REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE MOHAMED BOUDIAF - M'SILA

FACULTE DES SCIENCES

DEPARTEMENT DE MICROBIOLOGIE & BIOCHIMIE

N°:

DOMAINE : SCIENCES DE LA NATURE ET DE LA VIE

FILIERE : SCIENCE BIOLOGIQUE

OPTION : BIOCHIMIE APPLIQUEE

Mémoire présenté pour l'obtention

Du diplôme de Master Académique

Par : Boudjellal Farah

Berbache Amal

Telli Sabrina

Intitulé

Conception d'un modèle d'IA dédié à l'identification de métabolites inconnus en exploitant les données NMR

Soutenu devant le jury composé de :

Dr. Abdallah RAHALI	Université Mohamed Boudiaf M'sila	Président
Dr. Abdenassar HARRAR	Université Mohamed Boudiaf M'sila	Rapporteur
Dr. Mohammed Abdallah KHODJA	Université Mohamed Boudiaf M'sila	Co-Rapporteur
Dr. Ahlam FATMI	Université Mohamed Boudiaf M'sila	Examinatrice

Année universitaire : 2024 /2025

Dédicace

Merci, Seigneur, de m'avoir accordé la capacité d'écrire et de penser, ainsi que la patience nécessaire pour réaliser mes rêves et surmonter les obstacles. Je suis honoré de dédier cet humble ouvrage :

À mon père, pour son soutien et son encadrement tout au long de mon parcours universitaire.

À ma chère mère, je ne trouve pas les mots pour exprimer mon profond respect et ma gratitude pour vos sacrifices pour mon éducation.

Je demande à Dieu Tout-Puissant de vous accorder santé, bien-être et longue vie.

Je n'oublie pas non plus mon frère Houssam pour son soutien constant, notamment pour mes mémoires. Je le félicite également pour son mariage et vous souhaite du bonheur.

À mes frères Nabil et Ilyes, je vous souhaite également bonne chance et bonheur.

À mon âme sœur pour le grand soutien moral et les prières sincères.

À mon oncle et deuxième mère, merci encore pour votre affection et votre gentillesse.

À mes professeurs, pour leur expérience, leurs conseils avisés et leur précieuse patience, qui ont enrichi mes connaissances et façonné ma pensée.

À mes collègues qui m'ont accompagné avec sincérité.

À tous ceux qui m'aiment et que j'aime.

J'espère aussi que mon succès ne s'arrêtera jamais et

Que Dieu m'accordera la réussite dans tout ce que mon

Cœur désire.

Boudjellal Farah

Dédicace

Louange à Dieu par la grâce de qui les bonnes œuvres s'accomplissent et c'est avec Son aide que ce travail a pu être mené à bien.

Je dédie ce mémoire ***A mes parents****, qui m'ont soutenu à chaque étape et ont toujours été mon soutien et mon refuge, merci pour chaque mot d'encouragement et chaque prière dans le dos de l'invisible, chaque lettre de ce travail est le fruit de votre patience et de vos sacrifices.*

À mes frères ***Faycel*** *,****Haydar*** *,****kamel et Radouane****, ainsi qu'à mes chères sœurs* ***Souad, Rima*** *et* ***Khawla****, je suis reconnaissant pour leur soutien et leurs encouragements constants.*

À mes honorables professeurs, qui ont donné de leur temps et de leurs efforts pour nous enseigner et nous guider, et qui n'ont pas lésiné sur leurs connaissances et leur expertise.

A mes collègues ***Farah****, avec qui j'ai partagé la réalisation de ce projet scientifique*

À mes chères amies, ***Nour El Houda*** *et* ***Abir*** *vous avez été comme des sœurs que la résidence universitaire m'a offertes, merci d'avoir été l'épaule sur laquelle je peux m'appuyer quand le fardeau devient lourd.*

Enfin, à tous ceux qui m'ont soutenu par un mot gentil ou un encouragement sincère, à tous ceux qui ont cru en mes capacités ne serait-ce qu'un instant, je dédie ce modeste effort, en espérant qu'il sera une pierre à l'édifice de mon avenir et un pas vers la réalisation de mes ambitions. Que Dieu m'accorde le succès.

Berbache Amal

Dédicace

D’un cœur épuisé par la nostalgie, je commence mes mots de là-bas…
De Gaza, cette plaie ouverte, cette dignité persistante malgré le blocus,
De cette terre qui résiste et enseigne au monde entier des leçons de résilience,
Je dédie mon diplôme à chaque âme qui a reposé dans sa terre sacrée,
Aux martyrs, aux prisonniers, à ceux qui ont tenu bon sous les décombres et sont restés fidèles à l’espoir.

Ensuite, à ceux qui ont affronté la vie pour moi,
À ceux qui ont planté le rêve dans mon cœur et l’ont arrosé de leurs prières,
À ceux qui furent la lumière de mon chemin et mon ombre lorsque les jours m’alourdissaient.
Ce succès, je n’en ai été que le moyen,
Mais vous avez été la véritable finalité vers laquelle j’ai tendu de tout mon cœur.

À mes sœurs, Khouloud, Djhed, Rouaida,
Je suis la sœur aînée, celle qui a ouvert la voie,
Mais vous avez été la lumière derrière moi,
La force qui me poussait en avant, et la chaleur qui m’enveloppait.

En vous, j’ai découvert le vrai sens de la tendresse,
Et dans vos rires, la véritable signification de la sérénité.
Vous avez été la côte stable de mon âme lorsque tout vacillait,
Et vos prières ont veillé sur mon cœur,
Vos paroles ont essuyé la fatigue de mes jours.

À mon frère, Abdel Rahman,
Qui fut mon soutien à chaque instant,
Mon compagnon dans la difficulté avant la joie,
Ce diplôme ne saurait porter mon nom sans évoquer ton cœur, qui m’a précédée.

À ceux qui ont cru en moi, même lorsque je doutais de moi-même,
Ce diplôme ne m’appartient pas à moi seule,
Il est à chacun qui résiste à sa manière : par le savoir, la patience et l’amour.

Car c’est à Gaza que l’histoire a commencé…
Et c’est à elle que toutes les histoires reviennent.

Telli Sabrina

Remerciement

Avant tout, nous exprimons notre profonde gratitude à **Allah**, le Tout-Puissant, pour nous avoir accordé la force, la volonté et la patience nécessaires à l'accomplissement de ce travail.

Nous adressons nos remerciements les plus sincères à notre encadreur, **Dr. Abdenassar HARRAR**, pour nous avoir proposé ce sujet de mémoire, ainsi que pour sa disponibilité constante, son accompagnement rigoureux et ses conseils éclairés tout au long de ce projet. Sa patience, sa bienveillance et ses compétences scientifiques ont été d'une grande valeur. Nous lui témoignons notre profonde reconnaissance, et nous lui souhaitons santé et réussite.

Nous adressons nos plus sincères remerciements au **Dr. Abdallah RAHALI**, Président du jury, pour l'honneur qu'il nous a fait en acceptant de présider cette soutenance et pour l'intérêt attentif qu'il a porté à notre travail.

Nous exprimons également notre profonde gratitude au **Dr. Ahlam FATMI**, Examinatrice, pour avoir accepté d'évaluer ce mémoire et pour la considération et les observations précieuses qu'elle y a apportées.

Nous tenons à exprimer une reconnaissance particulière à **Dr. Mohammed Abdallah KHODJA**, dont l'expertise et les orientations méthodologiques ont marqué le point de départ décisif de cette recherche. Sa supervision attentive, ses remarques constructives et son soutien indéfectible ont grandement contribué à la réalisation de la partie pratique de ce projet. Qu'il trouve ici l'expression de notre profonde gratitude.

Nos remerciements vont également à l'équipe administrative du département de microbiologie et de biochimie, avec une mention spéciale à **Dr. Seifeddine DRIF**, chef de département, pour leur accompagnement, leur disponibilité et leur professionnalisme tout au long de notre parcours.

Enfin, nous remercions chaleureusement nos collègues de promotion pour leur esprit de collaboration, leur aide précieuse et leur soutien quotidien durant la période de stage. À chacun d'entre vous, merci.

Sommaire

ملخص

تُعدّ عملية التعرف الدقيق على المستقلبات تحدياً أساسياً في مجال الميتابولوميات المعتمدة على التحليل بالرنين المغناطيسي النووي (NMR).

في هذه الدراسة، قمنا بتطوير سلسلة معالجة تعتمد على التعلم الآلي الخاضع للإشراف باستخدام خوارزمية الغابة العشوائية (Random Forest) .

تهدف الدراسة الى التنبؤ بمعرّفات قاعدة بيانات HMDB استناداً فقط إلى بيانات وصفية تجريبية مستخلصة من تجارب NMR، تشمل المذيب، نوع النواة، عدد القمم، الرقم الهيدروجيني، درجة الحرارة، ومرجع الإزاحة الكيميائية. وقد خضع هذا المعطى، المستخرج من قاعدة بيانات HMDB، لمرحلة إعداد شملت الترميز التصنيفي وتوحيد القيم. أظهر تقييم النموذج على مجموعة اختبار غير مرئية دقة تصنيف بلغت 33%، ودقة إيجابية (Precision) بنسبة 50%، واسترجاعاً (Recall) بنسبة 33%، ودرجة F1 بلغت 39%، وخسارة لوجاريتمية (Log Loss) قدرها 1.48، مما يعكس تحديات التوزيع غير المتوازن للفئات والغموض الاحتمالي في التنبؤات. كما تم تطوير واجهة رسومية تفاعلية (GUI) لتمكين المستخدم من الحصول على تنبؤات فورية بناءً على بيانات وصفية يزودها.

تبرز هذه النتائج إمكانية استخدام البيانات الوصفية وحدها في تحديد هوية المستقلبات، وتمهّد الطريق أمام تحسينات مستقبلية من خلال دمج البيانات الطيفية وتقنيات التعلم الجماعي.

الكلمات المفتاحية: الرنين المغناطيسي النووي، التعلم الآلي، تحديد المستقلبات، الغابة العشوائية، البيانات الوصفية التجريبية، التعليق الأيضي

Abstract

Accurate identification of metabolites remains a critical challenge in NMR-based metabolomics.

In this study, we developed a supervised machine learning pipeline using a Random Forest.

The objective of this study is to classify and predict HMDB Database IDs directly from experimental NMR metadata. Including solvent, nucleus, peak count, pH, temperature, and chemical-shift reference. The dataset, curated from HMDB, was preprocessed via categorical encoding and feature scaling. Model evaluation on an unseen test set yielded an accuracy of 33%, precision of 50%, recall of 33%, F1 score of 39%, and a log loss of 1.48 highlighting both the challenges of class imbalance and the probabilistic uncertainty of predictions. A graphical user interface (GUI) was developed to facilitate real-time predictions from user-supplied metadata.

These results demonstrate the feasibility of using metadata alone for metabolite annotation and set the groundwork for future improvements through spectral integration and ensemble learning.

Keywords: NMR spectroscopy, Machine learning, Metabolite identification, Random Forest, Experimental metadata, Metabolomic annotation

Résumé

L’identification précise des métabolites demeure un défi majeur en métabolomique basée sur la RMN.

Dans cette étude, nous avons développé une chaîne d’analyse supervisée utilisant un classifieur Random Forest.

L'objectif de cette étude est de prédire les identifiants de la base de données HMDB à partir de métadonnées expérimentales RMN incluant le solvant, le noyau, le nombre de pics, le pH, la température et la référence de déplacement chimique. L’ensemble de données, agrégé à partir de HMDB, a été prétraité par encodage catégoriel et normalisation. L’évaluation du modèle sur un jeu de test indépendant a donné une exactitude de 33 %, une précision de 50 %, un rappel de 33 %, un score F1 de 39 % et une perte logarithmique (log loss) de 1,48, révélant des déséquilibres entre classes ainsi qu’une incertitude importante dans les prédictions probabilistes. Une interface graphique (GUI) a été développée afin de permettre des prédictions interactives à partir des métadonnées saisies par l’utilisateur.

Ces résultats démontrent la faisabilité d’une annotation métabolitique basée uniquement sur les métadonnées et ouvrent la voie à des améliorations futures intégrant les données spectrales et des approches par apprentissage ensembliste.

Mots-clés : Spectroscopie RMN, Apprentissage automatique, Identification des métabolites, Forêt aléatoire, Métadonnées expérimentales, Annotation métabolomique

Liste des abréviations

AI : Artificial Intelligence

CNN : Réseaux de Neurones convolutif

DA : Analyse Discriminante

DL : Deep Learning (Apprentissage profond)

GPU : Graphics Processing Unit

HCA : Analyse Hiérarchique de Clustering

HS-GC-IMS: Headspace Gas Chromatography-Ion Mobility Spectrometry

IA : Intelligence artificielle

ML : Machine Learning (Apprentissage automatique)

PCA : Analyse en Composantes Principales

RF: Radio fréquences

RNA : Réseau de neurones artificiels

RNN : Réseaux de Neurones récurent

SA : Stress Abiotique

Liste des figures

Liste des tableaux

Introduction

Introduction

La technologie de la résonance magnétique nucléaire (RMN) est cruciale dans les recherches métabolomiques et pharmacologiques, car elle offre aux scientifiques la possibilité de saisir plus finement les processus biochimiques liés à diverses situations physiologiques et pathologiques. C'est une méthode efficace et précise pour analyser des composés chimiques complexes (Lee et al., 2023). Elle constitue un outil essentiel pour l'identification des métabolites et la reconnaissance des signatures chimiques distinctives (Xia et al., 2008). En outre, la RMN est reconnue pour sa robustesse comparée à d'autres méthodes analytiques telles que la chromatographie en phase gazeuse (HS-GC-IMS), grâce à sa reproductibilité, sa quantification non destructive et son faible besoin de préparation d'échantillon. Globalement, ces recherches mettent en avant la valeur de la RMN dans l'examen des métabolites, ce qui contribue à une amélioration notable de la précision et de la rapidité des analyses dans les domaines chimiques et biologiques (Lee et al., 2023).

Avec la complexité croissante des données générées par les méthodes avancées d'analyse chimique, l'intelligence artificielle (IA), incluant l'apprentissage automatique (ML) et l'apprentissage profond (DL), occupe une place centrale dans la chimie analytique. Ces approches permettent le traitement rapide et précis de larges ensembles de données spectraux, en révélant des motifs cachés dans les données RMN, facilitant ainsi l'identification des métabolites et la découverte de composés nouveaux ou peu connus (Johnson & Tipirneni-Sajja, 2024). Les métabolites jouent un rôle central dans de nombreux processus biologiques et fournissent une empreinte métabolomique déterminante pour comprendre les voies biochimiques dans les organismes vivants. L'intégration de l'IA dans la RMN permet d'affiner l'analyse de ces données, renforçant ainsi les capacités de détection de maladies et l'exploration approfondie des processus biologiques (Galal et al., 2022).

L'objectif principal de ce mémoire est de développer un modèle d'intelligence artificielle capable de prédire automatiquement l'identifiant d'une base de données métabolomique (telle que HMDB) à partir des métadonnées issues de la spectroscopie RMN. Ce modèle est intégré dans une interface utilisateur interactive, conçue pour faciliter l'annotation rapide et fiable des métabolites inconnus, contribuant ainsi à l'amélioration des flux de travail en métabolomique computationnelle.

Partie bibliographique

Chapitre I : RMN

Et métabolites

Chapitre I. RMN et métabolites

I.1. Métabolites

I.1.1. Définition

Les métabolites sont généralement des produits transitoires ou stables au cours du processus naturel du métabolisme qui comprend les processus biochimiques de construction (anabolisme), de dégradation (catabolisme) et d'élimination (excrétion) des composés dans/depuis l'organisme. Le mécanisme de formation des métabolites implique des réactions chimiques en cascade induites par les enzymes. Ils peuvent être identifiés à l'aide de différentes stratégies, qu'il s'agisse d'approches ciblées ou non ciblées, ainsi que par diverses techniques, telles que la résonance magnétique nucléaire (James, 2017).

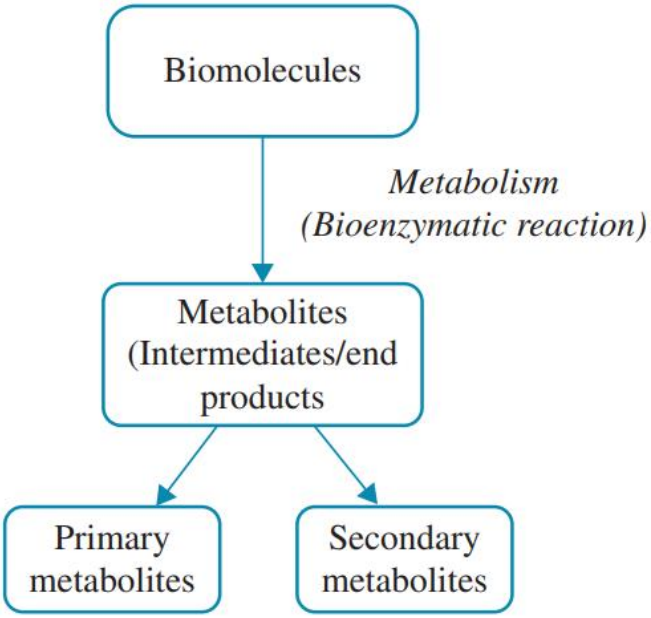

Figure I.1. Produits métaboliques des biomolécules et types de métabolites (James, 2017).

I.1.2. Classification

I.1.2.1. Métabolites primaires

Les métabolites primaires sont essentiels au fonctionnement normal des cellules végétales et sont directement impliqués dans divers processus biochimiques et physiologiques, tels que la photosynthèse et la respiration, fournissant l'énergie et les précurseurs requis pour la biosynthèse de nouvelles macromolécules nécessaires aux processus de développement des plantes. Les métabolites primaires comprennent les sucres (mono-, di- et tri saccharides), les polyols (sorbitol

et mannitol, par exemple) et les acides aminés tels que la proline, qui peuvent servir d'osmolytes et d'osmoprotecteurs dans les plantes soumises aux SA (Patel et al., 2020).

I.1.2.2. Métabolites secondaires

Les métabolites secondaires appartiennent à divers groupes chimiques, tels que les alcaloïdes, les terpènes et les composés phénoliques. Leur répartition dans le règne végétal est inégale, mais leur accumulation peut, dans certains cas, atteindre des niveaux élevés (Setyorini & Antarlina, 2022).

I.1.3. Fonctions

I.1.3.1. Métabolite primaire

- Lipides

Les lipides sont décrits comme de petites molécules hydrophobes ou amphipathiques facilement solubles dans les solvants organiques tels que le chloroforme, les éthers ou les alcools.

Les lipides assurent quatre fonctions essentielles chez les organismes. Tout d'abord, ils constituent les membranes cellulaires, où ils représentent entre 30 et 70 % du poids sec. Ensuite, les lipides servent de réserves énergétiques, principalement sous forme de triacylglycérols. Par ailleurs, certains lipides et leurs dérivés interviennent dans la signalisation cellulaire, ces molécules jouent un rôle clé dans divers processus physiologiques. Enfin, les lipides participent aux revêtements de surface des organismes. Chez les plantes, on retrouve des structures comme les cires de surface, la cutine et la subérine, qui protègent contre la déshydratation et les agressions extérieures (Li-Beisson et al., 2016).

- Les glucides

Les glucides forment une classe essentielle de molécules présentes chez tous les organismes vivants. Ils se trouvent principalement sous forme de polymères, tels que les oligosaccharides et les polysaccharides, souvent associés à des protéines ou des lipides. Grâce à leurs nombreux centres chiraux et à diverses modifications chimiques, comme l'acétylation, la méthylation, l'oxydation et la sulfonation, ils génèrent une grande diversité structurelle à partir de simples unités glucidiques (Garron & Cygler, 2010).

Les glucides, tels que le saccharose, le glucose et le fructose, jouent un rôle essentiel dans le métabolisme des plantes. Ils sont indispensables au métabolisme intermédiaire et respiratoire et servent de base à la synthèse des glucides complexes, comme l'amidon et la cellulose. De plus, ils fournissent les éléments de base nécessaires à la production des acides aminés, des acides gras et de nombreux autres composés présents dans les plantes.

Bien que les fonctions métaboliques des glucides soient bien établies, leur rôle dans la signalisation cellulaire suscite un intérêt croissant dans la recherche récente. En effet, ils peuvent moduler l'expression des gènes, à l'image des hormones. Toutefois, contrairement aux hormones, qui sont actives à de très faibles concentrations (nano- à micro molaires), les glucides sont présents en quantité bien plus élevée (milli molaire) en raison de leur implication directe dans le métabolisme cellulaire (Smeekens, 2000).

- Les protéines

Les protéines sont de grandes molécules biologiques composées d'une ou plusieurs chaînes d'acides aminés. Ces biopolymères présentent une valeur nutritionnelle élevée, une stabilisation, une élasticité ainsi qu'une capacité à protéger les cellules, les tissus et les organismes. Les protéines peuvent être classées en deux grandes catégories : les protéines fibreuses, généralement insolubles dans l'eau, et les protéines globulaires, qui sont en général solubles dans l'eau ou dans des solutions aqueuses acides ou basiques (Martins et al., 2018).

Les protéines possèdent des propriétés fonctionnelles importantes du point de vue de la technologie alimentaire ; leur caractère amphiphile et leur capacité à former des films interfaciaux contribuent à la création de systèmes stabilisants. En raison de leur composition et de leur structure, les protéines présentent différentes propriétés fonctionnelles dans divers aliments, qui peuvent être classées en trois groupes principaux :

Propriétés hydratantes ***:*** Ces propriétés dépendent directement des interactions entre les protéines et l'eau, telles que l'adhésion, la dispersion, la solubilité, la viscosité et la capacité de rétention d'eau.

Propriétés interfaciales : Elles incluent des processus tels que l'émulsification et la formation de mousses, qui reposent sur la capacité des protéines à interagir avec les interfaces entre les liquides et les gaz, ou entre différents liquides.

Propriétés gélifiantes : Ces propriétés sont expliquées par les interactions entre les protéines elles-mêmes et avec l'eau, permettant ainsi la formation de structures gélatineuses.

Ces propriétés des protéines jouent un rôle essentiel dans l'amélioration et la stabilisation des systèmes alimentaires, ce qui les rend indispensables dans de nombreuses applications technologiques alimentaires (Higuera-Barraza et al., 2016).

I.1.3.2. Métabolite Secondaire

- Polyphénols

Les composés phénoliques sont des constituants présents dans de nombreux aliments d'origine végétale tels que les fruits, légumes, noix, boissons dérivées des plantes (comme le thé et le vin), ainsi que dans des plantes médicinales traditionnelles comme le Ginkgo biloba. On les retrouve également dans une multitude de suppléments alimentaires à base de plantes.

Les plantes produisent des composés phénoliques en tant que métabolites secondaires, jouant un rôle crucial dans divers processus biologiques, tels que la croissance, la lignification, la pigmentation, la pollinisation et la défense contre les pathogènes, les prédateurs et les stress environnementaux .Outre leurs fonctions classiques d'antioxydants, les composés phénoliques peuvent exercer des effets modulatoires dans les cellules en agissant sélectivement sur différents composants de cascades de signalisation impliquant des kinases de protéines et des kinases lipidiques (Kumaraswamy et al., 2018).

Les flavonoïdes : sont les composés polyphénoliques les plus présents dans les végétaux. Leur structure chimique se compose de deux cycles aromatiques (A et B), portant plusieurs groupes phénoliques, reliés par une chaîne de trois atomes de carbone. Cette chaîne est souvent intégrée dans un hétérocycle contenant un atome d'oxygène. La présence de multiples fonctions phénoliques confère aux flavonoïdes des propriétés antioxydantes ou oxydantes. Grâce à leur abondance dans divers aliments et boissons, leur consommation est associée à des effets protecteurs contre de nombreuses affections chroniques, telles que l'athérosclérose, et, par extension, contre les accidents vasculaires cérébraux et coronariens qu'elle peut provoquer, ainsi que les maladies neurodégénératives (Choi et al., 2009).

Les flavonoïdes possèdent des propriétés antivirales, y compris contre le virus de l'immunodéficience humaine (VIH) et le virus de l'hépatite B. Ils présentent également des activités antivirales contre plusieurs autres virus. Récemment, les flavonoïdes ont suscité un intérêt croissant en raison de leurs diverses activités antivirales, notamment contre le virus de la grippe (Stoclet & Schini-Kerth, 2011). En plus les flavonoïdes jouent un rôle essentiel dans la production de la couleur des fleurs, en fournissant des teintes attrayantes pour les pollinisateurs des plantes (Harborne & Williams, 2000).

Les tanins : Le terme « tanin » a été inventé par Seguin1 pour décrire les substances présentes dans les extraits végétaux.

Tannin en tant que composés phénoliques solubles dans l'eau ayant un poids moléculaire compris entre 500 et 3000 D. Ces polyphénols contiennent un grand nombre de groupes

hydroxyles ou d'autres groupes fonctionnels (1 à 2 pour 100 D) et sont donc capables de former des liaisons croisées avec des protéines et d'autres macromolécules (Chung et al., 1998).

Les tanins sont des composés naturels présents dans de nombreuses plantes et ont montré un rôle dans la lutte contre le cancer. Ils agissent en inhibant la croissance des cellules cancéreuses en affectant les voies de signalisation cellulaire et les protéines liées au développement du cancer. Ils aident également à réduire les effets des substances cancérigènes et à favoriser l'apoptose (Kleszcz et al., 2025).

- Alcaloïdes

Les alcaloïdes sont des composés naturels contenant du carbone, de l'hydrogène, de l'azote et généralement de l'oxygène, que l'on trouve principalement dans les plantes, en particulier dans certaines plantes à fleurs (Hussain et al., 2018).

Les alcaloïdes jouent un rôle clé dans divers processus biologiques, notamment en influençant le métabolisme cellulaire. Ils agissent sur différents types de cellules, comme les fibroblastes dermiques, en modifiant les processus associés au vieillissement. Grâce à leurs effets, ils peuvent contribuer à maintenir l'équilibre de la peau et à prévenir les changements liés à l'âge (Abadie et al., 2021).

- Terpènes

Les terpènes constituent une famille très diversifiée de produits naturels synthétisés par les plantes. Cette famille compte environ 55 000 membres avec des structures chimiques différentes, présentant des applications pratiques potentielles. Les terpènes ont la formule chimique générale (C5H8) n, définie par l'isoprène en tant qu'unité.

Les terpènes et les terpénoïdes ont un grand potentiel pour traiter les maladies inflammatoires. Ces composés agissent simultanément sur plusieurs voies cellulaires, ce qui peut les rendre plus efficaces que les médicaments actuels (Del Prado-Audelo et al., 2021).

- Huiles essentielles

Les huiles essentielles constituent des mélanges complexes renfermant une grande diversité de composés chimiques, exclusivement d'origine végétale. Leur extraction repose le plus souvent sur la distillation à la vapeur d'eau appliquée à des matrices végétales.

Les huiles essentielles sont connues pour leurs nombreux effets sur la santé, tels que leurs propriétés antibactériennes et antivirales. Elles sont également connues pour soulager le stress et ont été utilisées dans de nombreux traitements tels que les troubles du sommeil, la maladie d'Alzheimer, les problèmes cardiovasculaires, le cancer et les douleurs liées à l'accouchement. En

outre, elles sont également connues pour leurs propriétés insectifuges et leur activité antioxydante et anti-inflammatoire (Farrar & Farrar, 2020).

I.1.4. Importance

I.1.4.1. Régulation du système immunitaire

Les métabolites influencent non seulement la santé intestinale, mais aussi la réponse immunitaire. Par exemple, certains métabolites dérivés des microbes intestinaux ont des effets immunomodulateurs, en régulant la réponse immunitaire locale et systémique. Cela peut influencer la susceptibilité à des infections, des maladies inflammatoires ou des troubles auto-immuns.

I.1.4.2. Communication entre l'intestin et le cerveau

Les métabolites jouent un rôle crucial dans la communication entre l'intestin et le cerveau, une interaction souvent désignée sous le nom d'axe intestin-cerveau. Ces molécules sont des produits intermédiaires ou finaux issus de la digestion bactérienne et peuvent exercer des effets bénéfiques ou nuisibles sur le cerveau, en fonction des niveaux spécifiques produits dans l'hôte. Les déséquilibres dans ces métabolites ont été liés à diverses maladies neurodégénératives et neuropsychiatriques, telles que la maladie d'Alzheimer, la dépression, et les troubles anxieux (Swer et al., 2023).

I.1.4.3. Diagnostic des maladies

Les métabolites jouent un rôle crucial dans le diagnostic de diverses maladies en reflétant l'état physiopathologique de l'organisme.

Tableau I.1. Récapitulatif des études sur les métabolites dans le contexte du cancer en 2012 (Dormoy & Massfelder, 2013).

Type de cancer	Métabolites d'intérêt identifiés
Œsophage	• Acide malonique • L-sérine
Estomac	• Acide 3-hydroxypropionique • Acide pyruvique
Colorectal	• L-alanine • Lactose glucuronique • L-glutamine

Vessie	• Glucose • Tyrosine • Phénylalanine
Sein	• Thréonine • Isoleucine • Glutamine • Acide linolénique
Rein	• Cinnamoylglycine • Glucose • Nicotinamide • Phénylpropionoylglycine • Valine

Ce tableau présente un récapitulatif des principales études menées en 2012 portant sur les métabolites identifiés dans divers types de cancers. Pour chaque localisation tumorale, des métabolites spécifiques ont été détectés dans les tissus, le sang ou l'urine des patients, suggérant leur potentiel en tant que biomarqueurs diagnostiques ou pronostiques. Ces molécules reflètent des altérations du métabolisme cellulaire propres à chaque type de cancer, et pourraient contribuer à une meilleure compréhension des mécanismes tumoraux ainsi qu'à l'optimisation des approches thérapeutiques.

I.2. RMN

I.2.1. Définition

La RMN repose sur l'interaction entre les moments magnétiques nucléaires et des champs magnétiques. Il s'agit d'une technique analytique utilisée pour étudier les propriétés magnétiques des noyaux atomiques dans les matériaux. Les noyaux contenant un nombre impair de protons ou de neutrons, tels que ^{1}H, ^{31}P, ^{13}C et ^{15}N, ont un spin égal à ½, ce qui permet leur analyse. Lorsqu'ils sont exposés à un champ magnétique externe, ces noyaux peuvent s'aligner soit parallèlement, soit antiparallèlement au champ, ce qui entraîne une séparation des niveaux d'énergie. En équilibre, les moments magnétiques nucléaires tendent à s'aligner avec le champ magnétique, générant une magnétisation macroscopique importante qui peut être exploitée pour analyser la structure moléculaire (Hanson, 2008).

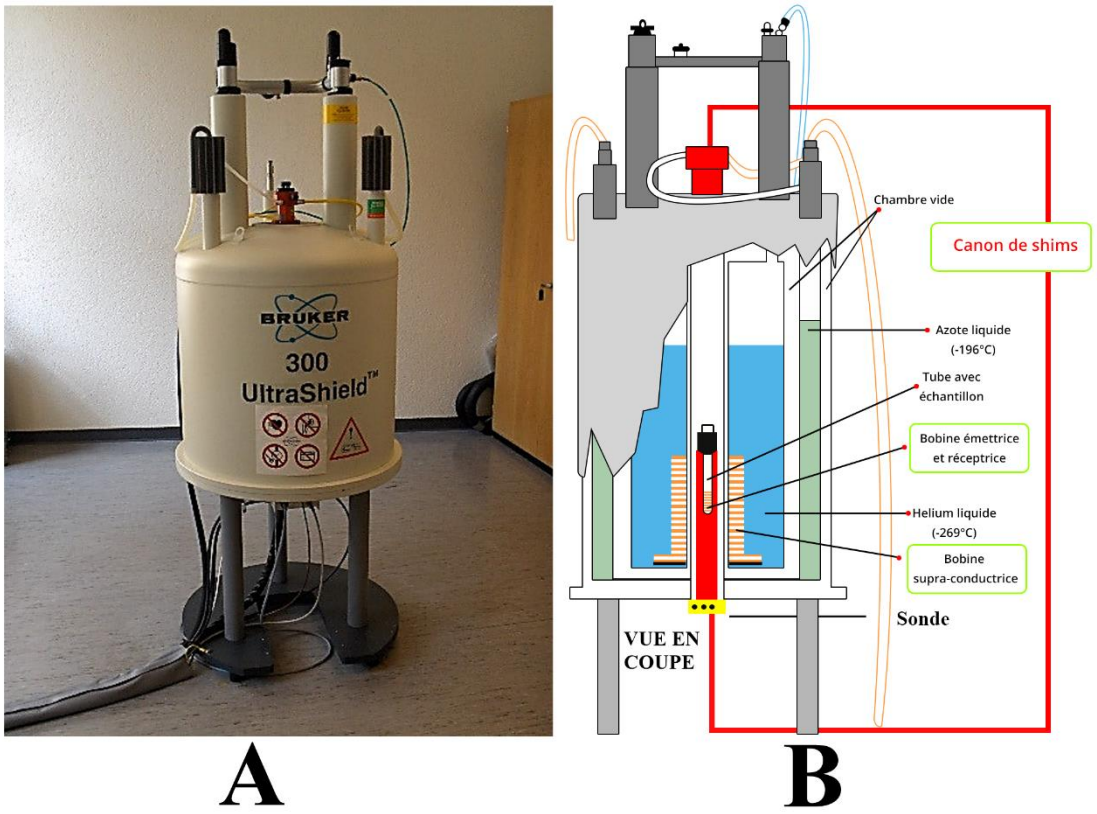

Figure I.2. Appareil de RMN (a) et Schéma de RMN (b) (Jonathan Clayden et al., 2013).

I.2.2. Principe

Les noyaux atomiques possédant un spin différent de zéro produisent un moment magnétique. Lorsque l'échantillon est exposé à un champ magnétique constant B_0, les niveaux d'énergie nucléaire se séparent selon l'effet Zeeman, ce qui conduit à la formation de deux états énergétiques distincts dans le cas des noyaux avec I = ½, où l'état avec le moment parallèle au champ magnétique est à une énergie plus faible, tandis que l'état opposé est à une énergie plus élevée. Les noyaux subissent un mouvement de rotation, où les moments magnétiques tournent autour de l'axe du champ magnétique à une fréquence de L'Armor selon la loi $\nu_0=\gamma B_0$, qui dépend de la constante gyromagnétique du noyau et de l'intensité du champ magnétique appliqué. Lorsque l'échantillon est exposé à une onde électromagnétique de fréquence égale à la fréquence de L'Armor, Les noyaux absorbent de l'énergie, ce qui provoque une inversion de spin (Spin Flip), c'est-à-dire que le noyau passe d'un état d'énergie plus faible à un état d'énergie plus élevé. En revenant à son état d'origine, de l'énergie est émise sous forme de signaux qui peuvent être mesurés et analysés (Lambert, 2018).

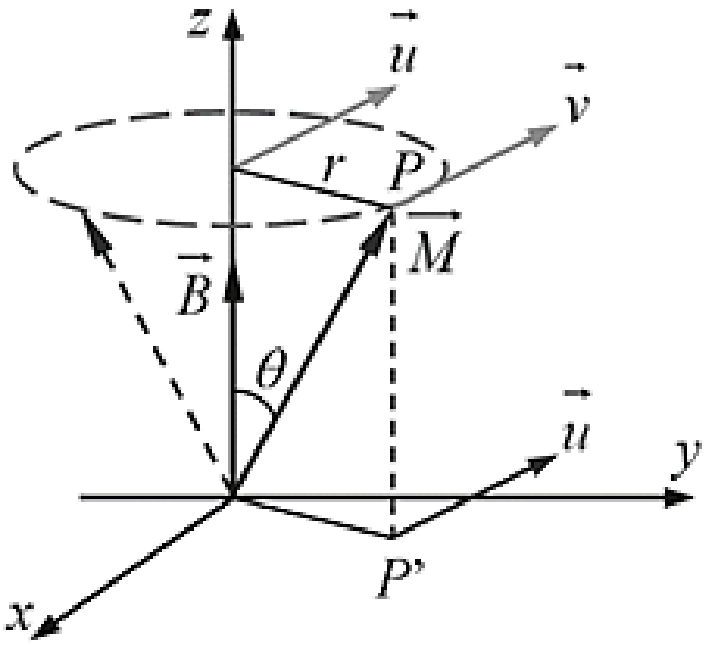

Figure I.3. Précession du moment magnétique orbital autour de la direction du champ magnétique (Schott et al., 2015)

Éléments représentés dans le schéma :

$\vec{B}$: vecteur du champ magnétique (champ magnétique appliqué).

$\vec{u}; \vec{v}$: vecteurs unitaires dans le plan horizontal (base mobile ou base de Frenet).

r : rayon de la trajectoire circulaire (orbite de la particule chargée).

θ : angle entre le vecteur du moment orbital et l'axe z.

$\vec{M}$: vecteur du moment magnétique orbital (ou moment cinétique orbital).

P: position de la particule sur son orbite.

p': projection orthogonale de P sur le plan xOy.

x, y, z: axes du repère cartésien.

Cercle en pointillés : trajectoire de précession du vecteur moment autour de l'axe du champ magnétique.

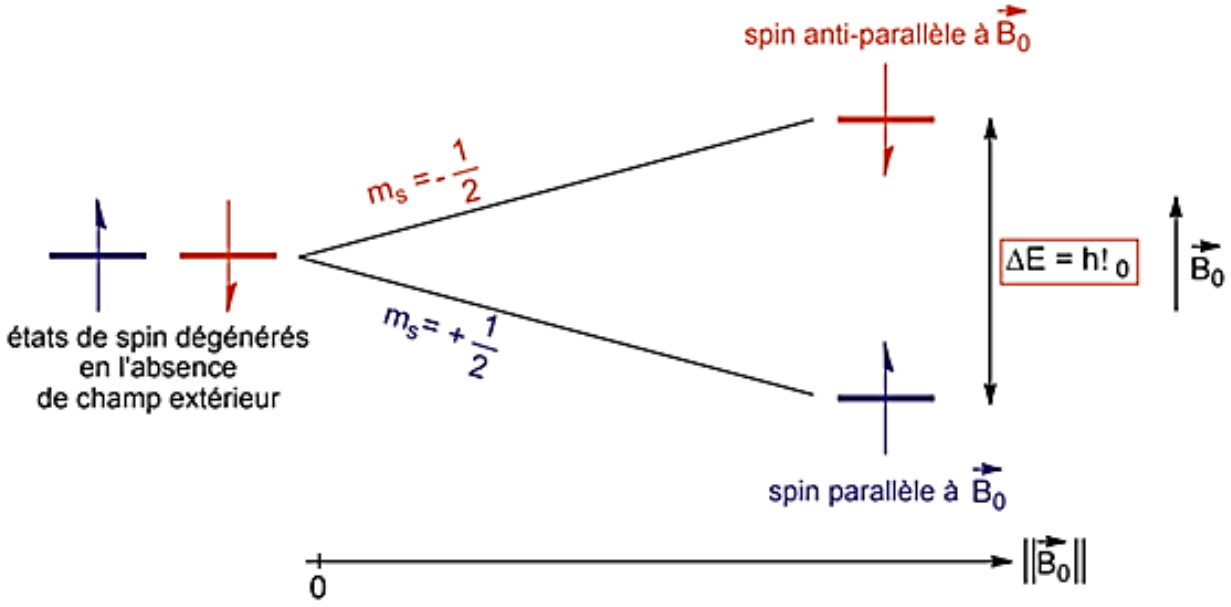

Figure I.4. Levée de dégénérescence des états de spin sous l'effet d'un champ magnétique extérieur (Schott et al., 2014).

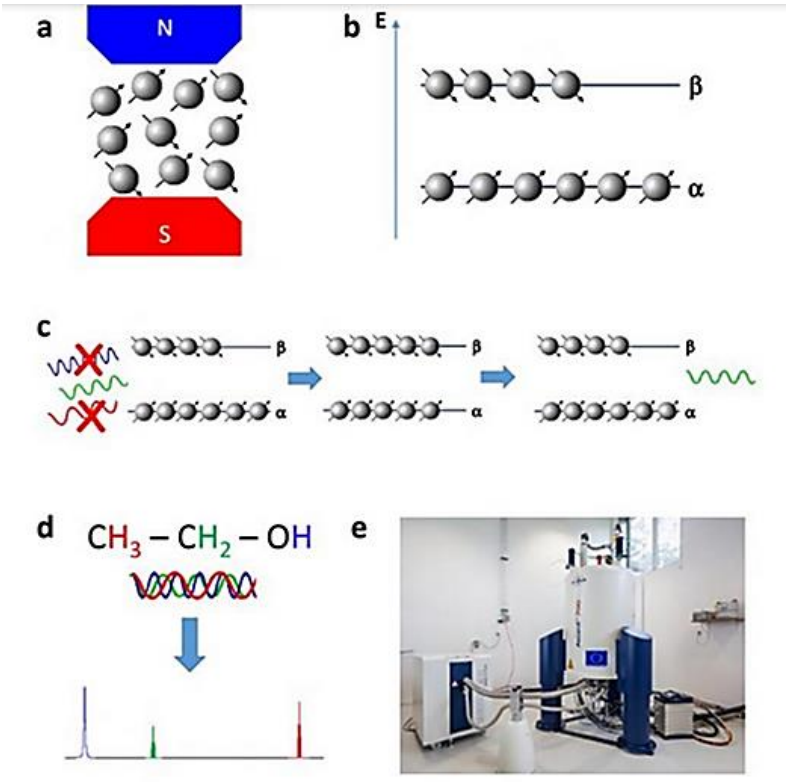

Figure I.5. Illustration du principe de la RMN (Gonçalves, 2018).

I.2.3. Types de RMN

I.2.3.1. RMN du proton (^{1}H-RMN)

Elle repose essentiellement sur les caractéristiques magnétiques du noyau d'hydrogène, étant donné que les atomes d'hydrogène (^{1}H) sont largement présents et associés à une multitude de groupes fonctionnels comme -OH, -NH, -CH_3, entre autres. Le proton (^{1}H) présent dans le noyau de l'hydrogène a un spin nucléaire (spin = ½), ce qui lui donne la capacité d'interagir avec un champ magnétique externe fort (B_0). Quand l'échantillon est exposé à ce champ (par exemple 14 Tesla), conformément à la mécanique quantique, le noyau peut adopter deux états d'énergie. L'un de ces états, nommé état d'énergie inférieur (α), correspond à lorsque l'orientation du moment magnétique du noyau est alignée avec le champ B_0. Dans un état de haute énergie (β), le moment magnétique s'oppose au champ. En les soumettant à des ondes radio d'une fréquence déterminée, les protons passent d'un niveau d'énergie plus bas à un niveau d'énergie plus élevé. Suite à l'absorption d'énergie et leur retour à l'état normal, les noyaux libèrent des signaux qui sont surveillés et examinés (J. Clayden et al., 2013; Kiemle et al., 2016; Rebstein & Soerensen, 2011).

Le déplacement chimique (δ), qui illustre l'influence de l'environnement chimique des protons sur leur fréquence d'absorption, constitue le centre névralgique de la ^{1}H-RMN. L'unités de mesure du déplacement est le parti par million (ppm). Les protons affichent des valeurs δ distinctes en fonction des atomes environnants : les protons situés à proximité d'atomes électronégatifs tels que l'oxygène ou les halogènes subissent une carence de protection électronique (désblindage), entraînant un glissement vers des valeurs δ supérieures(J. Clayden et al., 2013; Kiemle et al., 2016; Rebstein & Soerensen, 2011).

Un autre élément clé de l'analyse du spectre ^{1}H-RMN est le couplage spin-spin. Cela se manifeste par la séparation des signaux en multiplets, conformément à la règle (n+1), où n indique le nombre de protons adjacents. Cette division offre des détails concernant le nombre de protons voisins et leur arrangement au sein de la molécule. L'intégration joue aussi un rôle crucial, car la surface de chaque signal est proportionnelle au nombre de protons qui lui sont associés. Cela facilite le calcul du rapport numérique des protons dans chaque environnement chimique (Silverstein et al., 2016).

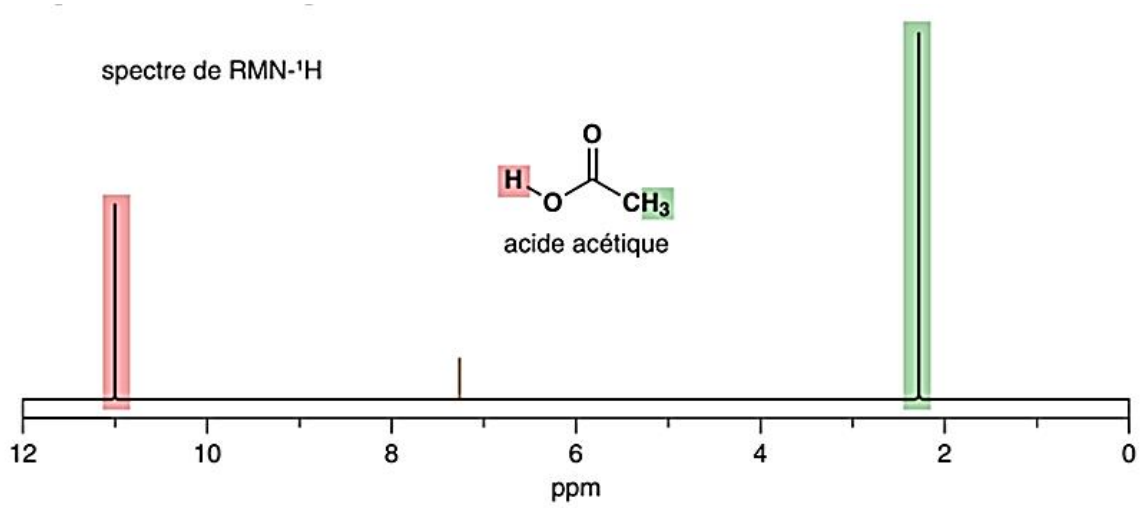

Figure I.6. Affiche le spectre de résonance magnétique nucléaire du proton (¹H-RMN) (Jonathan Clayden et al., 2013).

I.2.3.2. RMN du carbone-13 (¹³C-RMN)

Contrairement à la RMN du proton (¹H), la RMN avec le ¹³C est moins courante en raison de sa faible abondance naturelle. S'appuie sur l'analyse des noyaux d'atomes de carbone contenant l'isotope 13C, un isotope stable qui ne constitue qu'environ 1,1 % du carbone présent dans la nature. À l'inverse du proton 1H, le carbone 13 possède un spin (spin = ½), ce qui lui permet de résonner sous l'effet d'un champ magnétique puissant externe. Dans l'expérience de RMN du carbone-13, l'échantillon est positionné dans un champ magnétique extérieur et soumis à des ondes radio. La variation de l'état énergétique des noyaux de carbone 13 est influencée par leur environnement électronique, ce qui provoque l'émission de signaux distinctifs pour chaque atome de carbone chimiquement non équivalent au sein de la molécule. Par rapport au 1H-RMN, l'un des atouts majeurs du 13C-RMN est son vaste éventail de déplacement chimique, dont les valeurs δ se situent généralement entre 0 et 220 parties par million (ppm). Cela favorise une distinction plus précise des signaux sans superposition. On observe généralement la présence de carbonates saturés (comme le CH_3 et le CH_2) entre 0 et 50 ppm, des carbonates insaturés ou aromatiques entre 100 et 160 ppm, tandis que les carbonates carbonylés se retrouvent à des niveaux plus élevés, entre 160 et 220 ppm. Étant donné la rareté du 13C dans la nature, l'intensité des signaux est faible. Par conséquent, des techniques d'amélioration de la sensibilité, telles que le découplage des protons, sont utilisées, où les signaux des protons liés sont saturés, ce qui entraîne la disparition des divisions de liaison de spin entre 13C et 1H. Les signaux de carbone se manifestent par des pics distincts forts, ce qui simplifie l'analyse spectrale (Breitmaier et al., 1988).

I.2.3.3. RMN multidimensionnelle (2D/3D RMN)

Le RMN multidimensionnel fonctionne en observant les interactions de spin entre les noyaux atomiques d'une molécule, en organisant les données spectrales sur deux axes ou plus. Cela permet de détecter avec précision les relations structurelles et les voisinages entre les atomes. La phase cruciale est l'exécution d'une séquence d'impulsions radioélectriques successives avec des délais minutieusement déterminés, en stimulant les noyaux et en surveillant la manière dont les interactions de spin entre eux sont propagées. Suite à la collecte des données, on applique une transformée de Fourier bidimensionnelle ou tridimensionnelle pour analyser les signaux et les transposer en un spectre multi-axial, chaque axe représentant la fréquence d'un noyau spécifique. Des pics superposés pourraient suggérer une corrélation entre deux noyaux (Keeler, 2015).

Une des méthodes 2D-RMN les plus significatives qui symbolise ce concept est l'expérience COSY (spectroscopie de corrélation). Celle-ci rend possible la détection des interactions de spin entre les protons proches, facilitant ainsi l'identification des liens entre atomes d'hydrogène dans une molécule. La Spectroscopie par Effet Overhauser Nucléaire (NOESY), grâce à sa capacité à révéler les interactions stériques entre des noyaux qui ne sont pas nécessairement adjacents sur le plan chimique, mais qui se trouvent à proximité l'un de l'autre dans l'espace, s'avère être un instrument précieux pour analyser la structure en trois dimensions des molécules, notamment celle des protéines (Charrette et al., 2023; Jacobsen, 2016).

La HSQC (cohérence quantique hétéronucléaire unique) établit un lien direct entre les signaux de protons et les noyaux hétéronucléaires comme le carbone 13 ou l'azote 15, ce qui permet d'identifier aisément le type d'atomes reliés aux protons. Finalement, la HMBC (Heteronuclear Multiple Bond Correlation) sert à observer les interactions entre les protons et les noyaux hétéronucléaires sur plusieurs liaisons, facilitant ainsi la reconnaissance de liaisons à longue distance au sein d'une molécule (Charrette et al., 2023; Jacobsen, 2016).

I.2.3.4. RMN des noyaux multiples (multi-nucléaires)

Il offre la possibilité d'examiner des noyaux atomiques distincts des noyaux usuels comme l'hydrogène (^{1}H) et le carbone 13 (^{13}C). Cette méthode s'appuie sur le même fondement physique que la RMN, où l'échantillon est positionné dans un puissant champ magnétique. Des impulsions de radiofréquence (RF) spécifiques à chaque type de noyau magnétiquement actif avec un moment de spin nucléaire ($I \neq 0$) sont ensuite appliquées. En ajustant les fréquences de résonance de chaque noyau, il est possible d'obtenir un spectre RMN qui dépeint son environnement chimique et ses interactions à l'intérieur de la molécule. Cette technique permet d'analyser une diversité de noyaux, tels que le fluor-19 (^{11}F), le phosphore-31 (^{31}P), l'azote-15 ($^{1}\mu N$), le silicium-29 (^{21}Si), le bore-11 (^{11}B), et le lithium-7 (^{1}Li), entre autres. Chaque noyau a sa propre fréquence de résonance, ce qui

exige un ajustement précis de l'appareil selon le type de noyau examiné. Cette méthode est employée pour offrir des détails exacts sur les liaisons chimiques et les interactions entre divers noyaux, tout en facilitant également la détermination de la structure moléculaire et de la répartition spatiale des atomes dans les composés. Les RMN multinucléaires constituent un outil précieux dans plusieurs disciplines, y compris la chimie organique et inorganique, la biochimie, la science des matériaux ainsi que le secteur pharmaceutique. Il facilite l'examen de composés comprenant des éléments comme le fluor, le phosphore et le silicium, et est fréquemment employé dans l'analyse de protéines, d'acides nucléiques, de polymères et de composés inorganiques (Akoka, 2022).

I.2.3.5. RMN des solides

C'est un instrument utile pour examiner les solides sous l'angle de leur structure atomique et de leur organisation moléculaire. À l'opposé de la RMN traditionnelle, habituellement appliquée aux échantillons liquides, la RMN des solides fait face à des difficultés spécifiques liées aux caractéristiques des solides. En effet, dans les solides, les noyaux nucléaires se trouvent à des emplacements fixes et sont exposés à d'intenses interactions magnétiques qui engendrent une largeur et un flou notable des lignes spectrales. Ces interactions englobent les liaisons dipolaires entre les noyaux, l'impact de l'anisotropie du déplacement chimique (CSA) ainsi que les interactions quadrupolaires pour des noyaux possédant un spin nucléaire supérieur à 1/2. Afin de relever ces défis et d'optimiser la clarté spectrale, la RMN du solide utilise le mécanisme du Magic Angle Spinning (MAS). Dans ce processus, l'échantillon est positionné dans un rotor qui tourne à une vitesse élevée autour d'un angle de 54,74° par rapport au champ magnétique. Cela permet d'atténuer les interactions directionnelles et de minimiser la largeur du signal. On utilise également des méthodes de découplage dipolaire en envoyant des impulsions de radiofréquence aux noyaux d'hydrogène, ce qui permet d'obtenir des spectres fins semblables à ceux observés en état liquide (Akoka, 2022).

I.2.3.6. RMN avec relaxation (Temps de relaxation T_1 & T_2)

L'un des principes clés de la Résonance Magnétique Nucléaire (RMN) est la relaxation nucléaire, qui est cruciale pour saisir la dynamique moléculaire et la structure détaillée des matériaux. Suite à l'exposition des noyaux nucléaires à une impulsion de radiofréquence (impulsion RF), ces derniers transitionnent d'un état d'équilibre vers un état excité. L'état de retour à l'équilibre est appelé relaxation et est défini par deux intervalles de temps spécifiques : la durée de relaxation longitudinale T_1 et la durée de relaxation transversale T_2 (Keeler, 2015).

Le temps de relaxation longitudinale (T_1), aussi nommé temps de relaxation spin-réseau, décrit la durée nécessaire pour que la composante de l'aimantation nucléaire parallèle au champ

magnétique externe (B_0) revienne à 63 % de sa valeur initiale après stimulation. Ce phénomène illustre le transfert d'énergie entre les noyaux et l'environnement environnant, et sa valeur est influencée par des éléments tels que le déplacement des molécules, la viscosité de l'environnement et la température. Le temps de relaxation transverse (T_2), également nommé temps de relaxation spin-spin, définit la durée nécessaire pour que les éléments transverses de l'aimantation (dans le plan xy) cessent leur synchronisation due aux interactions réciproques entre les noyaux. Ceci entraîne une atténuation graduelle de la force du signal résonnant (Claridge, 1999).

La distinction majeure entre T_1 et T_2 tient au type d'interaction à l'origine de la perte de magnétisme ; alors que T_1 requiert un transfert d'énergie vers le réseau, T_2 découle d'influences réciproques entre les noyaux eux-mêmes sans échange énergétique avec l'environnement. D'un point de vue technique, ces durées influencent la largeur et la précision des lignes spectrales dans la RMN, un T_2 réduit conduisant à des lignes spectrales plus larges. L'examen de ces deux périodes est crucial pour comprendre la structure, la dynamique et le comportement d'interaction des molécules (Ibrahima, 2025; Sakho, 2025).

Chapitre II : Intelligence artificielle

Chapitre II. Intelligence artificielle

II.1. Généralités

II.1.1. Définition

L'intelligence artificielle (IA) désigne l'utilisation de systèmes informatiques et de technologies pour imiter des comportements intelligents et des processus de pensée similaires à ceux des êtres humains. Ce concept a été introduit par John McCarthy en 1956, qui l'a défini comme la science et l'ingénierie visant à concevoir des machines capables de pensée intelligente (Amisha et al., 2019).

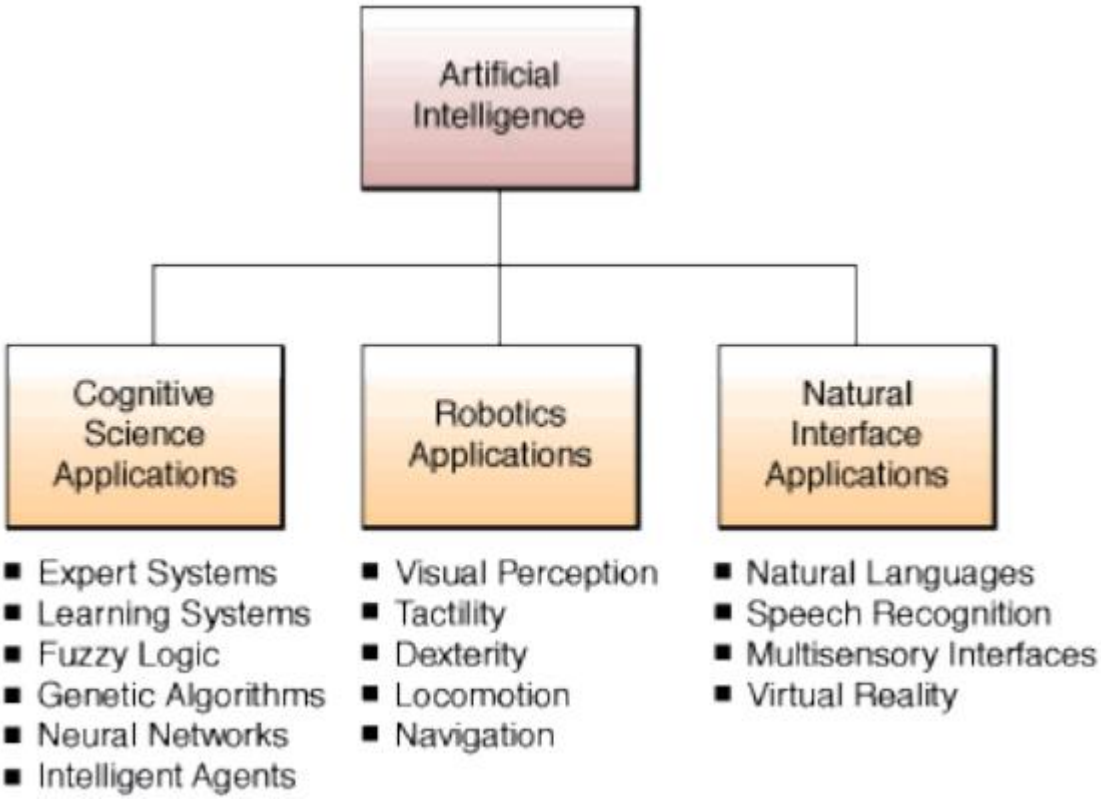

Figure II.1. Vue d'ensemble de l'intelligence artificielle (Strong, 2016).

II.1.2. Historique et évolution

Le concept d'utiliser des ordinateurs pour simuler un comportement intelligent et une pensée critique a été décrit pour la première fois par Alan Turing en 1950. Dans son livre Computers and Intelligence, il a décrit un test simple qui est devenu plus tard connu sous le nom de test de Turing pour déterminer si les ordinateurs sont capables de reproduire l'intelligence humaine (Greenhill & Edmunds, 2020).

Six ans plus tard, John McCarthy a défini le terme intelligence artificielle (IA) comme étant la science et l'ingénierie liées à la création de machines intelligentes (Hamet & Tremblay, 2017).

L'IA a commencé par une simple série de règles si-alors et a évolué au fil de plusieurs décennies pour inclure des algorithmes plus complexes qui fonctionnent de manière similaire au cerveau humain. Il existe de nombreux sous-domaines de l'IA, semblables aux spécialités en médecine (Ruffle et al., 2019).

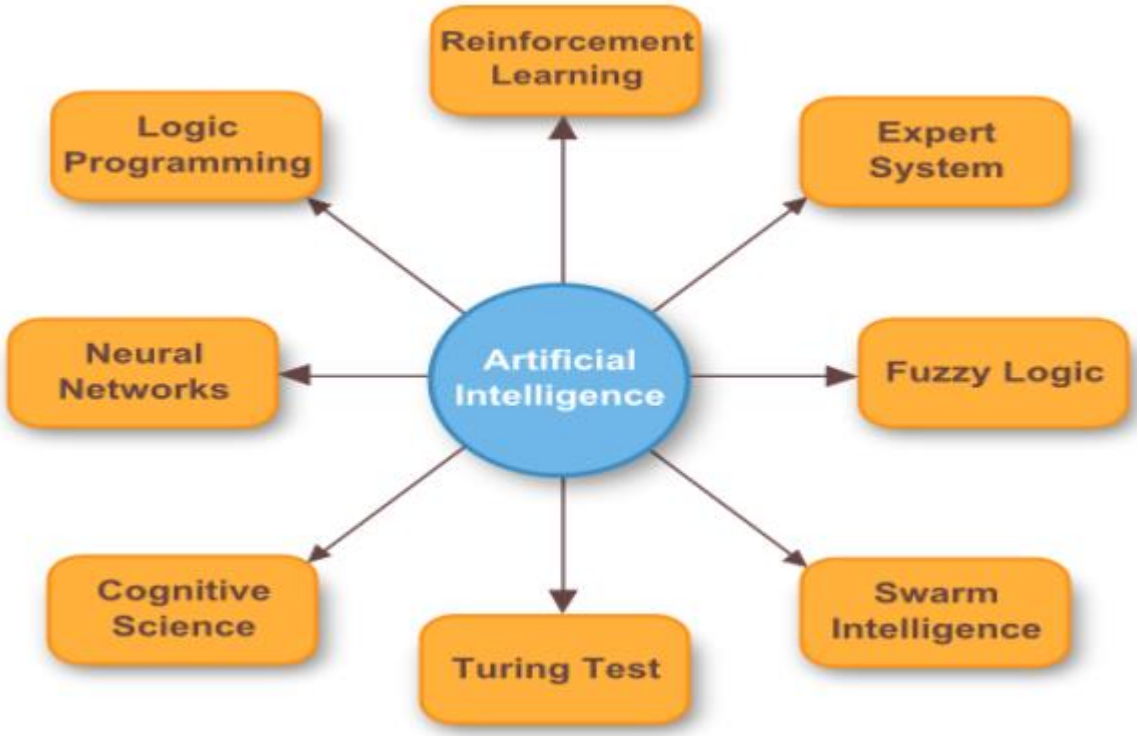

Figure II.2. Domaines de l'intelligence artificielle (Strong, 2016).

II.2. Technologies clés en IA

II.2.1. Apprentissage automatique (Machine Learning)

L'apprentissage automatique est une branche scientifique qui étudie la manière dont les ordinateurs acquièrent des connaissances à partir des données. Il se situe à la croisée des chemins entre les statistiques, qui cherchent à découvrir des relations dans les données, et l'informatique, qui se concentre sur la conception d'algorithmes de calcul performants. Ce croisement entre les mathématiques et l'informatique répond aux défis spécifiques liés à la création de modèles statistiques à partir de gigantesques ensembles de données, qui peuvent comporter des milliards, voire des trillions de points de données (Deo, 2015).

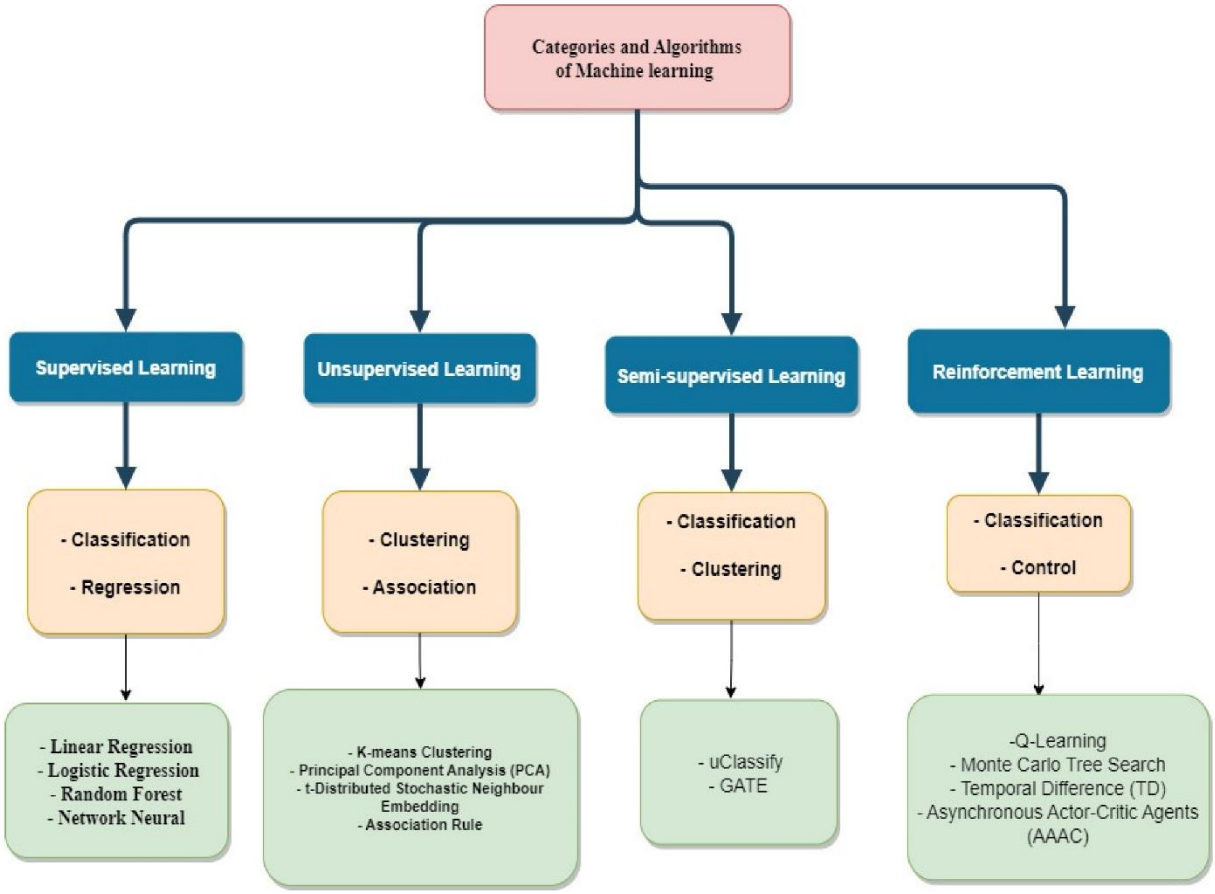

Figure II.3. Les différentes catégories d'apprentissage automatique et les algorithmes (Taye, 2023).

II.2.1.1. Branches de l'apprentissage automatique

II.2.1.1.1. Apprentissage supervisé

L'apprentissage supervisé est la tâche d'apprentissage automatique qui consiste à apprendre une fonction qui associe une entrée à une sortie sur la base d'exemples de paires entrée-sortie. Il déduit une fonction à partir de données d'apprentissage étiquetées constituées d'un ensemble d'exemples d'apprentissage. Les algorithmes d'apprentissage automatique supervisé sont ceux qui nécessitent une assistance externe. L'ensemble des données d'entrée est divisé en ensembles de données de formation et de test. L'ensemble de données de formation contient des variables de sortie qui doivent être prédites ou classées. Tous les algorithmes apprennent un certain type de modèles à partir de l'ensemble de données de formation et les appliquent à l'ensemble de données de test pour la prédiction ou la classification (Mahesh, 2020).

II.2.1.1.2. Apprentissage non supervisé

L'apprentissage non supervisé constitue une catégorie de méthodes d'apprentissage automatique dans laquelle les données d'entrée sont structurées de manière similaire à celles utilisées en apprentissage supervisé, c'est-à-dire sous forme de vecteurs caractéristiques. Toutefois, aucune étiquette ou sortie attendue n'est fournie. L'objectif principal est d'extraire des structures latentes ou des régularités intrinsèques au sein de l'ensemble de données, souvent complexes ou non triviales. Un exemple représentatif de ce type d'approche est le regroupement (clustering), qui consiste à partitionner les données en sous-ensembles homogènes, chaque groupe rassemblant des instances partageant des caractéristiques communes (Carleo et al., 2019).

II.2.1.1.3. Apprentissage semi supervisée

L'apprentissage semi-supervisé peut être considéré comme le "juste milieu" entre l'apprentissage supervisé et l'apprentissage non supervisé. Il est particulièrement utile pour les ensembles de données qui contiennent à la fois des données étiquetées et non étiquetées (c'est-à-dire que toutes les caractéristiques sont présentes, mais que toutes les caractéristiques n'ont pas de cibles associées). Cette situation se produit généralement lorsque l'étiquetage des images prend beaucoup de temps ou est trop coûteux. L'apprentissage semi-supervisé est souvent utilisé pour les images médicales, où un médecin peut étiqueter un petit sous-ensemble d'images et les utiliser pour former un modèle. Ce modèle est ensuite utilisé pour classer le reste des images non étiquetées de l'ensemble de données. L'ensemble de données étiquetées qui en résulte est ensuite utilisé pour former un modèle de travail qui devrait, en théorie, être plus performant que les modèles non supervisés.

II.2.1.1.4. Apprentissage renforcé

L'apprentissage par renforcement est la technique qui consiste à former un algorithme à une tâche spécifique pour laquelle aucune réponse n'est correcte, mais pour laquelle un résultat global est souhaité. Il s'agit sans doute de la tentative la plus proche de la modélisation de l'expérience d'apprentissage humaine, car elle apprend également à partir d'essais et d'erreurs plutôt qu'à partir de données seules (Choi et al., 2020).

II.2.2. L'apprentissage en profondeur 'Deep learning'

L'apprentissage en profondeur est un ensemble d'algorithmes d'apprentissage automatique qui tentent d'apprendre à plusieurs niveaux, correspondant à différents niveaux d'abstraction. Il utilise généralement des réseaux neuronaux artificiels. Les niveaux de ces modèles statistiques appris correspondent à des niveaux distincts de concepts, où les concepts de niveau supérieur sont définis

à partir des concepts de niveau inférieur, et où les mêmes concepts de niveau inférieur peuvent aider à définir de nombreux concepts de niveau supérieur (Deng & Yu, 2014).

L'apprentissage profond, également connu sous le nom de réseau neuronal profond, est un domaine de recherche nouveau et populaire qui donne des résultats impressionnants et se développe rapidement (Erickson et al., 2017).

II.2.2.1. Types des réseaux neuronaux

Le réseau neuronal est une structure composée de plusieurs couches cachées de neurones où la sortie d'un neurone d'une couche devient l'entrée d'un neurone de la couche suivante (Choi et al., 2020).

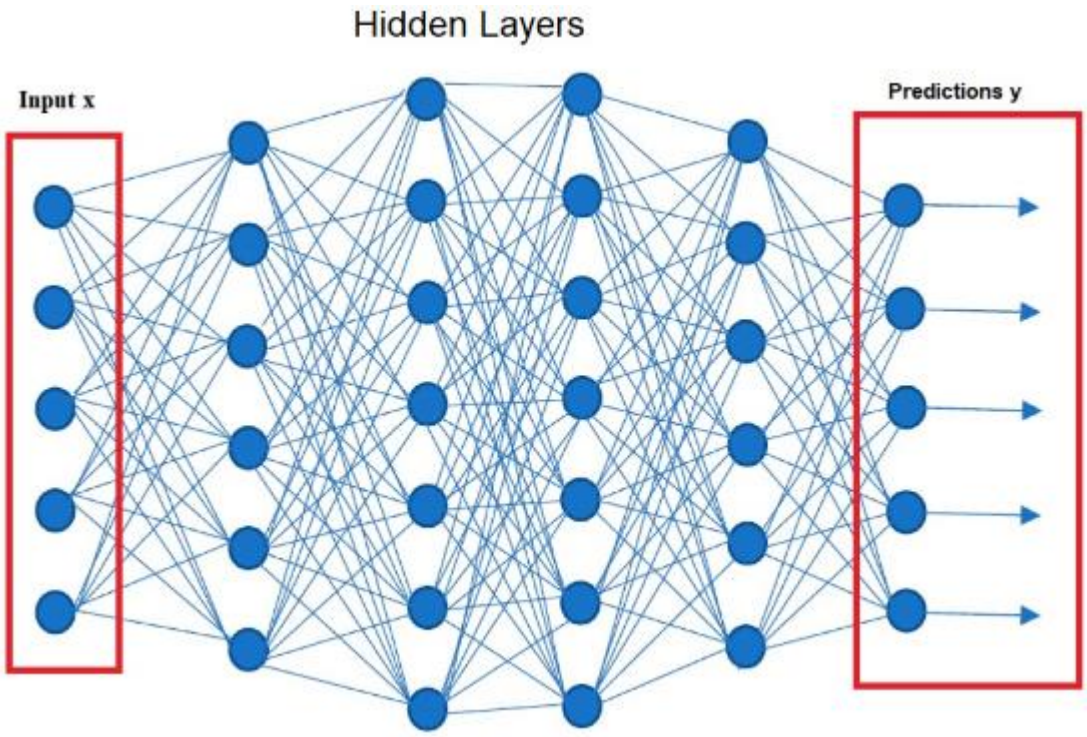

Figure II.4. Représentation d'un réseau neuronal (Taye, 2023).

II.2.2.1.1. Réseaux neuronaux convolutifs (CNN)

Les Réseaux neuronaux convolutifs sont un type de modèle d'apprentissage profond spécialement conçu pour le traitement des données en grille, telles que les images. Ils se composent de plusieurs couches de filtres convolutifs qui apprennent automatiquement à détecter des caractéristiques telles que les bords, les textures et les formes. Ces filtres sont appliqués à de petites parcelles des données d'entrée, et les cartes de caractéristiques résultantes passent par des couches supplémentaires pour détecter des modèles de plus haut niveau (Taye, 2023).

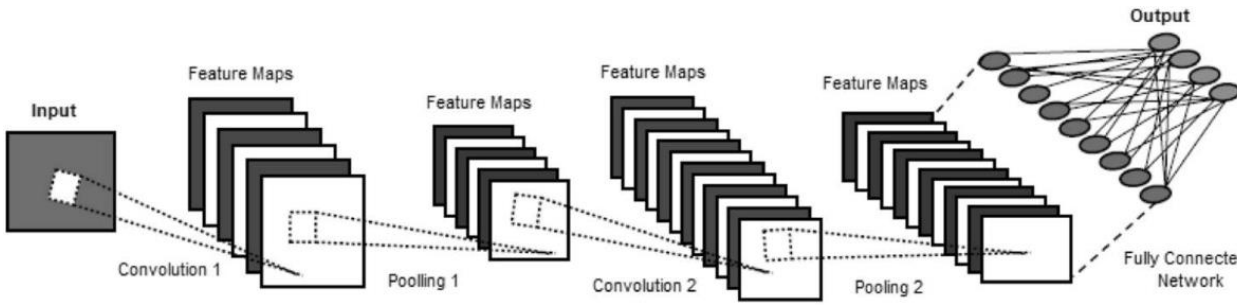

Figure II.5. Exemples de réseau neuronal convcolutif (CNN et ConveNet) comprenant plusieurs couches de convolution et de mise en commun (Sarker, 2021).

II.2.2.1.2. Réseaux neuronaux récurrents (RNN)

Les réseaux neuronaux récurrents sont conçus pour les données séquentielles, où l'ordre des points de données est important. Contrairement aux réseaux de type feedforward, les Réseaux neuronaux récurrents ont des boucles qui permettent aux informations de persister dans le temps. Cela les rend efficaces pour les tâches impliquant des séries de données temporelles, telles que la prédiction des prix des actions ou les prévisions météorologiques (Taye, 2023).

Tableau II.1. Avantages et inconvénients de l'apprentissage profond (Taye, 2023).

Avantages de l'apprentissage profond	Inconvénients de l'apprentissage profond
-La possibilité de générer de nouvelles caractéristiques à partir des données de formation existantes limitées. -Peut produire des résultats fiables et exploitables pour des tâches en utilisant des approches d'apprentissage non supervisées. - Il réduit le temps nécessaire à l'ingénierie des fonctionnalités, l'une des activités impliquées dans l'apprentissage de l'utilisation de l'apprentissage automatique. - La formation continue a rendu son architecture adaptable et capable de résoudre toute une série de problèmes.	- Il y a moins de possibilités d'amélioration dans le processus de formation, car l'ensemble du processus de formation dépend du flux constant de données. - Avec l'augmentation du nombre d'ensembles de données disponibles, l'entraînement informatique devient nettement plus coûteux. - La révision des fautes manque de transparence. Il n'y a pas d'étapes intermédiaires pour soutenir les revendications d'une faute particulière. Un algorithme entier est mis à jour pour résoudre le problème. - Pour la formation des ensembles de données, vous avez besoin de ressources coûteuses, de processeurs rapides et de GPU puissants.

II.2.2.2. Apprentissage profond Vs Apprentissage Automatique

Tableau II.2. Comparaison entre l'apprentissage profond et l'apprentissage automatique (Taye, 2023).

	Apprentissage automatique	Apprentissage profond
Intervention humaine	Pour obtenir des résultats, l'apprentissage automatique nécessite un engagement humain plus continu.	L'apprentissage en profondeur est plus difficile à mettre en œuvre au départ, mais nécessite peu d'interventions par la suite.
Matériel	Les programmes d'apprentissage automatique sont généralement moins complexes que les algorithmes d'apprentissage profond et peuvent souvent être exécutés sur des ordinateurs standard.	Les systèmes d'apprentissage profond exigent des ressources puissantes, d'où l'utilisation croissante des GPU, qui offrent une mémoire rapide et un parallélisme efficace permettant de réduire les délais de transfert de données.
Temps	Les systèmes d'apprentissage automatique sont rapides à mettre en place, mais leurs performances peuvent être limitées.	Les systèmes d'apprentissage en profondeur sont plus longs à configurer, mais peuvent fournir des résultats rapides, avec une qualité qui s'améliore avec plus de données.
Applications	L'apprentissage automatique est déjà utilisé dans des domaines comme le courrier électronique, la banque et la médecine.	L'apprentissage en profondeur permet de développer des systèmes complexes et autonomes comme les voitures autonomes et les robots chirurgicaux.
Approche	L'apprentissage automatique repose généralement sur des données organisées et des méthodes classiques comme la régression linéaire.	L'apprentissage profond utilise des réseaux neuronaux et est conçu pour traiter des volumes massifs de données non structurées.
Data	Les algorithmes d'apprentissage automatique nécessitent moins de données que ceux de l'apprentissage profond, mais la qualité des données y est plus essentielle.	Les algorithmes d'apprentissage profond ont besoin de grandes quantités de données, mais ils s'améliorent de façon autonome avec l'apport de nouvelles données.

Partie pratique

Chapitre III : Matériels et méthodes

Chapitre III. Matériel et méthodes

III.1. Matériels

III.1.1. Data set

La Human Metabolome Database (HMDB Version 5.0) est une base de données électronique en libre accès contenant des informations détaillées sur les métabolites de petites molécules présents dans le corps humain. Elle est destinée à des applications en métabolomique, en chimie clinique, à la découverte de biomarqueurs ainsi qu'à la formation générale. La base de données est conçue pour contenir ou référencer trois types de données : 1) des données chimiques, 2) des données cliniques, et 3) des données de biologie moléculaire/biochimie. Elle comprend 220 945 entrées métaboliques, incluant à la fois des métabolites hydrosolubles et liposolubles. De plus, 8 610 séquences protéiques (enzymes et transporteurs) sont associées à ces entrées. Chaque fiche *MetaboCard* contient 130 champs de données, dont 2/3 sont consacrés aux données chimiques/cliniques et 1/3 aux données enzymatiques ou biochimiques. De nombreux champs sont hyperliés à d'autres bases de données (KEGG, PubChem, MetaCyc, ChEBI, PDB, UniProt et GenBank) et à divers applets de visualisation des structures et des voies métaboliques.

La HMDB prend en charge des recherches complexes par texte, séquence, structure chimique, spectres MS et spectres NMR. Quatre bases de données supplémentaires, DrugBank, T3DB, SMPDB et FooDB, font également partie de la suite HMDB. DrugBank contient des informations équivalentes sur environ 2 832 médicaments et 800 métabolites médicamenteux ; T3DB fournit des données sur environ 3 670 toxines courantes et polluants environnementaux ; SMPDB comprend des diagrammes pour environ 132 335 voies métaboliques, médicamenteuses et pathologiques humaines, ainsi que 60 628 voies pour d'autres organismes ; enfin, FooDB fournit des informations comparables sur environ 70 000 composants alimentaires et additifs.

Le jeu de données principal utilisé pour la modélisation spectroscopique par apprentissage automatique a été extrait de HMDB_NMR_Spectra_Files_2023-07-01.zip, une archive compressée de 873 Mo issue de la Human Metabolome Database (HMDB), spécialisée dans la résonance magnétique nucléaire (RMN) appliquée aux métabolites humains. Après extraction, elle génère 260 138 fichiers XML, totalisant 19,6 Go. Chaque fichier correspond à un spectre RMN unique (1D ou 2D), accompagné de métadonnées physicochimiques et instrumentales détaillées, acquises dans des conditions expérimentales rigoureusement documentées.

Les fichiers RMN 1D contiennent des champs tels que la concentration de l'échantillon, la masse, le solvant utilisé, la température, le pH, la référence chimique, le type d'instrument, le noyau observé (ex. : ^{1}H) et la fréquence spectrale (ex. : 500 MHz), ainsi que le nombre de pics

détectés. En revanche, les fichiers RMN 2D décrivent deux noyaux d'observation (nucleus-x et nucleus-y, souvent $^{1}H/^{1}H$), un nombre de pics généralement plus élevé, et comportent des notes techniques précisant le type d'expérience (ex. : TOCSY, COSY) ainsi que l'état d'exportation (exported).

Chaque spectre est relié à une molécule via un identifiant unique HMDB (database-id ou structure-id), ce qui permet d'identifier précisément les entités représentées. Après extraction et analyse, le jeu de données couvre 13 538 molécules distinctes, chacune représentée par un ou plusieurs fichiers spectroscopiques. L'ensemble constitue une base robuste pour des applications telles que la prédiction spectrale, l'annotation automatique ou la modélisation de relations structure–spectre en chimioinformatique et en métabolomique computationnelle.

Tableau III.1. Paramètres de classement descriptifs du dataset.

N°	Paramètre	Description
1	Database-id	Identifiant principal du métabolite dans la HMDB (ex. : HMDB0000001).
2	Structure-id	Identifiant de la structure chimique. Il peut être équivalent au database-id.
3	Nucleus / nucleus-x, nucleus-y	Noyau ou paires de noyaux observés (ex. : ^{1}H pour 1D, $^{1}H/^{1}H$ pour 2D).
4	Frequency	Fréquence d'acquisition du spectre en MHz (ex. : 500 MHz).
5	Instrument-type	Type d'instrument utilisé pour l'analyse RMN (ex. : Varian, Bruker).
6	Sample-concentration	Concentration de l'échantillon (ex. : 50).
7	Sample-concentration-units	Unité de la concentration (ex. : mM).
8	Sample-mass	Masse de l'échantillon utilisé (ex. : 5.9).
9	Sample-mass-units	Unité de la masse (ex. : mg).
10	Solvent	Solvant employé pour la mesure RMN (ex. : Water, $CDCl_3$).
11	Sample-temperature	Température de l'échantillon pendant l'acquisition (ex. : 25).
12	Sample-temperature-units	Unité de température (ex. : Celsius).
13	Sample-ph	pH de l'échantillon analysé (ex. : 7.1).
14	Chemical-shift-reference	Référence de déplacement chimique utilisée (ex. : DSS, TMS).
15	Peak-counter	Nombre de pics dans le spectre (ex. : 15 pour 1D, 168 pour 2D).
16	Distinct-peaks	Nombre de pics distincts détectés (parfois identique à peak-counter).
17	Spectra-assessment	Évaluation qualitative du spectre (ex. : Excellent).
18	Sample-assessment	Évaluation de la qualité de l'échantillon (ex. : Excellent).
19	Collection-date	Date d'acquisition du spectre (ex. : 4/18/2006).
20	Created-at	Date de création du fichier XML (ex. : 12/4/2012 23:50).
21	Updated-at	Date de dernière mise à jour du fichier (ex. : 8/3/2022 21:58).
22	Notes	Commentaires techniques (ex. : "TOCSY. Unexported temporarily…"). Souvent présents pour les spectres 2D.
23	Exported	Statut d'exportation du fichier (TRUE / FALSE).
24	Searchable	Booléen indiquant si le spectre est indexable pour la recherche (TRUE).

III.1.2. Machine

Tous les travaux de prétraitement des données spectrales, de développement des modèles et d'entraînement des réseaux de neurones ont été réalisés sur une station de travail équipée d'un processeur Intel Core i7-8700K, de 32 Go de mémoire vive (RAM), d'un disque SSD NVMe de troisième génération, ainsi que d'un GPU NVIDIA GeForce RTX 3060, optimisé pour les charges de calcul intensives en apprentissage profond. Le système d'exploitation utilisé était Windows 10 Professionnel 64 bits, version 22H2. Cette configuration s'est révélée parfaitement adaptée aux tâches de modélisation avancée, telles que la prédiction multitâche des structures chimiques à partir de spectres RMN, démontrant à la fois performance, fluidité et fiabilité dans les flux de travail bioinformatiques et chimiométriques.

III.1.3. Logiciels et bibliothèques

Le développement du modèle d'intelligence artificielle et l'analyse des données spectrales ont été réalisés dans un environnement de programmation moderne, stable et adapté aux besoins des applications en bioinformatique, chimio-informatique et science des données. Les outils principaux utilisés sont les suivants :

- Anaconda (version 2.6.6) : utilisé comme plateforme de gestion d'environnements et de paquets, facilitant l'installation et la compatibilité entre les bibliothèques scientifiques.
- Python (version 3.13) : langage principal utilisé pour le traitement des données, l'analyse spectroscopique et le développement du modèle d'intelligence artificielle.

Plusieurs bibliothèques spécialisées ont été mobilisées pour le traitement, l'analyse et la modélisation des données, notamment :

- *Pandas* : bibliothèque essentielle pour la manipulation des données tabulaires (chargement, filtrage, traitement des colonnes).
- *Numpy* : utilisée pour les calculs numériques performants sur des vecteurs et matrices, notamment lors du binning des spectres RMN.
- *Pickle* : outil de sérialisation permettant d'enregistrer et de recharger efficacement les objets Python (modèles, encodeurs, configurations).
- *Scikit-learn (sklearn)* : bibliothèque d'apprentissage automatique utilisée pour les étapes de prétraitement, de normalisation et de transformation des jeux de données.
- *TensorFlow et tensorflow.keras* : infrastructure centrale pour la construction, l'entraînement et la validation des modèles de deep learning, avec une interface de haut niveau adaptée à la modélisation multitâche.

Cette configuration logicielle, couplée à une infrastructure matérielle robuste, a permis la mise en œuvre d'un pipeline analytique reproductible, modulable et adapté à l'extraction de connaissances à partir de grands ensembles de spectres RMN.

III.2. Méthodes

III.2.1. Préparation et nettoyage des données

La phase de préparation des données constitue une étape essentielle dans tout processus de modélisation en intelligence artificielle, en particulier lorsqu'il s'agit de données spectrales issues de fichiers XML riches, complexes et hétérogènes. Dans le cadre de cette étude, nous avons exploité l'archive NMR Spectra Files (XML) de la base de données HMDB, laquelle contient initialement 260 138 fichiers XML représentant des spectres RMN annotés.

Étant donné les contraintes de temps et de ressources, nous avons opté pour la constitution manuelle d'un sous-ensemble représentatif de 40 fichiers XML, ouverts un à un, puis fusionnés dans une table structurée. Ce travail a été réalisé de manière semi-automatisée, avec une inspection manuelle préalable et une normalisation rigoureuse des champs sélectionnés afin de garantir la cohérence et l'intégrité des données.

Parmi l'ensemble des attributs disponibles dans les fichiers XML, nous avons choisi de ne conserver que les champs présents systématiquement dans tous les fichiers, de manière à garantir l'homogénéité du jeu de données utilisé pour l'entraînement. Les attributs retenus sont les suivants : Id, Solvent, Peak-counter, Nucleus, Sample-Ph, Sample-Temperature, Chemical-Shift-Reference, Structure-Id, Database-Id.

Tous les autres attributs, bien que parfois informatifs (comme les commentaires, le type d'instrument, l'état de l'échantillon, etc.) ont été exclus manuellement, car leur présence était irrégulière ou incomplète dans les fichiers sélectionnés.

Les 40 spectres RMN extraits correspondent à 10 composés moléculaires uniques, chacun étant représenté par plusieurs fichiers expérimentaux variant selon les paramètres analytiques (type de noyau, solvant, nombre de pics, etc.). Ce choix méthodologique vise à conserver une diversité minimale tout en réduisant la complexité du prétraitement initial, rendant ainsi le jeu de données exploitable dans un cadre de modélisation supervisée.

III.2.2. Transformation des données

Les données ont été préparées en vue de l'apprentissage supervisé à l'aide des bibliothèques *Pandas*, *NumPy* et *Scikit-learn*. Les variables explicatives comprenaient des données numériques (peak-counter, sample-ph, sample-temperature, structure-id, id) et catégorielles (solvent, nucleus,

chemical-shift-reference). Les variables catégorielles ont été encodées par la méthode *one-hot encoding* (*pandas*.get_*dummies*), et les variables numériques ont été standardisées via *StandardScaler* de *Scikit-learn*. La variable cible (database-id) a été convertie en étiquettes numériques à l'aide de *LabelEncoder*. Les classes contenant moins de trois échantillons ont été filtrées pour garantir un apprentissage stable. L'ensemble de données a ensuite été divisé en un jeu d'entraînement (80 %) et un jeu de test (20 %) de manière stratifiée à l'aide de *train_test_split*.

III.2.3. Entraînement du Modèle

Le modèle a été entraîné à l'aide de l'algorithme *Random Forest*, implémenté dans *Scikit-learn*. Ce choix repose sur sa robustesse face aux données bruitées, sa capacité à gérer les variables catégorielles transformées, ainsi que sur sa performance en classification multi-classes.

Le classificateur a été configuré avec 300 arbres décisionnels (n_estimators=300) et une profondeur maximale de 20 (max_depth=20) afin d'assurer un bon compromis entre biais et variance. Le paramètre *class_weight*="*balanced_subsample*" a été activé pour compenser les déséquilibres dans les effectifs des classes. L'entraînement a été effectué sur les données prétraitées *(X_train, y_train*) via la méthode *fit*, après filtrage des classes sous-représentées et division stratifiée des données. Ce classificateur a pour objectif de prédire avec précision l'identifiant de la molécule (database-id) à partir des métadonnées expérimentales NMR.

III.2.4. Evaluation du modèle

L'évaluation du modèle *Random Forest* a été réalisée sur un jeu de test stratifié, en calculant plusieurs métriques de performance : accuracy, log loss, précision pondérée, rappel pondéré et F1-score pondéré, à l'aide de *Scikit-learn*. La fonction *predict_proba* a été utilisée pour estimer les probabilités de classes et calculer la log loss. Une matrice de confusion a été générée pour visualiser la répartition des erreurs de classification entre les classes cibles (identifiants database-id), accompagnée d'un rapport de classification détaillé. L'importance relative des dix variables les plus discriminantes a été visualisée à l'aide d'un graphique horizontal. Enfin, l'ensemble des scores a été résumé dans un graphique comparatif, et toutes les sorties (figures et valeurs) ont été enregistrées pour traçabilité.

III.2.5. Sauvegarde du Modèle

Le modèle entraîné (Random Forest), ainsi que les objets de prétraitement essentiels à savoir le *StandardScaler* pour la normalisation des données et le *LabelEncoder* pour l'encodage des classes ont été sauvegardés au format pickle (.pkl) à l'aide du module *pickle*. Ces fichiers permettent de réutiliser le modèle sans nécessiter un réentraînement. Tous les objets ont été stockés

dans un répertoire dédié (Model), assurant une organisation claire pour les étapes futures de prédiction et de déploiement.

III.2.6. Déploiement du Modèle (GUI)

Une interface graphique interactive a été développée à l'aide de la bibliothèque *Gradio* afin de permettre la prédiction directe de l'identifiant de base de données (database ID) à partir de métadonnées RMN saisies par l'utilisateur. L'interface accepte les champs suivants : solvant, noyau, nombre de pics, pH de l'échantillon, température, référence de déplacement chimique, identifiant de structure et identifiant d'échantillon. Les entrées sont transformées pour correspondre aux caractéristiques utilisées lors de l'entraînement (encodage one-hot et normalisation). Le modèle préalablement entraîné prédit la classe, laquelle est ensuite décodée en identifiant d'origine via l'encodeur de labels. L'interface fournit le résultat en temps réel dans un environnement convivial.

Chapitre IV : Résultats et discussion

Chapitre IV. Résultats et discussion

IV.1. Résultats

IV.1.1. Préparation et nettoyage des données

Un total de 40 fichiers NMR ont été intégrés dans un fichier CSV structuré. Les données brutes comprenaient des variables telles que le solvant, le noyau observé, le nombre de pics, la température et le pH (Tab. IV.1).

Tableau IV.1. Résumé des données métabolomiques après les étapes de prétraitement et de nettoyage.

Id	Solvent	Peak-counter	Nucleus	pH	T° C	Chemical-shift reference	Structure-id	Database-id
1024	Water	7	1H	7	25	DSS	9391	HMDB0000005
4754	D2O	7	1H	7.4	25	DSS	9391	HMDB0000005
4755	D2O	4	13C	7.4	25	DSS	9391	HMDB0000005
917	Water	2	1H	7	25	DSS	9391	HMDB0000005
1039	Water	45	1H	7	25	TMS	5112	HMDB0000030
4827	D2O	39	1H	7.4	25	DSS	5112	HMDB0000030
4828	D2O	85	13C	7.4	25	DSS	5112	HMDB0000030
932	DMSO	10	1H	7	25	TMS	5112	HMDB0000030
1047	Water	12	1H	7	25	DSS	8416	HMDB0000039
1124	Water	5	13C	7	25	DSS	8416	HMDB0000039
5025	D2O	12	1H	7.4	25	DSS	8416	HMDB0000039
5026	D2O	4	13C	7.4	25	DSS	8416	HMDB0000039
1048	Water	1	1H	7	25	DSS	9983	HMDB0000042
1126	Water	3	13C	7	25	DSS	9983	HMDB0000042
4746	D2O	1	1H	7.4	25	DSS	9983	HMDB0000042
1106	Water	1	1H	7	25	DSS	9983	HMDB0000042
1087	Water	7	1H	7	25	DSS	5881	HMDB0000108
1148	Water	3	13C	7	25	DSS	5881	HMDB0000108
4985	D2O	7	1H	7	25	DSS	5881	HMDB0000108
4986	D2O	2	13C	7	25	DSS	5881	HMDB0000108
1104	Water	1	1H	7	25	DSS	9144	HMDB0000134
1163	Water	3	13C	7	25	DSS	9144	HMDB0000134
4764	D2O	1	1H	7	25	DSS	9144	HMDB0000134
4765	D2O	2	13C	7	25	DSS	9144	HMDB0000134
1107	Water	1	1H	7	25	DSS	6820	HMDB0000142
1165	Water	2	13C	7	25	DSS	6820	HMDB0000142
4875	D2O	1	1H	7	25	DSS	6820	HMDB0000142
4876	D2O	1	13C	7	25	DSS	6820	HMDB0000142
1115	Water	12	1H	7	25	DSS	10055	HMDB0000156
1173	Water	5	13C	7	25	DSS	10055	HMDB0000156
4825	D2O	3	1H	7.4	25	DSS	10055	HMDB0000156
4826	D2O	4	13C	7.4	25	DSS	10055	HMDB0000156
1117	Water	24	1H	7	25	DSS	5125	HMDB0000158

1175	Water	7	13C	7	25	DSS	5125	HMDB0000158
4862	D2O	16	1H	7.7	25	DSS	5125	HMDB0000158
4863	D2O	16	1H	9.43	25	DSS	5125	HMDB0000158
1183	Water	6	1H	7	25	DSS	9029	HMDB0000202
1201	Water	4	13C	7	25	DSS	9029	HMDB0000202
5057	D2O	1	1H	7.4	25	DSS	9029	HMDB0000202
5058	D2O	3	13C	7.4	25	DSS	9029	HMDB0000202

IV.1.2. Transformation des données

Les variables catégorielles telles que "solvent", "nucleus" et "chemical-shift-reference" ont été converties en variables indicatrices via un encodage one-hot. Les variables numériques, notamment "peak-counter", "sample-ph", "sample-temperature", "structure-id" et "id", ont été normalisées à l'aide d'un StandardScaler. Enfin, les identifiants cibles ("database-id") ont été encodés numériquement à l'aide d'un LabelEncoder. Cette transformation a permis une vectorisation cohérente et compatible avec les algorithmes d'apprentissage supervisé (Tab. IV.2).

Tableau IV.2. Vue d'ensemble des données métabolomiques après transformation et encodage.

Id	Solvent	Peak-counter	Nucleus	pH	T° C	Chemical-shift reference	Structure-id	Database-id
1024	1	7	1H	7	25	1	9391	0000005
4754	0	7	1H	7.4	25	1	9391	0000005
4755	0	4	13C	7.4	25	1	9391	0000005
917	1	2	1H	7	25	1	9391	0000005
1039	1	45	1H	7	25	0	5112	0000030
4827	0	39	1H	7.4	25	1	5112	0000030

IV.1.3. Entraînement du modèle

Un modèle de forêt aléatoire (RandomForestClassifier) a été entraîné avec 300 arbres et une profondeur maximale de 20, sur des données stratifiées. Le modèle a montré une faible capacité de généralisation sur le jeu de test.

IV.1.4. Évaluation du modèle

L'évaluation sur le jeu de test a révélé des performances limitées du modèle Random Forest. L'accuracy atteint 33 %, avec une précision pondérée de 50 %, un rappel de 33 % et un F1-score de 39 %, accompagnés d'un log loss élevé de 1.48.

La matrice de confusion (Figure IV.2.) montre des prédictions dispersées, avec plusieurs erreurs d'attribution entre structures proches, traduisant une confusion entre certaines classes. Ces résultats soulignent un problème potentiel de représentation des données ou d'insuffisance d'échantillons pour certains identifiants.

Des optimisations seront nécessaires pour améliorer la capacité de généralisation, notamment par un enrichissement des données, un équilibrage des classes ou une exploration de modèles alternatifs.

Cette figure (Figure IV.1.) résume les principales métriques de performance obtenues lors de l'évaluation du modèle sur les données de test :

Accuracy (Exactitude) : 0.33, Cela signifie que 33 % des prédictions totales du modèle sont correctes, toutes classes confondues. Une valeur relativement faible, indiquant une capacité limitée à généraliser sur l'ensemble du jeu de test.

Precision (Précision) : 0.50, Cette métrique reflète la proportion de prédictions positives qui sont réellement correctes. Une précision de 50 % indique que la moitié des identifiants prédits sont exacts, ce qui suggère une qualité prédictive modérée.

Recall (Rappel) : 0.33, Le rappel mesure la capacité du modèle à détecter correctement les cas positifs parmi l'ensemble des cas réels. Une valeur de 0.33 indique que seul un tiers des identifiants réels ont été correctement identifiés.

F1 Score : 0.39, Il s'agit de la moyenne harmonique entre la précision et le rappel, offrant une mesure équilibrée entre les faux positifs et les faux négatifs. Un F1 score de 0.39 souligne une performance globale insuffisante.

Log Loss : 1.48, Cette métrique évalue la qualité de la probabilité prédite. Une valeur élevée comme 1.48 indique que le modèle attribue souvent des probabilités incorrectes ou trop incertaines, ce qui affecte la fiabilité des prédictions.

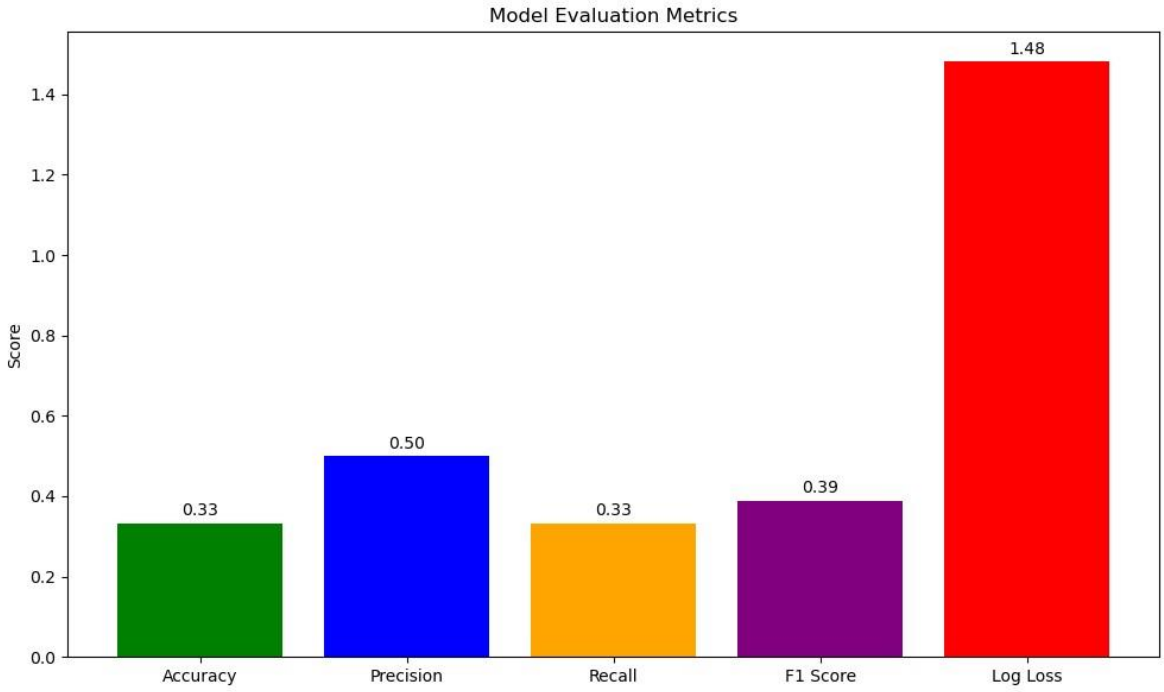

Figure IV.1. Métriques globales d'évaluation du modèle.

Ces résultats révèlent des performances globalement faibles, probablement dues à un déséquilibre entre les classes, une complexité des attributs, ou un surapprentissage. Des ajustements supplémentaires, comme l'optimisation des hyperparamètres, le rééchantillonnage ou l'enrichissement des données, pourraient améliorer la qualité du modèle.

Cette figure (Figure IV.2.) représente la matrice de confusion, qui compare les classes réelles (en ligne) aux classes prédites (en colonne). Les prédictions correctes apparaissent sur la diagonale principale, tandis que les erreurs se situent hors de cette diagonale.

Dans ce cas, plusieurs confusions notables sont observées, notamment pour les identifiants HMDB000030, HMDB000042 ou HMDB000156, qui sont fréquemment mal classés. Ces résultats peuvent indiquer :

- Une confusion inter-classes due à la similarité des profils NMR ou des métadonnées associées,
- Un déséquilibre dans le jeu d'entraînement, certaines classes étant sous-représentées,
- Ou une limite du modèle à distinguer certaines structures chimiques proches.

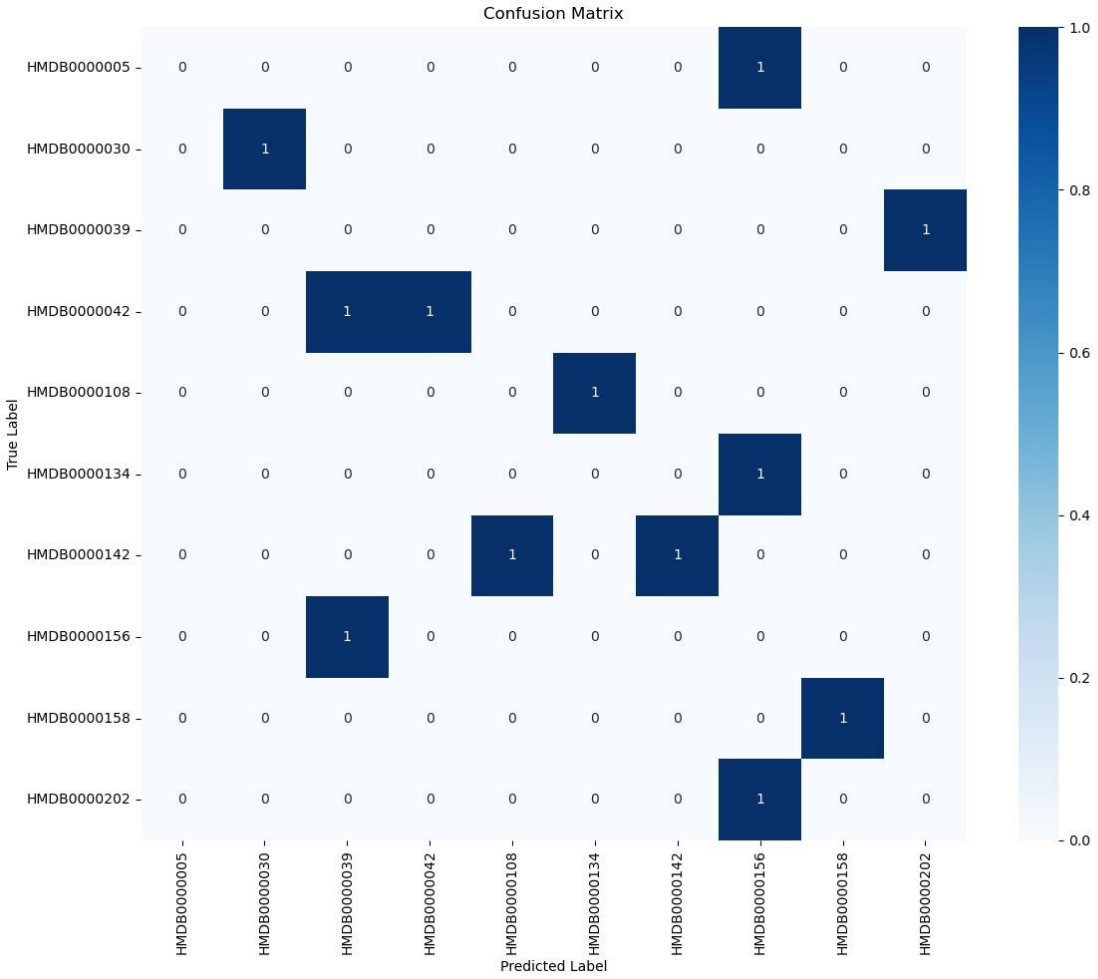

Figure IV.2. Matrice de confusion – Étiquettes réelles vs prédites.

Cette matrice met en évidence les points faibles du modèle, en particulier pour certaines classes spécifiques. Elle justifie l'exploration de techniques de rééchantillonnage ou l'utilisation de modèles plus sensibles aux données déséquilibrées.

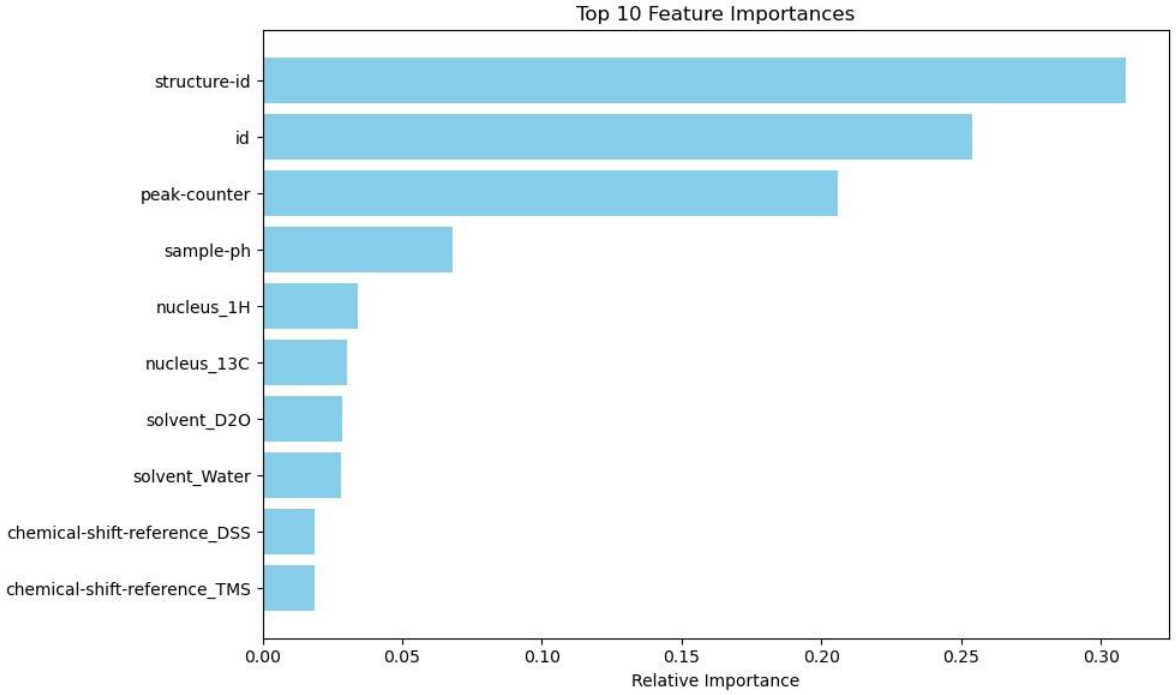

Figure IV.3. Analyse des importances des variables dans le modèle RMN pour l'annotation structurale.

Le graphique présente les dix variables les plus importantes dans le modèle d'apprentissage supervisé appliqué aux données de résonance magnétique nucléaire (RMN). Les résultats montrent que les identifiants structurels (structure-id), les identifiants globaux (id), et le nombre de pics détectés (peak-counter) sont les contributeurs dominants à la performance prédictive du modèle. Ces variables sont suivies, à moindre importance, par des facteurs expérimentaux tels que le pH de l'échantillon (sample-ph), le type de noyau observé (nucleus_1H, nucleus_13C), et les solvants utilisés (solvent_D2O, solvent_Water). Enfin, les références de décalage chimique (chemical-shift-reference_DSS, chemical-shift-reference_TMS) ont un impact relativement faible. Cette hiérarchisation reflète la prédominance des caractéristiques moléculaires et spectrales par rapport aux paramètres expérimentaux dans la capacité du modèle à discriminer les structures.

IV.1.5. Sauvegarde du modèle

Le modèle entraîné (RandomForestClassifier), le scaler de normalisation (StandardScaler) ainsi que l'encodeur de labels (LabelEncoder) ont été sauvegardés au format pickle (.pkl) pour garantir leur réutilisation dans des prédictions ultérieures. Tous les fichiers ont été exportés avec succès dans le répertoire dédié (Model/), assurant ainsi la portabilité et la reproductibilité du pipeline de prédiction.

IV.1.6. Déploiement du Modèle (GUI)

Une interface utilisateur interactive a été développée à l'aide de Gradio afin de permettre la prédiction de l'identifiant de base de données d'un métabolite à partir de métadonnées RMN. Cette interface prend en entrée sept attributs (solvent, nucleus, peak-counter, sample pH, sample temperature, chemical shift reference, structure ID, ID), puis affiche l'identifiant prédictif (Database ID) correspondant. L'outil s'est révélé intuitif et efficace, avec un retour de prédiction instantané, facilitant l'exploitation du modèle sans besoin de codage.

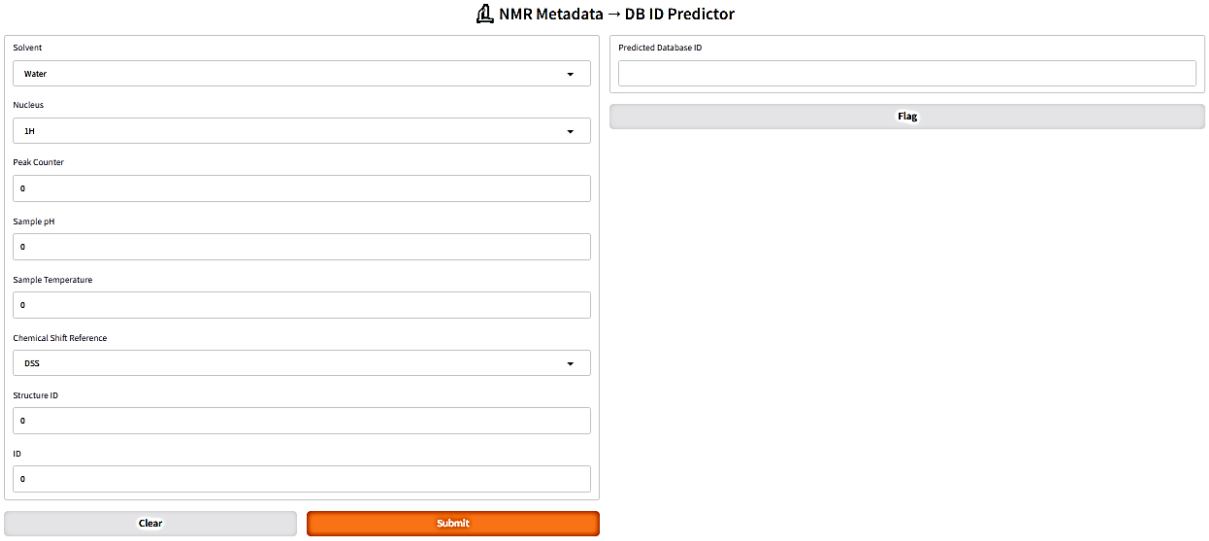

Figure IV.4. Interface graphique pour la prédiction de l'identifiant de base de données à partir des métadonnées RMN.

IV.2. Discussion

Nos travaux s'inscrivent dans un contexte où l'utilisation de méthodes de machine learning, notamment la Random Forest, pour l'analyse de données spectrales NMR montre un réel potentiel, bien que le type de données utilisé diffère.

Par exemple, Kuhn et al. (2008) décrit l'utilisation de la Random Forest pour prédire des déplacements chimiques (1H NMR) à partir de descripteurs numériques et catégoriels, atteignant une MAE entre 0.15 et 0.29 ppm. Leur approche démontre l'efficacité des classifieurs sur des données structurées issues de NMR, renforçant la pertinence de notre choix de modèle.

Nos résultats, avec une précision de 33 %, un F1-score de 39 % et un log loss de 1.48, restent significativement inférieurs aux performances rapportées par Fathi et al. (2014), qui ont atteint une précision de 94 % dans la classification entre patients atteints de la maladie de Crohn et sujets sains à partir de profils métaboliques obtenus par spectroscopie RMN 1H, en utilisant un modèle Random Forest. Cette différence notable s'explique par la nature fondamentalement différente de notre approche : alors que leur modèle exploite les spectres RMN bruts riches en information métabolique, notre méthodologie repose uniquement sur des métadonnées expérimentales (solvant, noyau, nombre de pics, pH, température et référence de décalage chimique), sans accès direct aux intensités spectrales.

D'autre part, Un autre travail pertinent est celui de Wenck et al. (2023), qui appliquent la méthode Surrogate Minimal Depth (SMD) au Random Forest pour investiguer les relations complexes entre variables dans des données RMN 1H de différentes espèces de truffes. Cette étude démontre que SMD permet d'identifier non seulement les variables pertinentes, mais aussi les interactions biologiquement significatives entre métabolites

Kim et al. (2023) ont développé DeepSAT, une architecture d'apprentissage profond capable d'apprendre la structure moléculaire à partir de spectres RMN HSQC bidimensionnels (2D). Leur approche atteint des taux de précision supérieurs à 90 %, démontrant ainsi l'avantage considérable d'utiliser directement les caractéristiques spectrales brutes comme entrées.

Un travail parallèle récent, mené par Li et al. (2024), utilise la spectroscopie NMR combinée à plusieurs algorithmes de machine learning, notamment Random Forest, pour prédire la progression du prédiabète vers le diabète sur un jeu de données UK Biobank (n = 13 489). Leur modèle atteint une AUC (Area Under the Curve) de 0,83 à 1 an, contre 0,759 sans métabolites, démontrant la grande valeur des variations métaboliques détectées par NMR dans des contextes cliniques prédictifs.

Plus récemment, Ivanova et al. (2025) ont démontré l'efficacité des approches basées sur l'apprentissage automatique, y compris la Random Forest, pour prédire des caractéristiques à partir de données 13C-NMR simulées à partir de SMILES, atteignant une accuracy autour de 72 %, bien supérieure à nos résultats. Cependant, leurs entrées étaient dérivées de spectres simulés et enrichies en informations structurales, contrairement à notre approche basée uniquement sur les métadonnées expérimentales.

Dans ce contexte, notre travail propose un pipeline innovant et léger permettant la classification directe des métabolites à partir des seules métadonnées expérimentales. Ce cas d'usage, encore peu exploré dans la littérature, constitue une preuve de concept pertinente pour l'annotation préliminaire rapide dans les situations où les données spectrales sont incomplètes, indisponibles ou coûteuses à acquérir. Par ailleurs, le développement d'une interface utilisateur graphique conviviale rend l'outil immédiatement accessible aux non-programmeurs, favorisant ainsi son intégration dans des flux de travail à haut débit ou semi-automatisés.

Cependant, l'écart observé entre nos résultats et ceux rapportés dans les études antérieures souligne la nécessité critique d'intégrer des descripteurs plus riches en information, tels que les profils spectraux complets, les empreintes moléculaires (molecular fingerprints) ou des descripteurs hybrides. De telles améliorations sont indispensables pour accroître à la fois la précision des prédictions et la capacité de discrimination entre les classes.

Conclusion

Conclusion

Cette étude met en évidence le potentiel de l'apprentissage automatique pour l'annotation automatique des métabolites à partir des métadonnées RMN. Malgré une performance modeste, notamment en raison du déséquilibre des classes et de la complexité des profils moléculaires, le modèle Random Forest a permis d'atteindre une prédiction cohérente dans plusieurs cas. L'intégration d'une interface graphique interactive facilite l'utilisation de l'outil dans un cadre exploratoire ou pédagogique. À l'avenir, l'enrichissement du modèle par des spectres RMN, des représentations moléculaires plus informatives et des techniques avancées de rééquilibrage pourrait significativement améliorer la performance et la robustesse du système.

Perspectives

Pour améliorer les performances du modèle, plusieurs perspectives concrètes peuvent être envisagées : l'augmentation du jeu de données afin de mieux gérer le déséquilibre observé, l'utilisation de techniques avancées telles que les modèles ensemblistes ou l'apprentissage profond, ainsi que l'intégration des données spectrales brutes (décalages chimiques et intensités).

Limitations de l'étude

Cette étude est limitée par la taille réduite du jeu de données et l'utilisation exclusive de métadonnées sans spectres, ce qui peut réduire la précision du modèle. De plus, le déséquilibre entre classes et le faible pouvoir discriminant de certaines variables ont probablement affecté les performances.

Références bibliographiques

Références bibliographiques

Abadie, S., Leduc, C., Lintner, K., & Bedos, P. (2021). The alkaloid centcyamine increases expression of klotho and lamin B1, slowing the onset of skin ageing in vitro and in vivo. *International Journal of Cosmetic Science*, *43*(5), 561-572. https://doi.org/10.1111/ics.12731

Akoka, S. (2022). *INTRODUCTION A LA RESONANCE MAGNETIQUE NUCLEAIRE - AVEC 204 ILLUSTRATIONS EN COULEURS*. ELLIPSES.

Amisha, Malik, P., Pathania, M., & Rathaur, V. K. (2019). Overview of artificial intelligence in medicine. *Journal of Family Medicine and Primary Care*, *8*(7), 2328-2331. https://doi.org/10.4103/jfmpc.jfmpc_440_19

Breitmaier, E., Bauer, G., & Cassels, B. K. (1988). *13C NMR spectroscopy : a working manual with exercises* ([2nd print.] ed.). Harwood Academic Press.

Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., Vogt-Maranto, L., & Zdeborová, L. (2019). Machine learning and the physical sciences. *Reviews of Modern Physics*, *91*(4), 045002. https://doi.org/10.1103/RevModPhys.91.045002

Charrette, H., Jacques, É., Rup-Jacques, S., & Sindt, M. (2023). *Synthèses chimiques: Des cristaux métalliques aux molécules organiques - Rappels de cours et exercices progressifs*. ELLIPSES. https://books.google.dz/books?id=KT7YEAAAQBAJ

Choi, H. J., Song, J. H., Park, K. S., & Kwon, D. H. (2009). Inhibitory effects of quercetin 3-rhamnoside on influenza A virus replication. *Eur J Pharm Sci*, *37*(3-4), 329-333. https://doi.org/10.1016/j.ejps.2009.03.002

Choi, R. Y., Coyner, A. S., Kalpathy-Cramer, J., Chiang, M. F., & Campbell, J. P. (2020). Introduction to Machine Learning, Neural Networks, and Deep Learning. *Transl Vis Sci Technol*, *9*(2), 14. https://doi.org/10.1167/tvst.9.2.14

Chung, K.-T., Yee, W. T., I., W. C.-., Yao-Wen, H., & and Lin, Y. (1998). Tannins and Human Health: A Review. *Critical Reviews in Food Science and Nutrition*, *38*(6), 421-464. https://doi.org/10.1080/10408699891274273

Claridge, T. (1999). *High-resolution NMR Techniques in Organic Chemistry*. Elsevier Science. https://books.google.dz/books?id=GUzwRTg5dfAC

Clayden, J., Greeves, N., & Warren, S. (2013). *Chimie organique*. De Boeck. https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=3327132

Clayden, J., Greeves, N., Warren, S., & Pousse, A. (2013). *Chimie organique : une approche orbitalaire*. De Boeck Supérieur. https://books.google.dz/books?id=BacsDwAAQBAJ

Del Prado-Audelo, M. L., Cortés, H., Caballero-Florán, I. H., González-Torres, M., Escutia-Guadarrama, L., Bernal-Chávez, S. A., Giraldo-Gomez, D. M., Magaña, J. J., & Leyva-Gómez, G. (2021). Therapeutic Applications of Terpenes on Inflammatory Diseases. *Front Pharmacol*, *12*, 704197. https://doi.org/10.3389/fphar.2021.704197

Deng, L., & Yu, D. (2014). Deep Learning: Methods and Applications. *Foundations and Trends® in Signal Processing*, *7*(3–4), 197-387. https://doi.org/10.1561/2000000039

Deo, R. C. (2015). Machine Learning in Medicine. *Circulation*, *132*(20), 1920-1930. https://doi.org/10.1161/CIRCULATIONAHA.115.001593

Dormoy, V., & Massfelder, T. (2013). La métabolomique au service de la médecine. *Med Sci (Paris)*, *29*(5), 463-468. 10.1051/medsci/2013295007

Erickson, B. J., Korfiatis, P., Akkus, Z., & Kline, T. L. (2017). Machine Learning for Medical Imaging. *RadioGraphics*, *37*(2), 505-515. https://doi.org/10.1148/rg.2017160130

Farrar, A. J., & Farrar, F. C. (2020). Clinical Aromatherapy. *Nurs Clin North Am*, *55*(4), 489-504. https://doi.org/10.1016/j.cnur.2020.06.015

Fathi, F., Majari-Kasmaee, L., Mani-Varnosfaderani, A., Kyani, A., Rostami-Nejad, M., Sohrabzadeh, K., Naderi, N., Zali, M. R., Rezaei-Tavirani, M., Tafazzoli, M., & Arefi-Oskouie, A. (2014). 1H NMR based metabolic profiling in Crohn's disease by random forest methodology. *Magnetic Resonance in Chemistry*, *52*(7), 370-376. https://doi.org/10.1002/mrc.4074

Galal, A., Talal, M., & Moustafa, A. (2022). Applications of machine learning in metabolomics: Disease modeling and classification. *Front Genet*, *13*, 1017340. https://doi.org/10.3389/fgene.2022.1017340

Garron, M.-L., & Cygler, M. (2010). Structural and mechanistic classification of uronic acid-containing polysaccharide lyases. *Glycobiology*, *20*(12), 1547-1573. https://doi.org/10.1093/glycob/cwq122

Gonçalves, O. (2018). *Altération des ovoproduits : de la métabolomique au contrôle en ligne*. ISTE Editions.

Greenhill, A. T., & Edmunds, B. R. (2020). A primer of artificial intelligence in medicine. *Techniques and Innovations in Gastrointestinal Endoscopy*, *22*(2), 85-89. https://doi.org/10.1016/j.tgie.2019.150642

Hamet, P., & Tremblay, J. (2017). Artificial intelligence in medicine. *Metabolism - Clinical and Experimental*, *69*, S36-S40. https://doi.org/10.1016/j.metabol.2017.01.011

Hanson, L. G. (2008). Is quantum mechanics necessary for understanding magnetic resonance? *Concepts in Magnetic Resonance Part A*, *32A*(5), 329-340. https://doi.org/https://doi.org/10.1002/cmr.a.20123

Harborne, J. B., & Williams, C. A. (2000). Advances in flavonoid research since 1992. *Phytochemistry*, *55*(6), 481-504. https://doi.org/10.1016/S0031-9422(00)00235-1

Higuera-Barraza, O. A., Del Toro-Sanchez, C. L., Ruiz-Cruz, S., & Márquez-Ríos, E. (2016). Effects of high-energy ultrasound on the functional properties of proteins. *Ultrason Sonochem*, *31*, 558-562. https://doi.org/10.1016/j.ultsonch.2016.02.007

Hussain, G., Rasul, A., Anwar, H., Aziz, N., Razzaq, A., Wei, W., Ali, M., Li, J., & Li, X. (2018). Role of Plant Derived Alkaloids and Their Mechanism in Neurodegenerative Disorders. *Int J Biol Sci*, *14*(3), 341-357. https://doi.org/10.7150/ijbs.23247

Ibrahima, S. (2025). *Physique Nucléaire 3 Radiopharmaceutiques Utilisés en Médecine Nucléaire*. ISTE Editions Ltd. https://public.ebookcentral.proquest.com/choice/PublicFullRecord.aspx?p=31967787

Ivanova, M. L., Russo, N., & Nikolic, K. (2025). Leveraging 13C NMR spectroscopic data derived from SMILES to predict the functionality of small biomolecules by machine learning: a case study on human Dopamine D1 receptor antagonists. https://doi.org/10.48550/arXiv.2501.14044

Jacobsen, N. E. (2016). *NMR Data Interpretation Explained: Understanding 1D and 2D NMR Spectra of Organic Compounds and Natural Products*. Wiley. https://books.google.dz/books?id=OxQWCgAAQBAJ

James, K. D. (2017). Chapter 19 - Animal Metabolites: From Amphibians, Reptiles, Aves/Birds, and Invertebrates. In S. Badal & R. Delgoda (Eds.), *Pharmacognosy* (pp. 401-411). Academic Press. https://doi.org/10.1016/B978-0-12-802104-0.00019-6

Johnson, H., & Tipirneni-Sajja, A. (2024). Explainable AI to Facilitate Understanding of Neural Network-Based Metabolite Profiling Using NMR Spectroscopy. *Metabolites*, *14*(6), 332. https://doi.org/doi:10.3390/metabo14060332

Keeler, J. (2015). *Comprendre la RMN*. PPUR Presses Polytechniques. https://books.google.dz/books?id=e2YGDQAAQBAJ

Kiemle, D. J., Silverstein, R. M., Webster, F. X., & Lafond, V. (2016). *Identification spectrométrique de composés organiques - 3ème édition*. De Boeck supérieur. https://books.google.dz/books?id=amPHDgAAQBAJ

Kim, H. W., Zhang, C., Reher, R., Wang, M., Alexander, K. L., Nothias, L.-F., Han, Y. K., Shin, H., Lee, K. Y., Lee, K. H., Kim, M. J., Dorrestein, P. C., Gerwick, W. H., & Cottrell, G. W. (2023). DeepSAT: Learning Molecular Structures from Nuclear Magnetic Resonance Data. *Journal of Cheminformatics*, *15*(1), 71. https://doi.org/10.1186/s13321-023-00738-4

Kleszcz, R., Majchrzak-Celińska, A., & Baer-Dubowska, W. (2025). Tannins in cancer prevention and therapy. *British Journal of Pharmacology*, *182*(10), 2075-2093. https://doi.org/10.1111/bph.16224

Kuhn, S., Egert, B., Neumann, S., & Steinbeck, C. (2008). Building blocks for automated elucidation of metabolites: Machine learning methods for NMR prediction. *BMC Bioinformatics*, *9*(1), 400. https://doi.org/10.1186/1471-2105-9-400

Kumaraswamy, R. V., Kumari, S., Choudhary, R. C., Pal, A., Raliya, R., Biswas, P., & Saharan, V. (2018). Engineered chitosan based nanomaterials: Bioactivities, mechanisms and perspectives in plant protection and growth. *Int J Biol Macromol*, *113*, 494-506. https://doi.org/10.1016/j.ijbiomac.2018.02.130

Lambert, J. B. M. E. (2018). *NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY : an introduction to principles, applications and ... experimental methods*. JOHN WILEY.

Lee, B. L., Rout, M., Mandal, R., & Wishart, D. S. (2023). Automated identification and quantification of metabolites in human fecal extracts by nuclear magnetic resonance spectroscopy. *Magnetic Resonance in Chemistry*, *61*(12), 705-717. https://doi.org/10.1002/mrc.5372

Li-Beisson, Y., Nakamura, Y., & Harwood, J. (2016). Lipids: From Chemical Structures, Biosynthesis, and Analyses to Industrial Applications. In Y. Nakamura & Y. Li-Beisson (Eds.), *Lipids in Plant and Algae Development* (pp. 1-18). Springer International Publishing. https://doi.org/10.1007/978-3-319-25979-6_1

Li, J., Yu, Y., Sun, Y., Fu, Y., Shen, W., Cai, L., Tan, X., Wang, N., Lu, Y., & Wang, B. (2024). Nuclear magnetic resonance-based metabolomics with machine learning for predicting progression from prediabetes to diabetes. *medRxiv*, 2024.2005.2014.24307378. https://doi.org/10.1101/2024.05.14.24307378

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, *9*(1), 381-386.

Martins, J. T., Bourbon, A. I., Pinheiro, A. C., Fasolin, L. H., & Vicente, A. A. (2018). Protein-Based Structures for Food Applications: From Macro to Nanoscale [Review]. *Frontiers in Sustainable Food Systems*, *Volume 2 - 2018*. https://doi.org/10.3389/fsufs.2018.00077

Patel, M. K., Kumar, M., Li, W., Luo, Y., Burritt, D. J., Alkan, N., & Tran, L.-S. P. (2020). Enhancing Salt Tolerance of Plants: From Metabolic Reprogramming to Exogenous Chemical Treatments and Molecular Approaches. *Cells*, *9*(11), 2492. https://doi.org/10.3390/cells9112492

Rebstein, M., & Soerensen, C. (2011). *Chimie avancée: préparation au bac et à la maturité*. Presses polytechniques et universitaires romandes. https://books.google.dz/books?id=nlXn72OKgTUC

Ruffle, J. K., Farmer, A. D., & Aziz, Q. (2019). Artificial Intelligence-Assisted Gastroenterology— Promises and Pitfalls. *Official journal of the American College of Gastroenterology | ACG*, *114*(3), 422-428. https://doi.org/10.1038/s41395-018-0268-4

Sakho, I. (2025). *Physique nucléaire 3: Radiopharmaceutiques utilisés en médecine nucléaire*. ISTE Editions Limited. https://books.google.dz/books?id=vTJPEQAAQBAJ

Sarker, I. H. (2021). Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Computer Science*, *2*(6), 420. https://doi.org/10.1007/s42979-021-00815-1

Schott, M.-A., Valentin, J., & Magadur, G. (2014). *Chimie : PCSI, option PC et PSI, MPSI, 1ère année : un accompagnement au quotidien : tout-en-un : cours, exercices corrigés*. De Boeck.

Schott, M. A., Valentin, J., Magadur, G., Clède, S., Lefevre, A. L., & Altmayer-Henzien, A. (2015). *Chimie PCSI/MPSI - 1re année: Tout-en-un*. De Boeck Supérieur. https://books.google.dz/books?id=C2MEDgAAQBAJ

Setyorini, D., & Antarlina, S. S. (2022). Secondary metabolites in sorghum and its characteristics. *Food Science and Technology*, *42*, e49822. https://doi.org/10.1590/fst.49822

Silverstein, R. M., Webster, F. X., Kiemle, D. J., Bryce, D. L., & Lafond, V. (2016). *Identification spectrométrique de composés organiques* (3e édition ed.). De Boeck supérieur. https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=3109278

Smeekens, S. (2000). SUGAR-INDUCED SIGNAL TRANSDUCTION IN PLANTS. *Annu Rev Plant Physiol Plant Mol Biol*, *51*, 49-81. https://doi.org/10.1146/annurev.arplant.51.1.49

Stoclet, J. C., & Schini-Kerth, V. (2011). Flavonoïdes alimentaires et santé humaine. *Annales Pharmaceutiques Françaises*, *69*(2), 78-90. https://doi.org/10.1016/j.pharma.2010.11.004

Strong, A. (2016). Applications of artificial intelligence & associated technologies. *Science [ETEBMS-2016]*, *5*(6), 64-67.

Swer, N. M., Venkidesh, B. S., Murali, T. S., & Mumbrekar, K. D. (2023). Gut microbiota-derived metabolites and their importance in neurological disorders. *Molecular Biology Reports*, *50*(2), 1663-1675. https://doi.org/10.1007/s11033-022-08038-0

Taye, M. M. (2023). Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions. *Computers*, *12*(5), 91. https://doi.org/10.3390/computers12050091

Wenck, S., Mix, T., Fischer, M., Hackl, T., & Seifert, S. (2023). Opening the Random Forest Black Box of 1H NMR Metabolomics Data by the Exploitation of Surrogate Variables. *Metabolites*, *13*(10), 1075. https://www.mdpi.com/2218-1989/13/10/1075

Xia, J., Bjorndahl, T. C., Tang, P., & Wishart, D. S. (2008). MetaboMiner – semi-automated identification of metabolites from 2D NMR spectra of complex biofluids. *BMC Bioinformatics*, *9*(1), 507. https://doi.org/10.1186/1471-2105-9-507