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 I 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 : Sciences biologiques
Spécialité : *Bioinformatique*

N° d'ordre :
N° de série :

Intitulé :

Annotation des SNPs génomiques pour la prédiction de la maladie génétique/héréditaire d'Alzheimer par Deep Learning

Présenté par : ABDELAZIZ Aya
NOUI Manel Ghosn El Ben

Le 19/06/2022

Jury d'évaluation :

Président : Dr CHEHILI Hamza (Université Frères Mentouri, Constantine 1).
Encadreur : Pr HAMIDECHI M. Abdelhafid (Université Frères Mentouri, Constantine 1).
Examinateur : Dr BOULAHROUF Khaled (Université Frères Mentouri, Constantine 1).

Année universitaire
2021 - 2022

REMERCIEMENTS

Nous remercions Dieu tout puissant de nous avoir données la santé, la force et la volonté d'entamer et de terminer ce mémoire de fin d'étude.

Nous exprimons nos profonds remerciements à notre Encadreur et Directeur de mémoire, le professeur Mohamed Abdelhafid HAMIDECHI, un superviseur très généreux, inlassablement dévoué à son équipe et engagé à sa réussite. Nous le remercions pour sa patience, sa disponibilité et surtout ses sages conseils qui ont contribué à notre réflexion et sans qui ce travail n'aurait pas pu être arrivé à bon port. Nous le remercions d'avoir eu la patience de répondre à nos questions et pour le temps qu'il nous a consacré tout au long de cette expérience combien enrichissante pour nous deux.

Nous sommes très heureuses et satisfaites de travailler avec vous durant cette expérience de mémoire de fin de cycle et surtout nous apprécions la richesse scientifique et la très grande qualité de votre encadrement, vos talents de Professeur qualifié, vos grandes qualités d'expérimentateur, votre incroyable capacité à trouver une solution à chaque problème et votre disponibilité malgré vos responsabilités académiques et professionnelles. Enfin, nous apprécions particulièrement l'esprit de rigueur, de qualité et d'honnêteté scientifique que vous n'avez cessé d'insuffler dans ce travail. Votre enthousiasme a été le moteur tout au long de notre travail.

Un grand merci également aux membres du jury. Le D^{r} CHEHILI Hamza, Vice-Recteur chargé de la formation supérieure de post-graduation, de l'habilitation universitaire et de la recherche scientifique, président de notre jury et ayant l'amabilité d'accepter d'expertiser et de valoriser ce modeste travail. Nous lui exprimons par le biais de cet aboutissement nos meilleurs remerciements et nos profonds respects vis-à-vis des enseignements et des efforts dévoués de sa part et de sa sincérité eu moment de l'apprentissage et en dehors de celui-ci durant toute la période de enseignements donnés à notre promotion de Bioinformatique. Vous avez été un enseignant modèle que les mots n'arriveront sans doute jamais à égaler votre valeur intrinsèque que nous tous apprécions chez vous.

Le D^{r} BOULAHROUF Khaled a fort aimablement accepté d'expertiser ce travail et ce en dépit des charges pédagogiques et des tâches administratives qu'il assure et qu'il exerce très consciencieusement. Nous le remercions pour l'intérêt qu'il a exprimé pour notre travail en acceptant d'évaluer notre mémoire de fin d'études et de l'enrichir par ses compétences avérées en tant qu'enseignant afin de lui apporter une valeur ajoutée positive pour développer encore plus le contenu et la vision de ce modeste mémoire de fin d'études.

Nous tenons à remercier particulièrement Monsieur ALIOUANE Salah Eddine. Le Pr HAMIDECHI M. A., notre Directeur de travail, a vu en lui la personne qui peut, sans le moindre doute, apporter une contribution pratique de ce travail de mémoire, car connaissant ses compétences avérées en Bioinformatique et ses engagements d'honneur à faire aboutir les travaux pour lesquels il s'est engagé. Il nous a fourni les outils et les explications nécessaires et utiles pour l'aboutissement et le succès de notre travail. Une aide précieuse voire généreuse, des explications claires et précises qui nous ont conduites à initier le chemin de la recherche scientifique.

Nous tenons à exprimer nos sincères remerciements à tous nos Enseignants surtout pour leurs compétences et pour avoir soutenu notre promotion de Bioinformatique durant la période de nos études de Tronc-Commun, de Licence et de Master.

Nos profondes pensées vont vers nos familles, et surtout à nos parents, qui nous ont encouragées à poursuivre nos études jusqu'à ce joyeux jour par lequel nous espérons leur rendre joie et bonheur comme simple récompense de leurs intérêts envers nous et de la qualité de l'éducation enseignée durant tout cet âge de formation sans relâche. Un énorme merci pour leur soutien indéfectible et leurs encouragements tout au long de l'élaboration de notre travail.

Enfin, Merci à tous ceux qui, de près ou de loin, ont contribué à la réalisation de ce travail.

RÉSUMÉ

La maladie d'Alzheimer (MA) est une maladie irréversible dans le cerveau qui provoque des troubles neurodégénératifs progressifs. La MA peut être détectée en dépistant des gènes spécifiques dans des chromosomes spécifiques responsables de cette maladie où se trouvent les mutations qui se produisent dans les SNPs sur ces gènes spécifiques. Notre travail a mis en exergue l'importance de l'utilisation des données de la BDD ADNI pour étudier et analyser la MA, en développant une approche bioinformatique d'apprentissage approfondi pour classer les deux stades de la maladie (MA et CN). L'objectif de cette approche est de développer un modèle d'annotation (ou un système de prédiction) afin de d'identifier le type de la maladie chez des individus ou d'estimer le stade de celle-ci, à l'aide d'un Réseau de Neurones Artificiel, comme résultats, nous avons classé les données génétiques fonctionnelles avec une précision de test qui a atteint 94%.

Mots clés : Maladie d'Alzheimer ; ADNI ; SNPs ; Apprentissage approfondi ; Réseau de Neurones Artificiel ; Annotation.

ABSTRACT

Alzheimer disease (AD) is an irreversible disease in the brain that causes progressive neurodegenerative disorders. AD can be detected by detecting specific genes on specific chromosomes that cause the disease and by finding mutations that occur in the SNPs of these specific genes. Our work emphasizes the importance of studying and analyzing AD using ADNI-DB data, and deep learning to classify the two stages of disease (AD and CN). The goal of this approach is to develop an annotation model (or predictive system) for identifying the disease type of an individual or for estimating the stage of the disease using artificial neural networks. As a result, we classified functional genetic data with test accuracy of 94%.

Key words : Alzheimer disease; SNPs ; ADNI ; Deep learning ; Artificial Neural Networks ; Annotation.

الملخص

مرض الزهايمر هو مرض غير رجعي على مستوي الدماغ والَّذِي يسبب اضطرابات تنكسيه عصبية تدريجية. مرض الزهايمر يمكن اكتشافه عن طريق تتبع جينات معينة على كروموسومات معينة مسؤولة عن هدا المرض اين توجد طفرات تحدث في تعدد أشكال النوكليوتيدات المفردة (SNPs) لهذه الجينات المحددة. عملنا يسلط الضوء على أهمية استعمال البيانات ADNI-DB، لدراسة و تحليل مرض الزهايمر، من خلال تطوير نهج في الاعلام الالي الحيوي بالتعلم العميق لتصنيف مرحلتي المرض (AD,CN). الهدف من هذا النهج هو تطوير نموذج (أو نظام تنبؤي) لتحديد نوع المرض للفراد أو لتقدير مرحلة المرض باستخدام الشبكات العصبية الاصطناعية. كنتيجة، قمنا بتصنيف البيانات الجينية الوظيفية بدقة اختبار تصل الى 92%.

الكلمات المفتاحية : مرض الزهايمر ؛ تعدد أشكال النوكليوتيدات المفردة ؛ SNPs ؛ ADNI-DB ؛ التعلم العميق ؛ الشبكات العصبية الاصطناعية.

LISTES DES FIGURES

LISTE DES TABLEAUX

ACRONYMES

- AI : Artificial Intelligence (Intelligence artificielle)
- ANN : Artificial Neural Network (Réseau de neurones artificiels)
- APOE : L'Apolipoprotein E
- APP : Protéine précurseur amyloïde
- CNN : Convolutional Neural Network (Réseau de neurones convolutionnels)
- DL : Deep Learning (Apprentissage Approfondi)
- GWAS : Genome-Wide Association Studies (Etudes d'associations à l'échelle du génome)
- HGP : Human Genome Project (Projet du génome humain)
- MA : Maladie d'Alzheimer
- MAPT : Microtubule Associated Protein tau (la protéine tau associée aux microtubules)
- ML : Machine Learning (Apprentissage Automatique)
- PS : Preseniline (PS1 et PS2)
- SNPs : Single Nucleotide Polymorphisme (Polymorphismes nucléotidiques uniques)
- TAU : Tubulin-Associated Unit
- TREM2 : Triggering receptor expressed on myeloid cells 2 (le récepteur déclencheur exprimé sur les cellules myéloïdes 2)

TABLE DES MATIÈRES

TABLE DES MATIÈRES

Introduction

INTRODUCTION

L’évolution de la recherche scientifique permet d’identifier les événements biomoléculaires responsables des maladies génétiques causées par des anomalies dans les gènes ou les chromosomes qui peuvent être transmis à la descendance (maladies héréditaires). Ces anomalies génétiques peuvent être détectées en annotant le génome et en déterminant le risque de développer une maladie et donc avoir la possibilité de prévoir des préventions selon le résultat de diagnostic précoce. Cette évolution est actuellement le premier défi dans le domaine biomédical qui donne ainsi les objectifs de produire des méthodes pertinentes pour prévenir le risque de développer une ou plusieurs des maladies génétiques voire héréditaires telles que la maladie d’Alzheimer (MA).

En tant que maladie complexe, la MA est une maladie génétique héréditaire qui peut être affectée par de multiples variantes génétiques. En général, les troubles de cette maladie posent un défi en termes de prédiction parce que l’information pathologique ne peut pas être facilement accessible. Ainsi la classification de la MA a récemment suscité beaucoup d’attention, car les progrès rapides des techniques bioinformatiques ont généré des modèles de risque pertinents associés au polymorphisme (polymorphisme génétique unique : SNPs) dans la population. Autrement dit, la médecine prédictive est un défi majeur pour la santé au XXIe siècle.

L'intelligence artificielle (AI) est un outil utile pour la médecine prédictive avec des progrès remarquables. L'utilisation d'un tel système permettrait de détecter la maladie plus tôt, de mieux identifier ses symptômes, de mieux comprendre son comportement et son évolution et d'ajuster plus précisément les traitements. En d'autres termes, les algorithmes d'apprentissage approfondi qui sont un domaine important d’AI, peuvent jouer un rôle clé dans la lutte contre la MA, en fonction de données génétiques qui sont une ressource précieuse pour prédire les maladies, diagnostiquer les pathologies ou améliorer le suivi des patients.

L’objectif de notre travail est d’apporter une contribution d’ordre pratique qui permettra à l’utilisateur une meilleure prédiction rapide et précise en se basant sur une approche de l’AI à savoir le Deep Learning (DL) ; en utilisant comme matière première des données biomoléculaires chromosomiques en relation étroite avec ladite maladie. La nature de ces données nucléiques est systématiquement basée sur les structures SNPs (Single Nucleotide Polymorphism) largement rencontrés dans les anomalies chromosomiques et ayant pour conséquence l’apparition de la mutation nouvelle causatrices de nouveau phénotypes non sauvages atteint de la maladie.

Pour ce faire, des données des SNPs seront utilisées dans ce travail permettant à la fois de réaliser l'apprentissage approfondi, le test et la validation du modèle proposé.

PARTIE 1 :

RECHERCHE BIBLIOGRAPHIQUE

CHAPITRE 1 :

LES MALADIES GÉNÉTIQUES

1. LE GÉNOME HUMAIN

La structure de la molécule d'ADN, découverte par James Watson et Francis Crick en 1953, a permis l'étude approfondie des gènes. Depuis, la génomique à étudier tous les gènes en tant que système, dans le but de connaître leurs interactions et leurs influences sur les processus biologiques et les réseaux physio-biochimique à l'intérieur des organismes (Sfar et Chouchane, 2008).

En révélant la structure et la composition de l'ADN, cette molécule est le support de l'information génétique qui donne aux gènes une réalité physique et objective par les deux mécanismes : la transcription et la traduction. Il existe une colinéarité entre la séquence de bases d'un gène et la séquence d'acides aminés d'une protéine (figure 1). La fonction de ces gènes est également soumise à un système de contrôle qui régule leur expression en fonction des besoins de la cellule (Sfar et Chouchane, 2008) du corps humain, qui contient environ 100 milliards de cellules, à l'intérieur de celles-ci se trouve le génome humain au sein du nucléus (Ridley, 1999).

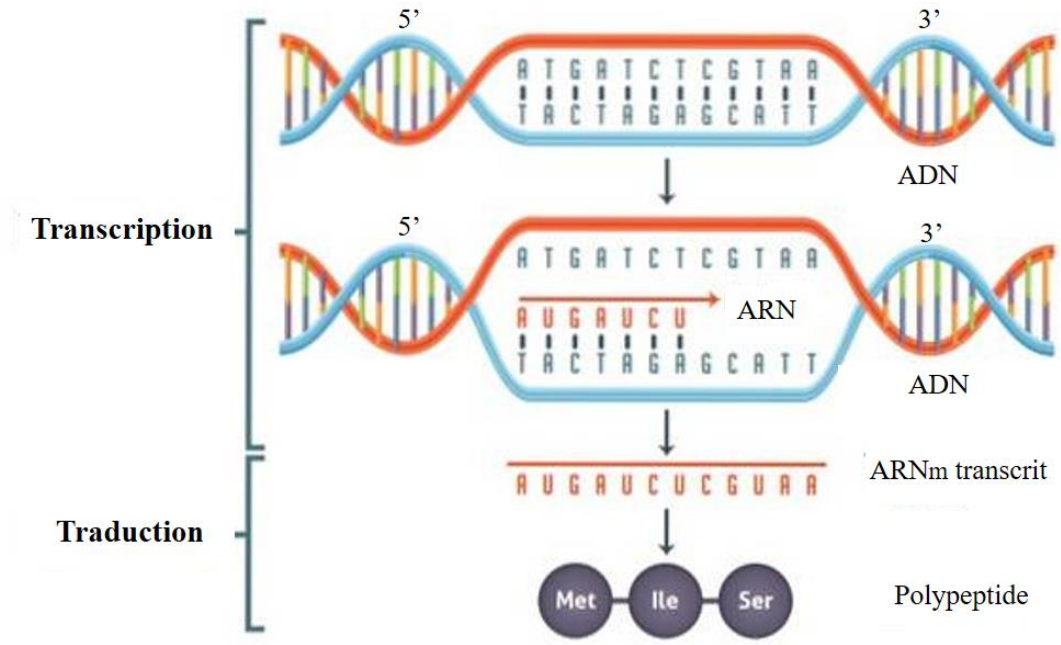


Figure 1 : Représentation schématique du dogme central de la biologie moléculaire[1]

Le terme génome a été inventé en 1920 et est décrit comme l'ensemble des chromosomes haploïdes qui forment la base matérielle d'une espèce (Goldman et Landweber, 2016). Le génome humain est souvent appelé la banque de données d'un individu. Sa transmission à travers les générations fournit le support primaire pour hériter des traits d'un organisme.

[1] https://fr.differbetween.com/

Il est divisé en 23 paires chromosomiques (figure 2). Vingt-deux paires sont numérotées par ordre de taille approximative (tableau 1), de la plus grande (numéro 1) à la plus petite (numéro 22) (Ridley, 1999).

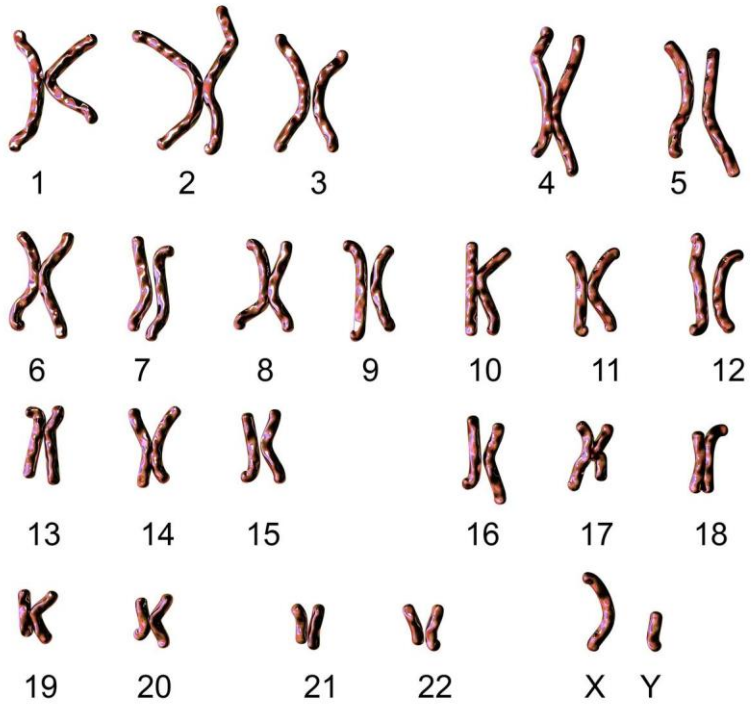


Figure 2 **:** L'ensemble des 23 paires de chromosomes du génome humain[2]

Tableau 1 **:** Caractéristiques des chromosomes (ergopix, 2012)

Chromosome	**Description**
1	– Contient plus de 3000 gènes et plus de 240 millions de paires de base. – C'est le plus grand chromosome humain et le dernier qui a été séquencé.
2	– Contient plus de 2500 gènes et plus de 240 millions de paires de base.
3	– Contient environ 1900 gènes et environ 200 millions de paires de base.
4	– Contient environ 1600 gènes et environ 190 millions de paires de base. – Le chromosome 4 contient de vastes régions, appelées déserts, dans lesquelles il n'y a aucun gène.
5	– Contient environ 1700 gènes et environ 180 millions de paires de base.
6	– Contient environ 1900 gènes et environ 170 millions de paires de base.
7	– Contient environ 1800 gènes et plus de 150 millions de paires de base.

[2] https://sante.lefigaro.fr/sante/analyse/cariotype/pour-quelles-indications

Chromosome	Description
8	– Contient plus de 1400 gènes et plus de 140 millions de paires de base. – Il porte un groupe de plus de 20 gènes qui sont impliqués dans la défense antimicrobienne.
9	– Contient plus de 1400 gènes et plus de 130 millions de paires de base.
10	– Contient plus de 1400 gènes et plus de 130 millions de paires de base.
11	– Contient environ 2000 gènes et plus de 130 millions de paires de base. – Il porte des hôtes de gènes impliqués dans diverses maladies. – Au moins 180 gènes impliqués dans la perception des odeurs se trouvent sur le chromosome.
12	– Contient plus de 1600 gènes et plus de 130 millions de paires de base.
13	– Contient environ 800 gènes et plus de 110 millions de paires de base.
14	– Contient environ 1200 gènes et plus de 100 millions de paires de base. – Il porte de nombreux gènes qui sont très importants pour notre système immunitaire.
15	– Contient environ 1200 gènes et environ 100 millions de paires de base.
16	– Contient environ 1300 gènes et environ 90 millions de paires de base.
17	– Contient plus de 1600 gènes et environ 80 millions de paires de base
18	– Contient plus de 600 gènes et plus de 70 millions de paires de base.
19	– Contient plus de 1700 gènes et plus de 60 millions de paires de base.
20	– Contient plus de 900 gènes et plus de 60 millions de paires de base.
21	– Contient plus de 400 gènes et plus de 40 millions de paires de base. – C’est le plus petit chromosome humain. Il est responsable du syndrome de Down.
22	– Contient plus de 800 gènes et plus de 40 millions de paires de base. – C’était le premier chromosome humain à avoir été séquencé.
X	– Contient plus de 1400 gènes et plus de 150 millions de paires de base. – Il porte de nombreux gènes qui sont importants pour le développement du cerveau humain.
Y	– Contient plus de 200 gènes et plus de 50 millions de paires de base. – Il porte très peu de gènes parce qu’il a perdu beaucoup d’ADN au cours de l’évolution humaine.

Il reste encore un long parcours pour comprendre pleinement la structure et la fonction du génome humain, mais l'obtention de sa séquence complète, grâce au projet du génome humain (Human Genome Project : HGP), donne aux scientifiques l'opportunité d'explorer les complexités de la biologie humaine, la promesse de nouvelles connaissances sur la physiologie de la santé, le diagnostic des maladies et le développement de médicaments (Korf, 2022) (Timpson et al, 2018), et fournit également une base de référence pour comparer et explorer les différences génétiques entre les individus (Marli, 2015).

− Projet du génome humain : Le Projet du génome humain était une coopération internationale à grande échelle qui a commencé dans les années 1990 (Korf, 2022).

En 2003, l'achèvement de la séquence du génome humain a marqué le début d'une nouvelle ère dans la recherche biomédicale. Il a stimulé des avancées technologiques dans les sciences de la vie, y compris le développement de technologies à haut débit pour détecter la variation et l'expression génétique (Sfar et Chouchane, 2008). Il a également accéléré et encouragé la recherche sur le décodage de la structure et de la fonctionnalité du génome, alimenté par les énormes avancées dans les technologies de génotypage (Tawfik et Spruit, 2018).

L'objectif principal du projet du génome humain était de fournir une séquence complète et précise de trois milliards de paires de bases d'ADN qui forment le génome humain ; ce qui aidera à mieux comprendre la biologie humaine (Hofker, Fu, et Wijmenga, 2014), d'identifier et de caractériser les gènes qui interviennent dans plusieurs maladies génétiques et pointer les gènes responsables (Sfar et Chouchane, 2008).

L'un des objectifs de la recherche génétique est d'utiliser l'information génomique (Hofker et al, 2014) pour identifier et révéler les variations génétiques associées aux maladies complexes courantes causées par une combinaison de multiples facteurs génétiques et environnementaux (Hofker et al, 2014), et leur prévalence (une mesure de l'état de santé d'une population, dénombrant le nombre de cas de maladies, à un instant donné ou sur une période donnée) dans différentes populations (Sfar et Chouchane, 2008).

2. ANNOTATION GÉNOMIQUE

L'annotation du génome est le processus d'identification de la structure et de la fonction des caractéristiques codées d'un génome (Dunn et al, 2019), par lequel l'information biologique est extraite, recueillie et affichée dans un format bien défini adapté aux requêtes, basé sur la séquence des acides nucléiques et des protéines (Aubourg et Rouzé, 2001).

L'annotation du génome eucaryote comporte deux étapes principales (Beyne, 2009) :

- L'annotation structurale : consiste principalement à identifier les éléments génomiques d'intérêt (cartographier les éléments du gène) (Aubourg et Rouzé, 2001 ; Beyne, 2009).
- L'annotation fonctionnelle : vise à attribuer des fonctions biochimiques et physiologiques aux produits génétiques déduits (Aubourg et Rouzé, 2001) des gènes identifiés lors de l'annotation structurale (Beyne, 2009).

L'annotation permet d'obtenir les connaissances sur le fonctionnement cellulaire ainsi que sur les mécanismes hypothétiques de son évolution et déduire ses caractéristiques fonctionnelles et physiologiques fondamentales (Beyne, 2009). Cette augmentation de la résolution et de la couverture des annotations du génome (des génotypes aux phénotypes) conduit à une compréhension précise de la biologie des espèces, l'identification des polymorphismes génétiques et les mutations responsables des maladies complexes (Abril et Castellano, 2019), c'est pourquoi l'annotation est souvent synonyme de prédiction (Aubourg et Rouzé, 2001).

3. POLYMORPHISMES GÉNETIQUES

Les facteurs génétiques et environnementaux sont les deux principaux facteurs qui font la différence dans le phénotype humain (Sripichai et Fucharoen, 2007).

Les variations des séquences d'ADN entre les individus, les groupes ou les populations sont connues sous le nom de polymorphisme génétique. Pour être considéré comme un polymorphisme, la fréquence de la variante dans une population donnée doit être >= à 1% ou considérée comme une mutation si elle se produit à une fréquence < à 1% (Sripichai et Fucharoen, 2007). Les polymorphismes résultent des types de mutations suivantes (Ismail et Essawi, 2012) :

Tableau 2 : Types de mutation provoqués par le polymorphisme.

Mutation	Définition
Les séquences répétées	Elles couvrent environ 40 % du génome humain (Genoscope). Deux types de séquences répétées sont distinguées : - Les séquences répétées en tandem : sont constituées de motifs adjacents similaires en taille et en composition - Les séquences répétées dispersées : se trouvent partout, dans les régions géniques, dans les régions intergéniques, et aussi dans les introns (Xu et al, 2015 ; Bérard, 2003).

Mutation	Définition
Substitution	Provoquée par le remplacement d'un nucléotide par un autre
Insertion	Provoquée par l'ajout d'une ou plusieurs nucléotides (séquences répétées) par rapport à la séquence initiale.
Délétion	Une suppression d'une ou plusieurs nucléotides de l'ADN.
Les recombinaisons	Un phénomène qui apparaît chez un individu ou trait génétique d'une cellule ou d'un gène. C'est un échange d'information génétique par une coupure d'ADN et l'ajout d'une nouvelle combinaison entre deux molécules d'ADN distinctes ou entre deux fragments d'une même molécule (Bérard, 2003).
Les polymorphismes nucléotidiques uniques (SNPs)	La forme la plus courante de polymorphisme dans le génome humain (Xu et al, 2005).

Il existe également d'autres types de polymorphismes :

- Synonymes qui n'ont aucun effet sur l'organisme et sont considérés comme sélectivement neutres puisque la substitution n'a aucune influence sur la séquence d'acides aminés de la protéine générée (mutation silencieuse),
- Non synonymes qui provoquent le changement de l'acide aminé codé (mutation non silencieuse),
- Mésense qui provoque des changements du codon qui modifie la protéine et
- Non-sens qui provoque un codon de terminaison d'être mal placé (figure 3) (Ismail et Essawi, 2012).

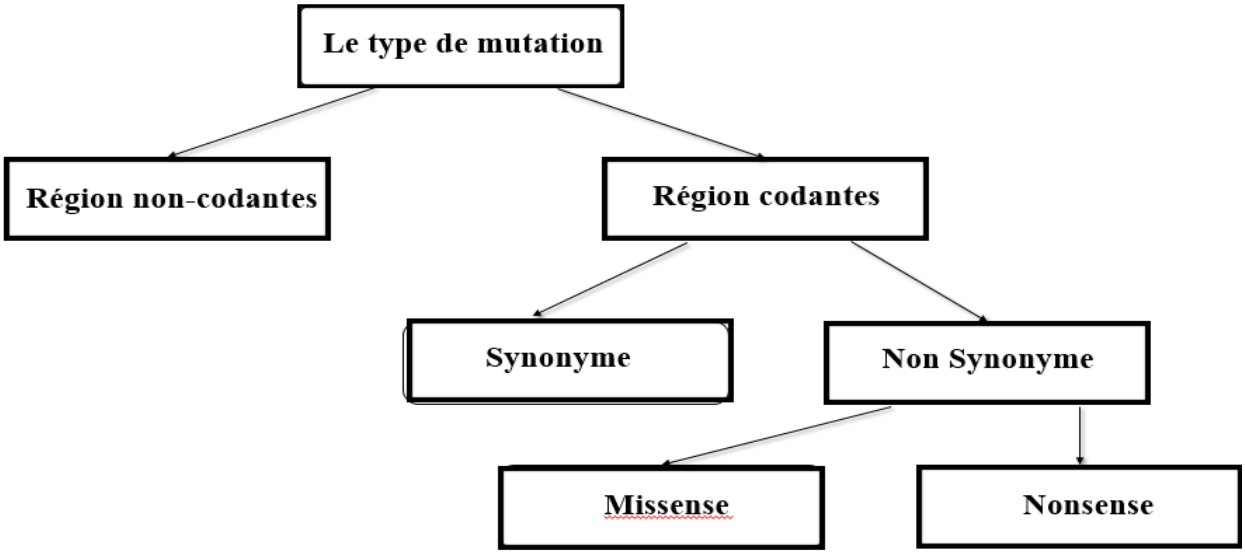


Figure 3 **:** Différents types de mutations

Si les variations dans la séquence d'ADN entre les individus sont liées à une maladie, elles sont communément appelées une mutation génétique et peuvent avoir un impact significatif sur la façon dont les gènes réagissent aux maladies, bactéries, virus, produits chimiques, médicaments et autres thérapies. La présence de certaines variations génétiques alléliques peut être considérée comme facteur causal dans les maladies héréditaires humaines. Par conséquent, le dépistage de ces allèles chez un individu peut permettre de détecter la prédisposition génétique des maladies (Sripichai et Fucharoen, 2007), et l'une des méthodes les plus utilisées est les SNP, qui est largement utilisés pour la recherche sur l'évolution humaine, les études d'association de maladies complexes (Xu et al, 2005).

3.1. Les SNPs : Un SNP est une variation d'une seule paire de base qui se produit à un site spécifique dans des régions codantes et non codantes du génome (figure 4). Ils se produit plus dans les régions non codantes (Sripichai et Fucharoen, 2007). Il ne provoque pas directement une maladie, mais augmente la prédisposition génétique des individus à une certaine maladie et peut affecter leurs réactions aux médicaments (Tawfik et Spruit, 2018).

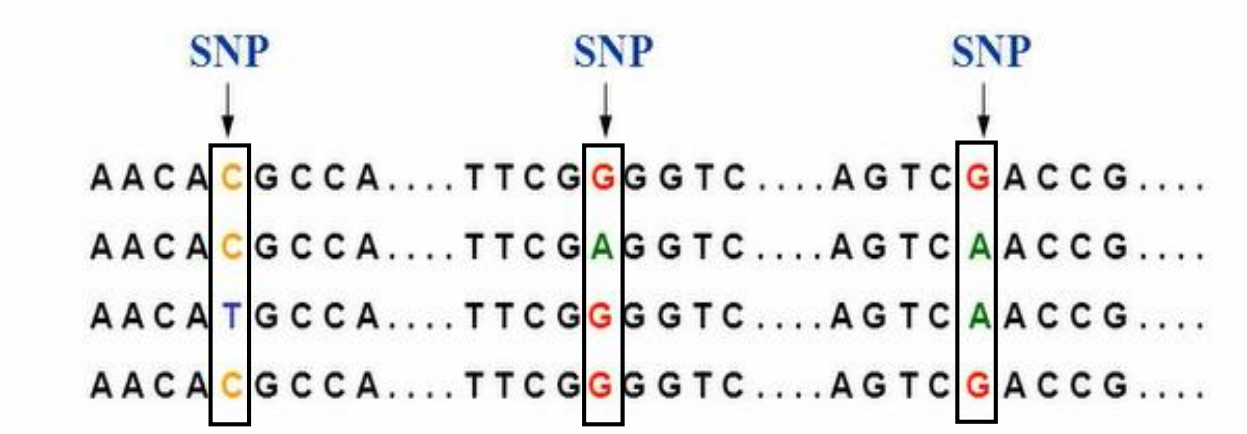


Figure 4 : Exemple d'un polymorphisme nucléotidique unique[3]

Les scientifiques sont particulièrement intéressés par les SNPs situés dans les séquences codantes (Sripichai et Fucharoen, 2007) parce qu'ils provoquent des changements dans la protéine produite (Mahdi, Nassiri, et Nasiri, 2013).

Actuellement, l'information sur les SNP est disponible dans des bases de données (Tawfik et Spruit, 2018) (tableau 3) qui fournissent des ressources pour l'analyse des associations qui cherchent à corréler les phénotypes avec les régions du génome humain (Sripichai et Fucharoen, 2007).

Tableau 3 : Exemples de bases de données

Base de données	**Lien URL**
NCBI dbSNP	http://www.ncbi.nlm.nih.gov/SNP/
SNP Curator	https://tdslab.nl/snpcurator/
Alliance of Genome Resources	https://www.alliancegenome.org/

3.2. Les études d'association SNP/maladie à l'échelle du génome : Les études visant à établir une relation entre un phénotype (habituellement une maladie) et une ou plusieurs régions du génome sont appelées études d'association (Sripichai et Fucharoen, 2007) telles que l'analyse à l'aide de l'étude d'association à l'échelle du génome (Genome-Wide Association Studies : GWAS) qui est l'une des avancées génétiques les plus importantes résultant de la disponibilité de séquences complètes du génome humain (Hofker et al, 2014), qui étudient la relation entre les différents traits phénotypiques et les SNPs (Tawfik et Spruit, 2018).

[3] https://slideplayer.com/slide/8066887/

Ces études comparent la fréquence des allèles chez les personnes atteintes et les témoins. Si un allèle spécifique d'un SNP est plus fréquent chez les atteintes que chez les témoins, l'SNP est lié à cette maladie, et l'allèle spécifique est l'allèle de risque de la maladie. La technique est basée sur le concept de variante-commune / maladie-commune, qui stipule que la variation génétique commune doit jouer un rôle important dans la maladie (Hofker et al, 2014).

Cela est possible en mesurant la fréquence de multiples SNPs dans deux populations avec des phénotypes différents et en détectant les SNP qui montrent une différence significative de fréquence. Si un facteur (Exemple de la maladie d'Alzheimer présenté dans le tableau 4) contribue à un risque accru de développer un phénotype, le facteur devrait être trouvé plus fréquemment chez les personnes atteintes de ce phénotype que chez les témoins non phénotypiques (Sripichai et Fucharoen, 2007).

Tableau 4 **:** Facteurs de risque de la maladie d'Alzheimer (Club Prévention Santé, 2021)

Maladie	**Facteurs**
Alzheimer	L'NIA () conclut généralement sept facteurs : - Faible niveau d'instruction 19% - Le tabagisme 14% - L'inactivité physique 13% - La dépression et l'anxiété 11% - L'hypertension 5% - L'obésité 2%, Le diabète 2% - Les facteurs de risque génétiques : Les personnes ayant les parents sont atteints sont plus susceptibles d'être touchées que celles qui n'ont pas d'antécédents génétiques familiaux.

Des progrès considérables ont été accomplis dans la compréhension des maladies complexes et la révolution de la médecine prédictive (Tak et Farnham, 2015) qui ne serait pas possible aujourd'hui sans les données génétiques recueillies à partir des GWAS en utilisant des puces contenant des millions des SNPs associés à de nombreuses maladies ou traits complexes (Tawfik et Spruit, 2018).

4. MÉDECINE PRÉDICTIVE

La médecine prédictive est une médecine diagnostique plutôt qu'une médecine thérapeutique, et sa logique ultime reste la prévention par l'utilisation des concepts de prédiction, de propension et de probabilité. Elle englobe des phénomènes connus depuis longtemps qu'un certain type d'individu a une probabilité plus élevée de développer une certaine maladie, et cherche à déterminer la probabilité de développer une pathologie donnée en étudiant les gènes (Claudia et François, 2007).

Le terme prédictif en médecine fait référence à l'identification du risque d'un individu de développer une maladie en fonction du profil génétique, il s'agit d'identifier des antécédents familiaux de maladie, d'une prédisposition génétique avant l'apparition de la maladie dans une population en bonne santé (Slim, Selvy, et Veziant, 2021). Ses approches s'appuient sur des bases d'expérience souvent suivie par des dizaines de milliers de personnes durant des années, selon lesquelles des facteurs de risque spécifiques peuvent être considérés comme étant associés à un risque accru de maladie (Pourtau, 2015).

Les maladies complexes comme le Cancer, le Diabète et l'Alzheimer représentent une menace énorme pour la santé humaine et ont fait l'objet d'études approfondies au cours des dernières décennies. Toutefois, la cause sous-jacente de ces maladies n'est pas encore clairement connue. Grâce au développement rapide de la technologie génomique, les méga-données sur les changements dans l'ADN tels que les SNPs permettent une caractérisation complète des maladies complexes et prédisent leur état (Li et al, 2018).

5. MALADIES GÉNÉTIQUES ET HÉRÉDITAIRES

Les maladies génétiques sont classées comme monogéniques, oligogéniques (causée par plusieurs mutations concomitantes chez le même malade par exemple la mutation A et la mutation B provoquent la maladie), polygéniques/multifactorielles (complexes ou les symptômes peuvent apparaître chez deux patients qui ont des mutations différentes par exemple la mutation A ou la mutation B provoque la maladie), ou chromosomiques. Cette classification est fondée sur la nature présumée ou connue du défaut génétique de la maladie sous-jacente (Iourov, Vorsanova, et Yurov, 2019).

Les maladies héréditaires sont causées par des aberrations des cellules germinales propres aux cellules du corps humain. Pour des centaines de ces maladies, ces aberrations germinales ont été identifiées et attribuées aux gènes responsables. Cependant, l'identification du gène responsable n'est souvent que le point de départ pour comprendre la base moléculaire de chaque

maladie. La relation génotype-phénotype entre le gène causal et le phénotype de la maladie est généralement complexe et de nombreuses maladies héréditaires doivent encore être clarifiées (Barshir et al, 2018).

La maladie d'Alzheimer (MA) est parmi les maladies héréditaires les plus courantes qui affectent un grand nombre de personnes, elle sera décrite dans la section suivante, en se concentrant sur leurs pathologies et leurs mutations génétiques.

5.1. L'Alzheimer : La maladie d'Alzheimer (MA) a été décrite pour la première fois en 1906 lors d'une conférence à Tubingen, en Allemagne, par Alois Alzheimer (Sanabria-Castro, Alvarado-Echeverría, et Monge-Bonilla, 2017). C'est une maladie neurodégénérative (un terme générique englobant différents troubles médicaux qui touchent principalement les neurones du cerveau humain) (Sügis et al, 2019), considérées comme la quatrième cause de décès chez les personnes âgées après le cancer, les maladies cardiaques et les maladies cérébrovasculaires (Hu et al, 2019).

Selon l'Organisation mondiale de la santé, environ 50 millions de personnes sont atteintes de démence (une sérieuse perte ou réduction des capacités cognitives suffisamment importante pour retentir sur la vie d'un individu), et il y a environ 10 millions de nouveaux cas de MA chaque année (Romero-Rosales et al, 2020). La MA est hautement héréditaire et on estime que 80 % de la responsabilité s'explique par des facteurs génétiques (Leonenko et al, 2019).

Les changements dans les premiers stades de la maladie, comme indiqué dans le tableau 5 peuvent commencer de 10-20 ans avant le diagnostic et produire des problèmes de mémoire chez les individus. Les patients peuvent éprouver des changements de personnalité et avoir de la difficulté à identifier les éléments (Ghazi, 2020), des troubles du langage et d'autres symptômes neuropsychiatriques qui progressent avec l'âge (Hu et al, 2019) et peuvent durer de deux à 10 ans. Enfin, au stade sévère de la MA, qui peut durer de un à cinq ans, la mort cellulaire se produit, laissant les patients incapables de communiquer, de reconnaître leur famille ou de prendre soin d'eux-mêmes (voir figure 5) (Ghazi, 2020).

Tableau 5 : Stades de la maladie d'Alzheimer depuis la phase Asymptomatique (stade 1) vers la phase sévère (stade 5) (Krolak-Salmon, 2020)

Stade 1 : La phase Asymptomatique (stade léger)
- Performances dans la norme - Pas de déclin ou de plainte signalée. - Pas de changement cognitif ou comportemental signalé par l'aidant ou objectivé - Aucun signe clinique perceptible mais des lésions commencent tout de même à se développer dans le cerveau. Cette phase peut durer plusieurs années.
Stade 2 : La phase Pré-démentielle
- Performances dans la norme - Déclin cognitif relatif. - Déclin cognitif subjectif ou, - Déclin cognitif objectivé par des tests longitudinaux ou, - Changement comportemental.
Stade 3 : La phase de démence légère
- Trouble neurocognitif léger, - Préservation de l'autonomie, - Une altération des capacités intellectuelles (langage, orientation, attention) et - Possible déclin fonctionnel léger.
Stades 4 : La phase de démence modérée
- Troubles somatiques peuvent s'additionner aux troubles cognitifs, - L'autonomie de la personne est alors compromise et cette dernière nécessite des aides afin de pouvoir rester à domicile.
Stade 5 : La phase de démence sévère
- Inexorablement vers une perte d'autonomie et une totale dépendance. - Les troubles cognitifs sont majeurs et ont un fort impact sur la vie quotidienne et les activités sociales. - Des troubles affectent la marche, la continence, ou encore la déglutition. - La dénutrition et la perte de poids sont de plus en plus invalidantes et peuvent entraîner des complications qui conduiront au décès

Figure 5 : Différents stades et symptômes de la MA (Ghazi, 2020)

5.1.1. Pathologie : La maladie d'Alzheimer est une atrophie du cortex cérébral dû à la lésion des neurones qu'il contient. Deux types de lésions principales sont observés dans le cortex (figure 6) :

- Les lésions extra neuronales : résultent de l'accumulation pathologique de la protéine amyloïde transmembranaire qui est clivée pour former un peptide amyloïde Aβ.
- Les lésions intra-neuronal : résultent de l'accumulation pathologique de la protéine TAU (Tubulin-Associated Unit), une protéine d'organisation, de stabilisation et de régulation de la dynamique des microtubules les accumulations intra-neuronales (Govaerts, Schoenen, et Bouhy, 2007), responsable des dégénérescences neurofibrillaire. Elle contribue au bon fonctionnement des neurones. Au cours de la maladie d'Alzheimer, la protéine Tau devient hyperphosphorylée. Elle ne peut donc plus accomplir sa fonction principale.
- En conséquence, les neurones fonctionnent mal et ils ne transporteront plus leurs produits de synthèse, tels que les neurotransmetteurs, dont l'acétylcholine. Ainsi, ce déficit biochimique entraîne des modifications de l'apparence des neurones (SIOU, 2021).

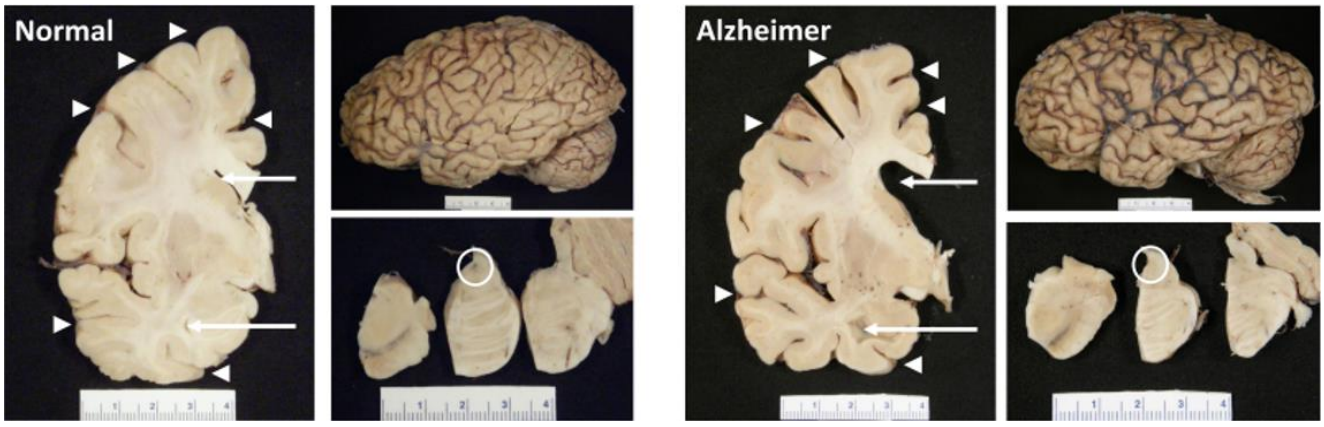


Figure 6 **:** Anatomie du cerveau d'Alzheimer ; Vue latérale d'un cerveau d'Alzheimer peut montrer l'élargissement des espaces sulcaux et le rétrécissement de gyri par rapport à un cerveau normal. Ceci peut être plus facilement observé dans les sections coronales comme indiqué par les pointes de flèche, et cette atrophie est souvent accompagnée d'un élargissement des cornes frontales et temporelles des ventricules latéraux comme mis en évidence par les flèches. En outre, la perte de neurones pigmentés dans le locus coeruleus est couramment observée dans le tegmentum pontin comme le montre le cercle ouvert. Aucune de ces caractéristiques n'est exclusive à la maladie d'Alzheimer (DeTure et Dickson, 2019).

5.1.2. Mutations génétiques dans la MA : La protéine précurseur amyloïde (APP), l'Apolipoprotein E (APOE), la preseniline (PS1 et PS2), la protéine tau associée aux microtubules (MAPT : microtubule associated protein tau) et le récepteur déclencheur exprimé sur les cellules myéloïdes 2 (TREM2 : Triggering receptor expressed on myeloid cells 2) sont les gènes de virulence définitifs (tableau 6) (Yi-Gang Chen, 2018). Des études ont révélé que ces protéines sont étroitement liées à la neurodégénérescence (Ceci est démontré par une augmentation de la protéine tau totale dans le liquide céphalo-rachidien (Krolak-Salmon, 2020)) et l'atrophie cérébrale montrée par l'imagerie structurale et à l'apoptose des cellules nerveuses (figure 6), et ses dommages fonctionnels peuvent entraîner des déséquilibres dans le cerveau, induisant les troubles du MA (Yi-Gang Chen, 2018).

Tableau 6 : Principales mutations rencontrées dans la MA (A. Armstrong, 2013 ; Liu et al, 2013)

Gène	**Chromosome**	**Mutation**	**Mécanisme**
APP (Amyloid Precursor protein)	21q21	73 mutations Exemples : - A → C - GAA (Glu) → AAA(Lys)	Ces mutations sont associées à la MA précoce familiale et à l'angiopathie amyloïde cérébrale (AAC) et modifient généralement le traitement de la sécrétase. L'hypothèse de cascade amyloïde (HCA) suggère que le dépôt de peptide Aβ est le premier événement pathologique en MA, conduisant à la formation de PS et NFT (sont des agrégats de protéine tau hyperphosphorylée qui sont généralement connus comme un biomarqueur primaire de la MA), suivie par la mort cellulaire et la démence (A. Armstrong, 2013).

Gène	Chromosome	Mutation	Mécanisme
APOE **(Apolipo-protein E)**	19q13.32	171 mutations Exemples : - G → C - GTG (Val) → ATG (Met)	Elle se connecte à une variété de récepteurs de surface des cellules pour fournir des lipides, ainsi que le peptide hydrophobe amyloïde, qui est supposé pour commencer des processus nocifs qui conduisent à une neurodégénérescence dans la MA. Les isoformes d'APOE ont des rôles variés dans le transport des lipides, le métabolisme du glucose, la signalisation neuronale, la neuroinflammation et la fonction mitochondriale (Liu et al, 2013).
PSEN1 **(Presenilin 1)**	14q24.2	349 mutations Exemples : - A → G - AAT(Asn) → AGT(Ser)	Une sous-unité de γ-secretase qui se trouvent sur la membrane du réticulum endoplasmique. Le PSEN1 peut favoriser un clivage de 42-γ-secretase spécifique du APP normal, ce qui entraîne une augmentation de l'accumulation d'espèces amyloïdes en raison d'une perte de fonction (A. Armstrong, 2013). Les mutations dans le PSEN1 ont été signalées et sont la cause la plus fréquente de l'apparition précoce de la MA.
PSEN2 **(Presenilin 2)**	1q42.13	87 mutations Exemples : - T → C - ACG (Thr) → ATG (Met)	Une sous-unité de γ-secretase, la protéase aspartyle responsable de la génération d'Aβ. Les gènes PSEN2 sont impliqués dans la régulation de l'expression des gènes par la perturbation de l'homéostasie calcique cellulaire ou les interactions avec la protéine coactivatrice transcriptionnelle liaison de réponse d'élément AMPc (CREB-binding) (A. Armstrong, 2013). Les mutations de Méssense dans PSEN2 sont une cause rare de l'apparition précoce de la MA.

Gène	Chromosome	Mutation	Mécanisme
MAPT **(microtubule associated protein tau)**	17q21.31	112 mutations Exemples : - G → A - CGC (Arg) → CAC (His)	Une protéine centrale dans la neuropathologie de la MA. Les mutations MAPT ne sont pas associées à la maladie d'Alzheimer familiale, mais peuvent provoquer une démence front temporale (FTD) et plusieurs autres tauopathies.
TREM2 (récepteur déclencheur exprimé sur les cellules myéloïdes 2)	6 p14	143 mutations Exemples : - C → T - GGT (Gly) → TGT (Cys)	Récepteur transmembranaire qui régule l'activité et la survie de la microgliale (sont les macrophages résidents du système nerveux central (SNC)). La variante TREM2 provoque la maladie de NasuHakola (NHD), une démence autosomique récessive rare d'apparition précoce, associée à la MA, à la démence front temporale (DFT), à la maladie de Parkinson (MP) et à la démence latérale amyotrophique sclérose de la moelle épinière (LAS).

CHAPITRE 2 : CONCEPTS D'INTELLIGENCE ARTIFICIELLE

L'histoire de l'intelligence artificielle n'est pas seulement une histoire de machines essayant de reproduire ou de remplacer un concept statique de l'intelligence humaine, mais un récit évolutif sur la façon dont nous percevons l'intelligence elle-même.

Stephanie Dick

1. INTELLIGENCE ARTIFICIELLE

Le système informatique a évolué, conduisant les machines à imiter le comportement humain et d'effectuer des tâches qui nécessitent l'intelligence humaine telles que la capacité de raisonner, de découvrir le sens, de généraliser ou d'apprendre de l'expérience passée. Ces processus intellectuels caractéristiques de l'homme sont ensuite exprimés, recueillis, et intégrés dans les machines. On parle de l'intelligence artificielle (Artificial Intelligence : AI) inventée en 1956 par McCarthy. Ce terme est difficile à définir et à de nombreuses définitions différentes en raison de son ampleur. Il englobe différentes technologies, par exemple, l'apprentissage automatique. En termes simples l'AI est l'ensemble de théories et de techniques dédiées au développement de programmes informatiques complexes capables de simuler certains traits de l'intelligence humaine (raisonnement, apprentissage …) (Lexcellent, 2019), donc l'AI vise à transférer la responsabilité de la prise de décision aux machines (Dick, 2019).

L'AI est un domaine à fort potentiel qui joue un rôle de plus en plus important dans la société d'aujourd'hui. Alors que les machines (algorithmes) précédentes n'effectuent que des activités manuelles, celles-ci étendent désormais leurs capacités aux tâches cognitives en créant des robots pour un large éventail de disciplines, notamment la robotique, le marketing, l'analyse commerciale, le traitement d'images et de vidéos, etc (Hussain, 2018).

L'AI apporte d'énormes contributions aux applications médicales et biologiques, de l'équipement médical au diagnostic et à la prédiction des maladies (Hussain, 2018).

Dans ce chapitre, nous allons présenter les bases de l'AI et ses différentes branches, en particulier le domaine de l'apprentissage automatique, qui comprend l'apprentissage approfondi. Ce dernier étant l'outil principal pour réaliser notre travail.

2. APPRENTISSAGE AUTOMATIQUE

L'apprentissage automatique, ou Machine Learning (ML) est une branche de l'AI, qui utilise des techniques de calcul basées sur des données et des informations historiques pour résoudre des problèmes (Berry, Mohamed, et Yap, 2020). Elle est basée sur le développement d'algorithmes qui améliore ses performances en apprenant des données (Zhang et Lu, 2021), et peuvent raisonner et planifier vers leur but sans aucune base de connaissances intégrée dans leur environnement (Schlegel et Uenal, 2021). En général, l'efficacité d'une solution d'un ML dépend de la nature, les caractéristiques des données et de la performance des algorithmes d'apprentissage (Sarker, 2021).

Nous vivons à l'ère des données massives (Big Data), où tout ce qui nous entoure est connecté à des sources de données et enregistré numériquement. Le monde électronique regorge de données de tous types qui nécessitent des analyses avancées, conduisant à l'utilisation du ML (Berry et al, 2020), telles que les données de l'Internet des objets (IoT), les villes intelligentes, les smartphones, les réseaux sociaux, les données d'entreprise, les données de cyber sécurité, les données de santé et bien d'autres (Sarker, 2021).

Les algorithmes d'apprentissage automatique sont appliqués à différents domaines tels que l'analyse prédictive, la prise de décision intelligente, les soins de santé, le traitement du langage naturel (NLP) et l'analyse des sentiments, la reconnaissance d'images, de parole et de modèle etc. (Sarker, 2021).

2.1. Types de données : Il existe plusieurs types de données. Les plus couramment utilisées sont les données structurées et non structurées.

2.1.1. Données structurées : Elles ont une structure bien définie, conforme à un modèle de données suivant un ordre standard. Ces données sont bien organisées, facilement accessibles et stockées dans un format tabulaire. Par exemple : noms, dates, adresses, numéros de carte de crédit, géolocalisation, etc (Sarker, 2021).

2.1.2. Données non structurées : Ce type de données n'a pas de format ou d'organisation prédéfinis ; ce qui rend la saisie, le traitement et l'analyse beaucoup plus difficiles. Par exemple, les données de capteurs électroniques, les courriels, les entrées de blogues, les wikis et les documents de traitement de texte, les fichiers audio, les vidéos, les images, etc. (Sarker, 2021).

2.2. Types de techniques du ML : Pour analyser intelligemment les données et développer des applications, les algorithmes d'apprentissage automatique sont la clé (Sarker, 2021). Ces algorithmes peuvent être catégorisés en quatre types basées sur les données qu'ils utilisent : supervisés, non supervisés, semi-supervisé et de renforcement (Ippolito, Ferguson, et Jenson, 2021). Plusieurs techniques informatiques sont employées dans le ML. La figure 7 représente les principales approches.

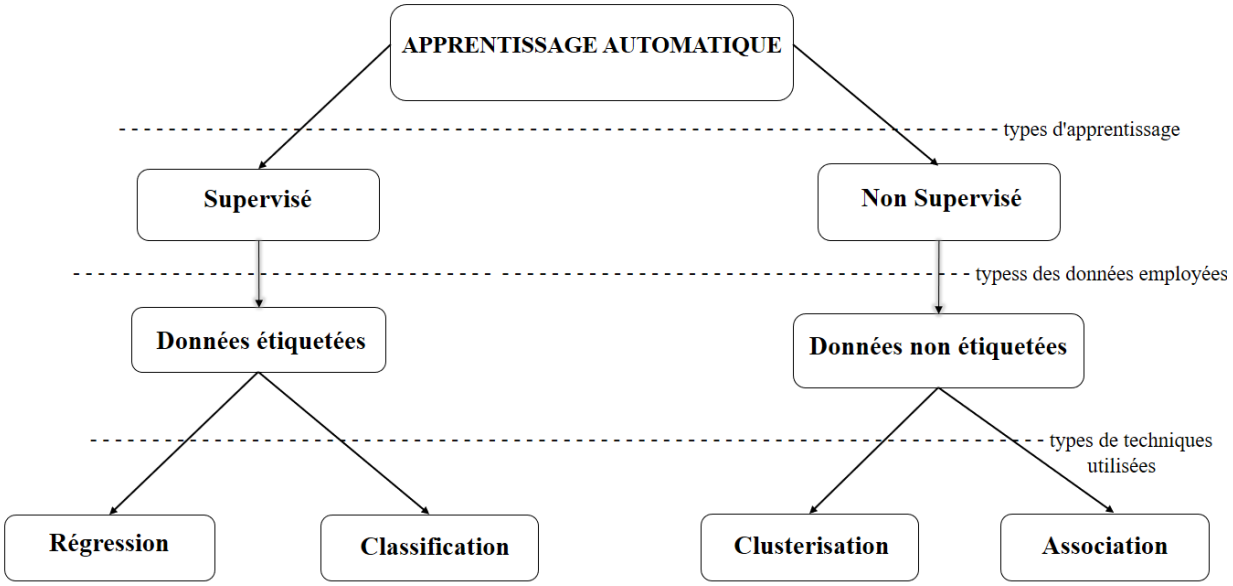


Figure 7 : Différentes techniques pour le ML

2.2.1. Apprentissage supervisé : Un paradigme d'apprentissage automatique qui implique l'étude de la tâche de mappage (le processus d'association des données source aux données cibles) des données d'entrée étiquetées aux données de sortie à partir d'échantillons de paires d'entrée-sortie (van Gerven et Bohte, 2017).

Ce système d'apprentissage développe un modèle prédictif (Ippolito et al, 2021), constituant d'une approche d'analyse et un type d'algorithme qui peut être modifié selon le type des données d'entrées (Rashidi et al, 2019).

Il s'agit d'une approche basée sur des tâches supervisées : la classification séparant les données et la régression montrant les liens entre celles-ci (Sarker, 2021).

- Classification : La classification est définie comme le processus d'attribution d'une ou de plusieurs catégories prédéfinies à chaque élément où la valeur de sortie est discrète. La classification binaire où chaque élément est classé dans l'une des deux catégories est mentionnée comme le type de classification le plus simple.

La classification binaire est élargie en une classification multiple en définissant plus de catégories (figure 8) (Jo, 2021). Ces méthodes sont utilisées, par exemple, pour identifier des objets ou des textes, faire des prédictions au niveau de l'image, comme la classification des maladies (Srinidhi, Ciga, et Martel, 2021).

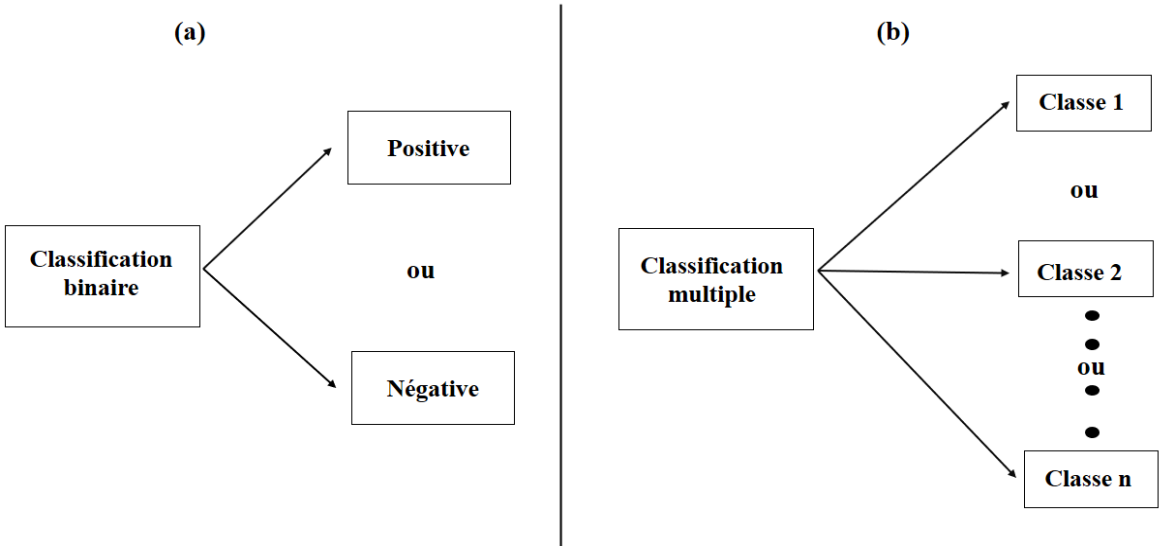


Figure 8 : Classification binaire (a) et multiple (b)

- Régression : La régression est définie comme le processus d'estimation d'une valeur de sortie basée sur de multiples facteurs. Dans la régression, la valeur de sortie est continue (Jo, 2021).

2.2.2. Non supervisé : L'apprentissage non supervisé analyse des ensembles de données non étiquetées, c'est-à-dire un processus axé sur les données et non sur les taches comme dans le cas de l'apprentissage supervisé (Sarker, 2021), il est appelé non supervisé parce que les algorithmes sont laissés à eux-mêmes pour regrouper l'information non triée en trouvant des similitudes, des différences et des modèles dans les données. Il est connu sous le nom d'algorithme auto-organisé ou adaptatif (Dike et al, 2018).

Les tâches d'apprentissage non supervisées les plus courantes sont le regroupement (Clustering), telles que la détection d'anomalies (Sarker, 2021), la classification d'images en deux groupes ou clusters en fonction de caractéristiques spécifiques telles que la couleur, la taille, la forme, etc (Dike et al, 2018).

- Clusterisation : Est défini comme le processus de segmentation d'un groupe d'éléments en sous-groupes dont chacun contient des éléments similaires les unes aux autres que

celles des autres groupes (Jo, 2021), utilisé dans le ML, l'analyse d'image, recherche médicale, recherche d'informations, modèles de reconnaissance et bio-informatique (Dike et al, 2018).

3. APPRENTISSAGE APPROFONDI

Avec l'amélioration significative de la puissance de calcul et l'avancement du Big Data, l'apprentissage approfondi est devenu l'un des algorithmes d'apprentissage automatique les plus performants. Il a connu un grand succès dans divers domaines, y compris la bio-informatique (Li et al, 2019).

L'apprentissage approfondi ou le Deep Learning (DL) est défini comme une sous-classe du ML au sein des technologies de l'AI (figure 9) (Woschank, Rauch, et Zsifkovits, 2020). Il offre des méthodes pour apprendre les représentations de données de manière supervisée et/ou non supervisée (Habimana et al, 2020) en utilisant des algorithmes appelés réseaux de neurones artificiels (Artificial Neural Network : ANN) inspirés du neurone biologique (Gupta et al, 2021).

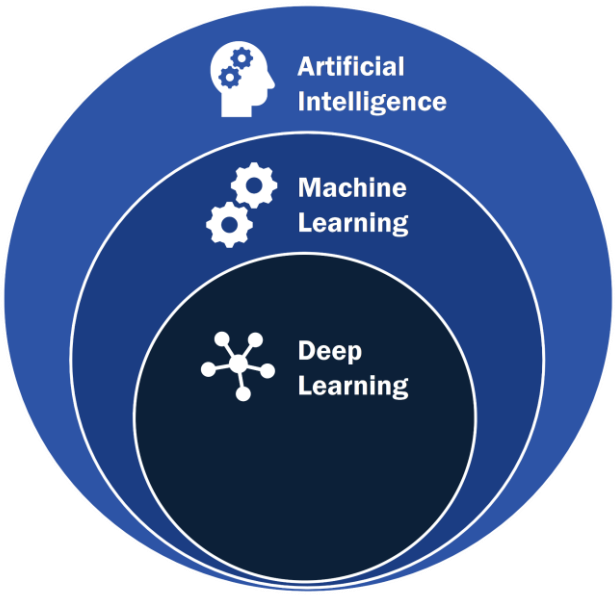


Figure 9 : Relation entre l'AI, ML et DL[4]

3.1. Réseau de neurones artificiels : Un réseau de neurones artificiels se compose d'unités de traitement appelées neurones artificiels ou perceptron. Un perceptron tente de reproduire la structure et le comportement du neurone naturel. Il se compose des nœuds d'entrées (dendrites), de poids (synapse) et une sortie (axone) (figure 10) (Kukreja, 2016).

[4] https://blog.oursky.com/2020/05/07/artificial-intelligence-ai-for-businesses-what-you-need-to-know-before-starting-an-ai-project/ai-vs-ml-vs-dl/

Un seul neurone n'est pas puissant, il génère une sortie avec une seule valeur numérique. La puissance d'un ANN émerge de la combinaison de nombreuses unités d'une manière appropriée, qui composent un système de calcul appelé un réseau de neurones artificiel (nombreuses couches non linéaires) (Dongare, Kharde, et Kachare, 2012) qui traitent l'information par leur changement d'état dynamique en réponse à une entrée externe. Chaque couche successive utilise la sortie de la couche précédente comme entrée (Shinde et Shah, 2018).

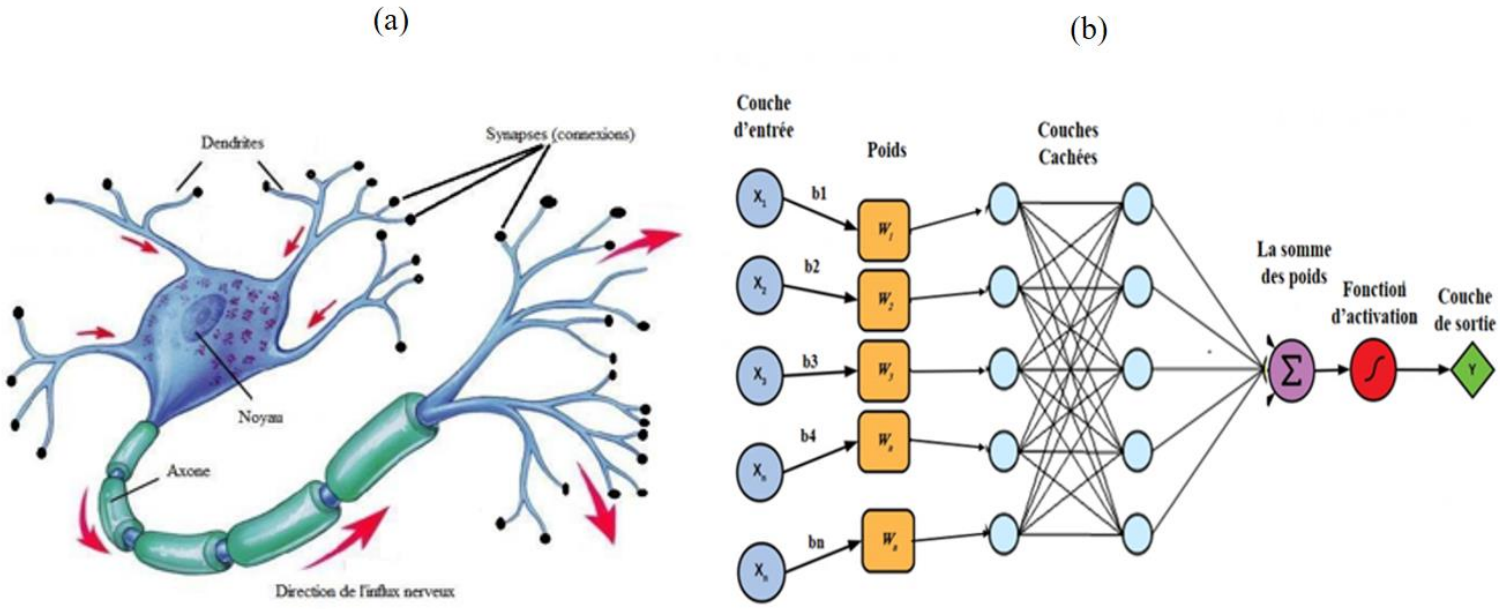


Figure 10 : Mise en correspondance entre le neurone biologique[5] (a) et le neurone artificiel[6] (b)

3.2. La structure d'un réseau de neurones artificiels : Chaque ANN est composée des quatre composantes suivantes (Buscema et al, 2018) :

Tableau 7 : Composants d'un réseau de neurones artificiels (Buscema et al, 2018)

Les composants d'un ANN		**Définition**
Le type et le nombre de nœuds et leurs propriétés correspondantes	Nœuds d'entrée	C'est le début du flux de travail d'un ANN. Il reçoit les données initiales dans le système et les transmettra au reste du réseau.
	Nœuds cachés	Les nœuds qui ne reçoivent et n'envoient leur signal qu'à d'autres nœuds à l'intérieur de l'ANN.
	Nœuds de sortie	Ils contiennent le résultat ou la sortie du problème. Ils donnent la sortie prévue qui sera comparée à la sortie réelle.

[5] www.researchgate.net
[6] https://starship-knowledge.com/neural-networks-perceptrons

Les composants d'un ANN		Définition
Les connexions	Adaptatifs	Ils changent selon l'équation d'apprentissage.
	Fixées	Ils restent à des valeurs fixes tout au long du processus d'apprentissage.
	Variables	elles changent de façon déterministe à mesure que les autres connexions changent.
Les couches	Monocouche	Tous les nœuds ont les mêmes propriétés.
	Multicouches	Les nœuds sont regroupés en classes fonctionnelles ; par exemple les nœuds qui partagent les mêmes fonctions de transfert de signal ou qui reçoivent le signal uniquement à partir de nœuds d'autres couches et les envoient uniquement à de nouvelles couches.
	Couche sensible aux nœuds	Chaque nœud est spécifique à la position qu'il occupe dans l'ANN; par exemple, les nœuds les plus rapprochés communiquent plus intensément que ceux qui sont plus éloignés.
Le flux du signal	Feed-Forward ANN	Le signal passe de l'entrée à la sortie de l'ANN en passant par tous les nœuds une seule fois.
	Feedback ANN	Le signal procède avec un feedback spécifique déterminé au préalable ou en fonction de la présence de conditions particulières.

3.3. Le fonctionnement d'un réseau de neurones artificiels : Les réseaux de neurones artificiels ont leur propre façon d'apprendre et de s'intégrer dans un réseau. Ce processus est illustré dans la figure 11 :

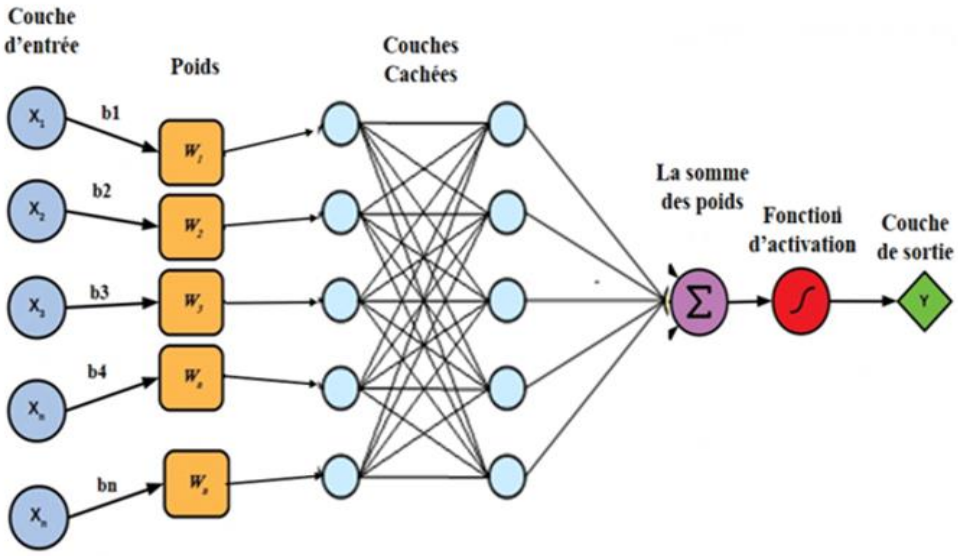


Figure 11 **:** Fonctionnement d'un réseau de neurones artificiels[7]

i. (x_1...x_n) sont des nœuds d'entrée formant la couche d'entrée, où chaque nœud correspond à une variable indépendante. Les nœuds d'entrée doivent avoir des valeurs comprises entre 0 et 1 ; par conséquent, les transformations sont utilisées pour convertir des variables catégoriques et continues indépendantes en des zéros et uns.
   Ils constituent la porte d'entrée pour le flux de l'information initiale qui servira à l'apprentissage (Kukreja, 2016).

ii. Ils seront pondérés puis additionnés dans une fonction de combinaison, afin d'améliorer l'efficacité de calcul des algorithmes d'apprentissage (Kukreja, 2016).

iii. (b_1…b_n) Un biais est un estimateur d'erreurs habituellement initialisée à 1, ajoutée au neurone avec les entrées, pour améliorer la précision du réseau (Kukreja, 2016).

iv. (W_0... W_n) sont des poids synaptiques utilisés pour pondérer chacune des variables d'entrée ; ce qui permet de quantifier leur pertinence par rapport à la fonctionnalité du neurone (Kukreja, 2016).

v. La somme des produits calculée (poids (W) x biais (b)) est passée en entrée à la fonction d'activation dans les couches cachées (Georgevici et Terblanche, 2019).

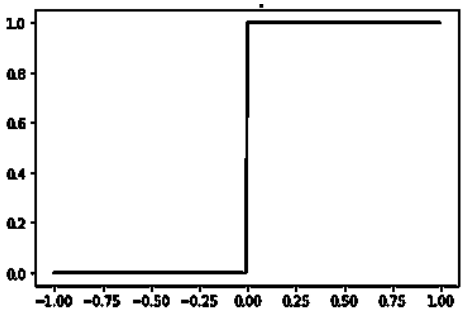
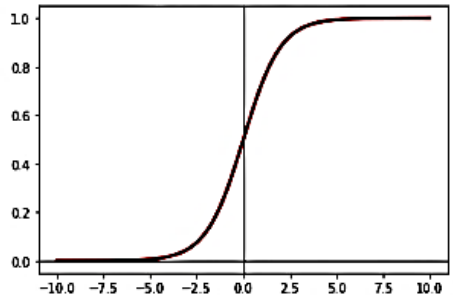
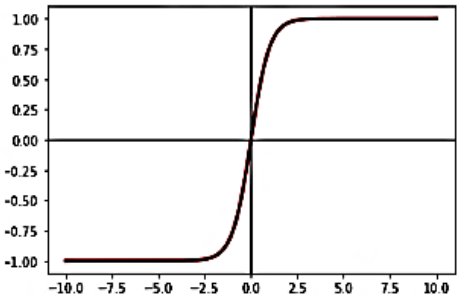
vi. La fonction d'activation est une fonction qui mappe les entrées à la sortie désirée, considérée comme un paramètre choisi pour optimiser les performances du modèle, par la détermination si le neurone est activé ou non (Kattenborn et al, 2021). Diverses fonctions sont utilisées pour l'activation. L'une des fonctions d'activation les plus couramment utilisées est la fonction sigmoïde.

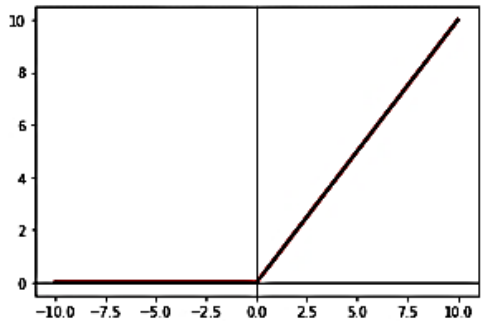
[7] https://starship-knowledge.com/neural-networks-perceptrons

Les autres fonctions utilisées sont la fonction Step (Threshold), la fonction tangente hyperbolique et la fonction d'unité linéaire rectifiée (ReLU) (tableau 8) (Kukreja 2016).

vii. La sortie est la prédiction du réseau de neurones en fonction des données d'entrées.

Tableau 8 : Différents fonctions d'activation (Kukreja, 2016)

La fonction d'activation	Description	L'équation	La courbe
Step ou Threshold (Seuil)	C'est une fonction d'activation simple, elle renvoie 1 (vrai) pour les valeurs qui sont au-dessus du seuil spécifié.	$\emptyset(x) = \begin{cases} 1, if\ x \geq 0.5 \\ 0, \end{cases}$	
Sigmoïde	C'est la plus utilisée pour les ANN qui ne doivent produire que des valeurs positives. Elle garantit que les valeurs restent dans un éventail relativement petit.	$S(x) = \frac{1}{1+e^{-x}}$	
Tangente Hyperbolique	Elle est utilisée pour les ANN qui doivent produire des valeurs comprises entre -1 et 1.	$\tanh x = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$	

Unité Linéaire Rectifiée (ReLU)	C'est une fonction linéaire non saturante. ReLU ne sature pas en -1, 0 ou 1. Elle produit l'entrée directement si elle est positive, sinon, elle produira zéro.	$RELU(x) = \begin{cases} 0 \ if \ x < 0 \\ x \ if \ x >= 0 \end{cases}$	

Une fois que l'information a traversé toutes les couches du neurone, le modèle génère une sortie qui est comparée à la vérité, basée sur la valeur de l'étiquette. Une erreur est calculée puis utilisée pour mettre à jour les poids dans un processus connu sous le nom de rétro-propagation, en utilisant une fonction mathématique appelée fonction de perte. Pendant les périodes d'apprentissage multiples, le modèle vise à minimiser l'erreur et à trouver la combinaison de valeurs de poids qui génère la valeur d'erreur la plus faible (Georgevici et Terblanche, 2019), selon la formule de fonction de perte (Andreassen et Nachman, 2020).

$$\text{loss}(f(x)) = -\sum_{i \in \mathbf{0}} \log f(x_i) - \sum_{i \in \mathbf{1}} \log(1 - f(x_i)),$$

Équation 1 : La formule de la fonction du perte (Andreassen et Nachman, 2020)

3.4. Processus du DL : Les étapes principales pour créer un modèle DL sont les suivantes :

3.4.1. Collecte des données : Étant donné un problème à résoudre, il faut récupérer (collecter) les données (par exemple des SNPs, des températures, des poids), les rassembler sous un format adéquat, afin qu'elles soient toutes contenues dans un seul tableau. La qualité et la quantité des données obtenues ont impact un direct sur le fonctionnement et le rendement du modèle.

3.4.2. Prétraitement des données : C'est un processus de nettoyage (suppression de données inutiles, répétées, incomplètes et manquantes, enrichissement par d'autres données, décomposition des données) dans le but d'acquérir des données utilisables et appropriées à l'apprentissage et à l'atteinte de l'objectif.

3.4.3. Choix du modèle : Une fois le format de données requis, concevoir la structure de l'ANN (le nombre et les types de couches, le nombre de neurones dans chaque couche, la fonction

d'activation requise, et quelques autres caractéristiques du réseau) (Moolayil, 2019) afin de choisir le modèle de la performance souhaitée.

Il existe différents modèles à utiliser selon :

- Le type, la quantité, la structure et la normalité des données.
- L'objectif : les algorithmes de prédiction (la régression logistique), le Clustering (K-means ou le voisin K), la classification (naïve bayes).

3.4.4. Entraînement du modèle : C'est un ensemble de règles pour résoudre un problème d'apprentissage, appelé un algorithme d'entraînement (Burney, Jilani, et Ardil, 2005). Il réduit la probabilité d'erreurs au fil du temps et évalue les données et garantit qu'elles sont adaptées à des résultats fiables. Les données sont séparées en :

- 80 % de ces données serviront à entraîner l'algorithme choisi,
- 10 % pour tester et vérifier la performance du résultat et
- 10 % pour valider le processus d'apprentissage (Movassagh et al, 2021).

3.4.5. Évaluation : L'apprentissage neuronal n'est considéré comme réussi que lorsque le système peut bien fonctionner sur des données de test sur lesquelles le système n'a pas été formé (Lahiri et Ghanta, 2009). Les principales façons de vérifier le rendement du modèle d'apprentissage approfondi sont les suivantes (Nighania, 2019) :

i. Matrice de confusion : Lors de l'exécution des prédictions de classification, il y a quatre types de résultats qui pourraient se produire :

- Les vrais positifs (VP) sont des classes prédites par le classificateur et qui appartiennent réellement à la vrai classe (exemple : le système prédira une séquence porteuse d'une maladie donnée alors que la séquence est réellement porteuse de ladite maladie).
- Les vrais négatifs (VN) sont les sorties prédites qui n'appartiennent pas à une classe et qui réellement n'appartiennent pas à cette classe (exemple : le système prédira une personne qui n'est pas malade alors qu'elle est réellement non malade).
- Les faux positifs (FP) se produisent lorsque la prédiction d'une sortie appartient à une classe alors qu'en réalité cette donnée appartienne à cette classe positive.
- Les faux négatifs (FN) se produisent lorsque la prédiction d'une sortie n'appartient à une classe alors qu'en réalité cette donnée appartienne à cette classe négative.

Ces quatre résultats sont tracés sur une matrice de confusion, qui est illustrée dans le tableau 9 :

Tableau 9 **:** Matrice de confusion (Nighania, 2019)

	Étiquettes observées	
Étiquettes Prédites	Vrais Positifs (VP)	Faux Positifs (FP)
	Faux Négatifs (FN)	Vrais Négatifs (VN)

Cette matrice est utilisée pour analyser le potentiel d'un algorithme de classification. Tous les éléments diagonaux dénotent des résultats correctement classés. Les résultats mal classés sont représentés sur les diagonales décalées de la matrice de confusion. Par conséquent, le meilleur classificateur aura une matrice de confusion avec seulement des éléments diagonaux et le reste des éléments mis à zéro. Une matrice de confusion génère des valeurs réelles et des valeurs prévues après le processus de classification. Une fois la matrice de confusion constituée, la performance des algorithmes de classification des données sera ensuite analysée à l'aide de paramètres de classification comme la sensibilité et la précision, qui seront calculés à l'aide des valeurs de la matrice (VP, VN, FP, FN) (Reddi et Eswar, 2021).

ii. Accuracy : Accuracy (exactitude) est décrite comme le nombre de données correctement prédites par rapport au nombre total de données. Elle est définie comme la somme des vrais négatifs et des vrais positifs combinés sur la somme des vrais positifs, des vrais négatifs, des faux positifs et des faux négatifs (Priyadarshini et Cotton, 2021) :

$$\text{Accuracy} = \frac{VP + VN}{VP + VN + FP + FN}$$

Équation 2 : La formule du calcul d'accuaracy (Priyadarshini et Cotton, 2021)

iii. Précision : Définie comme la fraction d'exemples pertinents (vrais positifs) parmi tous les exemples qui devaient appartenir à une certaine classe : Elle est décrite comme le nombre d'observations pertinentes (VP) par rapport aux observations récupérées. Compte tenu d'une matrice de confusion, la précision peut être calculée par les VP par rapport au nombre total de VP et de FP combinés (Priyadarshini et Cotton, 2021) :

$$\text{Précision} = \frac{VP}{VP + FP}$$

Équation 3 : La formule du calcul de précision (Priyadarshini et Cotton, 2021)

iv. Sensibilité : La sensibilité est représentée par le ratio d'événements positifs réels qui ont été prédits comme positifs. On l'appelle aussi le rappel. Compte tenu d'une matrice de confusion, la sensibilité peut être calculée par la valeur VP par rapport à la somme des valeurs VP et FN combinées (Priyadarshini et Cotton, 2021) :

$$\text{Sensibilité} = \frac{VP}{VP + FN}$$

Équation 4 : La formule du calcul de sensibilité (Priyadarshini et Cotton, 2021)

v. Spécificité : La spécificité est décrite comme le ratio de négatifs réels qui ont été prédits comme négatifs. Compte tenu d'une matrice de confusion, la spécificité serait calculée par les vrais négatifs par rapport à la somme des vrais négatifs et des faux positifs combinés (Priyadarshini et Cotton, 2021) :

$$\text{Spécificité} = \frac{VN}{VN + FN}$$

Équation 5 : La formule du calcul de spécificité (Priyadarshini et Cotton, 2021)

3.4.6. Réglage des paramètres : Une fois le modèle créé et évalué, son exactitude (accuracy) peut être améliorée et cela se fait en ajustant les paramètres présents dans le modèle. Les paramètres sont les variables du modèle que le programmeur décide généralement. À une valeur particulière du paramètre, l'exactitude sera le maximum. Le réglage des paramètres fait référence à la recherche de ces valeurs. Quelques exemples des paramètres sont les suivants (Lahiri et Ghanta, 2009) :

- Nombre de nœuds de couches cachées : un réseau avec peu de nœuds ne serait pas en mesure d'apprendre les relations de données correctement, alors qu'un réseau avec beaucoup de nœuds augmenterait la complexité du réseau et le temps d'exécution.
- Ajouter plus d'époques d'apprentissage : cela peut parfois mener à des exactitudes plus élevées.
- La mise à jour des valeurs de poids et biais (Lahiri et Ghanta, 2009).

3.4.7. Prédiction : Le DL consiste à utiliser les données pour répondre aux questions. La prédiction est l'étape où l'on peut obtenir des réponses à certaines questions. C'est le but de tout ce travail, où la valeur de DL est réalisée (Nighania, 2019).

PARTIE 2 :

MATÉRIEL ET MÉTHODES

1. MATÉRIEL

1.1. Données biologiques : Les données utilisées pour la suite de ce travail est un dataset des SNPs extrait le 25/12/2021. Ces données ont historiquement été enregistrées en 2014 sur Alzheimer's Disease Neuroimaging Initiative (ADNI[8] : site qui s'appuie sur des collaborations public-privé visant à déterminer les relations entre les biomarqueurs cliniques, cognitifs, d'imagerie, génétiques et biochimiques dans l'ensemble du spectre de la MA et à améliorer les essais cliniques pour la prévention et le traitement de la MA).

Le dataset téléchargé est composé de 2 dossiers (PLINK_format & snp_summary : ce dernier n'ayant pas utilisé) :

1.1.1. PLINK_format : Ce dossier contient trois fichiers (voir Tableau 10) :

Tableau 10 : Description du contenu du dossier PLINK_format[9]

Nom du fichier	Description	Taille
ADNI_plink.bim	Le fichier bim est un fichier texte contenant les informations sur les variants, il contient les noms des SNPs, leurs positions cartographiques et le numéro du chromosome sur lequel ils se trouvent.	368 MB
ADNI_plink.bed	Représentation primaire des appels génotypes aux variantes bialleliques. Doit être accompagnée de fichiers .bim et .fam.	2,80 GB
ADNI_plink.fam	Le fichier fam contient les informations phénotypiques des patients de l'étude.	21,8 KB

- Le fichier ADNI_plink.bim est composé des champs suivants : Chrom, SNP, Cm, Pos, A0, A1 et I (Tableau 11) :

[8] https://adni.loni.usc.edu/
[9] https://genome.ucsc.edu/FAQ/FAQformat.html#format1

Tableau 11 : Composants du fichier ADNI_plink.bim[10] (Mészáros, 2021)

Nom de colonne	Description
Chrom	Le numéro du chromosome où se trouve l'SNP (un entier; '0' indique qui est inconnu).
SNP	Identificateur du SNP (rsID)
Cm	Position en morgans ou centimorgans.
Pos	Position des paires de base sur le chromosome (Le comptage commence à partir du début du chromosome jusqu'au SNP spécifié).
A0	Allèle mineur : l'allèle dont on étudie les effets sur la maladie (de faible fréquence).
A1	Allèle majeur : l'allèle le plus commun au sein de la population et ne portant pas la mutation.
I	Index

- Le fichier ADNI_plink.bed : Dans ce fichier, les génotypes sont codés dans un format binaire, afin qu'il prenne un très petit espace dans le disque dur (Mészáros, 2021).

Les codes de génotypes binaires ont les significations suivantes[11] :

- 00 Homozygote pour le premier allèle (allèle mineur) dans le fichier .bim
- 01 Génotype manquant dans la base de donnée ADNI
- 10 Hétérozygote
- 11 Homozygote pour le deuxième allèle (allèle majeur) dans le fichier .bim

- Le fichier ADNI_plink.fam est composé des champs suivatns : FID, Within-family ID, Father, Mother, Gender, Trait et I (tableau 12) :

[10] https://plink.readthedocs.io/en/latest/plink_fmt/#bed
[11] https://zzz.bwh.harvard.edu/plink/

Tableau 12 : Composants du fichier ADNI_plink.fam[12] (Mészáros, 2021)

Nom de colonne	Description
FID	ID de la famille
Within-family ID	Appelé "ID individuel" et abrégé IDI. Le même IDI pourrait être utilisée avec un FID différent, il est utilisé pour avoir des IDI uniques pour chaque individu dans le dataset.
Father	ID du père ('0' si le père n'est pas dans le dataset)
Mother	ID de la mère ('0' si la mère n'est pas dans le dataset)
Gender	Code de sexe ('1' = homme, '2' = femme, '0' = inconnu)
Trait	Valeur du phénotype ('1' = témoin, '2' = cas, '-9'/'0' = données manquantes)
I	Index

1.2. Configuration de la machine : Ses caractéristiques sont détaillées dans le tableau suivant :

Tableau 13 : Caractéristiques de la machine utilisée pour le DL

Ordinateur	Caractéristiques
Processeur	Intel i5-4670K @3.40GHz (4 Cores, 4 Threads)
Mémoire installée RAM	8 Go DDR3 @ 2133Hz
Stockage	Western Digital Blue Desktop 1 To SATA 6Gb/s 64 Mo
Système d'exploitation	Windows 10 professionnel
Type de système	Système d'exploitation 64 bits

1.3. Outils et bibliothèques : Nous décrivons brièvement les outils et les bibliothèques utilisés pour réaliser le travail dans cette section :

1.3.1. Les outils : Le travail est réalisé en utilisant le langage de programmation Python, via Jupyter notebook de l'anaconda (tableau 14).

[12] https://plink.readthedocs.io/en/latest/plink_fmt/#fam

Tableau 14 **:** Principaux outils utilisés

Outil	Description
Python[13] 3.9.7	Python est un langage de programmation open source puissant et facile à apprendre. Il présente des structures de données de haut niveau efficaces et une approche de POO simple et efficace. C'est le langage idéal pour les scripts et le développement rapide d'application dans de nombreux domaines sur la plupart des plates-formes.
Anaconda[14] 4.12.0	Anaconda est libre et open source pour une distribution de données Python/R et une collection de plus de 7500 paquets open-source, qui comprend un gestionnaire de paquets et d'environnement. Anaconda est une distribution adaptée pour Windows, Linux et MacOS, et il offre un soutien communautaire gratuit.
Jupyter notebook[15] 6.4.8	Application Web open source qui permet de créer et de partager des documents Web pour l'informatique interactive dans tous les langages de programmation. Il offre une expérience simple, simplifiée et axée sur les documents.

1.3.2. Les bibliothèques : Les fonctions python utilisées proviennent de ces bibliothèques principales (tableau 15) :

Tableau 15 : Différents bibliothèques python utilisées

Bibliothèque	Description
Pandas[16] 1.4.2	Une bibliothèque open source qui fournit des structures de données (format tabulaires, telles que des données stockées dans des feuilles de calcul ou des bases de données) et des outils d'analyse de données haute performance (explorer, nettoyer et traiter vos données) et faciles à utiliser pour le langage de programmation Python.
Numpy[17] 1.19.5	C'est le paquet fondamental open source pour le calcul scientifique en Python. Il s'agit d'une bibliothèque Python qui fournit des structures de tableaux multidimensionnels, et un assortiment de routines pour les opérations rapides sur les tableaux, y compris mathématique, logique, manipulation de la forme,

[13] https://www.python.org/
[14] https://docs.anaconda.com/
[15] https://jupyter.org/
[16] https://pandas.pydata.org/docs/
[17] https://numpy.org/doc/stable/

	tri, sélection, algèbre linéaire de base, opérations statistiques de base, simulation aléatoire et bien plus encore.
Matplotlib[18] 3.5.2	Une bibliothèque communautaire complète, utiliser pour créer des visualisations statiques, animées et interactives en Python, qui rend les taches difficiles possibles.
Sklearn[19] 1.0.2	Une bibliothèque Python facilement accessible et puissante spécialisée dans les travaux de Data Science. Le Sklearn, est la bibliothèque la plus robuste pour le ML en Python. Elle fournit une sélection d'outils efficaces pour le ML et la modélisation statistique, notamment la classification, la régression, le prétraitement et la clusterisation via une interface cohérente en Python. Cette bibliothèque s'appuie sur NumPy, SciPy et Matplotlib.
Tensorflow[20] 2.6.0	Est une bibliothèque open-source de logiciels utilisé dans le DL par les réseaux de neurones et le déploiement de modèles de ML (Goldsborough 2016). Développé par Google en 2011 sous le nom de DistBelief, TensorFlow (Le nom vient de tableaux multidimensionnels connus sous le nom de tenseurs, qui sont couramment utilisés dans les réseaux de neurones) est officiellement sorti en 2017 gratuitement. La bibliothèque peut fonctionner sur plusieurs processeurs et GPU et est disponible sur plusieurs plates-formes, y compris mobile.

[18] https://matplotlib.org/
[19] https://www.data-transitionnumerique.com/scikit-learn-python/
[20] https://deepai.org/machine-learning-glossary-and-terms/tensorflow

2. MÉTHODES

L'ensemble du processus du travail est divisé en deux sections principales, la section du nettoyage des données et la section du DL :

2.1. Prétraitement des données : Plusieurs fonctions sont utilisées pour filtrer et nettoyer les données afin de les rendre plus adaptées. Nos principales étapes de nettoyage et prétraitement sont les suivantes :

2.1.1. Nettoyage du fichier fam :

- La lecture du fichier. fam avec fait apparaitre un phénotype manquant codé -9.

```
1 fam
```

	fid	iid	father	mother	gender	trait	i
0	1	014_S_0520	0	0	2	-9	0
1	2	005_S_1341	0	0	2	-9	1
2	4	012_S_0803	0	0	2	-9	2
3	5	018_S_0055	0	0	1	-9	3
4	6	027_S_0118	0	0	1	-9	4

Figure 12 : La lecture et l'affichage du fichier fam initial

- Chargement du fichier à partir de l'ADNI pour récupérer ce trait manquant.
- Sur ce fichier chaque patient présente deux phénotypes : un de son premier examen et un autre après 24 mois. Nous avons gardé le dernier état pour chaque patient car plus informatif.
- Réinitialisation de l'index après la suppression des doublants.
- Ajout des phénotypes de ce fichier au fichier fam en fonction des identifiants des patients (concaténation).
- La suppression des colonnes non nécessaires : fid, father, gender et i. Les colonnes restantes sont 'iid' et 'trait'.

```
1 fam = fam[["iid","trait"]]
2 fam
```

	iid	trait
0	014_S_0520	CN
1	005_S_1341	AD
2	012_S_0803	AD
3	018_S_0055	-9
4	027_S_0118	CN

Figure 13 : Le fichier fam final

- Enregistrement du fichier fam final sous format csv.

2.1.2. Nettoyage du fichier bim :

- Lecture du fichier bim.

```
1 bim
```

	chrom	snp	cm	pos	a0	a1	i
0	1	rs190291950	0.0	52144	A	T	0
1	1	rs140052487	0.0	54353	A	C	1
2	1	rs184233019	0.0	55852	C	G	2
3	1	rs184286948	0.0	61743	C	G	3
4	1	rs140556834	0.0	62162	A	G	4

Figure 14 : La lecture et l'affichage du fichier bim initial

- La suppression des colonnes non nécessaires : cm, pos, a0, a1 et i. Les colonnes restantes sont 'chrom' et 'snp'.

```
1 bim = bim[["chrom", "snp"]]
2 bim
```

	chrom	snp
0	1	rs190291950
1	1	rs140052487
2	1	rs184233019
3	1	rs184286948
4	1	rs140556834

Figure 15 : Le fichier bim final

- Enregistrement du fichier bim final sous format csv.

2.1.3. Nettoyage du fichier bed :

- Lecture de fichier bed.

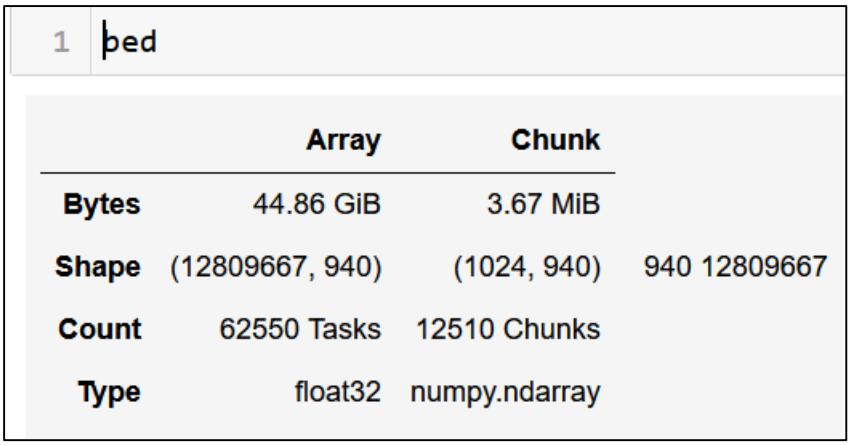
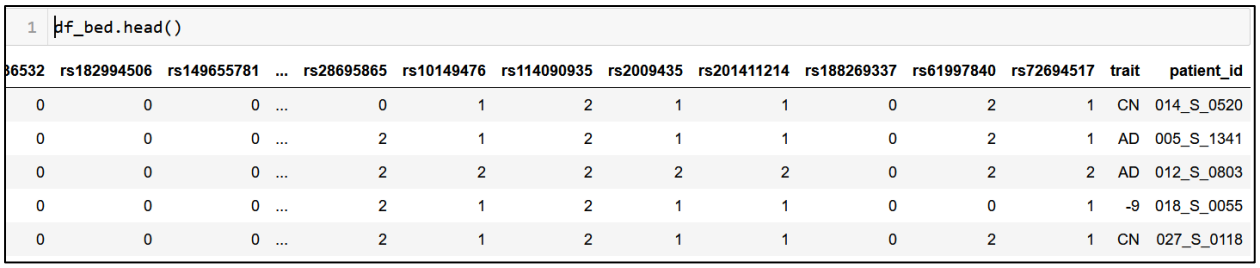


Figure 16 : La lecture et l'affichage du fichier bed initial

- Lecture des fichiers bim et fam csv.
- Ajout de noms de chromosomes et de SNP de bim à bed.
- Sélection des SNPs du chromosome 14.
- Remplacement des SNPs manquants par des zéros.
- Définition de la colonne SNP comme index.
- Ajout des phénotypes et des identifiants des patients à partir du fichier fam au fichier bed.



36532	rs182994506	rs149655781	...	rs28695865	rs10149476	rs114090935	rs2009435	rs201411214	rs188269337	rs61997840	rs72694517	trait	patient_id
0	0	0	...	0	1	2	1	1	0	2	1	CN	014_S_0520
0	0	0	...	2	1	2	1	1	0	2	1	AD	005_S_1341
0	0	0	...	2	2	2	2	2	0	2	2	AD	012_S_0803
0	0	0	...	2	1	2	1	1	0	0	1	-9	018_S_0055
0	0	0	...	2	1	2	1	1	0	2	1	CN	027_S_0118

Figure 17 : Aperçu du fichier bed après l'ajout des phénotypes et identifiants

- Définition de la colonne ID patient en tant qu'index.
- Suppression des patients présentant des phénotypes manquants (-9).
- Enregistrement du fichier bed final en format csv.

2.1.4. Réduction des colonnes (snp) du dataset en fonction du test Khi2 :

- Lecture du fichier bed précédent.
- Application du test khi2 pour réduire le nombre de colonnes, et pour sélectionner uniquement les SNPs significatifs quant à la présence ou absence du trait.
- Sélection des SNP avec khi2_score > 5,99 en utilisant la colonne des scores.

```
dataframe_reduced = dataframe_reduced[dataframe_reduced['khi2_score'] > 5.99] #valeur seuille
dataframe_reduced
```

10_S_0472	128_S_0266	...	073_S_2264	072_S_4462	009_S_2381	041_S_4014	031_S_4194	011_S_2274	033_S_4179	033_S_4176	099_S_2042	khi2_score
0	0	...	0	0	0	0	0	0	0	0	0	14.748882
0	0	...	0	0	0	0	0	0	0	0	0	8.730290
0	0	...	0	0	0	0	0	0	0	0	0	7.050664
0	0	...	0	0	0	0	0	0	0	0	0	8.730290
0	0	...	0	0	0	0	0	0	0	0	0	7.206476

Figure 18 : Aperçu du dataset après l'application du test du Khi^2 d'ajustement

- Enregistrement du fichier réduit du dataset final sous format csv.

2.2. Apprentissage :

2.2.1. Transformation des données : Conversion des valeurs du trait en nombres entier :

- Attribution de la valeur 0 à la place de la valeur CN (cognitivement normale).
- Attribution de la valeur 1 à la place des valeurs EMCI (déficience cognitive légère précoce), LMCI (déficience cognitive légère tardive), AD (maladie d'Alzheimer).

```
import numpy as np
conditions = [
    dataframe['trait'].eq('CN'),
    dataframe['trait'].eq('AD'),
    dataframe['trait'].eq('LMCI'),
    dataframe['trait'].eq('EMCI'),
            ]

choices = [0, 1, 1, 1]

dataframe['trait'] = np.select(conditions, choices, default=0)
dataframe.head()
```

Figure 19 : Code python effectuant la conversion des valeurs du trait

2.2.2. Répartition des données pour l'apprentissage, le test et l'évaluation : Les données (567 SNPs pour 767 individus) on était divisées en trois ensembles. Le premier représente 80% (soit 613 individus) des données. Elles ont servi à entraîner le modèle. La deuxième fraction du dataset représente les 10% (soit 77 individus) qui servira à tester le modèle après chaque époque, ce qui permettra au modèle d'optimiser son apprentissage, et enfin 10% (soit 77 individus) pour évaluer le modèle. La division a été effectuée aléatoirement en utilisant la fonction train_test_split de sklearn.

```python
from sklearn.model_selection import train_test_split

X = dataframe.drop('trait', axis=1)
y = dataframe['trait']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
X_test, X_eval, y_test, y_eval = train_test_split(X_test, y_test, test_size=0.5)
```

Figure 20 : Répartition des données via la fonction train_test_split

2.2.3. Construction du modèle : Le modèle défini est un modèle TensorFlow Keras séquentiel tf.keras.Sequential. Il se compose de deux couches cachées. La fonction de perte utilisée est binary_crossentropy étant donné que le problème traité est une classification binaire. Le résumé de la description du modèle est le suivant :

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 128)               72704

 dense_1 (Dense)             (None, 64)                8256

 dense_2 (Dense)             (None, 32)                2080

 dense_3 (Dense)             (None, 1)                 33

=================================================================
Total params: 83,073
Trainable params: 83,073
Non-trainable params: 0
_________________________________________________________________
```

Figure 21 : Récapitulation du réseau de neurones artificiel

```python
from tensorflow.keras import layers

model = tf.keras.Sequential([
    tf.keras.layers.Dense(128, activation='relu'),
    tf.keras.layers.Dense(64, activation='relu'),
    tf.keras.layers.Dense(32, activation='relu'),
    tf.keras.layers.Dense(1, activation='sigmoid')
])
model.compile(
    loss=tf.keras.losses.binary_crossentropy,
    optimizer=tf.keras.optimizers.Adam(learning_rate=0.03),
    metrics=[
        tf.keras.metrics.BinaryAccuracy(name='accuracy'),
        tf.keras.metrics.Precision(name='precision'),
        tf.keras.metrics.Recall(name='recall')
    ]
)
```

Figure 22 : Construction du modèle ANN

2.2.4. Apprentissage et évaluation du modèle : La fonction fit est utilisée pour lancer l'apprentissage :

```
history = model.fit(X_train,y_train, validation_data=(X_test, y_test), epochs=20)
```

Figure 23 : Utilisation de la fonction fit pour entrainer le modèle

Cela représente le début des époques d'apprentissage :

```
Epoch 1/20
20/20 [==============================] - 2s 35ms/step - loss: 1.2082 - accuracy: 0.6705 - precision: 0.7166 - recall: 0.85
10 - val_loss: 0.4648 - val_accuracy: 0.7532 - val_precision: 0.9286 - val_recall: 0.7091
Epoch 2/20
20/20 [==============================] - 0s 8ms/step - loss: 0.2859 - accuracy: 0.8646 - precision: 0.9012 - recall: 0.899
0 - val_loss: 0.1760 - val_accuracy: 0.9351 - val_precision: 0.9310 - val_recall: 0.9818
```

Figure 24 : Premières itérations de l'apprentissage du modèle

Après l'apprentissage, le modèle est évalué avec les données de validation :

```
evaluation = model.evaluate(X_eval, y_eval)
print(f"Evaluation accuracy = {evaluation[1]:.2f}")
```

2.2.5. Prédiction : Le modèle est maintenant prêt à être utilisé pour faire des prédictions :

```
predictions = model.predict(X_test)
prediction = pd.DataFrame(predictions)
prediction=prediction.apply(lambda x:round(x,2))
prediction
```

Figure 25 : Effectuation de la prédiction via la fonction predict

2.2.6. Enregistrement : Le modèle est enregistré via l'une des fonctions du TensorFlow.

```
model.save('adni_model.h5')
```

Figure 26 : Enregistrement du modèle

2.2.7. Visualisation des résultats : Les fonctions confusion_matrix, accuracy_score et classification_report de sklearn et roc_curve, auc et roc_auc_score de sklearn.metrics couplées à plt de matplotlib et sns de seaborn offrent, un excellent outil pour visualiser les résultats d'évaluation d'un modèle.

```
from sklearn.metrics import accuracy_score, precision_score, recall_score
print(f'Accuracy: {accuracy_score(y_test, prediction_classes):.2f}')
print(f'Precision: {precision_score(y_test, prediction_classes):.2f}')
print(f'Recall: {recall_score(y_test, prediction_classes):.2f}')
```

Figure 27 : Calcule des valeurs d'évaluation

```python
import seaborn as sns

cfm = confusion_matrix(y_test, prediction_classes)
ax = sns.heatmap(cfm, annot=True, cmap='Blues')

ax.set_title('Confusion Matrix for ADNI model\n\n');
ax.set_xlabel('\nPredicted Values')
ax.set_ylabel('real Values ');

## Ticket labels - List must be in alphabetical order
ax.xaxis.set_ticklabels(['CN','AD'])
ax.yaxis.set_ticklabels(['CN','AD'])

## Display the visualization of the Confusion Matrix.
plt.show()
```

Figure 28 : Code d'affichage de la matrice du confusion

```python
from sklearn.metrics import roc_curve
from sklearn.metrics import auc
from sklearn.metrics import roc_auc_score
import matplotlib.pyplot as plt
def plot_roc_curve(fpr, tpr):
    plt.plot(fpr, tpr, color='orange', label='ROC')
    plt.plot([0, 1], [0, 1], color='darkblue', linestyle='--')
    plt.xlabel('False Positive Rate')
    plt.ylabel('True Positive Rate')
    plt.title('Receiver Operating Characteristic (ROC) Curve')
    plt.legend()
    plt.show()
# Computing manually fpr, tpr, thresholds and roc auc
fpr, tpr, thresholds = roc_curve(y_test, predictions)
roc_auc = auc(fpr, tpr)
print("ROC_AUC Score : ",roc_auc)
print("Function for ROC_AUC Score : ",roc_auc_score(y_test, predictions))
optimal_idx = np.argmax(tpr - fpr)
optimal_threshold = thresholds[optimal_idx]
print("Threshold value is:", optimal_threshold)
plot_roc_curve(fpr, tpr)
```

Figure 29 : Code d'affichage de la courbe ROC

PARTIE 3 :

RÉSULTATS ET DISCUSSION

RÉSULTATS

A la fin de l'apprentissage, le modèle est testé dans le but de vérifier son efficacité. Le test a été fait sur 10% (soit 77 individus) des données mentionnées dans la partie méthodes et est réalisé grâce à la matrice de confusion représentée sur la figure 12.

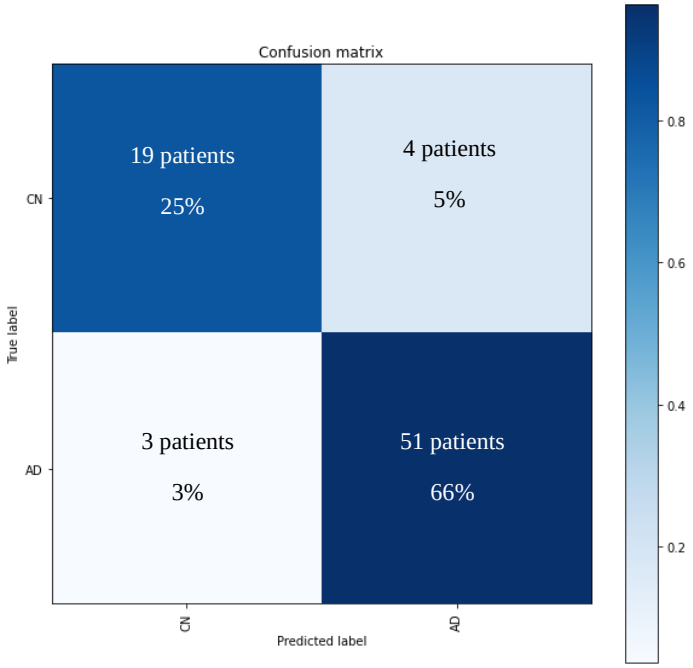


Figure 30 : Matrice de confusion du test du modèle d'ADNI

Cette matrice a permis de calculer les valeurs suivantes :

- Accuracy = 92.2%
- Précision = 93%
- Sensibilité = 91%
- ROC_AUC = 91.62%

La matrice de confusion est par la suite utilisée à nouveau pour évaluer le modèle en faisant des prédictions sur les 10% (77 individus) des données d'évaluation.

La matrice de confusion de l'évaluation a permis de calculer les valeurs suivantes :

- Accuracy = 92.2%
- Précision = 92.9%
- Sensibilité = 96%

Le résultat obtenu par la prédiction est délimité par zéro et un (tableau 18), comme suit :

- Si la valeur de prédiction est inférieure à 0.5, c'est-à-dire la prédiction est négative et donc le patient est CN.

- Si la valeur de prédiction est supérieure ou égale à 0.5, c'est-à-dire la prédiction est positive et donc le patient est AD.

Tableau 16 **:** Table des phénotypes réels contre ceux prédits

	real_phenotype	predicted_phenotype
100_S_0296	1	1
137_S_0994	1	1
128_S_2151	1	1
123_S_4170	1	0
029_S_1318	1	1
941_S_4365	0	0
126_S_0680	0	0
073_S_4360	1	1
013_S_0240	1	1
037_S_4214	1	1

L'accuracy, précision, sensibilité et le score ROC_AUC sont calculées à l'aide de la matrice de confusion :

- La matrice de confusion : est représentée dans la figure 13, sous forme d'un tableau 2×2. Le nombre de lignes et de colonnes est en fonction du nombre de classes (deux classes). Les lignes correspondent aux valeurs réelles d'une classe tandis que les colonnes indiquent les valeurs prédites. La matrice de confusion nous aide à visualiser si le modèle est confus ou bien performant dans la discrimination entre les deux classes. La matrice de confusion de notre modèle est la suivante :

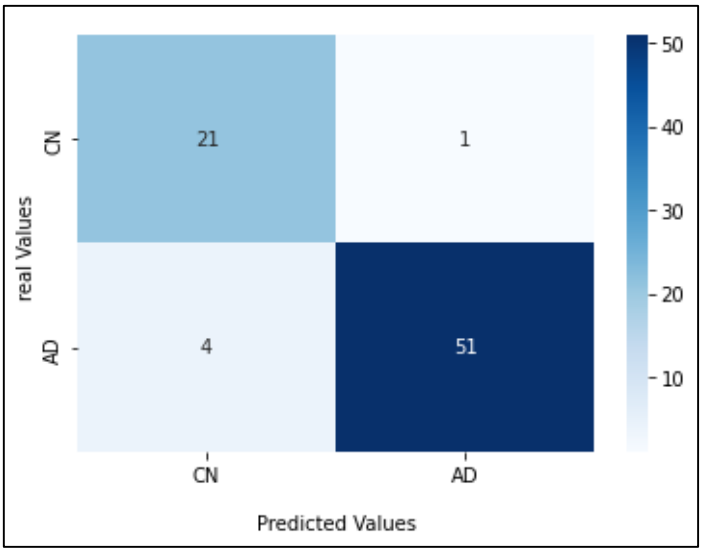


Figure 31 : Matrice de confusion de validation du modèle d'ADNI

La matrice est calculée à l'aide des données de validation (77 individus), dans lesquelles 55 individus appartiennent à la classe AD et 22 à la classe CN. Le modèle prédit 51 AD (Vrais positifs) et 21 CN (Vrais négatifs) correctement, ce qui laisse 4 AD (Faux négatifs) et 1 CN (Faux positifs).

- Zone sous la courbe AUC et caractéristique de fonctionnement du récepteur ROC : La courbe de caractéristique de fonctionnement du récepteur (ROC) est une mesure d'évaluation des problèmes de classification binaire. C'est une courbe de probabilité qui trace le taux de vrais positifs par rapport au taux de faux positifs. La zone sous la courbe (AUC) est la mesure de la capacité d'un classificateur à distinguer les classes et est utilisée comme résumé de la courbe ROC. Plus l'AUC est élevée (dans notre cas il est égale à 0.91 %, voir figure 14), plus le modèle est efficace pour distinguer les classes positives des classes négatives.

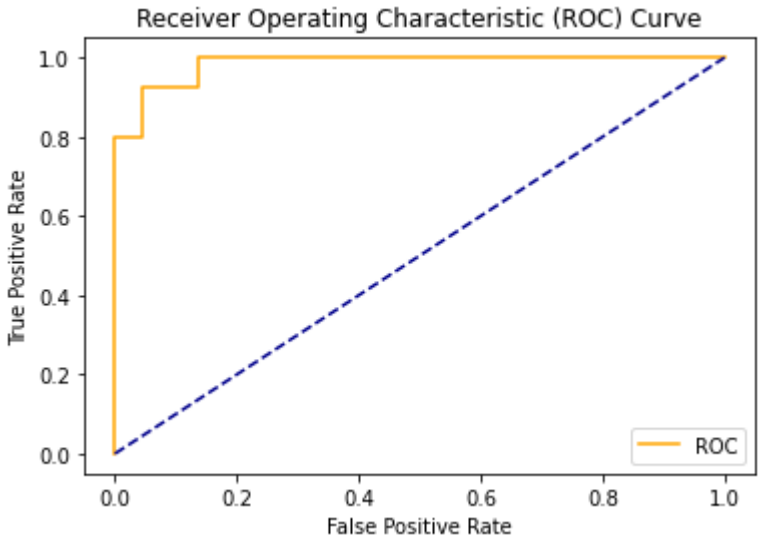


Figure 32 **:** Presentation graphique de la courbe de caractéristique de fonctionnement du récepteur

DISCUSSION

Dans notre travail nous avons identifié 567 SNPs localisés dans le chromosome 14 avec des positions bien précises responsables de la MA de 767 échantillons témoins. A l'aide de notre modèle ANN génétique prédictif adapté, nous somme dans la capacité d'indiquer la présence des SNPs chez un individu. Ce modèle serait suffisant pour informer de la présence ou de l'absence de la MA pour un patient donné.

L'importance de ce travail est la prévention de la MA, de la découverte des nouveaux mécanismes (de nouvelles interactions génétiques impliqués dans la MA ça donne un dépistage pour des nouvelles sites thérapeutiques), d'améliorer la compréhension des phénomènes d'héritabilité de celle-ci.

La méthode proposée a réussi à classer les personnes non malades et malades et a montré son efficacité avec une précision et une accuracy de 92% et un score ROC_AUC de 91%. Les résultats obtenus durant ce travail démontrent que notre modèle ANN pourrait jouer un rôle efficace dans la prédiction et la classification de la MA. La comparaison entre les différents travaux récents peut se faire selon différents critères présentés dans le tableau 19 tenant compte des références suivantes :

- Alzheimer's Disease Classification Through Imaging Genetic Data With IGnet : Ce travail vise à prédire la maladie d'Alzheimer à l'aide de l'imagerie et des données génétiques, ils se sont concentrés sur les 8946 SNPs sur le chromosome 19 et les IRM du cerveau, les deux données sont extraites de l'ADNI (Wang et al, 2022).
- Deep learning-based identification of genetic variants (application to Alzheimer's disease classification) : ils ont divisé le génome entier en fragments d'une taille optimale et ont ensuite exécuté le CNN (réseau de neurones convolutionnel) sur chaque fragment pour sélectionner des fragments associés au phénotype. Ensuite à l'aide d'un test d'association par fenêtre coulissante (SWAT), ils ont exécuté le CNN sur les fragments sélectionnés pour calculer les scores d'influence du phénotype (PIS) et identifier les SNP associés au phénotype basés sur PIS. Enfin, ils ont exécuté CNN sur tous les SNP identifiés pour développer un modèle de classification (Jo et al, 2022).
- Multimodal deep learning models for early detection of Alzheimer's disease stage : Ils ont utilisé l'ensemble de données d'ADNI qui contient les SNPs (808 patients),

l'imagerie par IRM (503 patients) et les données des tests cliniques et neurologiques (2004 patients). Ils ont utilisé les CNNs pour les données d'imagerie (Venugopalan et al, 2021).

Tableau 17 : Comparaison avec d'autre travaux similaires

	Description des données utilisées	Accuracy	Précision	Sensibilité	ROC_AUC
Notre travail	567 SNPs sur le chromosome 14 pour 767 individus	92%	92%	96%	91%
(Wang et al, 2022)	8946 SNPs sur le chromosome 19 et les IRM du cerveau	78%	77,78%	_	92%
(Jo et al, 2022)	SNPs 981 individus; (CN=650 et AD=331)	75%	_	_	82%
(Venugopalan et al, 2021)	SNPs (808 individus), IRM (503 individus) et les données des tests cliniques et neurologiques (2004 individus).	86%	66%	67%	_

Notre travail présente encore des limites surtout en termes de nombre de données SNPs qui restent un autre avis insuffisant. Le DL étant une approche nécessite d'une masse plus importante que la nôtre pour des résultats plus fiables et plus précis donc plus crédible.

CONCLUSION

CONCLUSION

Notre étude a confirmé la faisabilité et surtout l'importance de prédire la MA à partir des données génomiques SNPs, en utilisant un modèle de DL qui prend en charge les SNPs du dataset ADNI, les filtres à l'aide du test d'ajustement du khi^2, et utilise un ANN comme classificateur de ces SNPs. Nous avons validé notre algorithme d'apprentissage en utilisant 77 personnes du dataset ADNI ; ce qui a produit un classificateur dont l'accuracy était nettement meilleure par rapport à d'autres travaux similaires (Wang et al, 2022 ; Jo et al, 2022 ; Venugopalan et al., 2021).

De plus, n'a porté que sur les SNPs localisés dans le chromosome 14, qui est un petit nombre par rapport au nombre des SNPs provenant d'une analyse du génome entier. Le fait que notre étude ait produit des résultats statistiquement significatifs, malgré ces limites statistiques en nombre et en qualité, démontre le potentiel de cette approche de DL dans le contexte de la prédiction des maladies.

Dans nos travaux futurs, nous prévoyons d'appliquer cette méthode à des ensembles de données plus importantes et plus crédibles par l'utilisation des SNPs de tous les chromosomes et sur un échantillon humain plus grand, afin d'identifier de nouveaux éventuels SNPs liés à la MA et élaborer par voie de conséquence des modèles de classification plus précis et plus crédibles .

RÉFÉRENCES BIBLIOGRAPHIQUES

RÉFÉRENCES

A. Armstrong, Richard. 2013. « Review Article What Causes Alzheimer's Disease? » *Folia Neuropathologica* 3:169-88. doi: 10.5114/fn.2013.37702.

Abril, Josep F., et Sergi Castellano. 2019. « Genome Annotation ». P. 195-209 in *Encyclopedia of Bioinformatics and Computational Biology*, édité par S. Ranganathan, M. Gribskov, K. Nakai, et C. Schönbach. Oxford: Academic Press.

Andreassen, Anders, et Benjamin Nachman. 2020. « Neural Networks for Full Phase-Space Reweighting and Parameter Tuning ». *Physical Review D* 101(9):091901. doi: 10.1103/PhysRevD.101.091901.

Anon. 2022. « Korf: Human Genetics and Genomics ». Consulté 6 avril 2022 (https://www.korfgenetics.com/).

Aubourg, Sébastien, et Pierre Rouzé. 2001. « Genome Annotation ». *Plant Physiology and Biochemistry* 39(3):181-93. doi: 10.1016/S0981-9428(01)01242-6.

Barshir, Ruth, Idan Hekselman, Netta Shemesh, Moran Sharon, Lena Novack, et Esti Yeger-Lotem. 2018. « Role of Duplicate Genes in Determining the Tissue-Selectivity of Hereditary Diseases » édité par G. Gibson. *PLOS Genetics* 14(5):e1007327. doi: 10.1371/journal.pgen.1007327.

Berry, Michael W., Azlinah Mohamed, et Bee Wah Yap, éd. 2020. *Supervised and Unsupervised Learning for Data Science*. Cham: Springer International Publishing.

Beyne, Emmanuelle. s. d. « Règles de cohérence pour l'annotation génomique: développement et mise en oeuvre in silico et in vivo ». 200.

Burney, Syed Muhammad Aqil, Tahseen Ahmed Jilani, et Cemal Ardil. 2005. « A Comparison of First and Second Order Training Algorithms for Artificial Neural Networks ». 1:7.

Buscema, Paolo Massimo, Giulia Massini, Marco Breda, Weldon A. Lodwick, Francis Newman, et Masoud Asadi-Zeydabadi. 2018. « Artificial Neural Networks ». P. 11-35 in *Artificial Adaptive Systems Using Auto Contractive Maps*. Vol. 131, *Studies in Systems, Decision and Control*. Cham: Springer International Publishing.

DeTure, Michael A., et Dennis W. Dickson. 2019. « The Neuropathological Diagnosis of Alzheimer's Disease ». *Molecular Neurodegeneration* 14(1):32. doi: 10.1186/s13024-019-0333-5.

Dick, Stephanie. 2019. « Artificial Intelligence ». *Harvard Data Science Review*. doi: 10.1162/99608f92.92fe150c.

Dike, Happiness Ugochi, Yimin Zhou, Kranthi Kumar Deveerasetty, et Qingtian Wu. 2018. « Unsupervised Learning Based On Artificial Neural Network: A Review ». P. 322-27 in *2018 IEEE International Conference on Cyborg and Bionic Systems (CBS)*. Shenzhen: IEEE.

Dongare, A. D., R. R. Kharde, et Amit D. Kachare. 2012. « Introduction to Artificial Neural Network ». 2(1):6.

Dunn, Nathan A., Deepak R. Unni, Colin Diesh, Monica Munoz-Torres, Nomi L. Harris, Eric Yao, Helena Rasche, Ian H. Holmes, Christine G. Elsik, et Suzanna E. Lewis. 2019. « Apollo: Democratizing Genome Annotation » édité par A. E. Darling. *PLOS Computational Biology* 15(2):e1006790. doi: 10.1371/journal.pcbi.1006790.

ergopix. 2012. « List of Chromosomes | Chromosome Walk ». Consulté 1 juin 2022 (https://www.chromosomewalk.ch/en/list-of-chromosomes/).

Georgevici, Adrian Iustin, et Marius Terblanche. 2019. « Neural Networks and Deep Learning: A Brief Introduction ». *Intensive Care Medicine* 45(5):712-14. doi: 10.1007/s00134-019-05537-w.

van Gerven, Marcel, et Sander Bohte. 2017. « Editorial: Artificial Neural Networks as Models of Neural Information Processing ». *Frontiers in Computational Neuroscience* 11:114. doi: 10.3389/fncom.2017.00114.

Goldman, Aaron David, et Laura F. Landweber. 2016. « What Is a Genome? » édité par W. F. Doolittle. *PLOS Genetics* 12(7):e1006181. doi: 10.1371/journal.pgen.1006181.

Goldsborough, Peter. 2016. « A Tour of TensorFlow ».

Govaerts, L., J. Schoenen, et D. Bouhy. s. d. « PATHOGéNIE DE LA MALADIE D'ALZHEIMER : les mécanismes moléculaires et cellulaires ». *Rev Med Liege* 8.

Gupta, Rohan, Devesh Srivastava, Mehar Sahu, Swati Tiwari, Rashmi K. Ambasta, et Pravir Kumar. 2021. « Artificial Intelligence to Deep Learning: Machine Intelligence Approach for Drug Discovery ». *Molecular Diversity* 25(3):1315-60. doi: 10.1007/s11030-021-10217-3.

Habimana, Olivier, Yuhua Li, Ruixuan Li, Xiwu Gu, et Ge Yu. 2020. « Sentiment Analysis Using Deep Learning Approaches: An Overview ». *Science China Information Sciences* 63(1):111102. doi: 10.1007/s11432-018-9941-6.

Hofker, Marten H., Jingyuan Fu, et Cisca Wijmenga. 2014. « The Genome Revolution and Its Role in Understanding Complex Diseases ». *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 1842(10):1889-95. doi: 10.1016/j.bbadis.2014.05.002.

Hu, Yang, Tianyi Zhao, Tianyi Zang, Ying Zhang, et Liang Cheng. 2019. « Identification of Alzheimer's Disease-Related Genes Based on Data Integration Method ». *Frontiers in Genetics* 9:703. doi: 10.3389/fgene.2018.00703.

Hussain, Kamal. s. d. « Artificial Intelligence and Its Applications Goal ». 05(01):5.

Iourov, Ivan Y., Svetlana G. Vorsanova, et Yuri B. Yurov. 2019. « Pathway-Based Classification of Genetic Diseases ». *Molecular Cytogenetics* 12(1):4. doi: 10.1186/s13039-019-0418-4.

Ippolito, Marco, John Ferguson, et Fred Jenson. 2021. « Improving Facies Prediction by Combining Supervised and Unsupervised Learning Methods ». *Journal of Petroleum Science and Engineering* 200:108300. doi: 10.1016/j.petrol.2020.108300.

Ismail, Somaia, et Mona Essawi. 2012. « Genetic Polymorphism Studies in Humans »: *Middle East Journal of Medical Genetics* 1(2):57-63. doi: 10.1097/01.MXE.0000415225.85003.47.

Jo, Taeho. 2021. *Machine Learning Foundations: Supervised, Unsupervised, and Advanced Learning*. Cham: Springer International Publishing.

Jo, Taeho, Kwangsik Nho, Paula Bice, Andrew J. Saykin, et For The Alzheimer's Disease Neuroimaging Initiative. 2022. « Deep Learning-Based Identification of Genetic Variants: Application to Alzheimer's Disease Classification ». *Briefings in Bioinformatics* 23(2):bbac022. doi: 10.1093/bib/bbac022.

Kattenborn, Teja, Jens Leitloff, Felix Schiefer, et Stefan Hinz. 2021. « Review on Convolutional Neural Networks (CNN) in Vegetation Remote Sensing ». *ISPRS Journal of Photogrammetry and Remote Sensing* 173:24-49. doi: 10.1016/j.isprsjprs.2020.12.010.

Krolak-Salmon, P. 2020. « Physiopathologie de la maladie d'Alzheimer : le rôle central de la plaque amyloïde et de la protéine Tau ». *NPG Neurologie - Psychiatrie - Gériatrie* 20(120):120S2-6. doi: 10.1016/S1627-4830(20)30195-1.

Kukreja, Harsh. 2016. « AN INTRODUCTION TO ARTIFICIAL NEURAL NETWORK ». 1(5):4.

Lahiri, S. K., et K. C. Ghanta. 2009. « Artificial Neural Network Model with Parameter Tuning Assisted by Genetic Algorithm Technique: Study of Critical Velocity of Slurry Flow in Pipeline ». *Asia-Pacific Journal of Chemical Engineering* n/a-n/a. doi: 10.1002/apj.403.

Leonenko, Ganna, Rebecca Sims, Maryam Shoai, Aura Frizzati, Paola Bossù, Gianfranco Spalletta, Nick C. Fox, Julie Williams, the GERAD consortium, John Hardy, et Valentina Escott-Price. 2019. « Polygenic Risk and Hazard Scores for Alzheimer's Disease Prediction ». *Annals of Clinical and Translational Neurology* 6(3):456-65. doi: 10.1002/acn3.716.

Lexcellent, Christian. 2019. « Artificial Intelligence ». P. 5-21 in *Artificial Intelligence versus Human Intelligence*, *SpringerBriefs in Applied Sciences and Technology*. Cham: Springer International Publishing.

Li, Xiong, Liyue Liu, Juan Zhou, et Che Wang. 2018. « Heterogeneity Analysis and Diagnosis of Complex Diseases Based on Deep Learning Method ». *Scientific Reports* 8(1):6155. doi: 10.1038/s41598-018-24588-5.

Li, Yu, Chao Huang, Lizhong Ding, Zhongxiao Li, Yijie Pan, et Xin Gao. 2019. « Deep Learning in Bioinformatics: Introduction, Application, and Perspective in the Big Data Era ». *Methods* 166:4-21. doi: 10.1016/j.ymeth.2019.04.008.

Liu, Chia-Chen, Takahisa Kanekiyo, Huaxi Xu, et Guojun Bu. 2013. « Apolipoprotein E and Alzheimer Disease: Risk, Mechanisms and Therapy ». *Nature Reviews Neurology* 9(2):106-18. doi: 10.1038/nrneurol.2012.263.

Mahdi, Kooshyar Mohammad, Mohammad Reza Nassiri, et Khadijeh Nasiri. 2013. « Hereditary Genes and SNPs Associated with Breast Cancer ». *Asian Pacific Journal of Cancer Prevention* 14(6):3403-9. doi: 10.7314/APJCP.2013.14.6.3403.

Mészáros, Gábor. s. d. *Chapter 6 Genotype files in practice | Genomics Boot Camp*.

Moolayil, Jojo. 2019. « An Introduction to Deep Learning and Keras ». P. 1-16 in *Learn Keras for Deep Neural Networks*. Berkeley, CA: Apress.

Movassagh, Ali Akbar, Jafar A. Alzubi, Mehdi Gheisari, Mohamadtaghi Rahimi, Senthilkumar Mohan, Aaqif Afzaal Abbasi, et Narjes Nabipour. 2021. « Artificial Neural Networks Training Algorithm Integrating Invasive Weed Optimization with Differential Evolutionary Model ». *Journal of Ambient Intelligence and Humanized Computing*. doi: 10.1007/s12652-020-02623-6.

Nighania, Kartik. 2019. « Various Ways to Evaluate a Machine Learning Models Performance ». *Medium*. Consulté 30 mars 2022 (https://towardsdatascience.com/various-ways-to-evaluate-a-machine-learning-models-performance-230449055f15).

Pourtau, Lionel. s. d. « L'arrivée de la médecine prédictive, quelle autonomie du sujet après le dépistage d'une prédisposition du cancer ? » 11.

Priyadarshini, Ishaani, et Chase Cotton. 2021. « A Novel LSTM–CNN–Grid Search-Based Deep Neural Network for Sentiment Analysis ». *The Journal of Supercomputing* 77(12):13911-32. doi: 10.1007/s11227-021-03838-w.

Reddi, Sivaranjani, et G. V. Eswar. 2021. « Chapter 9 - Fake News in Social Media Recognition Using Modified Long Short-Term Memory Network ». P. 205-27 in *Security in IoT Social*

Networks, *Intelligent Data-Centric Systems*, édité par F. Al-Turjman et B. D. Deebak. Academic Press.

Ridley, Matt. 1999. *Genome: The Autobiography of a Species in 23 Chapters*. 1st U.S. ed. New York: HarperCollins.

Romero-Rosales, Brissa-Lizbeth, Jose-Gerardo Tamez-Pena, Humberto Nicolini, Maria-Guadalupe Moreno-Treviño, et Victor Trevino. 2020. « Improving Predictive Models for Alzheimer's Disease Using GWAS Data by Incorporating Misclassified Samples Modeling » édité par J. Gwak. *PLOS ONE* 15(4):e0232103. doi: 10.1371/journal.pone.0232103.

Sanabria-Castro, Alfredo, Ileana Alvarado-Echeverría, et Cecilia Monge-Bonilla. 2017. « Molecular Pathogenesis of Alzheimer's Disease: An Update ». *Annals of Neurosciences* 24(1):46-54. doi: 10.1159/000464422.

Sarker, Iqbal H. 2021. « Machine Learning: Algorithms, Real-World Applications and Research Directions ». *SN Computer Science* 2(3):160. doi: 10.1007/s42979-021-00592-x.

Schlegel, Dennis, et Yasin Uenal. s. d. « A Perceived Risk Perspective on Narrow Artificial Intelligence ». 15.

Sfar, S., et L. Chouchane. 2008. « Le projet génome humain : programme fédérateur de la médecine génomique ». *Pathologie Biologie* 56(3):170-75. doi: 10.1016/j.patbio.2007.12.001.

Shinde, Pramila P., et Seema Shah. 2018. « A Review of Machine Learning and Deep Learning Applications ». P. 1-6 in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*. Pune, India: IEEE.

Slim, K., M. Selvy, et J. Veziant. 2021. « Innovation conceptuelle : la médecine 4P et la chirurgie 4P ». *Journal de Chirurgie Viscérale* 158(3):S13-18. doi: 10.1016/j.jchirv.2021.01.001.

Srinidhi, Chetan L., Ozan Ciga, et Anne L. Martel. 2021. « Deep Neural Network Models for Computational Histopathology: A Survey ». *Medical Image Analysis* 67:101813. doi: 10.1016/j.media.2020.101813.

Sripichai, Orapan, et Suthat Fucharoen. 2007. « Genetic Polymorphisms and Implications for Human Diseases ». 90(2):5.

Tak, Yu Gyoung, et Peggy J. Farnham. 2015. « Making Sense of GWAS: Using Epigenomics and Genome Engineering to Understand the Functional Relevance of SNPs in Non-Coding Regions of the Human Genome ». *Epigenetics & Chromatin* 8(1):57. doi: 10.1186/s13072-015-0050-4.

Tawfik, Noha S., et Marco R. Spruit. 2018. « *The SNPcurator* : Literature Mining of Enriched SNP-Disease Associations ». *Database* 2018. doi: 10.1093/database/bay020.

Timpson, Nicholas J., Celia M. T. Greenwood, Nicole Soranzo, Daniel J. Lawson, et J. Brent Richards. 2018. « Genetic Architecture: The Shape of the Genetic Contribution to Human Traits and Disease ». *Nature Reviews Genetics* 19(2):110-24. doi: 10.1038/nrg.2017.101.

Venugopalan, Janani, Li Tong, Hamid Reza Hassanzadeh, et May D. Wang. 2021. « Multimodal Deep Learning Models for Early Detection of Alzheimer's Disease Stage ». *Scientific Reports* 11(1):3254. doi: 10.1038/s41598-020-74399-w.

Wang, Jade Xiaoqing, Yimei Li, Xintong Li, et Zhao-Hua Lu. 2022. « Alzheimer's Disease Classification Through Imaging Genetic Data With IGnet ». *Frontiers in Neuroscience* 16:846638. doi: 10.3389/fnins.2022.846638.

Woschank, Manuel, Erwin Rauch, et Helmut Zsifkovits. 2020. « A Review of Further Directions for Artificial Intelligence, Machine Learning, and Deep Learning in Smart Logistics ». *Sustainability* 12(9):3760. doi: 10.3390/su12093760.

Xu, H., S. G. Gregory, E. R. Hauser, J. E. Stenger, M. A. Pericak-Vance, J. M. Vance, S. Zuchner, et M. A. Hauser. 2005. « SNPselector: A Web Tool for Selecting SNPs for Genetic Association Studies ». *Bioinformatics* 21(22):4181-86. doi: 10.1093/bioinformatics/bti682.

Zhang, Caiming, et Yang Lu. 2021. « Study on Artificial Intelligence: The State of the Art and Future Prospects ». *Journal of Industrial Information Integration* 23:100224. doi: 10.1016/j.jii.2021.100224.

Année universitaire : 2021-2022	**Présenté par :** ABDELAZIZ Aya NOUI Manel Ghosn El Ben

L'annotation des SNPs génomiques pour la prédiction de la maladie génétique/héréditaire d'Alzheimer par Deep Learning

Mémoire pour l'obtention du diplôme de Master en Bioinformatique

La maladie d'Alzheimer (MA) est une maladie irréversible dans le cerveau qui provoque des troubles neurodégénératifs progressifs. La MA peut être détectée en dépistant des gènes spécifiques dans des chromosomes spécifiques responsables de cette maladie où se trouvent les mutations qui se produisent dans les SNP sur ces gènes spécifiques. Notre travail a mis en exergu l'importance de l'utilisation des données de la BDD ADNI pour étudier et analyser la MA, en développant une approche bioinformatique d'apprentissage approfondi pour classer les deux stades de la maladie (MA et CN). L'objectif de cette approche est de développer un modèle d'annotation (ou un système de prédiction) afin de d'identifier le type de la maladie chez des individus ou d'estimer le stade de celle-ci, à l'aide d'un Réseau de Neurones Artificiels, comme résultats, nous avons classé les données génétiques fonctionnelles avec une précision de test qui a atteint 94%.

Mots-clés : Maladie d'Alzheimer ; ADNI ; SNPs ; Apprentissage approfondi ; Réseau de Neurones Artificiel ; Annotation.

Président : Dr CHEHILI Hamza (Université Frères Mentouri, Constantine 1).

Encadreur : Pr HAMIDECHI M. Abdelhafid (Université Frères Mentouri, Constantine 1).

Examinateur : Dr BOULAHROUF Khaled (Université Frères Mentouri, Constantine 1).