. En effet, lors de l'analyse de microarrays, on mesure souvent des dizaines de milliers de gènes à partir de seulement quelques échantillons, ce qui peut entraîner des risques de détection de relations fausses ou spéculatives. Un autre défi majeur des données biologiques est que les informations fournies par les microarrays se limitent à la quantification de la concentration d'ARNm, ignorant les interventions potentielles ou les changements environnementaux survenant après la phase de transcription (Dillenburg, 2017).

L'une des puces à ADN les plus fréquemment utilisées est la puce GeneChip® Human Genome U133, connue pour ses nombreuses gènes. En utilisant des cartouches individuelles ou des plaques multi-puces, cette technologie garantit une analyse précise et complète des niveaux d'expression génique, sans compromettre la qualité des résultats. (Data Sheet GeneChip® Human Genome U133 Arrays) . (*GeneChip™ Human Genome U133 Plus 2.0 Array*, s. d.)

## 3.3. Le rôle du profilage de l'expression génique dans la recherche sur le cancer :

Le profilage de l'expression génique est important dans la recherche sur le cancer en offrant une compréhension approfondie des mécanismes moléculaires sous-jacents à cette maladie complexe. (*Profilage de l'expression génique*, s. d.)

L'analyse des schémas d'expression génique dans les tissus cancéreux offre aux chercheurs un outil puissant pour comprendre les mécanismes sous-jacents de la maladie. Cette approche permet de classifier les tumeurs en sous-types moléculaires distincts et ouvre une voie importante et plus précis pour le diagnostic, le pronostic et le traitement des patients atteints de cancer.

Le profilage de l'expression génique aide à identifier de nouveaux biomarqueurs prédictifs de la réponse aux traitements anticancéreux, ouvrant ainsi la voie à une médecine plus personnalisée et ciblée. (Capp, 2011)

En résumé, le profilage de l'expression génique représente un outil puissant pour décrypter les mécanismes biologiques du cancer et orienter le développement de thérapies innovantes.

# CHAPITRE 2 : INTRODUCTION A L'INTELLIGENCE ARTIFICIELLE ET A L'APPRENTISSAGE PROF

# Introduction :

L'intelligence artificielle (IA) et l'apprentissage profond (deep learning) représentent deux des avancées technologiques les plus significatives du XXIe siècle. Ils ont révolutionné de nombreux secteurs, notamment celui de la recherche biomédicale, en ouvrant de nouvelles perspectives pour le diagnostic, le traitement et la prévention des maladies. Dans le cadre de ce mémoire intitulé "Prédiction de profil de cancer à partir des données d'expression des gènes basée sur le Deep learning", le deuxième chapitre se propose d'explorer les fondements et les applications de l'IA et du deep learning, en mettant l'accent sur leur rôle crucial dans le domaine biomédical.

Ce chapitre explore les fondements et les applications de l'intelligence artificielle (IA) et de l'apprentissage profond (deep learning) dans le domaine biomédical. Il commence par une introduction aux concepts clés de l'IA, en abordant sa définition, son historique et les techniques principales comme les réseaux de neurones artificiels. Ensuite, le chapitre examine les applications spécifiques de l'IA en médecine, notamment dans le diagnostic, le traitement personnalisé et la surveillance des patients.

Enfin, une attention particulière est accordée aux concepts et architectures de l'apprentissage profond. L'apprentissage profond, avec ses réseaux neuronaux convolutifs, représente une avancée majeure dans le traitement des données complexes : en bioinformatique et en oncologie. Les techniques de deep learning permettent de prédire et de diagnostiquer des maladies comme le cancer en analysant des données d'expression des gènes, offrant ainsi des outils puissants pour la médecine de précision.

# 2. Fondamentaux de l'intelligence artificielle

## 2.1 Définition d'IA

L'intelligence artificielle (IA) représente un domaine majeur de l'informatique qui vise à émuler les capacités cognitives humaines à travers l'application d'algorithmes dans un environnement informatique dynamique. Son objectif principal est de permettre aux ordinateurs de simuler la pensée et les actions humaines. Trois éléments essentiels sont nécessaires pour soutenir le développement de l'IA :

1- Des systèmes informatiques performants.

2- Une gestion efficace de données structurées.

3- Des algorithmes d'IA sophistiqués (code).

Pour se rapprocher le plus possible du comportement humain, l'IA nécessite une quantité importante de données ainsi qu'une capacité de traitement élevée. (*Intelligence artificielle : définition et utilisations | NetApp*, s. d.)

Ce champ d'étude englobe une variété de techniques, facilitées par les progrès constants dans les capacités de calcul des ordinateurs, une meilleure compréhension des processus naturels associés à l'intelligence, et les avancées des chercheurs dans les sciences fondamentales. (Mathivet, 2017)

L'IA, en tant que domaine complexe et en évolution constante, est difficile à définir précisément en raison de sa nature étendue. Par exemple, des technologies allant des simples algorithmes de recommandation utilisés par des plateformes comme Netflix jusqu'aux systèmes sophistiqués de conduite autonome développés par des entreprises telles que Tesla, sont toutes considérées comme relevant de l'IA. Cette diversité rend ce domaine à la fois fascinant et mystérieux, avec des définitions qui évoluent au fur et à mesure des avancées technologiques (Voir la figure 6). (*Intelligence Artificielle : Définition, histoire, enjeux*, s. d.-a)

***Figure 6 :*** *L'AI atteint un niveau d'intelligence supérieur à celui de l'homme* (*Intelligence Artificielle : Définition, histoire, enjeux*, s. d.-a)

### 2.1.1 Historique de l'AI :

L'histoire de l'intelligence artificielle (IA) débute en 1943 avec la publication de l'article "A Logical Calculus of Ideas Immanent in Nervous Activity" par Warren McCullough et Walter Pitts, marquant ainsi le début de l'étude des réseaux de neurones artificiels. En 1950, Marvin Minsky et Dean Edmonds, deux étudiants de Harvard, créent Snarc, le premier ordinateur utilisant un réseau de neurones. Cette même année, Alan Turing publie le célèbre Turing Test, établissant ainsi les bases de l'IA en posant la question fondamentale de la reproduction de l'intelligence humaine par les machines. (*Intelligence Artificielle : Définition, histoire, enjeux*, s. d.-a).

Le terme "intelligence artificielle" est officiellement utilisé pour la première fois en 1956 lors de la conférence "Dartmouth Summer Research Project on Artificial Intelligence", organisée par John McCarthy, considérée comme le point de départ officiel de l'IA.(*Intelligence Artificielle : Définition, histoire, enjeux*, s. d.-a)

Les avancées dans le domaine se sont poursuivies au fil des années. En 1959, Arthur Samuel introduit le terme "Machine Learning" chez IBM, posant ainsi les fondements de cette branche

de l'IA. En 1989, Yann Lecun développe le premier réseau de neurones capable de reconnaître des chiffres manuscrits, ouvrant ainsi la voie au deep learning. (Mathivet, 2017)

En 1997, l'événement marquant de l'histoire de l'IA survient lorsque le système Deep Blue d'IBM bat le champion du monde d'échecs Gary Kasparov, démontrant la capacité des machines à surpasser les humains dans certains domaines. (*Intelligence Artificielle : Définition, histoire, enjeux*, s. d.-a)

Aujourd'hui, grâce au développement continu des technologies telles que le deep learning et le machine learning, on distingue généralement trois types d'IA :

***Tableau 2** : Les differences entre l'intelligence artificielle te l'intelligence naturelle (forte/faible)*

| L'intelligence artificielle générale | L'intelligence artificielle forte | L'inteligence artificielle faible |
|---|---|---|
| •L'IA connue sous le nom d'IA profonde, capable d'accomplir toute activité mentale comme un être humain ou un animal.<br>•Certains scientifiques considèrent que l'existence d'une IA générale, comme évoquée dans GPT-4, est encore hypothétique.<br>•De nombreux chercheurs en intelligence artificielle estiment que les avancées dans les réseaux de neurones pourraient conduire à la création d'une IA générale. | •L'IA forte, ou superintelligence, c'est un modèle démontre une conscience propre et fait référence à des connaissances philosophiques.<br>•Les chercheurs en IA considèrent que l'IA forte est actuellement impossible à créer et que la notion de conscience et de sentiments ne peut émerger dans des systèmes mathématiques basés sur la manipulation de symboles et de calculs. | • L'intelligence artificielle faible : IA étroite, est la dernière catégorie d'IA.<br>• Ce type d'IA est capable d'accomplir une seule tâche de manière presque parfaite.<br>• Contrairement à d'autres formes d'IA, l'IA faible n'a pas besoin de supervision humaine pour fonctionner.<br>• Elle est largement utilisée dans divers secteurs d'activité pour accélérer les processus. |

(*Intelligence Artificielle : Définition, histoire, enjeux*, s. d.-b)

### 2.1.2 Concepts clés en intelligence artificielle

L'intelligence artificielle (IA) est devenue un sujet de discussion majeur dans de nombreux médias spécialisés, notamment en ce qui concerne les technologies de Machine Learning (ML) et de Deep Learning, qui reposent sur l'utilisation de réseaux de neurones artificiels. Ces

avancées technologiques ouvrent la voie au développement d'applications à la fois publiques et professionnelles. (*L'IA - concepts et histoire - MetalBlog*, s. d.)

## Apprentissage automatique :

## Définition de l'APA :

Le Machine Learning, également appelé apprentissage automatique, constitue une branche essentielle de l'intelligence artificielle. Il repose sur l'utilisation de données et d'algorithmes pour détecter des schémas récurrents dans les données et imiter le processus d'apprentissage humain afin d'améliorer la précision des prédictions (Voir la figure 7). (*#Intelligence artificielle en #imagerie cardiovasculaire – Révolution actuelle et future « maismaismedicina*, s. d.)

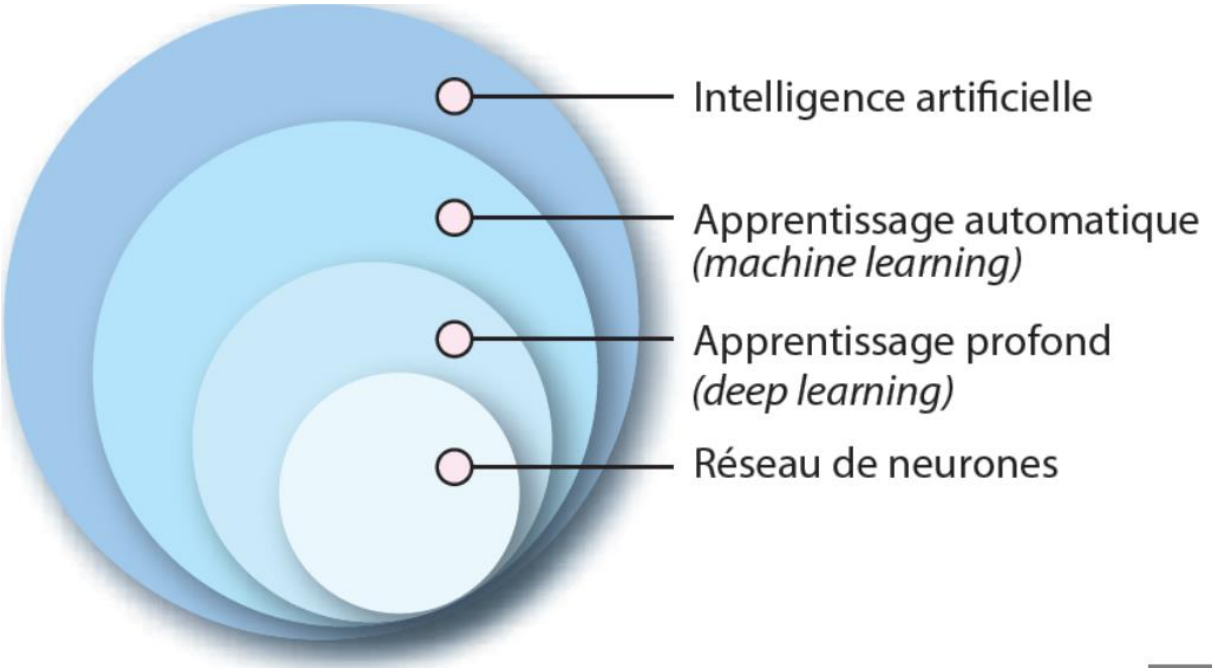


***Figure 7** : Représente les trois niveaux : intelligence artificielle, apprentissage automatique et apprentissage profond*

*(#Intelligence artificielle en #imagerie cardiovasculaire – Révolution actuelle et future « maismaismedicina, s. d.)*

## Fonctionnement et principes fondamentaux :

Les données utilisées peuvent être sous forme de nombres, de mots, d'images ou de statistiques, sont stockées numériquement et servent de base à l'apprentissage des algorithmes.

En analysant ces données pour identifier des patterns significatifs, les algorithmes de Machine Learning peuvent être formés pour effectuer des tâches telles que la classification, la prédiction ou d'autres analyses.

L'un des principaux avantages du Machine Learning est sa capacité à apprendre de manière autonome à partir des données, permettant ainsi une amélioration continue de la précision des prédictions au fil du temps. De plus, une fois que les algorithmes sont entraînés, ils peuvent généraliser leurs connaissances pour traiter de nouvelles données et résoudre des problèmes complexes de manière efficace.

(Robert, 2020)( (Aliouane & BENDAHMANE Abdelhafedh, s. d.)

## Processus de l'apprentissage automatique :

Dans l'apprentissage automatique on deux phases : la phase d'apprentissage et la phase de prédiction.

- La phase d'apprentissage :

La phase d'apprentissage, première étape du processus, est également appelée phase d'entraînement. Elle implique la conception d'un système par l'estimation d'un modèle à partir de l'analyse des données. L'objectif principal de cette phase est de comprendre la logique du modèle à intégrer et de déterminer l'algorithme de transformation requis. Cette progression se traduit par un affinement des prédictions à mesure que le processus d'apprentissage de la machine avancé. (Robert, 2020)

(Aliouane & BENDAHMANE Abdelhafedh, s. d.)

- La phase de prédiction :

Après avoir développé une compréhension approfondie de la logique et de l'algorithme du problème, la machine est en mesure d'identifier les objectifs spécifiques d'une situation donnée. (Robert, 2020)

## Différents champs d'applications de l'apprentissage automatique :

L'apprentissage automatique est une technologie omniprésente dans notre quotidien, souvent utilisée pour traiter le BIG DATA dans plusieurs domaines. Des outils tels que Google Maps, Google Assistant, Alexa, et bien d'autres encore en sont des exemples notables. Voici un aperçu des applications les plus en vogue de cette technologie dans le monde réel : (Voir la figure 8) ((*Applications of Machine Learning - Javatpoint*, s. d.) (Aliouane & BENDAHMANE Abdelhafedh, s. d.)

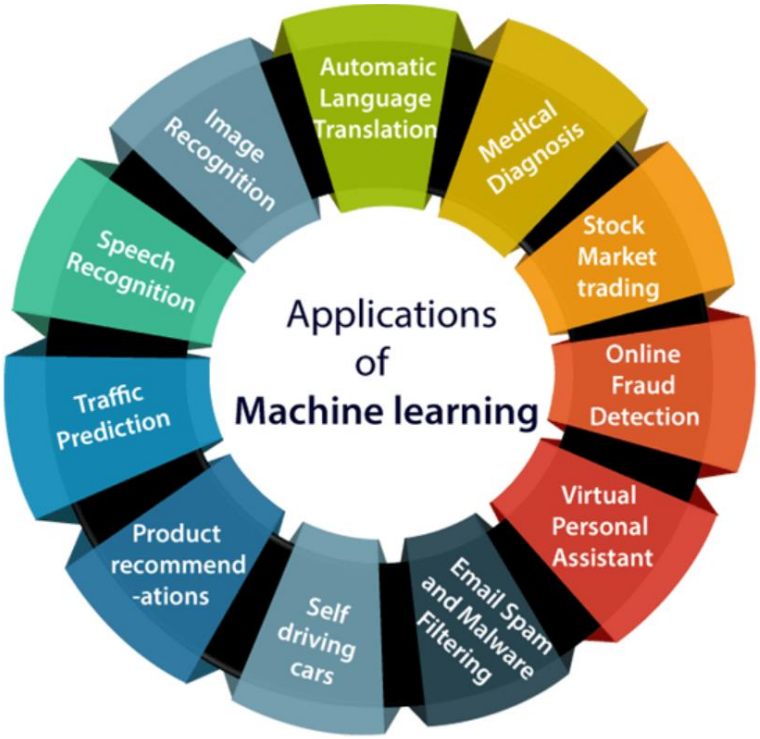


***Figure 8 :*** *Les domaines de l'apprentissage automatique*

## Réseaux neuronaux :

Les réseaux de neurones, également désignés sous les noms de réseaux de neurones artificiels (ANN) ou de réseaux de neurones simulés (SNN), sont des éléments fondamentaux de l'apprentissage automatique, notamment des algorithmes d'apprentissage en profondeur. Inspirés par le fonctionnement du cerveau humain, ces réseaux reproduisent le processus de transmission de signaux entre les neurones biologiques. (*Que sont les réseaux neuronaux ? | IBM*, s. d.)

Dans le domaine de l'apprentissage automatique, les réseaux de neurones s'appuient sur des ensembles de données d'entraînement pour améliorer leur précision au fil du temps. Une fois que ces algorithmes d'apprentissage sont suffisamment ajustés, ils deviennent de puissants outils pour l'informatique et l'intelligence artificielle. Ils permettant de classifier et de regrouper les données de manière très efficace. Des tâches telles que la reconnaissance vocale ou la reconnaissance d'image peuvent être accomplies en quelques minutes seulement, là où des experts humains auraient besoin de plusieurs heures.( (*Que sont les réseaux neuronaux ? | IBM*, s. d.)

Il est cependant important de noter que les réseaux de neurones ne représentent qu'une partie des nombreux outils et approches utilisés dans les algorithmes de Deep Learning (DL). En effet, le réseau neuronal peut être intégré comme composant dans divers algorithmes de DL pour traiter des entrées de données complexes dans un espace que les ordinateurs peuvent comprendre (Voir la figure 9). (Aliouane & BENDAHMANE Abdelhafedh, s. d.)

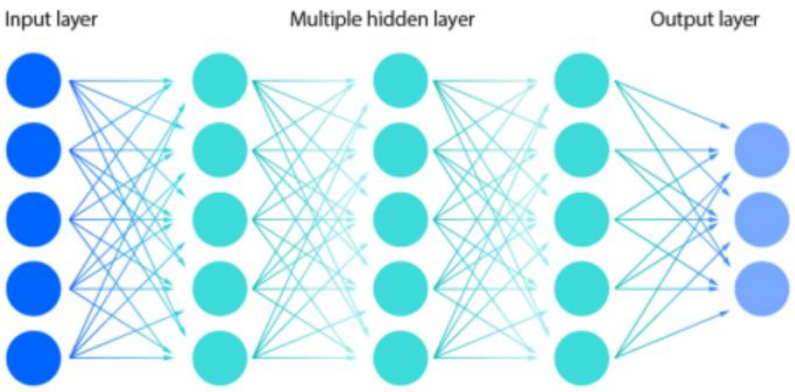


***Figure 9 :*** *Représente un réseau neuronal profond*

(*Que sont les réseaux neuronaux ? | IBM*, s. d.)

## 2.2 Applications de l'intelligence artificielle dans la recherche biomédicale :

### 2.2.1 Aperçu de l'IA dans les soins de santé :

L'intelligence artificielle (IA) est au cœur de la médecine du futur, elle est devenue un outil puissant dans le domaine de la recherche biomédicale, offrant des solutions innovantes pour la compréhension, le diagnostic, le traitement et la prévention des maladies, chirurgie assistée par ordinateur, anticipation d'une épidémie, triage des patients, Voici quelques-unes des applications les plus remarquables de l'IA dans ce domaine : (*5 applications de l'IA dans le domaine de la santé*, s. d.)

### 2.2.2 Applications spécifiques dans la recherche biomédicale :

Grâce à l'analyse de vastes quantités de données médicales, l'intelligence artificielle (IA) aide les médecins à diagnostiquer les maladies. L'IA peut analyser des images, des analyses, des dossiers médicaux et des informations sur les patients (symptômes, habitudes de vie) pour diagnostiquer plus précisément et plus rapidement, prédire l'évolution d'une maladie plus efficacement. Une fois ce problème résolu, l'IA permettra une médecine plus personnalisée et précise. (« Quatre applications de l'IA dans le domaine de la santé », 2021)

La reconnaissance d'images : l'IA optimise l'interprétation de l'imagerie médicale, offrant ainsi un avantage significatif dans le domaine du diagnostic radiologique. Cette technologie permet aux praticiens de gagner du temps, en les déchargeant de tâches fastidieuses et en leur permettant de se concentrer sur des aspects plus humains de leur pratique. En effet, les algorithmes d'IA sont capables de trier efficacement de vastes ensembles de données, guidant ainsi les radiologues vers les images les plus pertinentes à examiner en priorité, celles pouvant indiquer la présence de pathologies graves. (« Quatre applications de l'IA dans le domaine de la santé », 2021)

Personnalisation du traitement : en analysant les données génétiques et cliniques des patients on identifie les traitements les plus efficaces et les mieux adaptés. Cette approche individualisée permet une médecine plus personnalisée et prédictive, réduisant ainsi les risques d'effets

secondaires indésirables et améliorant les résultats thérapeutiques(« Quatre applications de l'IA dans le domaine de la santé », 2021)

Surveillance Médicale Intelligente et Suivi Personnalisé : Les dispositifs portables et les applications de santé connectée utilisent l'intelligence artificielle pour surveiller en continu les paramètres physiologiques des patients : la fréquence cardiaque, la glycémie et l'activité physique. Avec une surveillance constante, il est possible de détecter précocement les problèmes de santé et de gérer de manière proactive les maladies chroniques. (« Quatre applications de l'IA dans le domaine de la santé », 2021)

## 2.3 Concepts de l'apprentissage profond :

### 2.3.1 Aperçu des techniques d'apprentissage profond :

Avec les avancées technologiques et la croissance exponentielle des données, de nouvelles méthodes de calcul plus sophistiquées ont gagné en popularité. Cette évolution est alimentée à la fois par la demande croissante des consommateurs pour des produits de meilleure qualité et par la volonté des entreprises d'optimiser l'utilisation de leurs ressources. Dans ce contexte, l'intérêt pour le domaine de l'apprentissage automatique a été ravivé, attirant l'attention des chercheurs et des entreprises. (*Introduction to Deep Learning*, s. d.) (*Deep Learning Tutorial*, 2023)

L'apprentissage automatique, situé à l'intersection des statistiques, des mathématiques et de l'informatique, consiste à créer et à étudier des algorithmes capables d'améliorer leur propre performance de manière itérative. Initialement axé sur le développement de l'intelligence artificielle, ce domaine s'est progressivement recentré sur des tâches spécifiques en raison des limitations théoriques et technologiques de l'époque. Aujourd'hui, la plupart des algorithmes d'apprentissage automatique se concentrent sur l'optimisation des fonctions, bien que leurs solutions ne parviennent pas toujours à expliquer les tendances sous-jacentes dans les données ni à fournir le niveau d'inférence souhaité. (*An Introduction To Deep Learning*, s. d.)

Face à ces défis, l'apprentissage profond a émergé comme une réponse prometteuse. Ce sous-domaine de l'apprentissage automatique se concentre sur la création d'algorithmes capables de

comprendre et d'apprendre des niveaux d'abstraction élevés et bas des données, dépassant souvent les capacités des algorithmes traditionnels. Inspirés par diverses sources de connaissances telles que la théorie des jeux et la neuroscience, les modèles d'apprentissage profond imitent parfois la structure du système nerveux humain.

En permettant une représentation hiérarchique des données, les modèles d'apprentissage profond offrent une solution plus souple et généralisée aux problèmes complexes. Par exemple, dans le domaine de la reconnaissance d'images, ces modèles peuvent identifier des caractéristiques de bas niveau telles que les cils, puis les combiner pour reconnaître des entités plus complexes comme les visages ou les personnes. Cette capacité à apprendre des niveaux multiples de complexité ouvre la voie à des applications plus intelligentes, telles que l'autocorrection basée sur les schémas de parole individuels.

En résumé, l'apprentissage profond représente une avancée significative dans le domaine de l'intelligence artificielle, offrant des possibilités d'innovation dans une variété de domaines, de la vision par ordinateur à la traduction automatique. Son architecture, caractérisée par des couches d'unités non linéaires qui traitent les données à différents niveaux d'abstraction, marque une évolution importante dans notre capacité à modéliser et à comprendre des phénomènes complexes (Voir la figure 10).( (Beysolow II, s. d.)

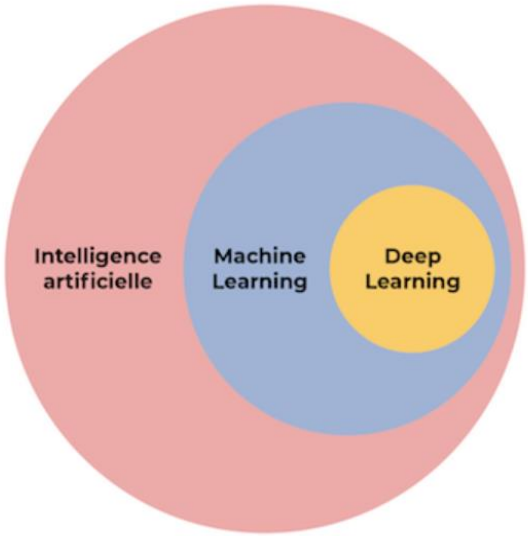


***Figure 10 :*** *Représente la relation entre l'IA, l'apprentissage automatique et l'apprentissage profond*

## 2.3.2 Architectures d'apprentissage profond :

L'architecture de l'AP est basée sur le domaine des réseaux neuronaux existe différents et plusieurs types, chaque type est développer pour un objectif particulier. Nous nous concentrerons ici sur le réseau convolutif, qui est l'un de ces types particuliers. (*Introduction à l'apprentissage profond (deep learning) de l'intelligence artificielle —*, s. d.)

(Boukhris Brahim & lBoughaba Mohammed Boukhris Brahim, s. d.) (Aliouane & BENDAHMANE Abdelhafedh, s. d.)

## Réseaux neuronaux convolutifs :

Dans le domaine de l'intelligence artificielle, les réseaux de neurones artificiels (RNA) ont révolutionné la façon dont les machines traitent et analysent l'information. Parmi les différents types de RNA, les réseaux de neurone convolutifs ou réseau de neurones à convolution, (CNN : Convolutional Neural Network). (*Qu'est-ce qu'un réseau de neurones convolutifs ?*, 2024) (Boukhris Brahim & lBoughaba Mohammed Boukhris Brahim, s. d.)

Un CNN est un type de réseau de neurones artificiels acycliques, composé de neurones disposés en trois dimensions : largeur, hauteur et profondeur.

Chaque neurone dans une couche donnée est connecté uniquement à une petite région de la couche précédente, ce qui permet aux CNN d'extraire efficacement des caractéristiques importantes des données d'entrée grâce à des opérations de convolution et de pooling.

(Beysolow II, s. d.).

Cette architecture permet aux CNN de démontré une grande efficacité dans une variété d' applications de vision par ordinateur qui réalise un apprentissage des tâches utilisé dans la reconnaissance faciale et le traitement des images, l'analyse médicale et l'analyse des pixels. Leur capacité à apprendre des caractéristiques pertinentes directement à partir des données brutes a considérablement amélioré les performances par rapport aux approches traditionnelles, ouvrant ainsi la voie à de nouvelles avancées dans le domaine de l'intelligence artificielle (Voir la figure 11).(Aliouane & BENDAHMANE Abdelhafedh, s. d.) (*Que signifie Réseau neuronal convolutif?*, s. d.)

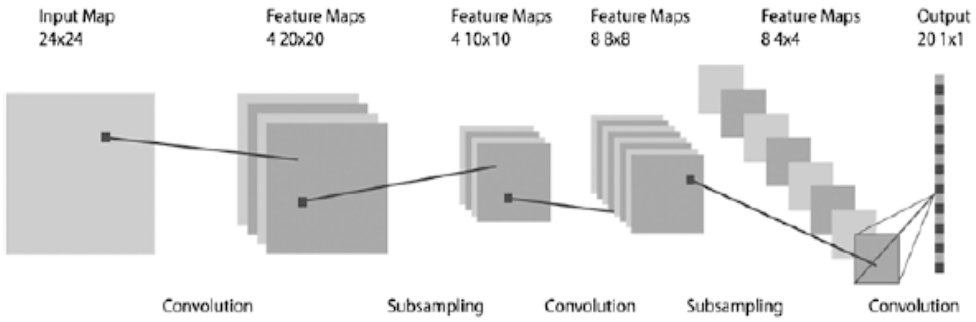


***Figure 11 :*** *Image qui représente les couches du CNN*

***Tableau 3 :*** *Représente la description de différents types de couches CNN*

| Couche de convolution (CONV): | Couche de pooling (POOL): | Couche de correction (ReLU): | Couche fully-connected |
|---|---|---|---|
| • sont les plus importantes, car elle est la composante principale du CNN. Composée de plusieurs neurones, Chaque neurone applique une opération de convolution à l'image d'entrée, ce qui permet d'extraire des caractéristiques locales de l'image.(salah)<br>• Donc son rôle est de détecter les caractéristiques des données reçues en entrée. | • Le pooling est une opération importante dans les CNNs placé entre deux couches de convolution, vise à réduire la dimension spatiale des images ce qui permet de diminuer le nombre de paramètres et de calculs nécessaires. Son rôle est d'optimiser l'efficacité et la robustesse des réseaux, tout en contribuant à leur performance globale. | • ReLU (Rectified Linear Units) sont des fonctions d'activation, introduisant des fonctions non-linéaires.<br>• ReLU se caractérise par sa rapidité d'apprentissage et sa contribution à la précision de la généralisation. | • La couche fully-connected représentent l'étape finale des CNNs et assurent le raisonnement de haut niveau du réseau**.** Ce type de couche permet au réseau d'intégrer les informations extraites par les couches convolutives et d'effectuer un raisonnement plus abstrait.<br>• Le calcul des fonctions d'activation dans les FC s'effectue par multiplication matricielle et décalage de biais, simplifiant le processus. |

(Boukhris Brahim & lBoughaba Mohammed Boukhris Brahim, s. d.)

(Aliouane & BENDAHMANE Abdelhafedh, s. d.)

## 2.4 Exemples d'applications d'apprentissage profond en bioinformatique et en prédiction du cancer :

L'apprentissage profond a innové le domaine de la bio-informatique en offrant des solutions innovantes à des problèmes complexes liés à l'analyse de données biologiques.(*A Review on the Application of Deep Learning in Bioinformatics*, s. d.)

Ses capacités d'apprentissage automatique à partir d'analyses de données complexes et de reconnaissance de motifs en font un outil puissant pour explorer de nouvelles recherches et développer de nouvelles applications dans divers domaines de la bio-informatique, en particulier dans la prédiction du cancer. (*A Review on the Application of Deep Learning in Bioinformatics*, s. d.)

### 2.4.1 Apprentissage profond en bioinformatique :

Des modèles ont entraîné des avancées remarquables dans divers domaines, tels que l'identification de mutations génétiques et la prédiction de la structure des protéines.

(*A Review on the Application of Deep Learning in Bioinformatics*, s. d.)

Par exemple : l'analyse des séquences en bio-informatique

L'analyse des séquences d'acides aminés et d'ADN constitue le socle fondamental de la bioinformatique, permettant de traiter et d'interpréter des données complexes pour décrypter le langage des acides aminés et des bases nucléiques. (Jsobel, 2014)

L'avènement de l'apprentissage profond a considérablement transformé cette discipline en offrant des outils d'une précision et d'une puissance exceptionnelle pour explorer les secrets cachés des séquences biologiques. (*Bioinformatique : Analyse des séquences de protéines | Techniques de l'Ingénieur*, s. d.)

La bioinformatique s'est distinguée de la modélisation numérique en mettant l'accent sur l'analyse de séquences plutôt que sur les simulations numériques, traçant ainsi sa propre voie dans le domaine de la recherche biologique.

Historiquement, l'identification des séquences de protéines a précédé celle de l'ADN, bien que plus complexe. Les protéines, constituées de 20 acides aminés aux propriétés physico-chimiques diverses, ont été examinées en premier lieu par rapport à l'ADN, qui est composé de seulement 4 nucléotides.(*Analyse des séquences de protéines*, s. d.)

La comparaison, la classification et la compréhension de la relation entre la séquence et la fonction des biomolécules ont joué un rôle moteur dans l'innovation en bioinformatique. Ces objectifs ont stimulé le développement d'outils et d'algorithmes visant à analyser et à interpréter ces données cruciales pour la compréhension du monde vivant. (*Analyse des séquences de protéines*, s. d.)

### 2.4.2 Apprentissage profond pour la prédiction et le diagnostic du cancer :

L'usage du deep learning en oncologie se concentre principalement sur la reconnaissance d'images, en particulier grâce aux réseaux CNN. Plusieurs exemples illustrent cette application dans le diagnostic du cancer. Par exemple, la détection des tumeurs cutanées est facilitée par l'apprentissage profond sur une base de données de photographies de la peau, » comme démontré dans une étude menée par Esteva et al. en 2017. » (Wan Zhu 1,2,*, et al., 2020)

Un algorithme de type CNN a réussi à distinguer les lésions malignes (carcinomes et mélanomes) des lésions bénignes (croissance non cancéreuse dans le corps.) avec une efficacité comparable à celle de 22 pathologistes. De même, l'application du deep learning par CNN s'étend à la reconnaissance d'images en radiologie médicale » comme le montre une étude sur des mammographies pour la détection de tumeurs mammaires menée par Ribli et al. en 2018. » (*Vaincre le cancer NRB*, s. d.)

Malgré l'importance croissante du deep learning dans le diagnostic du cancer en tant que méthode de détection visuelle, subsistent encore des zones d'ombre concernant l'efficacité de ces algorithmes à repérer des événements parfois imperceptibles à l'œil humain.

# PARTIE 02 : MATÉRIEL ET MÉTHODES

# 1. Matériel

## 1.1. Dataset utilisé

Le Dataset que j'ai utilisé dans cette recherche, provient de la page de BioStudies homepage qui est une source de données biologique de hautes qualités. Ce dataset comprend plusieurs fichiers, chacun à un rôle nécessaire dans cette étude.

Description des fichiers :

***Tableau 4 :*** *La description des fichiers du Dataset*

| Fichiers : | Taille : | Type : | Description : |
|---|---|---|---|
| 1. processedMatrix.Aurora.july2015.txt | 25.75 Go | Données traitées | - Contient les données traitées de l'étude.<br>- C'est la base de l'analyse, contient les informations essentielles sur les échantillons et les variables étudiées. |
| 2. E-MTAB-3732.idf.txt | 2 Ko | Format de conception de l'investigation (IDF) | - Contient le format de conception de l'investigation, fournissant des informations sur la conception expérimentale et les métadonnées associées à l'étude. |
| 3. E-MTAB-3732.sdrf.txt | 10.1 Mo | Format de relation entre les échantillons et les données (SDRF) | - Contient les labels nécessaires de l'analyse.<br>- Il est crucial pour l'identification et la classification des données dans l'analyse. |

## 1.2. Environnement de Travail

L'étude a été réalisée sur un ordinateur équipé d'un processeur Intel Ryzen 5 3600x, 32 Go de RAM et une carte graphique NVIDIA GeForce GTX 1060 à 6 Go de VRAM, fonctionnant sous le système d'exploitation Windows 10. Cette configuration a été choisie pour assurer une capacité de calcul suffisante pour les tâches d'entraînement du modèle de deep learning.

***Tableau 5 :*** *Les caractéristiques de l'ordinateur utilisé lors de l'apprentissage profond*

| | |
|---|---|
| Processeur | Processeur Intel Ryzen 5 3600x, 32 Go |
| RAM et une carte graphique | NVIDIA GeForce GTX 1060 à 6 Go de VRAM |
| Système d'exploitation | Windows 10 |
| Type de système | Système d'exploitation 64 bits |
| Version du système | 1903/ 18362.476 |

## 1.3. Logiciels et Bibliothèques

Le projet a été développé en Python 3.8, un langage de programmation largement utilisé en bioinformatique pour sa robustesse et ses nombreuses bibliothèques dédiées à la science des données et au machine learning. L'environnement de développement a été géré à l'aide d'Anaconda, une distribution Python spécialisée pour la science des données. Le travail a été réalisé dans Jupyter Notebook, un outil interactif qui permet de combiner code, visualisations et texte descriptif dans un même document.

- **Python :**

Python, créé par Guido van Rossum et publié pour la première fois en 1991, largement reconnu comme le langage de programmation le plus populaire et le plus rapide. Il est utilisé dans des domaines comme la Data Science et le Machine Learning et Web developement. (*Python □ : Focus sur le langage le plus populaire*, s. d.) (*What Is Python Used For? A Beginner's Guide | Coursera*, s. d.)

La syntaxe simple de Python le rend un langage facile à apprendre pour le débutant et aussi pour les développeurs expérimentés, tandis que sa capacité à gérer les modules et les packages simplifie et facilite la création de programmes complexes. (*Python □ : Focus sur le langage le plus populaire*, s. d.)

Python est un langage multiplateforme, ce qui signifie qu'il fonctionne sur divers systèmes d'exploitation tels que Windows, macOS et Linux, ce qui en fait un choix idéal pour que les développeurs travaillant sur différents environnements.

(*What Is Python Used For? A Beginner's Guide | Coursera*, s. d.)Tandis que sa nature open source bénéficie de la contribution continue de la communauté de développeurs, qui créent et

partagent régulièrement des bibliothèques et des fonctionnalités pour enrichir son écosystème. (*What is Python? | Teradata*, s. d.)

- **Anaconda :**

Anaconda est un outil en distribution **libre et open source** destinée à la programmation Python et R. Il est utilisé en sciences de données, en intelligence artificielle ou Machine Learning parce qu'il contient plusieurs packages fondamentaux tel que Python, Numpy, Pandas….etc. (*Anaconda pour Python - Présentation et installation | Jedha*, s. d.)

Ce logiciel permet la collection et la transformation des données grâce à ces outils. (*Anaconda : le guide complet pour l'installer et bien démarrer en Python*, s. d.)

- **Jupyter notebook:**

Jupyter Notebook, une application Web open source, permet de créer et partager des documents informatiques, équations, visualisations et texte. Jupyter offre un environnement puissant pour l'exploration et la communication des données. (Python, s. d.)

Les bibliothèques utilisées incluent, pandas 1.2.4 pour la gestion et l'analyse des données, numpy 1.19.5 pour les opérations numériques, et scikit-learn 0.24.1 pour les outils de prétraitement des données et les métriques de performance. Le modèle de deep learning a été construit et entraîné en utilisant TensorFlow 2.4.1 et Keras 2.4.3, des frameworks populaires pour le développement de modèles de réseaux de neurones. Les visualisations ont été réalisées avec seaborn 0.11.1 et matplotlib 3.3.4.

***Tableau 6 :*** *Définitions de bibliothèques utilisées*

| Bibliothèques : | Définitions : |
|---|---|
| • **Biopython :** | Biopython, une suite d'outils open source en langage Python, c'est le package de bioinformatique le plus populaire. (Aliouane & BENDAHMANE Abdelhafedh, s. d.) Il offre aux chercheurs en bioinformatique un environnement complet pour le développement de bibliothèques et d'applications sur l'analyse de données biologiques : séquences d'ADN et de protéines, structures 3D (fichiers PDB). |
| • **NumPy :** | NumPy est un outil indispensable en bioinformatique, crée pour permettre le calcul numérique avec Python. Il offre des fonctionnalités permettant de traiter et d'analyser efficacement les données complexes. NumPy facilite notamment les calculs vectoriels et matriciels, ce qui est essentiel pour manipuler les données biologiques et effectuer des opérations mathématiques avancées.<br>Grâce à sa simplicité d'utilisation et à sa puissance, NumPy est devenu un pilier dans le domaine de la bioinformatique, aidant les chercheurs à résoudre des problèmes complexes et à extraire des informations précieuses à partir des données biologiques.(*Introduction to NumPy*, s. d.) |
| • **Pandas :** | Pandas est une bibliothèque fabriquée pour le langage de programmation python. Pandas permet à python de de charger, d'aligner, de manipuler, d'analyser ou de fusionner les données. (*Pandas : la bibliothèque Python dédiée à la Data Science*, s. d.) |
| • **Matplotlib :** | Matplotlib est une bibliothèque puissante et largement utilisée en Python pour la visualisation de données. Elle offre la possibilité de créer une variété impressionnante de visualisations, allant des tracés simples aux histogrammes, diagrammes à barres, et bien d'autres types de graphiques, le tout avec seulement quelques lignes de code. Matplotlib est également utilisée dans divers contextes, que ce soit sur des serveurs d'application web, des shells ou des scripts Python. Ainsi, la maîtrise de Matplotlib est devenue une compétence essentielle pour tout Data Scientist.(*What Is Matplotlib In Python? How to use it for plotting? - ActiveState*, s. d.) |

# 2. Méthodes

## 2.1. Préparation des Données

Les données ont été téléchargées sous forme de fichier CSV intitulé "blood_cancer_labeled_5000.csv". Le fichier a été chargé dans un DataFrame pandas, une inspection initiale des étiquettes des catégories a révélé la nécessité d'exclure certains termes non associés au cancer, tels que "multiple_sclerosis_or_clinically_isolated_syndrome", "sepsis", "burn", "obesity", "acute_megakaryocytic_leukaemia__Down_syndrome", "AML_TLS", et "ATL". Les échantillons de ces catégories ont été filtrés, et les catégories avec moins de 20 échantillons ont également été exclues pour garantir une représentativité statistique adéquate. Les étiquettes restantes ont été binarisées en deux classes : "cancer" et "normal".

1) Lecture et Conversion de Données CSV :

Le code lit un fichier CSV et le charge dans un DataFrame pandas en utilisant la bibliothèque pandas (voir figure 12).

```
df = pd.read_csv("blood_cancer_labeled_5000.csv")
df
```

***Figure 12 :*** *Représente lecture et conversion de Données CSV*

La fonction df = pd.read_csv("blood_cancer_labeled_5000.csv") pour lire le fichier CSV et le charger dans un DataFrame.

2) Calcule et affichage de la Répartition des Catégories dans la Colonne 'label' :

Le code calcule et affiche la distribution des catégories dans la colonne label du DataFrame df (voir figure 13).

```
category_counts = df['label'].value_counts()
# pd.set_option('display.max_rows', None)
category_counts
```

***Figure 13 :*** *Code pour calculer et afficher la Répartition des Catégories dans la Colonne 'label'*

La fonction df['label'].value_counts() pour compter le nombre d'occurrences de chaque catégorie dans la colonne label du DataFrame df.

La fonction category_counts pour afficher le contenu de category_counts, qui est une série Pandas contenant les comptes des catégories dans la colonne label.

Le résultat est une catégorie unique de la colonne label et chaque valeur est le nombre d'occurrences de cette catégorie (Voir la figure 14).

```
normal                                                    1823
multiple_sclerosis_or_clinically_isolated_syndrome         237
chronic_lymphocytic_leukaemia                              194
septic_shock                                               163
acute_monocytic_leukaemia                                  103
kidney_transplant_recipient                                 82
sepsis                                                      78
burn                                                        74
obesity                                                     61
acute_myeloid_leukaemia                                     53
multiple_myeloma                                            52
acute_graft_versus_host_disease                             49
T_cell_leukaemia                                            40
acute_lymphoblastic_leukaemia                               39
myeloma                                                     39
systemic_inflammatory_response_syndrome                     38
erythroleukaemia                                            34
B_ALL                                                       34
Job_s_syndrome                                              34
Kawasaki_disease                                            31
clinically_isolated_syndrome                                29
meningococcal_sepsis                                        29
ATL                                                         28
moderate_chronic_allograft_nephropathy                      27
autism_spectrum_disorder                                    27
```

***Figure 14 :*** *Répartition des Catégories*

3) Les données des étiquettes doivent être converti, par conséquent on utilise la fonction df['label'] = df['label'].apply(lambda x: 'cancer' if x != 'normal' else x) qui convertit les étiquettes en catégories binaires : ‘cancer’ et ‘normal’ (voir figure 15).

```
df['label'] = df['label'].apply(lambda x: 'cancer' if x != 'normal' else x)
df
```

***Figure 15 :*** *Conversion des étiquettes*

## 2.2. Architecture du Modèle

Un modèle de réseau de neurones convolutif 1D (Conv1D) a été choisi pour son aptitude à capturer les motifs locaux dans les données d'expression génique. Le modèle commence par une couche Reshape pour ajuster les dimensions d'entrée à (5000, 1), suivie de deux couches Conv1D avec respectivement 16 et 32 filtres, des tailles de kernel de 9 et 7, et des activations ReLU. Chaque couche de convolution est suivie d'une couche de pooling (MaxPooling1D) pour réduire la dimensionnalité spatiale. Le modèle se termine par une couche Flatten pour aplatir les caractéristiques extraites, une couche dense avec 256 unités et une activation ReLU, une couche Dropout avec un taux de 0.7 pour éviter le surapprentissage, et une couche de sortie dense avec une activation sigmoïde pour la classification binaire (voir la figure 16).

```
Model: "sequential_10"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
reshape_10 (Reshape)         (None, 5000, 1)           0
_________________________________________________________________
conv1d_20 (Conv1D)           (None, 5000, 16)          160
_________________________________________________________________
batch_normalization_30 (Batc (None, 5000, 16)          64
_________________________________________________________________
activation_30 (Activation)   (None, 5000, 16)          0
_________________________________________________________________
max_pooling1d_20 (MaxPooling (None, 1666, 16)          0
_________________________________________________________________
conv1d_21 (Conv1D)           (None, 1666, 32)          3616
_________________________________________________________________
batch_normalization_31 (Batc (None, 1666, 32)          128
_________________________________________________________________
activation_31 (Activation)   (None, 1666, 32)          0
_________________________________________________________________
max_pooling1d_21 (MaxPooling (None, 555, 32)           0
_________________________________________________________________
flatten_10 (Flatten)         (None, 17760)             0
_________________________________________________________________
dense_20 (Dense)             (None, 256)               4546816
...
```

***Figure 16 :*** *Représente l'architecture du modèle de réseau neuronal convolutif 1D (Conv1D) pour la prédiction du cancer du sang*

## 2.3. Entraînement du Modèle

L'entraînement du modèle a été effectué en utilisant un GPU pour accélérer le processus. Les données ont été divisées en ensembles d'entraînement (80%) et de test (20%) en utilisant la méthode train_test_split de scikit-learn (voir figure 17).

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
```

***Figure 17 :*** *utilisation de la fonction train_test_split pour diviser les données*

Les données ont été normalisées avec MinMaxScaler pour s'assurer que toutes les caractéristiques sont sur une échelle comparable. Le modèle a été compilé avec l'optimiseur Adamax et un taux d'apprentissage initial de 0.001. La fonction de perte utilisée est la binary crossentropy, et la métrique d'évaluation est l'accuracy. Le modèle a été entraîné pendant 40 époques avec un batch size de 8 (Voir la figure 18).

```
history = model.fit(
    X_train, y_train,
    epochs=40,
    batch_size=8,
    validation_data=(X_test, y_test),
)
```

***Figure 18 :*** *Code utilisé pour l'entrainement du model*

## 2.4. Évaluation du Modèle

Les performances du modèle ont été évaluées en termes de précision, de matrice de confusion et de rapport de classification (précision, rappel, F1-score). La matrice de confusion a été normalisée et visualisée à l'aide des bibliothèques seaborn et matplotlib pour une interprétation plus claire des résultats (Voir la figure 19).

```
train_pred = model.predict(X_train)
test_pred = model.predict(X_test)

train_pred_binary = (train_pred > 0.5).astype(int).flatten()
test_pred_binary = (test_pred > 0.5).astype(int).flatten()

y_train_binary = y_train.flatten()
y_test_binary = y_test.flatten()

train_acc = accuracy_score(y_train_binary, train_pred_binary)
test_acc = accuracy_score(y_test_binary, test_pred_binary)

print("train-acc = " + str(train_acc))
print("test-acc = " + str(test_acc))

cm = confusion_matrix(y_test_binary, test_pred_binary)
```

***Figure 19 :*** *Code utilisé pour l'évaluation des performances du modèle de prédiction.*

# PARTIE 3 : RÉSULTATS ET DISCUSSION

# 1. Résultats

L'entraînement du modèle de réseau de neurones convolutif 1D (Conv1D) a montré des performances prometteuses pour la prédiction du cancer du sang à partir des données d'expression génique. Les données ont été divisées en ensembles d'entraînement (80%) et de test (20%), permettant une évaluation rigoureuse du modèle.

## 1.1. Précision du Modèle

Le modèle a atteint une précision de 99.45% sur l'ensemble d'entraînement et de 98.8% sur l'ensemble de test indiquant une bonne capacité de généralisation du modèle aux nouvelles données (Voir la figure 20).

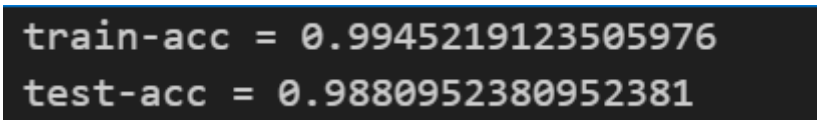


***Figure 20 :*** *Résultat des prédictions du model appliqué sur les données d'entrainement et les données du test*

## 1.2. Matrice de Confusion

La matrice de confusion pour l'ensemble de test a été utilisée pour évaluer plus finement les performances du modèle. Les valeurs de la matrice de confusion montrent un taux élevé de vraies prédictions positives et négatives, avec peu de faux positifs et de faux négatifs (voir figure 21).

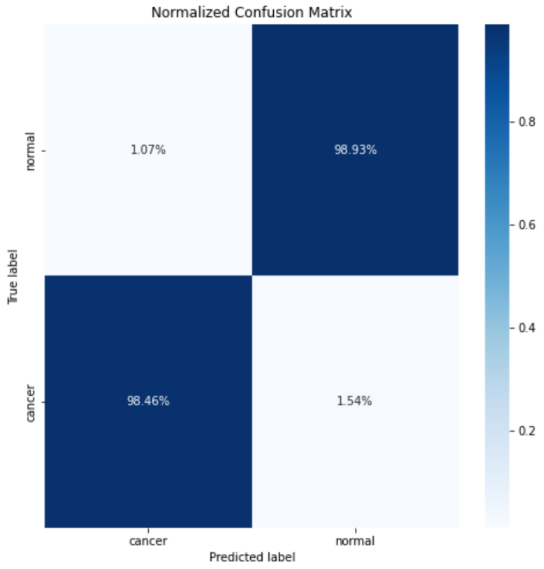


***Figure 211 :*** *Matrice de confusion représentant les performances du model.*

La matrice de confusion montre que le modèle a correctement prédit 98.93% des échantillons portant l'étiquette "normal" et 98.46% des échantillons portant l'étiquette "Cancer". Il y a eu 1.07% de faux positifs (échantillons prédits comme "Cancer" alors qu'ils étaient "normal") et 1.54% de faux négatifs (échantillons prédits comme "normal " alors qu'ils étaient "Cancer").

## 1.3. Rapports de Classification

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| cancer | 0.97 | 0.98 | 0.98 | 65 |
| normal | 0.99 | 0.99 | 0.99 | 187 |
| accuracy | | | 0.99 | 252 |
| macro avg | 0.98 | 0.99 | 0.98 | 252 |
| weighted avg | 0.99 | 0.99 | 0.99 | 252 |

***Figure 22 :*** *Représente les rapports de classification*

# 2. Discussion

Les résultats indiquent que le modèle Conv1D est efficace pour prédire le cancer du sang à partir des données d'expression génique. La haute précision observée sur les ensembles de validation et de test suggère que le modèle est bien entraîné et peut généraliser aux nouvelles données. Toutefois, les faux positifs et faux négatifs, bien que peu nombreux, nécessitent une attention particulière.

## 2.1. Interprétation des Résultats

La précision élevée (98.8%) sur l'ensemble de test indique une forte performance prédictive, mais les faux positifs et négatifs montrent qu'il y a encore de la marge pour l'amélioration. Les faux positifs pourraient entraîner des alarmes inutiles et des traitements inutiles, tandis que les faux négatifs pourraient conduire à des manques de diagnostic de cancer. Cela souligne l'importance d'un modèle encore plus robuste et sensible, en particulier pour les applications cliniques.

## 2.2. Améliorations Potentielles

- Augmentation des Données : En intégrant davantage de données d'expression génique et en diversifiant les sources de données, le modèle pourrait bénéficier d'une meilleure représentativité des différentes variations génétiques.
- Optimisation des Hyperparamètres : Une recherche plus approfondie des hyperparamètres pourrait améliorer la performance du modèle. Des techniques comme la recherche en grille (grid search) ou la recherche bayésienne (bayesian optimization) pourraient être utilisées pour trouver les paramètres optimaux.
- Ensemble Learning : Combiner plusieurs modèles de machine learning (ensemble learning) pourrait renforcer la robustesse du modèle et réduire les erreurs de prédiction.
- Feature Engineering Avancé : L'extraction de caractéristiques avancées et l'utilisation de techniques de réduction de dimensionnalité comme PCA (Principal Component Analysis) ou LDA (Linear Discriminant Analysis) pourraient aider à améliorer la précision du modèle.

## 2.3. Validations Futures

Il serait bénéfique de valider le modèle sur des jeux de données indépendants pour évaluer sa performance réelle en situation clinique. La collaboration avec des laboratoires cliniques pour tester le modèle sur des échantillons réels pourrait fournir des insights précieux et aider à affiner encore plus le modèle.

# CONCLUSION

# CONCLUSION :

En résumé, ce mémoire met en évidence le potentiel significatif des méthodes de deep learning pour la prédiction du cancer et démontre l'efficacité des réseaux de neurones convolutifs pour l'analyse des données d'expression génique. Les résultats obtenus montrent que ces modèles peuvent fournir des prédictions précises, ouvrant ainsi de nouvelles opportunités pour le diagnostic et le traitement personnalisé des cancers.

Cependant, des études supplémentaires sont nécessaires pour améliorer encore la précision de ces modèles et explorer leur application à d'autres types de données biomédicales. Il est également crucial de se concentrer sur l'incorporation de ces modèles dans des systèmes cliniques pour une utilisation pratique et efficace.

Les perspectives futures incluent l'intégration de plus de données pour enrichir les modèles, l'optimisation des techniques d'apprentissage pour une meilleure performance, et la validation sur des échantillons cliniques réels afin de renforcer la fiabilité et l'applicabilité du modèle. De plus, l'exploration de nouvelles architectures de réseaux neuronaux et l'utilisation de techniques d'ensemble learning pourraient offrir des améliorations supplémentaires. Enfin, une collaboration accrue avec des professionnels de santé pourrait faciliter l'adoption de ces technologies dans les pratiques cliniques quotidiennes, augmentant ainsi leur impact sur le diagnostic et le traitement du cancer.

# RÉFÉRENCES BIBLIOGRAPHIQUES

# REFERENCES


*5 applications de l'IA dans le domaine de la santé*. (s. d.). Consulté 31 mai 2024, à l'adresse https://www.cscience.ca/5-applications-de-lia-en-sante/

*A Review on the Application of Deep Learning in Bioinformatics*. (s. d.). Consulté 31 mai 2024, à l'adresse https://www.researchgate.net/publication/342875026_A_Review_on_the_Application_of_Deep_Learning_in_Bioinformatics

Aliouane, S. E., & BENDAHMANE Abdelhafedh. (s. d.). *Nouvelle approche de prédiction des classes protéiques issues d'un séquençage NGS par Deep Learning*. Université Frères Mentouri Constantine 1.

*An Introduction To Deep Learning*. (s. d.). Simplilearn.Com. Consulté 31 mai 2024, à l'adresse https://www.simplilearn.com/tutorials/deep-learning-tutorial/introduction-to-deep-learning

*Anaconda : Le guide complet pour l'installer et bien démarrer en Python*. (s. d.). Consulté 31 mai 2024, à l'adresse https://www.data-transitionnumerique.com/anaconda-python/

*Anaconda pour Python—Présentation et installation | Jedha*. (s. d.). Consulté 31 mai 2024, à l'adresse https://www.jedha.co/formation-python/ananconda-python

*Analyse des séquences de protéines*. (s. d.). Techniques de l'Ingénieur. Consulté 31 mai 2024, à l'adresse https://www.techniques-ingenieur.fr/base-documentaire/biomedical-pharma-th15/sante-numerique-et-connectee-42628210/bioinformatique-bio7050/analyse-des-sequences-de-proteines-bio7050niv10002.html

*Aperçu de la transcription (leçon) | Khan Academy*. (s. d.). Consulté 31 mai 2024, à l'adresse https://fr.khanacademy.org/_render

*Applications of Machine Learning—Javatpoint*. (s. d.). Consulté 31 mai 2024, à l'adresse https://www.javatpoint.com/applications-of-machine-learning

Beysolow II, T. (s. d.). *Introduction to Deep Learning Using R A Step-by-Step Guide to Learning and Implementing Deep Learning Models Using R /*.

*Bioinformatique : Analyse des séquences de protéines | Techniques de l'Ingénieur*. (s. d.). Consulté 31 mai 2024, à l'adresse https://www.techniques-ingenieur.fr/base-documentaire/biomedical-pharma-th15/sante-numerique-et-connectee-42628210/bioinformatique-bio7050/analyse-des-sequences-de-proteines-bio7050niv10002.html

Boukhris Brahim & lBoughaba Mohammed  Boukhris Brahim. (s. d.). *L`apprentissage profond (Deep Learning) pour la classification et la recherche d'images par le contenu* [UNIVERSITE KASDI MERBAH OUARGLA]. https://dspace.univ-ouargla.dz/jspui/bitstream/123456789/17195/1/Boughaba_Boukhris.pdf

Capp, J.-P. (2011). Chapitre 6. Le rôle de l'expression aléatoire des gènes dans la genèse du cancer. In *Le hasard au cœur de la cellule* (p. 174-210). Éditions Matériologiques. https://www.cairn-sciences.info/le-hasard-au-coeur-de-la-cellule--9782919694341-page-174.htm

Choudhuri, S. (2014). *Bioinformatics for Beginners : Genes, Genomes, Molecular Evolution, Databases and Analytical Tools*. Elsevier.

Contrôle de la transcription. (2021, novembre 13). *Biologie cellulaire et génétique du Développement*. https://bcgdevelop.fr/controle-de-la-transcription/

*Cours : Théorie de la traduction*. (s. d.). Consulté 31 mai 2024, à l'adresse http://foad.ugb.sn/course/view.php?id=590

*Deep Learning Tutorial*. (2023, avril 10). GeeksforGeeks. https://www.geeksforgeeks.org/deep-learning-tutorial/

Dillenburg, F. C. (2017). *An approach for analyzing and classifying microarray data using gene co-expression networks cycles*. https://lume.ufrgs.br/handle/10183/171353

*DNA Microarrays and Gene Expression : From Experiments to Data Analysis and Modeling - Baldi, Pierre; Hatfield, G. Wesley: 9780521800228 - AbeBooks*. (s. d.). Consulté 31 mai 2024, à l'adresse https://www.abebooks.fr/9780521800228/DNA-Microarrays-Gene-Expression-Experiments-0521800226/plp

*Expression génétique : Définition & étapes | StudySmarter*. (s. d.). Consulté 31 mai 2024, à l'adresse https://www.studysmarter.fr/resumes/biologie/svt/expression-genetique/

*Gene Expression*. (s. d.). Consulté 31 mai 2024, à l'adresse https://www.genome.gov/genetics-glossary/Gene-Expression

*GeneChip™ Human Genome U133 Plus 2.0 Array*. (s. d.). Consulté 31 mai 2024, à l'adresse https://www.thermofisher.com/order/catalog/product/900466

*Intelligence artificielle : Définition et utilisations | NetApp*. (s. d.). Consulté 31 mai 2024, à l'adresse https://www.netapp.com/fr/artificial-intelligence/what-is-artificial-intelligence/

*Intelligence Artificielle : Définition, histoire, enjeux*. (s. d.-a). Consulté 31 mai 2024, à l'adresse https://datascientest.com/intelligence-artificielle-definition

*Intelligence Artificielle : Définition, histoire, enjeux*. (s. d.-b). Consulté 31 mai 2024, à l'adresse https://datascientest.com/intelligence-artificielle-definition

*#Intelligence artificielle en #imagerie cardiovasculaire – Révolution actuelle et future « maismaismedicina*. (s. d.). Consulté 31 mai 2024, à l'adresse https://maismaismedicina.wordpress.com/2020/06/25/intelligence-artificielle-en-imagerie-cardiovasculaire-revolution-actuelle-et-future/

*Introduction à l'apprentissage profond (deep learning) de l'intelligence artificielle—*. (s. d.). Consulté 31 mai 2024, à l'adresse https://culturesciencesphysique.ens-lyon.fr/ressource/IA-apprentissage-Rousseau.xml

*Introduction to Deep Learning*. (s. d.). Consulté 31 mai 2024, à l'adresse https://www.linkedin.com/pulse/introduction-deep-learning-ilija-mihajlovic

*Introduction to NumPy*. (s. d.). Consulté 31 mai 2024, à l'adresse https://www.w3schools.com/python/numpy/numpy_intro.asp

Jsobel. (2014, février 19). Il était une fois... L'analyse de séquences d'ADN. *Bioinfo-fr.net*. https://bioinfo-fr.net/il-etait-une-fois-lanalyse-de-sequences-dadn

*Les bases de la biologie moléculaire—Principe de la traduction*. (s. d.). Consulté 31 mai 2024, à l'adresse https://www.supagro.fr/ress-tice/ue1-ue2_auto/Bases_Biologie_Moleculaire_v2/co/_gc_principe_traduction.html

*Les étapes de la transcription (leçon) | Khan Academy*. (s. d.). Consulté 31 mai 2024, à l'adresse https://fr.khanacademy.org/science/biologie-a-l-ecole/x5047ff3843d876a6:bio-6e-

annee-sciences-de-base/x5047ff3843d876a6:bio-6-1h-adn-arn-et-expression-d-un-gene/a/stages-of-transcription

*L'IA - concepts et histoire—MetalBlog*. (s. d.). Consulté 31 mai 2024, à l'adresse https://metalblog.ctif.com/2023/10/23/lintelligence-artificielle-concepts-et-histoire/

*L'information génétique et la molécule d'ADN - 2nde—Cours SVT - Kartable*. (s. d.). Consulté 31 mai 2024, à l'adresse https://www.kartable.fr/ressources/svt/cours/linformation-genetique-et-la-molecule-dadn/18908

Marti, J., Piquemal, D., Manchon, L., & Commes, T. (2002). Étude des transcriptomes par analyse sérielle de l'expression des gènes. *Journal de la Société de Biologie*, *196*(4), Article 4. https://doi.org/10.1051/jbio/2002196040303

Mathivet, V. (2017). *L'intelligence artificielle pour les développeurs : Concepts et implémentations en C#*. Éditions ENI.

*Pandas : La bibliothèque Python dédiée à la Data Science*. (s. d.). Consulté 31 mai 2024, à l'adresse https://datascientest.com/pandas-python-data-science

*Profilage de l'expression génique*. (s. d.). Consulté 31 mai 2024, à l'adresse https://french.longdom.org/scholarly/gene-expression-profiling-journals-articles-ppts-list-104.html

Puce à ADN. (2024). In *Wikipédia*. https://fr.wikipedia.org/w/index.php?title=Puce_%C3%A0_ADN&oldid=211940572

*Puce à ADN : pourquoi et pour qui ?* (s. d.). Consulté 31 mai 2024, à l'adresse https://www.revmed.ch/revue-medicale-suisse/2010/revue-medicale-suisse-237/puce-a-adn-pourquoi-et-pour-qui

*Python : Focus sur le langage le plus populaire*. (s. d.). Consulté 31 mai 2024, à l'adresse https://datascientest.com/python-tout-savoir

Python, R. (s. d.). *Jupyter Notebook : An Introduction – Real Python*. Consulté 31 mai 2024, à l'adresse https://realpython.com/jupyter-notebook-introduction/

*Quantification et expression d'un gène*. (s. d.). Biomnigene. Consulté 31 mai 2024, à l'adresse https://www.biomnigene.fr/fr/vos-besoins/quantification/quantification-et-expression-d-un-gene.html

Quatre applications de l'IA dans le domaine de la santé. (2021, janvier 19). *Calmedica*. https://www.calmedica.com/quatre-applications-de-lia-dans-le-domaine-de-la-sante/

*Que signifie Réseau neuronal convolutif? - Definition IT de LeMagIT*. (s. d.). LeMagIT. Consulté 31 mai 2024, à l'adresse https://www.lemagit.fr/definition/Reseau-neuronal-convolutif

*Que sont les réseaux neuronaux ? | IBM*. (s. d.). Consulté 31 mai 2024, à l'adresse https://www.ibm.com/fr-fr/topics/neural-networks

*Qu'est-ce qu'un réseau de neurones convolutifs ? | IBM*. (2024, mai 29). https://www.ibm.com/fr-fr/topics/convolutional-neural-networks

Robert, J. (2020, novembre 18). Machine Learning : Définition, fonctionnement, utilisations. *Formation Data Science | DataScientest.com*. https://datascientest.com/machine-learning-tout-savoir

*Séquençage d'ARN - Séquençage de nouvelle génération—GENEWIZ*. (s. d.). Consulté 31 mai 2024, à l'adresse https://www.genewiz.com/fr-FR/Public/Services/Next-Generation-Sequencing/RNA-Seq/

*Traduction ADN : Cours et explications │ StudySmarter*. (s. d.). StudySmarter FR. Consulté 31 mai 2024, à l'adresse https://www.studysmarter.fr/resumes/biologie/svt/traduction-adn/

*Transcription ADN : cours et explications │ StudySmarter*. (s. d.). Consulté 31 mai 2024, à l'adresse https://www.studysmarter.fr/resumes/biologie/svt/transcription-adn/

*Vaincre le cancer NRB : L'essor de l'apprentissage profond dans le diagnostic du cancer*. (s. d.). Consulté 31 mai 2024, à l'adresse https://www.vaincrelecancer-nrb.org/nos-actualites/articles/2020/lessor-de-lapprentissage-profond-dans-le-diagnostic-du-cancer.html

Wan Zhu 1,2,*, Longxiang Xie 1,y, Jianye Han 3, & Xiangqian Guo 1,*. (2020, février 4). The Application of Deep Learning in Cancer Prognosis Prediction. *5 March 2020,* 19.

*What Is Matplotlib In Python ? How to use it for plotting ? - ActiveState*. (s. d.). Consulté 31 mai 2024, à l'adresse https://www.activestate.com/resources/quick-reads/what-is-matplotlib-in-python-how-to-use-it-for-plotting/

*What is Python ? | Teradata*. (s. d.). Consulté 31 mai 2024, à l'adresse https://www.teradata.com/glossary/what-is-python

*What Is Python Used For ? A Beginner's Guide | Coursera*. (s. d.). Consulté 31 mai 2024, à l'adresse https://www.coursera.org/articles/what-is-python-used-for-a-beginners-guide-to-using-python

| Soutenu le : | Présenté par : |
|---|---|
| 10/06/2024 | BENCHAOUI Fella Mounira |

**Thème :**

**Prédiction de profil de cancer à partir des données d'expression des gènes basée sur le Deep learning**

**Mémoire Présenté en vue de l'obtention du Diplôme de Master en :**

**Bioinformatique**

**Domaine : Science de la nature et la vie**

**Département de Biologie Appliquée**


Ce mémoire présente le développement d'un modèle d'apprentissage profond basé sur un réseau de neurones convolutif unidimensionnel (Conv1D) pour prédire le cancer du sang à partir de données d'expression génique. Le dataset comprend des données d'expression génique, annotées par plusieurs types de cancer du sang et des cas normaux.

Le travail a suivi plusieurs étapes clés : préparation des données, conception et entraînement du modèle Conv1D, évaluation des performances et interprétation des résultats. Les données ont été divisées en ensembles d'entraînement (80%) et de test (20%). Le modèle a atteint une précision de 99.45% sur l'ensemble d'entraînement, 98.8% sur l'ensemble de test, montrant une bonne capacité de généralisation.

La matrice de confusion a révélé une forte proportion de prédictions correctes avec peu de faux positifs et de faux négatifs. Les résultats ont été analysés pour identifier les gènes les plus significatifs dans la classification des différents types de cancer du sang.

Cette étude montre que les réseaux de neurones convolutifs unidimensionnels peuvent efficacement classer les données d'expression génique, offrant un outil puissant pour le diagnostic et la recherche sur les cancers hématologiques.



**Mots clés :** intelligence artificielle, apprentissage profond, réseau de neurones convolutif unidimensionnel, expression génique, classification, cancer du sang, biomédecine.


**Jury d'évaluation :**

**Président du jury : Pr. HAMIDECHI M. Abdelhafid**

**Encadreur : Dr. DAAS.S**

**Examinateur : Pr. BELLIL Ines**