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 : SCINCES DE LA NATURE ET DE LA VIE

FILIERE : SCIENCE BIOLOGIQUE

OPTION : BIOCHIMIE APPLIQUÉE

Mémoire présenté pour l'obtention

Du diplôme de Master Académique

Par : Benali Hind Safa

Intitulé

Essai de développement d'un modèle d'IA pour la prédiction de l'identité de métabolites inconnus par analyse de données analytiques

Soutenu devant le jury composé de :

Mme. Samia Bouaziz	Université Mohamed Boudiaf M'sila	Président
Mr. Abdenassar Harrar	Université Mohamed Boudiaf M'sila	Rapporteur
Mr. Seifeddine Drif	Université Mohamed Boudiaf M'sila	Examinateur

Année universitaire : 2023 /2024

Dédicace

Je dédie ce mémoire à la mémoire précieuse de mon frère Toufik, qui nous a quittés bien trop tôt, en hommage à sa vie et à l'empreinte indélébile qu'il a laissée dans nos cœurs et nos vies, je t'aime mon frère.

À ***mes parents***, pour leur soutien indéfectible, leur amour inconditionnel et leurs encouragements constants tout au long de ce parcours académique.

À mon bras droit ***Moustapha,*** qui m'a soutenu tout au long de cette année et inchalah pour toute la vie.

À mes frères ***Mohamed*** et ***Khaled***, ainsi qu'à ma chère sœur ***Imene*** et son mari ***Malik***, je suis reconnaissant pour leur soutien et leurs encouragements constants.

À mes professeurs, pour leur expertise, leurs conseils éclairés et leur patience précieuse qui ont enrichi mes connaissances et façonné ma réflexion.

À mes amis, pour leur soutien moral, leurs encouragements chaleureux et leur présence réconfortante qui ont rendu ce chemin moins ardu et plus enrichissant.

Je remercie chaleureusement ***HAMDAOUI Amine***, le designer talentueux, pour avoir réalisé plusieurs designs inspirants pour ma présentation. Son expertise créative et son sens esthétique ont grandement enrichi la qualité visuelle de mon travail.

Ainsi, à ma collègue ***Anfal***, je tiens à te remercier sincèrement pour avoir partagé avec moi l'essentiel de tes connaissances en intelligence artificielle. Ton expertise et ta clarté ont grandement enrichi ma compréhension du sujet.

Et aussi Monsieur ***MEBARKIA ayoub***, Je tiens à vous exprimer ma profonde gratitude pour nous avoir fourni l'arme indispensable à la réalisation de ce travail dans un délai très court. Votre soutien a été déterminant pour le succès de notre projet.

À tous ceux qui ont croisé ma route durant cette aventure, je vous adresse ma profonde gratitude pour votre contribution à la réalisation de ce mémoire.

Ce travail est dédié à toutes les personnes qui ont été des piliers indispensables dans mon parcours académique et personnel.

Remerciement

On remercie dieu le tout puissant de nous avoir donné la santé et la volonté d'entamer et de terminer ce mémoire.

Tout d'abord, je tiens à exprimer ma plus profonde gratitude à mon directeur de thèse, ***Dr. HARRAR abdenassar*** , pour son encadrement exceptionnel, ses conseils avisés et son soutien indéfectible tout au long de cette recherche. Son expertise, sa patience et ses encouragements constants ont été essentiels pour mener à bien ce travail. Sa vision scientifique et sa rigueur m'ont guidé à chaque étape de ce projet, et je suis profondément reconnaissant pour les connaissances et les compétences que j'ai pu acquérir grâce à lui.

Je remercie également les membres de mon jury, ***Dr. BOUAZIZ samia***, ***Dr. DRIF siefeddine*** et ***Dr. AHNIA hadjira***, pour leur temps précieux, leurs critiques constructives et leurs suggestions enrichissantes. Leurs remarques pertinentes ont permis d'affiner et de perfectionner cette thèse, et leur soutien a été une source de motivation supplémentaire. Leur engagement à évaluer ce travail avec rigueur et bienveillance est grandement apprécié.

Je tiens à exprimer ma sincère gratitude au ***Dr. KADRI Said***, dont l'expertise et les conseils avisés ont été le point de départ crucial de ce travail. Son soutien inestimable et sa vision éclairée ont joué un rôle fondamental dans la réalisation de cette étude.

Je souhaite exprimer ma sincère reconnaissance envers le personnel administratif du département de microbiologie et biochimie, en particulier notre chef de département ***Dr. RAHALI abdallah***, pour leur soutien indéfectible et leur professionnalisme exemplaire. Leur dévouement constant et leur aide précieuse ont été des piliers essentiels au bon déroulement de ce travail.

La gratitude envers les autres est la clé de la gratitude envers Dieu. Je vous suis profondément reconnaissante d'avoir rendu possible la réalisation de cet objectif. Tous mes remerciements vous sont adressés.

Sommaire

ملخص

تطوير نماذج الذكاء الاصطناعي التي يمكنها التنبؤ بهوية الجزيئات الحيوية غير المعروفة في العينات البيولوجية المعقدة من خلال تحليل البيانات يمكن أن يُحدث ثورة في أبحاث metabolomics. يُعد التعرف الدقيق على الجزيئات الحيوية أمرًا أساسيًا لفهم العمليات البيولوجية وآليات الأمراض وتأثيرات الأدوية. غالبًا ما تواجه الطرق التقليدية صعوبات في التعامل مع تعقيد وحجم البيانات، مما يجعل التعرف الفعال والدقيق على المستقلبات مهمة صعبة. تهدف هذه الدراسة إلى تطوير ونشر نموذج قوي قائم على الذكاء الاصطناعي للتعرف على المستقلبات، متغلبين على القيود التي تواجهها الطرق التقليدية وتوفير أداة سهلة الاستخدام للتطبيقات العملية. قمنا بتدريب نموذج شبكات عصبية متعددة المخرجات باستخدامTensorFlow ، متبعين عملية شاملة تتضمن تحميل مجموعة بيانات معالجة، ترميز التسميات لأعمدة الاسم والصيغة الجزيئية، معالجة وتوسيع الاطياف لضمان طول موحد، دمج ميزات متنوعة مثل الوزن الجزيئي والكتلة الدقيقة وعدد الاطياف، توسيع الميزات باستخدامStandardScaler ، تقسيم البيانات إلى مجموعات تدريب (80%) واختبار (20%)، وتعريف وتدريب نموذج الشبكة العصبية. لتطبيق النموذج، قمنا بتطوير واجهة مستخدم رسومية (GUI) باستخدام .Tkinter تم تدريب نموذج الشبكة العصبية بنجاح دون حدوث زيادة في الحمل على النظام؛ ومع ذلك، حقق دقة منخفضة جدًا، وكانت جميع التنبؤات غير صحيحة. نشتبه في أن السبب هو طريقة معالجة الاطياف وتحويلها إلى صيغة رقمية. تتمثل القيود في هذه الدراسة في الوقت الكبير المطلوب لتحديد الخوارزمية الصحيحة لمعالجة وتحويل الاطياف بدقة، تدريب النموذج بفعالية، ودمج النتائج في واجهة المستخدم الرسومية. ينبغي على الأبحاث المستقبلية التركيز على تحسين معلمات النموذج، وتحسين تقنيات زيادة البيانات، ودمج أساليب تصور متقدمة لتحسين أداء النموذج وتجربة المستخدم.

الكلمات المفتاحية: الذكاء الاصطناعي، التعرف على المستقلبات، الشبكات العصبية، التحليل الطيفي الكتلي، معالجة البيانات، تحليل المستقلبات.

Abstract

Developing AI models that can predict the identity of unknown metabolites in complex biological samples through analytical data analysis could revolutionize metabolomics research. Accurate identification of metabolites is essential for understanding biological processes, disease mechanisms, and the effects of drugs. Traditional methods often struggle with the complexity and volume of data, making efficient and accurate metabolite identification a challenging task. This study aims to develop and deploy a robust AI-driven model for metabolite identification, overcoming the limitations of traditional methods and providing an accessible tool for practical applications. We trained a multi-output neural network model using TensorFlow, following a comprehensive process involving loading a processed dataset, encoding labels for Name and Formula columns, preprocessing and padding peaks to ensure uniform length, combining various features such as MW, ExactMass, and Num Peaks, scaling features using StandardScaler, splitting data into training (80%) and test (20%) sets, and defining and training the neural network model. For deployment, we developed a graphical user interface (GUI) using Tkinter. The neural network model successfully trained without system overload; however, it achieved very low accuracy, with all predictions being incorrect. We suspect the cause to be the method of preprocessing and converting the peaks to numerical format. The limitation of this study is the substantial amount of time required to identify the correct algorithm to preprocess and convert peaks accurately, train the model effectively, and integrate the findings into the GUI. Future research should focus on optimizing hyperparameters, improving data augmentation techniques, and integrating advanced visualization methods to enhance model performance and user experience.

Keywords: AI, Metabolite Identification, Neural Networks, Mass Spectrometry, Data Preprocessing, Metabolomics Analysis.

Résumé

Développer des modèles d'IA capables de prédire l'identité de métabolites inconnus dans des échantillons biologiques complexes grâce à l'analyse de données analytiques pourrait révolutionner la recherche en métabolomique. L'identification précise des métabolites est essentielle pour comprendre les processus biologiques, les mécanismes des maladies et les effets des médicaments. Les méthodes traditionnelles rencontrent souvent des difficultés en raison de la complexité et du volume des données, rendant l'identification des métabolites à la fois efficace et précise une tâche ardue. Cette étude vise à développer et déployer un modèle robuste d'identification de métabolites basé sur l'IA, surmontant les limitations des méthodes traditionnelles et fournissant un outil accessible pour des applications pratiques. Nous avons entraîné un modèle de réseau de neurones multi-sorties en utilisant TensorFlow, suivant un processus complet incluant le chargement d'un ensemble de données traité, l'encodage des étiquettes des colonnes Name et Formula, le prétraitement et le remplissage des pics pour assurer une longueur uniforme, la combinaison de diverses caractéristiques telles que MW, ExactMass et Num Peaks, la mise à l'échelle des caractéristiques à l'aide de StandardScaler, la division des données en ensembles d'entraînement (80%) et de test (20%), et la définition et l'entraînement du modèle de réseau de neurones. Pour le déploiement, nous avons développé une interface graphique utilisateur (GUI) utilisant Tkinter. Le modèle de réseau de neurones a été entraîné avec succès sans surcharge du système; cependant, il a obtenu une très faible précision, avec toutes les prédictions incorrectes. Nous soupçonnons que la cause en soit la méthode de prétraitement et de conversion des pics en format numérique. La limitation de cette étude réside dans le temps considérable nécessaire pour identifier l'algorithme correct pour prétraiter et convertir les pics avec précision, entraîner le modèle efficacement et intégrer les résultats dans la GUI. Les recherches futures devraient se concentrer sur l'optimisation des hyperparamètres, l'amélioration des techniques d'augmentation des données et l'intégration de méthodes de visualisation avancées pour améliorer les performances du modèle et l'expérience utilisateur.

Mots-clés : IA, Identification des Métabolites, Réseaux de Neurones, Spectrométrie de Masse, Prétraitement des Données, Analyse Métabolomique.

Liste des abréviations

CNN : Réseaux de Neurones convolutif.

DL : apprentissage profond.

DNN : Réseaux de Neurones profond.

FIR : infrarouge lointain.

IA : Intelligence Artificielle.

LSTM : Mémoire à long terme et à court terme.

MIR : infrarouge moyen.

ML : apprentissage automatique.

NIR : infrarouge proche.

RNA : Réseaux de Neurones Artificiels.

RNN : Réseaux de Neurones récurent.

Liste des figures

Liste des tableaux

Introduction

Introduction

Dans le domaine de la chimie des produits naturels et des études de métabolomique, l'identification sans ambiguïté des composés chimiques revêt une grande importance et d'un intérêt exceptionnel, par exemple pour le diagnostic et la prédiction des maladies ainsi que pour l'authentification des aliments (Medina et al., 2019). L'identification des petites molécules, est principalement réalisée par spectroscopie de résonance magnétique nucléaire (RMN) et/ou spectrométrie de masse (MS), toutes deux étant également les principales méthodes analytiques en métabolomique (Markley et al., 2017). Les méthodes MS ou les méthodes hybrides telles que la MS couplée à la chromatographie en phase gazeuse (GC) ou à la chromatographie liquide (LC) présentent des avantages évidents en termes de haute sensibilité ainsi que la possibilité de calculer la formule chimique de la molécule d'intérêt dans le cas des spectromètres de masse à haute résolution (Gathungu et al., 2020).

La croissance de l'échelle et de la complexité des données biologiques a stimulé l'adoption croissante de l'apprentissage automatique et de ses dérivés, tels que l'apprentissage profond, en biologie, afin de développer des modèles informatifs et prédictifs des processus biologiques sous-jacents (Greener et al., 2022). L'IA a été fondamentale dans la gestion des données hétérogènes et dans l'analyse avancée des spectres complexes dans des domaines tels que la spectroscopie et la chromatographie (Cardoso Rial, 2024).

L'une des difficultés majeures rencontrées dans le domaine de l'identification des métabolites est l'accès à des données issues de techniques analytiques complexes, ce qui rend l'identification ardue en raison du temps requis et des erreurs potentielles introduites par l'intervention humaine. Par exemple, les techniques telles que la MS et la RMN fournissent des données riches mais complexes à interpréter.

Dans ce travail, nous proposons l'introduction de l'IA comme une solution prometteuse pour automatiser une partie du processus d'identification des métabolites. Cette approche permet une analyse rapide et précise des vastes ensembles de données générées par des techniques analytiques complexes telles que la MS et la RMN. L'IA est capable d'identifier des motifs et des relations complexes qui pourraient échapper à une analyse humaine traditionnelle. En intégrant des algorithmes d'apprentissage automatique et d'apprentissage profond, nous visons à améliorer la précision et l'efficacité de l'identification des métabolites, tout en réduisant la dépendance à l'égard d'une expertise humaine exclusive. Ce processus pourrait accélérer significativement la découverte et la validation des composés bioactifs dans le cadre de la recherche en métabolomique.

Partie bibliographique

Chapitre I : Métabolites et techniques analytiques (spectroscopiques, chromatographiques)

Chapitre I. Métabolites et techniques analytiques

I.1. Définition des métabolites

Les métabolites peuvent être utilisés comme biomarqueurs et sont de petites molécules dérivées du processus métabolique trouvé dans différents milieux biologiques, tels que des échantillons de tissus, des cellules ou des bio fluides. Ils peuvent être identifiés en utilisant différentes stratégies, des expériences ciblées ou non ciblées, et par différentes techniques, telles que la chromatographie liquide haute performance, la spectrométrie de masse ou la résonance magnétique nucléaire (Donatti et al., 2020).

I.2. Déférents classes de métabolites

I.2.1. Métabolites primaires :

Les métabolites primaires sont essentiels pour la croissance et le développement des plantes car ils sont impliqués dans divers processus physiologiques et biochimiques (Kumar et al., 2017). Les métabolites primaires comprennent différentes classes de métabolites tels que les sucres, les acides gras et les acides aminés, jouant des rôles vitaux tels que les osmolytes et les osmoprotecteurs dans les plantes soumises à des stress biotiques et abiotiques (Patel et al., 2020).

I.2.2. Métabolites secondaires :

Les métabolites secondaires jouent un rôle crucial dans la protection des plantes contre divers stress environnementaux. On estime qu'environ 100 000 métabolites secondaires ont été rapportés dans différentes espèces végétales et sont classés en plusieurs groupes, notamment les composés azotés, les terpènes, les thiols et les composés phénoliques (Zagorchev et al., 2013).

I.3. Techniques analytiques

I.3.1. Techniques Spectroscopiques

Il existe une grande tendance à l'application de multiples techniques instrumentales pour la caractérisation complète des denrées alimentaires ou des produits naturels associés. Les techniques spectrométriques offrent généralement un aperçu complet et rapide de la composition et des propriétés des produits en déterminant des biomolécules spécifiques telles que les sucres, les minéraux, les polyphénols, les composés volatils, les acides aminés, les acides organiques, etc. Ce numéro spécial vise avant tout à renforcer les avancées de l'application des techniques spectrométriques telles que la chromatographie en phase gazeuse couplée à la spectrométrie de masse (GC-MS), la spectrométrie de masse à ratio isotopique (IRMS), la résonance magnétique nucléaire (RMN) (Karabagias, 2020).

I.3.1.1. Spectroscopie de masse (MS)

La spectroscopie de masse détermine les fragments ou ions caractéristiques qui résultent de la décomposition des molécules organiques. Le principe de base implique le bombardement d'un composé organique avec un faisceau d'électrons pour générer des ions chargés positivement. Ensuite, l'ion moléculaire est fragmenté en utilisant l'énergie des électrons pour rompre les liaisons et produire des espèces chargées positivement ou des ions fragmentaires. Les ions positifs ou ions fragmentaires formés sont ensuite accélérés et déviés dans un chemin circulaire à l'aide d'un champ magnétique, puis focalisés sur le détecteur ou la plaque photographique en fonction de leur masse et de leur charge. Chaque ion représente une ligne distincte sur la plaque et est enregistré sous forme d'intensité de pic. La déviation des ions est basée sur la charge, la masse et la vitesse, la séparation des ions est basée sur le rapport masse/charge (m/z) et la détection est proportionnelle à l'abondance des ions (Strathmann & Hoofnagle, 2011).

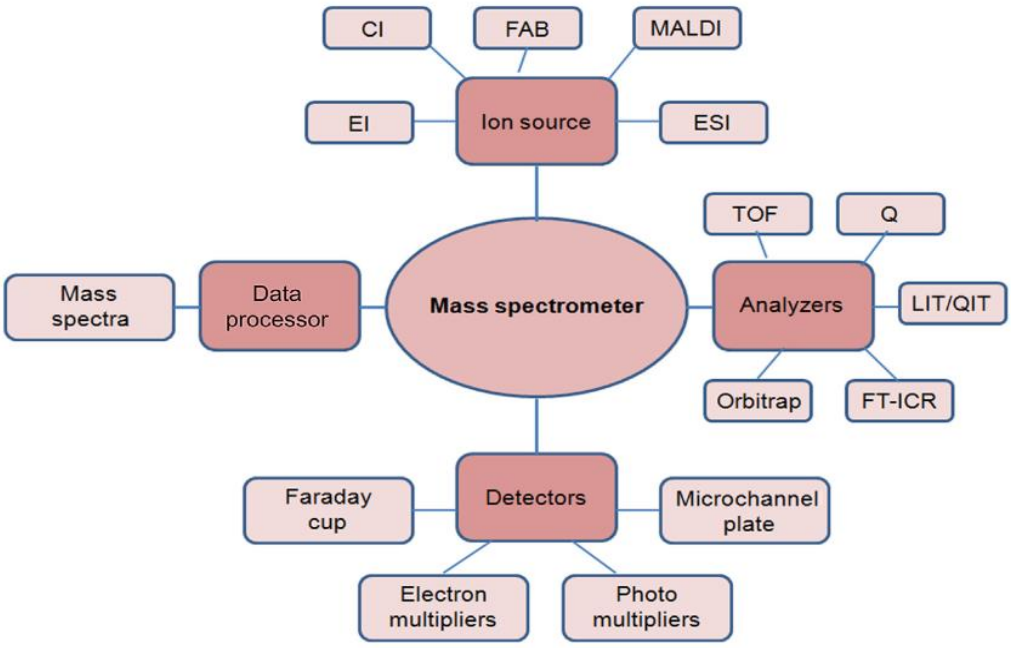

Figure I.1. Composants d'un spectromètre de masse (Rajawat & Jhingan, 2019)**.**

I.3.1.2. Résonance magnétiques nucléaires (RMN)

La RMN repose sur l'interaction des moments magnétiques nucléaires avec les champs magnétiques. Les noyaux avec un nombre impair de protons ou de neutrons, tels que ^{1}H, ^{31}P, ^{13}C et ^{15}N, ont un spin de ½, rendant leur étude pratique. Dans un champ magnétique, ces noyaux peuvent s'orienter parallèlement ou antiparallèlement au champ, avec des énergies différentes. À l'équilibre, les moments magnétiques nucléaires tendent à s'aligner avec le champ, créant une magnétisation macroscopique (Hanson, 2008).

I.3.1.3. Spectroscopie infrarouge (IR)

La spectroscopie infrarouge (IR) est une technique analytique largement appliquée pour les études de structure moléculaire. La lumière IR peut être divisée en trois sous-régions : l'IR lointain (FIR), l'IR moyen (MIR) et l'IR proche (NIR). Parmi ces sous-régions, la région MIR du spectre électromagnétique se situe dans des longueurs d'onde entre 2500 nm et 25000 nm (nombre d'onde entre 4000 cm^{-1} et 400 cm^{-1}). Elle est typiquement utilisée pour la détermination de structure moléculaire et la confirmation de composés organiques. Cela est dû au fait que les fréquences des radiations dans cette région correspondent en énergie aux fréquences de vibration des liaisons dans les molécules organiques. Un spectre IR est généré en mesurant l'absorption de la radiation IR par un échantillon en fonction des fréquences de radiation, ce qui est réalisé par un spectromètre IR. La bande d'absorption d'une liaison spécifique est appelée la bande caractéristique de cette liaison (Thompson, 2018).

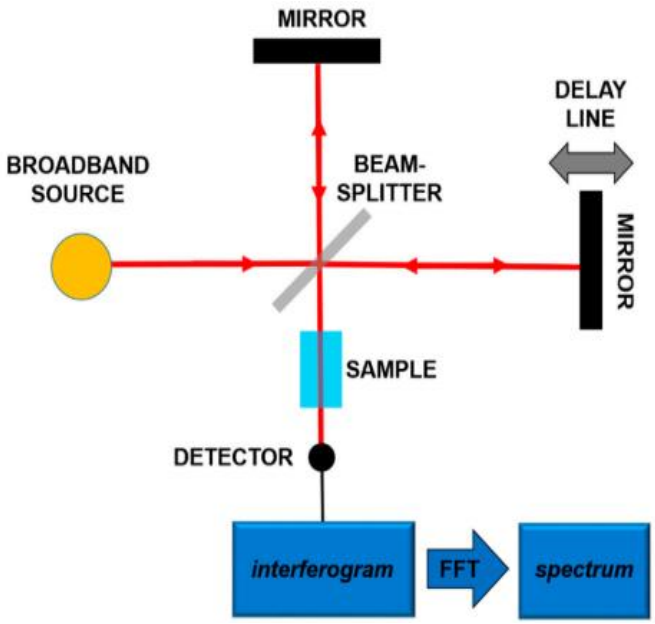

Figure I.2. Principe de la spectroscopie infrarouge à transformée de Fourier (FT-IR) (Patrizi et al., 2019).

I.3.2. Chromatographiques

La chromatographie représente une technique de séparation des constituants présents dans une diversité de mélanges, jouant un rôle essentiel dans l'analyse pour l'identification et la quantification de composés au sein d'échantillons variés. Son principe fondamental réside dans les équilibres de concentration qui se manifestent lorsqu'un composé est mis en contact avec deux phases non miscibles. Dans ce processus, une phase, appelée stationnaire, est immobilisée dans une colonne ou fixée sur un support, tandis que l'autre phase, appelée mobile, se déplace en contact avec la première. La présence de plusieurs composés entraîne des vitesses de déplacement différentes, ce qui induit leur séparation. Ce procédé hydrodynamique a engendré une méthode analytique instrumentale dotée d'un vaste domaine d'application (Rouessac et *al*., 2004).

I.3.2.1. Chromatographie liquide (LC)

La chromatographie liquide est une technique de séparation dans laquelle la phase mobile est un liquide et l'élution peut être réalisée dans une colonne remplie de phase stationnaire, dans un capillaire ou sur une surface plane. Lorsque l'échantillon est introduit, les composants se répartissent entre la phase mobile et la phase stationnaire ; l'introduction de phase mobile supplémentaire transporte les molécules solutés dans une série continue de transferts entre les deux phases. La séparation peut se produire en fonction d'un ou plusieurs des interactions suivantes : adsorption/partition, échange ionique et exclusion de taille. De plus, si la phase stationnaire est fonctionnalisée avec des espèces ou des groupes chimiques spécifiques (par exemple, des sites actifs chiraux), une interaction d'affinité peut se produire. Les types d'élution peuvent être isocratiques, lorsque la composition de la phase mobile ne change pas pendant la séparation, ou en gradient, lorsque la composition de la phase mobile change pendant le temps (Degano, 2019).

I.3.2.2. Chromatographie en phase gazeuse (CPG)

La chromatographie en phase gazeuse est une technique analytique largement utilisée pour séparer et analyser les composés gazeux et volatils. En chromatographie en phase gazeuse sur phase solide, un adsorbant solide est utilisé comme phase stationnaire et la séparation se fait par un processus d'adsorption, tandis qu'en chromatographie en phase gazeuse sur phase liquide, la phase stationnaire est constituée d'une fine couche de liquide non volatil liée à un support solide et la séparation se fait par le processus de partition. La chromatographie en phase gazeuse sur phase liquide est la technique la plus couramment utilisée. L'échantillon à séparer est d'abord converti en vapeurs et ainsi mélangé avec la phase mobile gazeuse. Les composants d'un échantillon qui sont plus solubles dans la phase stationnaire se déplacent plus lentement, tandis que les composants moins solubles dans la phase stationnaire se déplacent plus rapidement. Les composants sont ainsi séparés en fonction de leur coefficient de partition (Kaur & Sharma, 2018).

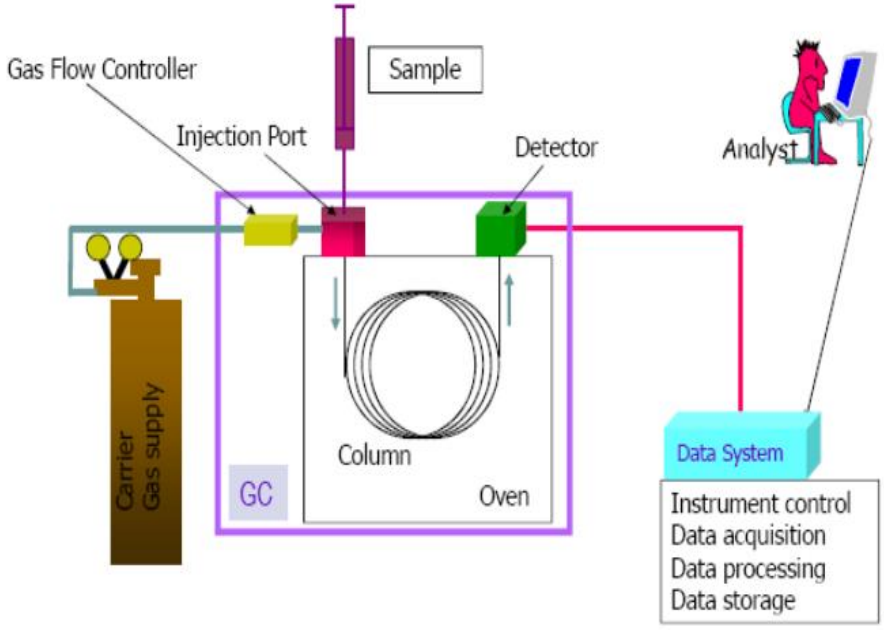

Figure I.3. Diagramme schématique de la chromatographie en phase gazeuse (Al-Bukhaiti & Al-Farga, 2017).

I.3.2.3. Chromatographie en phase liquide couplé spectroscopie de masse (LCMS)

Un système LC-MS typique est une combinaison de HPLC avec MS utilisant une interface (source d'ionisation). L'échantillon est séparé par LC, et les espèces d'échantillon séparées sont pulvérisées dans une source d'ionisation à pression atmosphérique, où elles sont converties en ions en phase gazeuse. L'analyseur de masse est ensuite utilisé pour trier les ions en fonction de leur rapport masse/charge, et le détecteur compte les ions émergeant de l'analyseur de masse et peut également amplifier le signal généré par chaque ion. En conséquence, un spectre de masse (un graphique du signal ionique en fonction du rapport masse/charge) est créé, qui est utilisé pour déterminer la nature élémentaire ou isotopique d'un échantillon, les masses des particules et des molécules, et pour élucider les structures chimiques des molécules (Korfmacher, 2005).

I.3.2.4. Chromatographie en phase gazeuse couplé spectroscopie de masse (GCMS)

En GC-MS, la partie GC implique l'injection d'échantillons dans une colonne de séparation, où ils sont transportés par un gaz porteur inerte (généralement de l'hélium ou de l'hydrogène). Le mélange gazeux est séparé en ses composants en fonction des interactions relatives entre les analytes et le revêtement intérieur de la colonne. Les molécules avec des interactions de colonne (par exemple, des points d'ébullition ou un poids moléculaire plus bas) sortent de la colonne avant les composés à interactions plus élevées. La MS est ensuite utilisée pour détecter les composés sortants. Une source d'ions est utilisée pour produire des ions d'analyte. L'ionisation par impact électronique est une technique couramment utilisée, qui consiste à bombarder les analytes avec un

faisceau d'électrons. Cela est suivi par un analyseur de masse, qui trie les ions fragmentés par leur rapport masse/charge (Beale et al., 2016). Enfin, les ions atteignent un détecteur (par exemple, un multiplicateur d'électrons), qui mesure les fragments de masse ionisés sous forme de signal électrique (Xu et al., 2016).

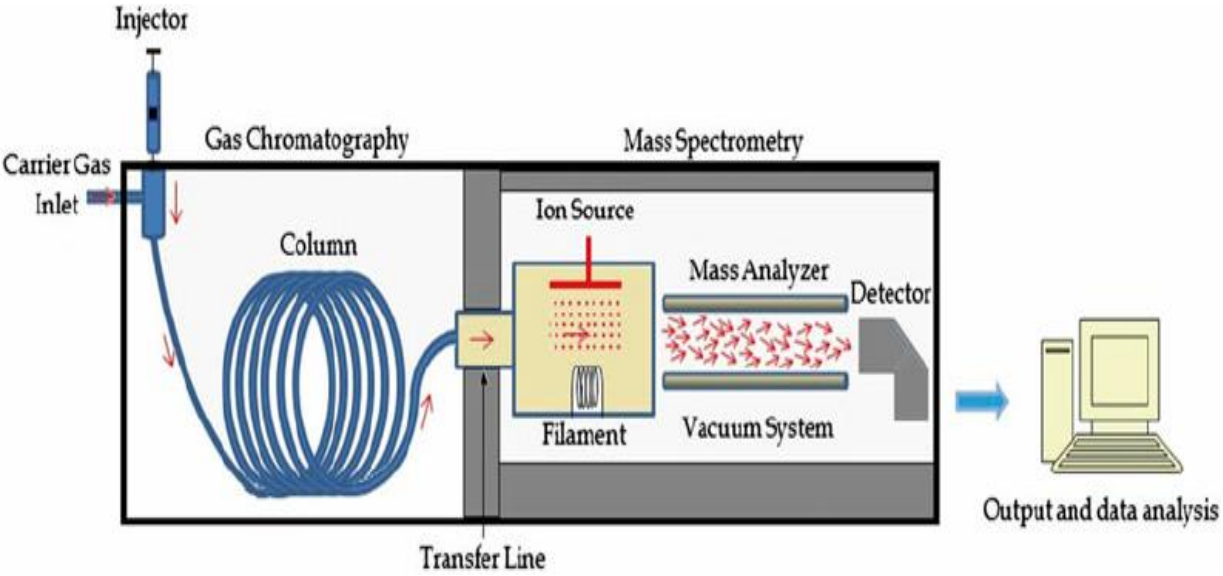

Figure I.4. Schéma des instruments de la technique chromatographie phase gazeuse couplée avec spectroscopie de masse (Emwas et al., 2015).

Chapitre II : Intelligence artificielle

Chapitre II. Intelligence artificielle

II.1. Généralités

II.1.1. Définition de l'intelligence artificielle

L'intelligence artificielle (IA) est le terme utilisé pour décrire l'utilisation d'ordinateurs et de technologies afin de simuler un comportement intelligent et une pensée critique comparables à ceux d'un être humain. John McCarthy a d'abord décrit le terme IA en 1956 comme la science et l'ingénierie de la création de machines intelligentes (Amisha et al., 2019).

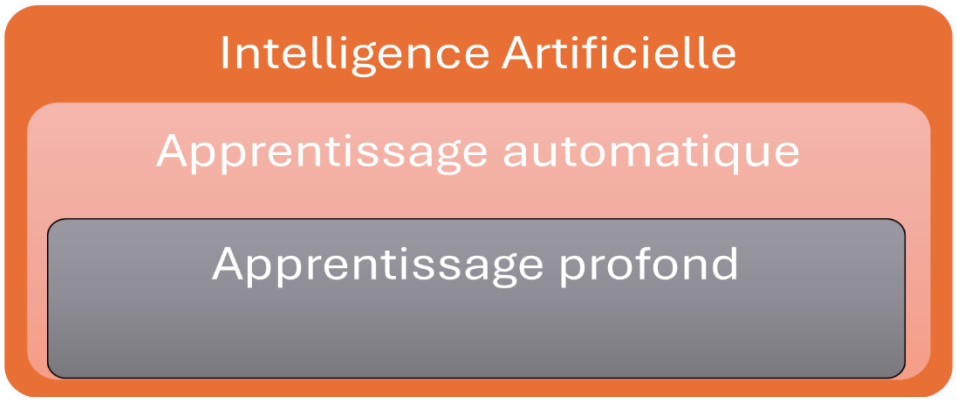

Figure II.1. Paradigmes d'intelligence artificielle.

II.1.2. Historique et évolution de l'IA

Alan Turing (1950) a été l'un des fondateurs des ordinateurs modernes et de l'IA. Le "test de Turing" était basé sur le fait que le comportement intelligent d'un ordinateur est sa capacité à atteindre des performances cognitives équivalentes à celles d'un être humain (Mintz & Brodie, 2019). Les années 1980 et 1990 ont vu un regain d'intérêt pour l'IA. Des techniques d'intelligence artificielle telles que les systèmes experts flous, les réseaux bayésiens, les réseaux neuronaux artificiels et les systèmes intelligents hybrides ont été utilisées dans différents contextes cliniques en santé. En 2016, la plus grande part des investissements dans la recherche en IA était consacrée aux applications de santé par rapport à d'autres secteurs (cbinsights,2017).

L'intelligence artificielle progresse de manière extraordinaire à un rythme sans précédent, atteignant et dépassant les capacités humaines dans de nombreuses tâches auparavant considérées comme inaccessibles aux machines, telles que la traduction de langues, la composition musicale, la détection d'objets, les diagnostics médicaux, la programmation de logiciels, et bien d'autres encore (Buttazzo, 2023).

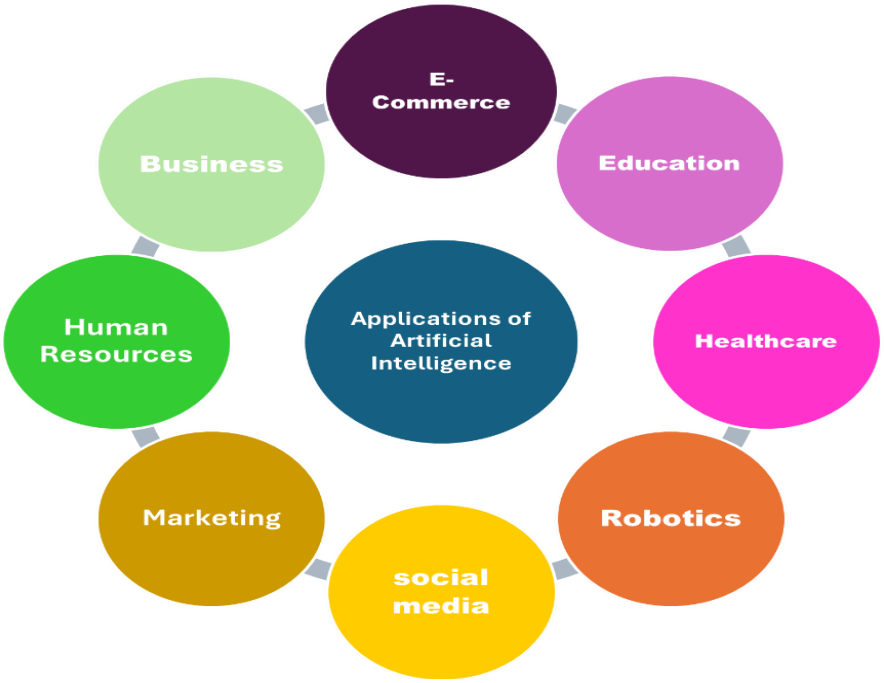

Figure II.2. Importance et applications de l'IA dans divers domaines.

II.2. Technologies clés en IA

II.2.1. L'apprentissage automatique (Machine Learning)

L'apprentissage automatique (ML) est un sous-ensemble de l'intelligence artificielle impliquant des algorithmes qui, contrairement aux règles d'experts, peuvent définir leurs propres règles à partir de données d'entrée grâce à un entraînement itératif et à des améliorations, sans programmation humaine explicite (Jiang et al., 2022).

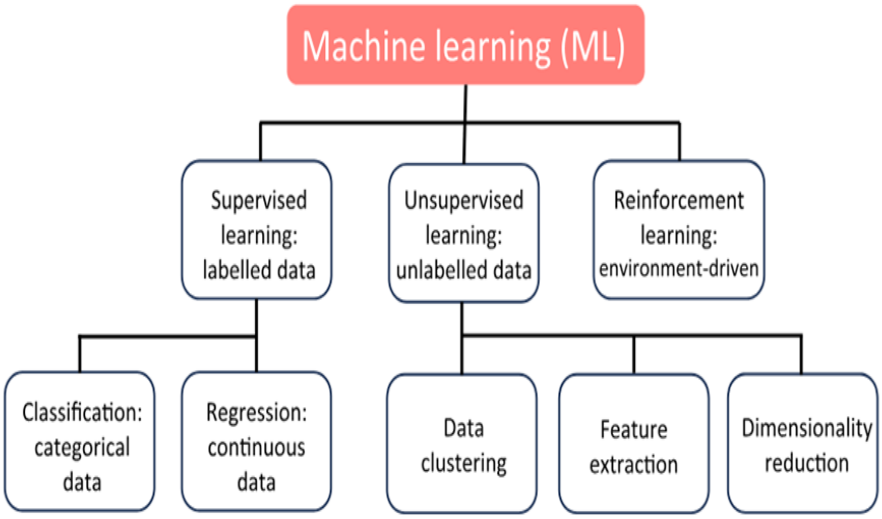

Figure II.3. Les subdivisions et les applications de l'apprentissage automatique (Theodosiou & Read, 2023)**.**

II.2.1.1. Branches de l’apprentissage automatique

II.2.1.1.1. Apprentissage supervisé

La tâche de l'apprentissage supervisé est d'apprendre une fonction qui mappe une entrée vers une sortie en se basant sur des paires d'entrée-sortie collectées. Plus précisément, il existe deux principaux types d'algorithmes d'apprentissage supervisé : les algorithmes de régression (comme les réseaux neuronaux artificiels, la régression bayésienne et la régression par vecteurs de support) et les algorithmes de classification (comme les arbres de décision, les forêts aléatoires et les machines à vecteurs de support) (Zhang & Moon, 2021).

II.2.1.1.2. Apprentissage non supervisé

En apprentissage automatique, les modèles d'apprentissage non supervisé n'ont pas accès à des exemples étiquetés pour l'entraînement afin de découvrir les similitudes sous-jacentes entre les points de données en fonction de leurs caractéristiques communes et d'une métrique pour mesurer la similarité. Il est important de noter que aucune supervision explicite n'est fournie pendant le processus d’entraînement (Molavian et al., 2023).

II.2.1.1.3. Apprentissage semi supervisée

L'apprentissage semi-supervisé est une catégorie avancée de techniques d'apprentissage automatique qui traitent des ensembles de données partiellement étiquetés. Ces méthodes cherchent à apprendre la corrélation entre les caractéristiques et les étiquettes en utilisant à la fois des données étiquetées et non étiquetées, permettant ainsi d'obtenir des performances comparables à l'apprentissage supervisé tout en nécessitant moins d'efforts humains pour annoter les données (Lindley et al., 2024).

II.2.1.1.4. Apprentissage renforcé

Une classe de méthodes avancées particulièrement notable est l'apprentissage par renforcement. L'apprentissage par renforcement adopte une approche complètement différente des méthodes d'apprentissage supervisé et non supervisé. Au lieu de prédire l'étiquette en utilisant un ensemble donné de caractéristiques, il vise à explorer un environnement en naviguant de manière itérative à travers celui-ci. Les algorithmes d'apprentissage par renforcement se composent de deux parties : une qui explore l'environnement, appelée l'acteur, et une qui évalue les actions de l'acteur, appelée le critique (Lindley et al., 2024).

II.2.1.2. Avantages et inconvénient de l'apprentissage automatiques

- **Avantages :**

- Automatisation des tâches répétitives : Le Machine Learning permet d'automatiser des tâches répétitives et chronophages, ce qui améliore l'efficacité opérationnelle des entreprises.
- Analyse prédictive : Il peut analyser de grandes quantités de données pour identifier des tendances et faire des prédictions précises, utile dans divers domaines tels que la finance, la santé et le marketing.
- Amélioration continue : Grâce à des algorithmes d'apprentissage, les systèmes de Machine Learning peuvent s'améliorer de manière autonome à partir de nouvelles données, augmentant leur précision et leur performance au fil du temps.
- Personnalisation : Il permet de fournir des expériences personnalisées aux utilisateurs, comme des recommandations de produits sur les sites de commerce électronique ou des contenus adaptés sur les plateformes de streaming.
- Détection des fraudes : Le Machine Learning est efficace pour détecter des anomalies et des fraudes en temps réel, en analysant les modèles de comportement et les transactions suspectes.

- **Inconvénients :**

- Dépendance aux données : La qualité et la quantité des données d'entraînement sont cruciales. Des données biaisées ou insuffisantes peuvent entraîner des modèles inefficaces ou injustes.
- Complexité et coût : La mise en place de systèmes de Machine Learning nécessite des compétences techniques spécialisées et peut être coûteuse en termes de ressources et de temps.
- Interprétabilité : Les modèles de Machine Learning, en particulier les réseaux de neurones profonds, peuvent être des "boîtes noires", rendant difficile l'explication des décisions prises par le modèle.
- Sécurité et confidentialité : L'utilisation de données sensibles pour entraîner les modèles soulève des préoccupations en matière de confidentialité et de sécurité des informations.
- Dynamisme des données : Les modèles doivent être régulièrement mis à jour avec de nouvelles données pour rester précis et pertinents, ce qui peut nécessiter un suivi et un entretien continus.

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

L'apprentissage profond est un élément fondamental de l'apprentissage automatique qui utilise une matrice cognitive multicouche au sein d'un réseau neuronal profond pour examiner et interpréter les données entrantes. L'objectif de l'apprentissage profond est de développer un réseau neuronal qui reconnaît de manière autonome les motifs afin d'améliorer le processus de détection des caractéristiques (Rodrigues et al., 2021).

'Deep' est un terme technique qui fait référence au nombre de couches dans un réseau de neurones artificiels (RNA). Il existe trois types de couches :

- **Couche d'entrée** : reçoit les données d'entrée.
- **Couche de sortie** : produit le résultat du traitement.
- **Couche cachée** : extrait les motifs dans les données (Jakhar & Kaur, 2020).

II.2.2.1. Types de réseaux de neurones profonds

- Réseau de neurones profond (Deep Neural Networks DNN)

Un réseau neuronal profond (DNN) est un modèle d'apprentissage automatique adopté pour représenter des relations d'entrée-sortie complexes (Das et al., 2020). L'architecture DNN relève de l'apprentissage supervisé et comprend différents types de couches cachées, d'opérations matricielles, d'opérations de convolution et d'accélérateurs. Les DNN nécessitent de nombreuses couches cachées, ce qui les rend les plus robustes (Gbadago et al., 2021). Dans un DNN, le processus d'entraînement implique des passes avant et arrière. Les DNN comprennent des couches entièrement connectées ainsi que des fonctions d'activation, des suppressions aléatoires et une normalisation par lots (Zhu et al., 2022).

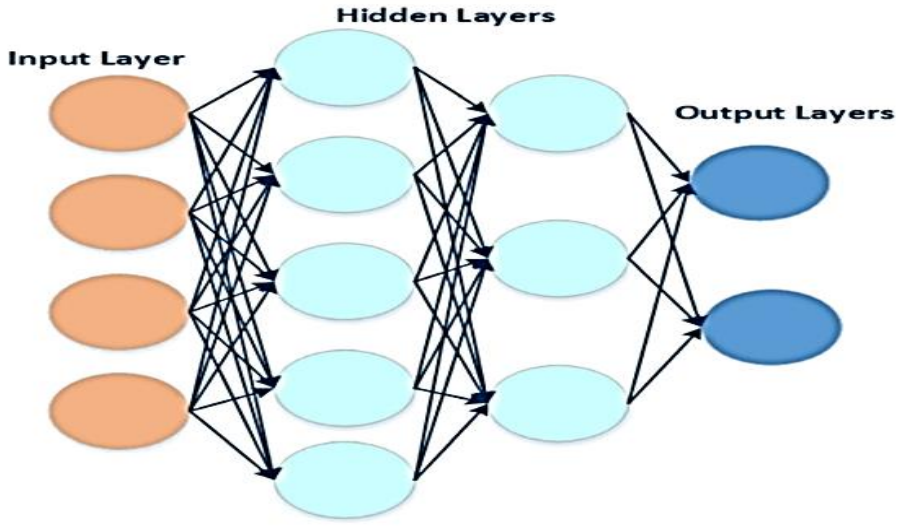

Figure II.4. Architecture d'un réseau neuronal profond (Mahum et al., 2022).

- Réseaux de neurones convolutifs (Convolutional Neural Networks CNN)

Les Réseaux Neuronaux Convolutionnels (CNN) sont une classe de modèles d'apprentissage automatique principalement utilisés dans les tâches de vision par ordinateur et peuvent atteindre des performances semblables à celles des humains en apprenant de l'expérience (Celeghin et al., 2023).

Dans lesquels les entrées sont passées à travers des couches cachées infinies, qui analysent l'image. L'image est décomposée en une collection de pixels et chaque nœud ou 'neurone' se voit attribuer une caractéristique comme la couleur, la taille et la forme, et enfin la sortie est générée (Nichols et al., 2019).

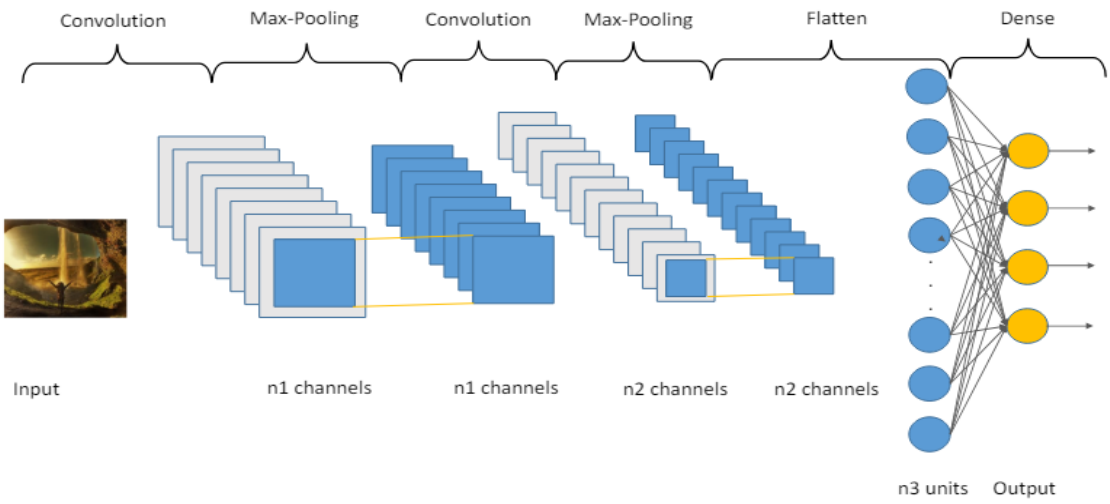

Figure II.5. Représentation des réseaux de neurones convolutifs CNN (García-Ordás et al., 2020).

- Réseaux de neurones récurrents (Recurrent Neural Networks RNN)

Les réseaux neuronaux récurrents (RNN) sont des architectures de réseaux neuronaux dotées d'un état caché et qui utilisent des boucles de rétroaction pour traiter une séquence de données qui informe finalement la sortie finale. Par conséquent, les modèles RNN peuvent reconnaître les caractéristiques séquentielles des données et aider à prédire le prochain point de données probable dans la séquence de données. En exploitant la puissance du traitement des données séquentielles, les cas d'utilisation des RNN sont généralement liés aux modèles de langage ou à l'analyse de données de séries chronologiques. Cependant, plusieurs architectures de RNN populaires ont été introduites dans le domaine, commençant par Simple RNN et LSTM jusqu'aux RNN profonds, et appliquées dans différents contextes expérimentaux (Das et al., 2023).

II.2.2.2. Champs d'application

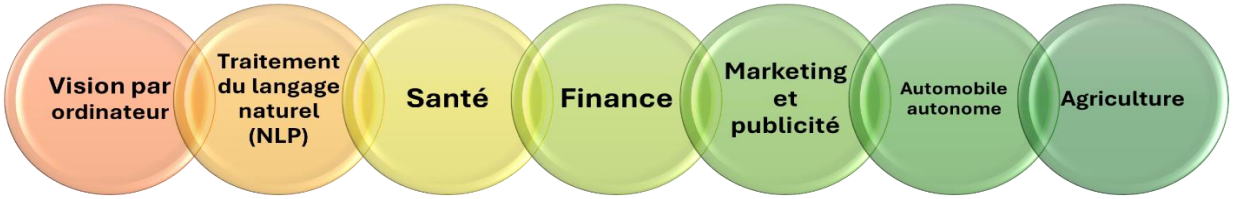

Figure II.6 : Champs d'application de Deep Learning.

II.2.2.3. Avantages et l'inconvénients de Deep Learning :

- **Avantages** :
- Performance exceptionnelle : Les réseaux de neurones profonds (DNN) peuvent atteindre des niveaux de précision extrêmement élevés dans des tâches complexes comme la reconnaissance d'images et la compréhension du langage naturel.
- Capacité à traiter de grandes quantités de données : Deep Learning est particulièrement efficace pour analyser et interpréter de vastes ensembles de données, permettant de découvrir des modèles et des insights cachés.
- Apprentissage automatique des caractéristiques : Contrairement à d'autres techniques de machine learning, le deep learning peut automatiquement extraire et apprendre des caractéristiques pertinentes des données brutes, éliminant ainsi le besoin de prétraitement manuel.
- Flexibilité : Les modèles de deep learning peuvent être appliqués à une large gamme de problèmes, y compris la vision par ordinateur, le traitement du langage naturel, et bien plus encore.
- Amélioration continue avec plus de données : Les performances des modèles de deep learning s'améliorent souvent avec la disponibilité de plus de données d'entraînement, ce qui permet une meilleure généralisation.

- **Inconvénients :**

- Besoin de grandes quantités de données : Pour atteindre des niveaux de performance élevés, les modèles de deep learning nécessitent souvent d'énormes quantités de données annotées, ce qui peut être coûteux et laborieux à obtenir.
- Complexité et opacité : Les modèles de deep learning sont souvent des "boîtes noires", rendant difficile l'interprétation et l'explication des décisions du modèle.
- Risque de surapprentissage : Les modèles de deep learning peuvent facilement surapprendre les données d'entraînement si ceux-ci ne sont pas suffisamment diversifiés ou si des techniques de régularisation ne sont pas utilisées.

- Maintenance et mise à jour difficiles : En raison de leur complexité, les modèles de deep learning peuvent être difficiles à mettre à jour et à maintenir, surtout lorsqu'ils doivent être déployés à grande échelle.

II.2.2.4. Deep Learning Vs Machine Learning

Complexité et Couches :

ML : Utilise des algorithmes plus simples avec moins de couches.

DL : Emploie des réseaux neuronaux avec plusieurs couches, permettant de modéliser des représentations de données complexes.

Exigences en Données :

ML : Efficace avec des ensembles de données plus petits mais peut nécessiter une extraction de caractéristiques intensive.

DL : Nécessite de grands ensembles de données pour bien fonctionner et peut extraire automatiquement des caractéristiques.

Puissance de Traitement :

ML : Généralement moins intensif en termes de calcul.

DL : Demande des ressources computationnelles importantes et nécessite souvent des GPU ou du matériel spécialisé.

Extraction de Caractéristiques :

ML : Repose sur l'extraction manuelle de caractéristiques par des experts du domaine.

DL : Capable d'extraire automatiquement des caractéristiques à partir de données brutes.

Interprétabilité :

ML : Les modèles comme les arbres de décision et la régression linéaire sont plus interprétables et plus faciles à comprendre.

DL : Les réseaux neuronaux profonds sont souvent considérés comme des boîtes noires, ce qui les rend plus difficiles à interpréter.

Partie pratique

Chapitre III : Matériels et méthodes

Chapitre III. Matériel et méthodes

III.1. Matériel

III.1.1. Collection des données

Les données de cette étude proviennent de multiples bases de données et de collaborateurs, garantissant une collecte exhaustive d'informations sur les métabolites. Ces ensembles de données, obtenus à partir de bases de données réputées et de collaborateurs estimés, fournissent une base solide pour le développement du système d'identification des métabolites piloté par l'IA. La diversité et l'exhaustivité des informations sur les métabolites visent à permettre d'entraîner le modèle d'IA sur une grande variété de données, conduisant à des résultats d'identification plus précis et fiables.

III.1.1.1. Base de données

III.1.1.1.1. Human Metabolome Database (HMDB V5.0)

C’est une base de données électronique librement accessible contenant des informations détaillées sur les métabolites de petites molécules trouvés dans le corps humain. Elle est destinée à être utilisée dans les applications de la métabolomique, de la chimie clinique, de la découverte de biomarqueurs et de l'éducation générale.

La base de données est conçue pour contenir ou lier trois types de données : des données chimiques, des données cliniques et des données de biologie moléculaire/biochimie. La base de données contient 220 945 entrées de métabolites, incluant à la fois des métabolites hydrosolubles et liposolubles. De plus, 8 610 séquences protéiques (enzymes et transporteurs) sont liées à ces entrées de métabolites. Chaque entrée MetaboCard contient 130 champs de données, avec les deux tiers des informations consacrées aux données chimiques/cliniques et l'autre tiers aux données enzymatiques ou biochimiques. De nombreux champs de données sont hyperliés à d'autres bases de données (KEGG, PubChem, MetaCyc, ChEBI, PDB, UniProt et GenBank) et à une variété d'applets de visualisation de structures et de voies métaboliques.

La base de données HMDB prend en charge des recherches étendues de texte, de séquences, de structures chimiques, ainsi que des requêtes spectrales en MS et NMR. Quatre bases de données supplémentaires, DrugBank, T3DB, SMPDB et FooDB, font également partie de la suite de bases de données HMDB. DrugBank contient des informations équivalentes sur environ 2832 médicaments et 800 métabolites de médicaments, T3DB contient des informations sur environ 3670 toxines courantes et polluants environnementaux, SMPDB contient des diagrammes de voies pour environ 132 335 voies métaboliques humaines, de médicaments et de maladies ainsi que pour

environ 60628 voies pour d'autres organismes, tandis que FooDB contient des informations équivalentes sur environ 7000 composants alimentaires et additifs alimentaires (https://hmdb.ca/).

Nous nous sommes particulièrement intéressés à des ensembles de données spécifiques notamment :

- MS-MS Spectra Files Experimental (2023-07-01)

Cet ensemble de données contient des spectres de spectrométrie de masse en tandem (MS-MS) obtenus expérimentalement. Il fournit des données spectrales détaillées essentielles pour l'identification et l'analyse des métabolites à travers les motifs de fragmentation observés dans des conditions expérimentales. Ces données sont cruciales pour valider les spectres prédits et améliorer la précision de l'identification des métabolites.

- MS-MS Spectra Files Predicted (2023-07-01)

Cet ensemble de données comprend des spectres MS-MS prédits basés sur des modèles computationnels. Les spectres prédits offrent des motifs de fragmentation théoriques pour les métabolites, ce qui est précieux pour identifier les métabolites dépourvus de spectres expérimentaux. Cet ensemble de données aide à étendre la portée de l'identification des métabolites en fournissant des spectres de référence pour l'appariement computationnel.

- GC-MS Spectra Files Experimental (2023-07-01)

Cet ensemble de données comprend des spectres de chromatographie en phase gazeuse-spectrométrie de masse (GC-MS) obtenus expérimentalement. Il inclut des données spectrales complètes issues d'analyses expérimentales, fournissant des informations précieuses sur la composition chimique et les motifs de fragmentation des métabolites dans des conditions de chromatographie en phase gazeuse. Ces données sont cruciales pour l'identification et l'analyse précises des métabolites.

- GC-MS Spectra Files Predicted (2023-07-01)

Similaire aux spectres MS-MS prédits, cet ensemble de données contient des spectres GC-MS prédits, comprenant des motifs de fragmentation théoriques pour les métabolites tels que prédits par des modèles computationnels. Les spectres GC-MS prédits sont essentiels pour identifier les métabolites qui n'ont pas été analysés expérimentalement, élargissant ainsi la portée de l'identification des métabolites.

- All Metabolites (2021-11-17)

Cet ensemble de données complet comprend des informations détaillées sur tous les métabolites enregistrés dans la HMDB au 17 novembre 2021. Il englobe une vaste gamme de

données, y compris les noms des métabolites, les formules chimiques, les poids moléculaires et d'autres attributs pertinents. Cet ensemble de données sert de référence fondamentale pour l'identification et l'analyse des métabolites.

- Metabolite Structures (2021-11-02)

Cet ensemble de données fournit des informations structurelles détaillées sur les métabolites, y compris leurs structures moléculaires et leur stéréochimie. Les données sont essentielles pour comprendre les propriétés chimiques et les comportements des métabolites. Les informations structurelles aident à l'identification et à la caractérisation précises des métabolites, soutenant diverses applications analytiques et de recherche.

III.1.1.1.2. MassBank

La base de données MassBank comprend des spectres de masse méticuleusement sélectionnés, obtenus dans diverses conditions expérimentales. Les données spectrales de haute qualité garantissent des résultats fiables et reproductibles dans l'identification des métabolites. L'ensemble de données couvre une vaste gamme de métabolites, fournissant des informations spectrales étendues. Cette couverture complète est essentielle pour identifier et caractériser une grande diversité de métabolites présents dans les échantillons biologiques.

Chaque spectre de masse dans la base de données MassBank est minutieusement annoté avec des informations sur le composé, y compris sa structure chimique, son poids moléculaire et ses motifs de fragmentation. Ces annotations aident à l'interprétation et à l'identification précises des métabolites.

Les données spectrales dans MassBank respectent des formats standardisés, facilitant l'intégration et la comparaison avec d'autres ensembles de données. Cette standardisation assure la cohérence et la fiabilité dans l'identification des métabolites à travers différentes études.

Les données incluses dans MassBank sont soumises à des processus rigoureux de validation et de contrôle de qualité. Cela garantit que les spectres sont précis et d'haute-fidélité, réduisant ainsi la probabilité d'identifications erronées.

MassBank est largement utilisé dans diverses applications de recherche et d'analyse, y compris la métabolomique, la découverte de médicaments et l'analyse environnementale. Sa base de données robuste soutient des investigations scientifiques variées en fournissant des références spectrales de masse fiables (https://massbank.eu/MassBank/).

Dans MassBank, Nous nous sommes particulièrement intéressés à MassBank_NIST.msp, contient des données spectrales de masse complètes et est conçu pour être compatible avec la

bibliothèque spectrale de masse NIST (National Institute of Standards and Technology). Le dataset MassBank_NIST.msp inclut des spectres de masse détaillés et annotés pour une large gamme de métabolites, ce qui en fait une ressource précieuse pour l'identification précise des métabolites.

L'ensemble de données comprend des spectres de masse de haute qualité obtenus dans diverses conditions expérimentales. Ces spectres sont minutieusement sélectionnés pour garantir la fiabilité et la reproductibilité de l'identification des métabolites.

Chaque spectre est annoté avec des informations essentielles telles que le nom du métabolite, le m/z du précurseur, le type de précurseur, le type d'instrument et l'énergie de collision. Ces annotations facilitent l'identification et l'analyse précises.

Les données sont formatées selon les normes NIST, assurant la compatibilité avec divers outils analytiques et logiciels utilisés en recherche métabolomique. Cette standardisation aide à l'intégration harmonieuse des données MassBank dans les workflows de recherche plus larges.

L'ensemble de données englobe une grande variété de métabolites, fournissant des informations spectrales étendues qui soutiennent des études métabolomiques complètes. MassBank_NIST.msp est largement utilisé dans des applications de recherche et d'analyse, y compris la découverte de médicaments, l'analyse environnementale et la recherche clinique. Son dataset robuste sert de référence pour identifier les métabolites inconnus et valider les résultats expérimentaux.

Le fichier MassBank_NIST.msp est une partie intégrante de cette étude, contribuant des données spectrales de masse de haute qualité et bien annotées, essentielles pour le développement et la validation du système d'identification des métabolites piloté par l'IA.

III.1.1.2. Données Collaboratives

En plus des ensembles de données de la HMDB et de MassBank, des données collaboratives ont été obtenues du Centre de Recherche en Biotechnologie (CRBt) de Constantine, du Centre de Recherche Scientifique et Technique en Analyses Physico-chimiques (CRAPC) de Bou Ismaïl, et du Centre de Recherche Scientifique et Technique sur les Régions Arides (CRSTRA) de Biskra. Ces centres de recherche ont contribué des données et des spectres de métabolites précieux, enrichissant ainsi la diversité de l'ensemble de données.

III.1.2. Machine

Le modèle d'identification des métabolites assisté par intelligence artificielle a été développé et exécuté à l'aide d'une station de travail haute performance. Cette station de travail est équipée d'un processeur Intel(R) Xeon(R) CPU E5-2698 v4 fonctionnant à une fréquence de base de 2,20

GHz, avec une fréquence Turbo maximale de 3,60 GHz. Ce processeur comporte 20 cœurs et 40 threads, ainsi qu'un cache SmartCache de 50 Mo. Avec une dissipation thermique (TDP) de 135W, ce processeur offre une capacité de traitement suffisante pour les calculs intensifs.

La mémoire de la station de travail est constituée de 64 Go de RAM DDR4. Ce type de mémoire, SDRAM DDR4, fonctionne à une vitesse de 2400 MHz et est configurée en quadri-canal (quad-channel). Cette capacité de mémoire permet de charger et de manipuler de grands ensembles de données de manière fluide, sans goulots d'étranglement de performance.

En complément, la station de travail est équipée d'un disque SSD PC401 NVMe de SK hynix, offrant des performances de stockage rapides et fiables pour le chargement et le traitement des données volumineuses nécessaires à l'entraînement du modèle. Ce SSD utilise l'interface PCIe 3.0 et présente des vitesses de lecture allant jusqu'à 1700 Mo/s et des vitesses d'écriture allant jusqu'à 700 Mo/s, ce qui assure une réactivité optimale et des temps de transfert de données réduits.

Cette configuration matérielle a fourni la puissance de calcul et la capacité de mémoire nécessaires pour traiter les grands ensembles de données et effectuer des calculs complexes impliqués dans le processus de développement du modèle. Le processeur multi-cœur a facilité le traitement parallèle, ce qui est essentiel pour gérer efficacement des volumes de données importants et entraîner des modèles complexes. De plus, la RAM DDR4 ample a assuré un chargement et une manipulation des données sans interruption, garantissant ainsi des performances optimales tout au long du développement du modèle.

III.1.3. DataSet

III.1.3.1. Choix du dataset

Suite à la collecte des données, nous avions accès à sept bases de données distinctes : la première, MS-MS Spectra Files Experimental; la deuxième, MS-MS Spectra Files Predicted (2023-07-01); la troisième, GC-MS Spectra Files Experimental (2023-07-01); la quatrième, GC-MS Spectra Files Predicted (2023-07-01); la cinquième, All Metabolites (2021-11-17); la sixième, Metabolite Structures (2021-11-02); et la septième, MassBank_NIST.msp.

Les bases de données 1 à 4 contiennent des informations numériques sur la séparation des métabolites telles que le rapport masse/charge, l'intensité, le poids moléculaire et le temps de rétention, mais ne contiennent pas de noms ni de formules chimiques. La base de données 5 inclut des informations sur les noms, les formules chimiques et une description détaillée des métabolites, tandis que la base de données 6 contient des informations sur la structure chimique des métabolites. Cependant, ces deux bases de données ne sont pas riches en valeurs numériques.

Notre problématique de recherche était de déterminer s'il est possible de former un modèle en utilisant différentes bases de données. Plus précisément, nous nous demandions si nous pouvions commencer l'apprentissage avec une base de données contenant des paramètres numériques, continuer avec une autre qui inclut les noms et les formules chimiques des molécules, et terminer avec une qui contient les structures moléculaires.

Après une investigation approfondie, nous avons découvert qu'il n'est pas faisable d'utiliser différents attributs pour la création du modèle. La cohérence des attributs à travers les bases de données est essentielle pour un apprentissage efficace du modèle. Par conséquent, nous avons sélectionné la septième base de données, MassBank_NIST.msp, comme notre jeu de données principal.

Ce choix a été motivé par plusieurs avantages : elle inclut simultanément les noms, les formules et les attributs numériques, et contient un grand nombre de molécules avec des données de haute qualité. Ainsi, MassBank_NIST.msp était le choix optimal pour notre modèle d'identification de métabolites piloté par IA en raison de son jeu de données complet et de haute qualité. Il est également important de noter que nous n'avons pas pu utiliser les données fournies par nos collaborateurs des centres de recherche, car l'extension des fichiers de leurs bases de données était inconnue. En conséquence, nous n'avons pas pu extraire les données de ces fichiers.

III.1.3.2. Description du dataset

Le dataset MassBank_NIST.msp consiste en 117 733 entrées de données représentant des mesures spectrales et chimiques pour diverses molécules, et 20 paramètres descriptifs ou attributs associés à chaque entrée (Tab. III.1). Le dataset est assez complet, avec des informations chimiques et spectrales détaillées pour un grand nombre de métabolites. Il inclut les noms, formules, poids moléculaires, masses exactes, et pics formatés, ce qui en fait une ressource riche pour l'identification de métabolites pilotée par l'IA.

Le dataset comprend divers types de molécules classées comme suit :

III.1.3.2.1. Métabolites Primaires

Ce sont des molécules essentielles au métabolisme de base, nécessaires pour les processus biologiques fondamentaux tels que la production d'énergie et la synthèse de biomolécules. Exemples : Pyrophen, Zearalenone et Roquefortine M.

III.1.3.2.2. Métabolites Secondaires

Ces molécules ne sont pas directement impliquées dans la croissance, le développement ou la reproduction de l'organisme, mais elles jouent des rôles importants dans la défense contre les

pathogènes, la signalisation et d'autres interactions écologiques. Exemples : Decarestrictine F Isomarticin et Trichoverroidin.

III.1.3.2.3. Lipides

Les lipides sont des molécules organiques hydrophobes ou amphiphiles qui jouent des rôles clés dans le stockage de l'énergie, la structure des membranes cellulaires et la signalisation cellulaire. Exemples : Lanosterol, Campesterol et Beta-sitosterol.

III.1.3.2.4. Composés Phénoliques

Ces molécules contiennent un ou plusieurs groupes phénol et sont largement réparties dans le règne végétal. Elles sont connues pour leurs propriétés antioxydantes et leur rôle dans la défense des plantes. Exemples : Quercetin, Kaempferol et Myricetin.

III.1.3.2.5. Nucléotides et Dérivés

Les nucléotides sont les blocs de construction des acides nucléiques comme l'ADN et l'ARN. Ils jouent également des rôles clés dans le métabolisme énergétique et la signalisation cellulaire. Exemples : Adenosine, Cytidine et Uridine.

III.1.3.2.6. Peptides et Protéines

Ce sont des chaînes d'acides aminés qui fonctionnent comme des éléments structuraux, des enzymes, des hormones et des anticorps dans les organismes vivants. Exemples : Penicillin G, Beauvericin et 3-Acetyldeoxynivalenol.

Cette classification met en évidence la diversité et la richesse des données disponibles dans le dataset MassBank_NIST.msp, ce qui le rend particulièrement utile pour l'entraînement de modèles d'identification de métabolites.

Tableau III.1. Paramètres de classement descriptifs du dataset

N°	Paramètre	Description
1	Name	Contient les noms des métabolites. Presque toutes les entrées ont cette information.
2	Synon	Synonymes pour les métabolites. Environ la moitié des entrées possèdent cette information.
3	DB#	Identifiants de la base de données pour les métabolites.
4	InChIKey	Identifiants chimiques uniques pour la plupart des entrées.
5	InChI	Identifiants chimiques internationaux, presque complets.

6	SMILES	Chaînes d'entrée de ligne de la structure moléculaire simplifiée, presque complètes.
7	Precursor_type	Informations sur les ions précurseurs, disponibles pour la majorité des entrées.
8	Spectrum_type	Type de spectre (e.g., MS2), disponible pour presque toutes les entrées.
9	PrecursorMZ	Rapport masse/charge de l'ion précurseur, disponible pour la plupart des entrées.
10	Instrument_type	Type d'instrument utilisé pour la mesure.
11	Instrument	Instruments spécifiques utilisés, presque complets.
12	Ion_mode	Mode d'ionisation (e.g., positif ou négatif), disponible pour presque toutes les entrées.
13	Collision_energy	Énergie de collision utilisée pendant la spectrométrie de masse, disponible pour la majorité des entrées.
14	Formula	Formules chimiques des métabolites.
15	MW	Poids moléculaire des métabolites.
16	ExactMass	Masse exacte des métabolites.
17	Comments	Commentaires supplémentaires, presque complets.
18	Splash	Codes Splash pour l'identification unique des spectres.
19	Num Peaks	Nombre de pics dans le spectre.
20	Peaks	Comprend le rapport masse/charge (m/z) et l'intensité (abondance) des ions.

III.2. Méthodes

III.2.1. Préparation des données

Dans l'analyse des métabolites, les données de spectrométrie de masse (MS) sont souvent stockées au format MSP (Mass Spectral Profile), qui contient des informations détaillées sur chaque molécule, y compris son nom, des informations sur le précurseur, et les rapports masse/charge (m/z) avec les intensités correspondantes. Pour faciliter l'analyse des données, en particulier lors de l'utilisation de modèles d'apprentissage automatique, il est essentiel de convertir ces données en un format structuré et facilement accessible tel que le CSV. Cette section décrit le

processus de conversion des données MSP en format CSV, en soulignant les avantages et la méthodologie employée.

La conversion du format MSP en CSV est cruciale pour plusieurs raisons. Les fichiers CSV sont largement pris en charge par divers outils et logiciels d'analyse de données, rendant les données plus accessibles pour les chercheurs et les analystes. Le format CSV fournit une manière structurée de représenter les données, facilitant ainsi leur manipulation et analyse. De plus, les modèles d'apprentissage automatique et les analyses statistiques nécessitent souvent des données d'entrée sous forme tabulaire, ce que les fichiers CSV fournissent facilement. Enfin, les fichiers CSV peuvent gérer efficacement de grands ensembles de données et peuvent être traités par de nombreux langages de programmation et outils.

Nous avons utilisé un script qui s'appuie sur les bibliothèques *csv* et *tqdm* pour convertir efficacement les données MSP en format CSV, fournissant ainsi une représentation structurée et accessible des données de spectrométrie de masse. La bibliothèque csv gère la lecture et l'écriture des fichiers CSV, tandis que la bibliothèque *tqdm* offre une visualisation de la progression. Le script garantit que les données sont aplaties en un seul vecteur de caractéristiques pour chaque molécule, facilitant ainsi une analyse plus approfondie et les applications d'apprentissage automatique.

Ce script a été exécuté dans un environnement Python 3.8 à l'aide de l'Anaconda Prompt, avec les bibliothèques nécessaires installées via *pip*. Cet environnement permet de gérer efficacement de grands ensembles de données et d'assurer une compatibilité avec divers outils d'analyse. La conversion des données MSP en format CSV permet une manipulation et une analyse faciles, répondant aux exigences des modèles d'apprentissage automatique et des analyses statistiques.

III.2.2. Nettoyage des Données

Une fois les données converties en format CSV, la prochaine étape consiste à nettoyer les données pour s'assurer de leur qualité et de leur intégrité. Le nettoyage des données implique la suppression des valeurs aberrantes, le traitement des valeurs manquantes et la correction des erreurs éventuelles. Nous avons utilisé des bibliothèques Python telles que *pandas* pour effectuer le nettoyage des données de manière efficace et systématique. Les doublons dans les données peuvent fausser les résultats de l'analyse. Pour les supprimer, nous avons utilisé des fonctions de pandas comme *drop_duplicates()*. Les valeurs manquantes ont été gérées en utilisant la méthode de propagation avant (*forward fill*) pour garantir que toutes les données soient complètes et cohérentes. Ce processus de nettoyage permet de s'assurer que les données sont prêtes pour l'analyse et l'entraînement des modèles d'apprentissage automatique.

III.2.3. Transformscation des Données

Pour préparer les données pour l'analyse et l'entraînement du modèle, il est souvent nécessaire de les transformer en utilisant des techniques telles que la normalisation ou la standardisation. Ces transformations sont essentielles pour garantir que les algorithmes d'apprentissage automatique puissent traiter efficacement les données et produire des résultats précis.

Nous avons utilisé des bibliothèques Python comme *scikit-learn* pour normaliser et encoder les données. La normalisation des données est une étape cruciale qui permet de mettre les différentes caractéristiques sur une échelle comparable, facilitant ainsi l'apprentissage des modèles. Nous avons utilisé la classe *StandardScaler* de *scikit-learn* pour normaliser les données, ce qui permet de centrer les données autour de la moyenne et de les mettre à l'échelle en fonction de l'écart-type.

Les variables catégorielles, telles que *Name* et *Formula*, ont été encodées en valeurs numériques pour être utilisables par les algorithmes d'apprentissage automatique. Nous avons utilisé la classe *LabelEncoder* de *scikit-learn* pour convertir ces colonnes catégorielles en entiers, assurant ainsi que les modèles peuvent traiter ces informations de manière efficace.

Enfin, nous avons transformé les *Peaks* en extrayant les valeurs m/z et intensité, puis en les aplatissant dans un seul vecteur de caractéristiques pour chaque molécule. Cette transformation garantit que chaque molécule est représentée de manière cohérente et complète, facilitant l'analyse et l'entraînement des modèles d'apprentissage automatique.

Ces étapes de transformation des données sont essentielles pour préparer un jeu de données de haute qualité, prêt à être utilisé pour des analyses approfondies et l'entraînement de modèles prédictifs robustes.

III.2.4. Choix du Modèle

Le choix du modèle d'apprentissage automatique dépend de la nature des données et des objectifs de l'analyse. Pour la classification ou la régression, des modèles tels que la régression logistique, les forêts aléatoires ou les réseaux de neurones peuvent être utilisés.

Pour évaluer les caractéristiques des données et sélectionner le modèle approprié, nous avons utilisé un script qui nous permet de comprendre la distribution des données et les relations entre les différentes caractéristiques.

Tout d'abord, nous avons chargé le jeu de données traité à partir du chemin spécifié en utilisant *pandas*. Cela nous permet d'accéder aux données dans un format structuré et d'effectuer des analyses exploratoires. Ensuite, en affichant les statistiques de base avec *dataset.describe()*, nous

obtenons des informations résumées sur les caractéristiques numériques du jeu de données, telles que la moyenne, l'écart-type, les valeurs minimales et maximales, et les quartiles. Cela nous aide à identifier les caractéristiques importantes et leur dispersion.

En traçant des histogrammes pour chaque caractéristique, nous pouvons visualiser la distribution des valeurs. Les histogrammes nous aident à comprendre la forme de la distribution, à identifier les valeurs aberrantes et à détecter les asymétries éventuelles. Pour cette visualisation, nous avons utilisé la fonction *hist* de *pandas*.

La matrice de corrélation, générée par *dataset.corr()*, montre les relations entre les différentes caractéristiques. En traçant une carte thermique (*heatmap*) de cette matrice à l'aide de *seaborn*, nous pouvons visualiser la force et la direction des corrélations entre les variables. Cela nous permet d'identifier les caractéristiques qui sont fortement corrélées et celles qui sont indépendantes, ce qui est crucial pour la sélection des caractéristiques et la réduction de la dimensionnalité. La fonction *heatmap* de *seaborn* a été utilisée pour cette visualisation.

Ces analyses exploratoires des données nous aident à mieux comprendre les caractéristiques du jeu de données et à sélectionner le modèle d'apprentissage automatique le plus approprié pour nos besoins. Grâce à ces visualisations et analyses statistiques, nous pouvons prendre des décisions éclairées sur les transformations supplémentaires des données et les choix de modèles pour maximiser la performance et la précision des prédictions.

III.2.5. Entraînement du Modèle

Pour entraîner le modèle d'apprentissage automatique, nous avons suivi plusieurs étapes détaillées en utilisant un script qui implique le chargement d'un jeu de données traité, l'encodage des étiquettes, le prétraitement et le *padding* des pics, la combinaison des caractéristiques, la mise à l'échelle des caractéristiques, la division des données, et l'entraînement d'un modèle de réseau de neurones multi-sorties avec *TensorFlow*.

III.2.5.1. Chargement du Jeu de Données

Nous avons chargé le jeu de données traité à partir du chemin spécifié en utilisant *pandas*. Cela nous permet d'accéder aux données dans un format structuré.

III.2.5.2. Encodage des Étiquettes

Les étiquettes des colonnes *Name* et *Formula* ont été encodées en valeurs numériques en utilisant des encodeurs de labels (*LabelEncoder*). Ces étiquettes sont ensuite transformées pour faciliter l'entraînement du modèle.

III.2.5.3. Prétraitement et Padding des Pics

Les pics de chaque molécule ont été prétraités et remplis (*padded*) pour garantir une longueur uniforme. La fonction *preprocess_peaks* convertit les représentations de chaînes des pics en listes de valeurs numériques, incluant les m/z et les intensités.

III.2.5.4. Combinaison des Caractéristiques

Les caractéristiques combinées comprenaient les valeurs de MW (mass weight), ExactMass, et Num Peaks, ainsi que les données des pics prétraités. Nous avons utilisé *np.hstack* pour combiner ces caractéristiques en une seule matrice de caractéristiques.

III.2.5.5. Mise à l'Échelle des Caractéristiques

Les caractéristiques ont été mises à l'échelle en utilisant StandardScaler de *scikit-learn*. Cela permet d'assurer que toutes les caractéristiques sont sur une échelle comparable, améliorant ainsi la performance et la convergence du modèle.

III.2.5.6. Division des Données

Les données ont été divisées en ensembles d'entraînement et de test en utilisant *train_test_split*, avec 80% des données pour l'entraînement et 20% pour les tests.

III.2.5.7. Définition et Entraînement du Modèle

Nous avons défini et entraîné un modèle de réseau de neurones multi-sorties en utilisant *TensorFlow*. Le modèle comporte plusieurs couches denses avec des activations *ReLU*, des couches de dropout pour éviter le surapprentissage, et des couches de normalisation par lot (*BatchNormalization*).

III.2.5.8. Sauvegarde des Données Traitées et des Modèles

Nous avons sauvegardé les données traitées avec des pics numériques, les encodeurs et le scaler pour une utilisation ultérieure dans une interface graphique utilisateur (GUI). Les noms des caractéristiques ont également été sauvegardés pour une référence future. Ces étapes détaillées garantissent un entraînement efficace du modèle, préparant les données et les modèles pour des analyses futures et des déploiements en production.

III.2.6. Évaluation du Modèle

Après l'entraînement du modèle, il est crucial d'évaluer sa performance pour s'assurer qu'il généralise bien sur de nouvelles données. L'évaluation du modèle implique l'utilisation de diverses métriques pour mesurer la précision, la capacité de généralisation et la robustesse du modèle. Voici les étapes que nous avons suivies pour évaluer notre modèle :

- Prédictions sur les Données de Test

Nous avons utilisé le modèle entraîné pour faire des prédictions sur l'ensemble de test. Cela nous permet de comparer les prédictions du modèle avec les valeurs réelles pour évaluer sa performance.

- Conversion des Prédictions

Les sorties du modèle sont des probabilités pour chaque classe. Nous avons utilisé *np.argmax* pour convertir ces probabilités en classes prédictives.

- Calcul des Métriques de Précision

La précision est une métrique couramment utilisée pour évaluer la performance des modèles de classification. Elle mesure la proportion de prédictions correctes parmi toutes les prédictions effectuées.

- Rapport de Classification

Le rapport de classification fournit des métriques détaillées telles que la précision, le rappel, et le score F1 pour chaque classe. Cela nous aide à comprendre la performance du modèle sur chaque classe individuellement.

- Matrice de Confusion

La matrice de confusion est une autre méthode pour évaluer la performance d'un modèle de classification. Elle montre le nombre de prédictions correctes et incorrectes, classées par chaque classe.

Ces étapes nous permettent de comprendre en profondeur la performance du modèle, en identifiant les points forts et les faiblesses. En évaluant le modèle de manière exhaustive, nous pouvons apporter les ajustements nécessaires pour améliorer sa précision et sa capacité de généralisation avant de le déployer en production.

III.2.7. Optimisation et Réglage Fin

Pour améliorer la performance du modèle, il est souvent nécessaire d'optimiser les hyperparamètres et d'ajuster finement le modèle. Nous avons utilisé des techniques avancées pour trouver les meilleurs paramètres et ainsi améliorer la précision et la robustesse du modèle.

Nous avons commencé par utiliser *GridSearchCV* pour explorer différentes combinaisons d'hyperparamètres. *GridSearchCV* permet de tester systématiquement toutes les combinaisons possibles de paramètres fournis dans une grille, afin de déterminer les meilleurs réglages pour notre modèle de réseau de neurones multi-sorties.

Ensuite, nous avons utilisé la validation croisée pour évaluer la performance du modèle sur plusieurs sous-ensembles des données d'entraînement. La validation croisée assure que le modèle généralise bien et n'est pas surajusté.

Une fois les meilleurs paramètres identifiés grâce à *GridSearchCV*, nous avons réentraîné le modèle en utilisant ces paramètres optimisés. Ce processus permet de maximiser la performance et la précision du modèle.

Enfin, nous avons analysé les courbes d'apprentissage pour visualiser la performance du modèle sur les ensembles d'entraînement et de validation au cours des époques. Les courbes d'apprentissage nous aident à identifier les problèmes de sous-ajustement ou de sur-ajustement, nous permettant ainsi d'ajuster le modèle en conséquence.

Ces étapes d'optimisation et de réglage fin permettent de maximiser la performance du modèle, assurant ainsi qu'il est prêt pour être déployé en production. L'analyse approfondie des courbes d'apprentissage et l'utilisation de techniques de validation croisée et de recherche en grille garantissent que le modèle est à la fois précis et robuste.

III.2.8. Déploiement du Modèle

Le déploiement du modèle en production permet de l'utiliser dans des applications réelles, facilitant son accès pour les utilisateurs finaux. Pour ce faire, nous avons développé une interface graphique utilisateur (GUI) et préparé les fichiers nécessaires pour garantir une utilisation fluide et intuitive du modèle.

Nous avons utilisé la bibliothèque *Tkinter* pour créer une interface graphique conviviale permettant aux utilisateurs de charger des fichiers CSV, d'exécuter des prédictions et de visualiser les résultats.

III.2.8.1. Chargement des Fichiers Nécessaires

Nous avons chargé le modèle, le scaler et les encodeurs de labels nécessaires pour la transformation des données et les prédictions. Ces fichiers incluent le modèle de réseau de neurones, le scaler pour la normalisation des données et les encodeurs pour les noms et formules moléculaires.

III.2.8.2. Fonction de Prétraitement des Pics

Nous avons développé une fonction pour traiter et remplir (pad) les pics. Cette fonction convertit les pics en liste de valeurs numériques, incluant les m/z et les intensités, puis les remplit pour garantir une longueur uniforme.

III.2.8.3. Fonction de Prédiction

La fonction de prédiction utilise les entrées de l'utilisateur pour faire des prédictions sur le nom et la formule de la molécule en utilisant le modèle chargé. Les caractéristiques de l'utilisateur sont transformées, mises à l'échelle, et utilisées pour faire des prédictions avec le modèle.

III.2.8.4. Fonction de Téléchargement de Fichier

Cette fonction permet de charger un fichier CSV contenant les données nécessaires et de faire des prédictions pour chaque ligne du fichier. Les valeurs des colonnes pertinentes sont extraites et utilisées pour faire des prédictions avec le modèle.

III.2.8.5. Fonction d'Explication des Caractéristiques

Une fonction pour afficher des explications sur les différentes caractéristiques utilisées dans le modèle. Cette fonction montre une boîte de dialogue avec des informations sur chaque caractéristique.

III.2.8.6. Configuration de la GUI

La configuration de la GUI a été faite en utilisant *Tkinter* pour créer une interface utilisateur conviviale avec des champs d'entrée et des boutons pour les différentes fonctionnalités. L'interface permet de saisir les valeurs des caractéristiques, de charger un fichier CSV, et d'obtenir des prédictions.

Ces étapes finales permettent de déployer le modèle en production, le rendant accessible et utilisable par les utilisateurs finaux à travers une interface intuitive et conviviale. Le développement de la GUI et la préparation des fichiers nécessaires assurent que les utilisateurs peuvent facilement charger des données, exécuter des prédictions et obtenir des résultats en temps réel.

Chapitre IV : Résultats et discussion

Chapitre IV. Résultats et discussion

IV.1. Préparation des données

L'analyse des métabolites par spectrométrie de masse a nécessité la conversion des fichiers MSP en format CSV pour faciliter l'analyse des données et l'utilisation de modèles d'apprentissage automatique. Le fichier CSV généré contient diverses informations pertinentes sur chaque molécule, incluant les noms, les synonymes, les formules chimiques, les poids moléculaires (MW), les masses exactes, les types de précurseurs, les types de spectres, les intensités des pics, le nombre de pics, et plus encore (Tab. VI.1).

La préparation des informations par conversion des fichiers MSP en format CSV a permis d'obtenir un ensemble de données structuré et facilement accessible, prêt pour des analyses avancées et l'application de modèles d'apprentissage automatique. Cette étape est fondamentale pour la réussite de l'analyse des métabolites et le développement de solutions d'identification automatique des métabolites.

IV.2. Nettoyage des données

Le jeu de données nettoyé comprend 117 732 entrées et six colonnes essentielles : *Name*, *Formula*, *MW*, *ExactMass*, *Num Peaks*, et *Peaks*. Chaque colonne est entièrement remplie, garantissant un jeu de données complet pour une analyse ultérieure. L'attribut *PrecursorMZ* a été supprimé en raison d'une quantité excessive de données manquantes.

De plus, des attributs non numériques tels que *Synon*, *DB#*, *InChIKey*, *InChI*, *SMILES*, *Spectrum_type*, *Instrument_type*, *Instrument, Ion_mode*, *Collision_energy*, *Comments*, et *Splash* ont été supprimés manuellement. Cela a permis de simplifier le jeu de données pour se concentrer sur les valeurs numériques pertinentes. Le processus de nettoyage complet garantit que le jeu de données est fiable et prêt pour une analyse précise et l'entraînement de modèles d'apprentissage automatique (Tab. VI.2).

IV.3. Transformation des données

Après le nettoyage des données, nous avons identifié les colonnes *Name*, *Formula* et *Peaks* comme nécessitant une transformation supplémentaire pour être exploitables par les algorithmes d'apprentissage automatique. Nous avons donc ajouté au jeu de données nettoyé des attributs numériques dérivés de ces colonnes. Après cette transformation, le jeu de données comprend désormais les colonnes suivantes : *Name, Formula, MW, ExactMass, Num Peaks, Peaks, EncodedName, EncodedFormula, ScaledEncodedName, ScaledEncodedFormula,* et *ProcessedPeaksNumerical*.

Les colonnes *EncodedName* et *EncodedFormula* contiennent les valeurs encodées des *noms* et *formules* en utilisant la classe *LabelEncoder* de *scikit-learn*, permettant une interprétation numérique standardisée des attributs catégoriels. Les valeurs encodées ont ensuite été normalisées, comme le montrent les colonnes *ScaledEncodedName* et *ScaledEncodedFormula*, grâce à la classe *StandardScaler*. Cette normalisation centre les données autour de la moyenne et les met à l'échelle en fonction de l'écart-type, garantissant que les différentes caractéristiques sont comparables et équilibrées lors de l'entraînement des modèles. Enfin, les valeurs *m/z* et *intensité* des *Peaks* ont été extraites et transformées en un seul vecteur de caractéristiques pour chaque molécule, ce qui est représenté par la colonne *ProcessedPeaksNumerical*. Cette transformation assure une représentation cohérente et complète des données spectrales, facilitant l'analyse et l'entraînement des modèles d'apprentissage automatique (Tab. VI.3).

Tableau IV.1. Vue d'ensemble des données métabolomiques préparées après la conversion des fichiers MSP en format CSV.

Name	Pyrophen	Decarestrictine F	Roquefortine A
Synon	N-[(1S)-1-(4-methoxy-6-oxopyran-2-yl)-2-phenylethyl]acetamide	(1S,3R,8Z,10R)-3-methyl-4,11-dioxabicyclo[8.1.0]undec-8-ene-5,7-dione	[(6aR,9S,10R,10aR)-7,9-dimethyl-6,6a,8,9,10,10a-hexahydro-4H-indolo[4,3-fg]quinoline-10-yl] acetate
DB#	MSBNK-AAFC-AC000854	MSBNK-AAFC-AC000767	MSBNK-AAFC-AC000552
InChIKey	VFMQMACUYWGDOJ-AWEZNQCLSA-N	MXRJZFNJVFPSQN-NQRYBKARSA-N	GJSSYQDXZLZOLR-QMHBMSAFSA-N
InChI	InChI=1S/C16H17NO4/c1-11(18)17-14(8-12-6-4-3-5-7-12)15-9-13(20-2)10-16(19)21-15/h3-7,9-10,14H,8H2,1-2H3,(H,17,18)/t14-/m0/s1	InChI=1S/C10H12O4/c1-6-4-9-8(14-9)3-2-7(11)5-10(12)13-6/h2-3,6,8-9H,4-5H2,1H3/b3-2-/t6-,8-,9+/m1/s1	InChI=1S/C18H22N2O2/c1-10-9-20(3)15-7-12-8-19-14-6-4-5-13(16(12)14)17(15)18(10)22-11(2)21/h4-6,8,10,15,17-19H,7,9H2,1-3H3/t10-,15+,17+,18+/m0/s1
SMILES	CC(=O)N[C@@H](CC1=CC=CC=C1)C2=CC(=CC(=O)O2)OC	C[C@@H]1C[C@H]2[C@H](O2)/C=C\C(=O)CC(=O)O1	C[C@H]1CN([C@@H]2CC3=CNC4=CC=CC(=C34)[C@H]2[C@@H]1OC(=O)C)C
Precursor_type	[M+H]+	[M+H]+	[M+H]+
Spectrum_type	MS2	MS2	MS2
PrecursorMZ	288.1225	197.0803	299.1749
Instrument_type	LC-ESI-ITFT	LC-ESI-ITFT	LC-ESI-ITFT
Instrument	Q-Exactive Orbitrap Thermo Scientific	Q-Exactive Orbitrap Thermo Scientific	Q-Exactive Orbitrap Thermo Scientific
Ion_mode	POSITIVE	POSITIVE	POSITIVE
Collision_energy	30(NCE)	10(NCE)	30(NCE)
Formula	C16H17NO4	C10H12O4	C18H22N2O2
MW	287	196	298
ExactMass	287.11575	196.07355	298.16813
Comments	Parent=288.1225	Parent=197.0803	Parent=299.1749
Splash	splash10-004j-1940000000-b1bd14eb30f6afd2739e	splash10-0a4i-0900000000-6881f47b296a28ed74f6	splash10-000b-0090000000-61fb8ffc68987b8be791
Num Peaks	8	4	6
Peaks	[91.0542, 245.0, 125.0233, 999.0, 154.0499, 80.0, 155.0577, 355.0, 185.0961, 349.0, 200.107, 45.0, 229.0859, 142.0, 246.1125, 734.0]	[85.0284, 54.0, 137.0597, 54.0, 155.0703, 999.0, 197.0808, 323.0]	[144.0808, 37.0, 168.0808, 32.0, 196.1121, 82.0, 208.1121, 54.0, 239.1543, 877.0, 299.1754, 999.0]

Tableau IV.2. Vue d'ensemble des données métabolomiques après le processus de nettoyage.

Name	Pyrophen	Decarestrictine F	Roquefortine A
Formula	C16H17NO4	C10H12O4	C18H22N2O2
MW	287	196	298
ExactMass	287.11575	196.07355	298.16813
Num Peaks	8	4	6
Peaks	[91.0542, 245.0, 125.0233, 999.0, 154.0499, 80.0, 155.0577, 355.0, 185.0961, 349.0, 200.107, 45.0, 229.0859, 142.0, 246.1125, 734.0]	[85.0284, 54.0, 137.0597, 54.0, 155.0703, 999.0, 197.0808, 323.0]	[144.0808, 37.0, 168.0808, 32.0, 196.1121, 82.0, 208.1121, 54.0, 239.1543, 877.0, 299.1754, 999.0]

Tableau IV.3. Vue d'ensemble des données métabolomiques après transformation et encodage.

Name	Pyrophen	Decarestrictine F	Roquefortine A
Formula	C16H17NO4	C10H12O4	C18H22N2O2
MW	287	196	298
ExactMass	287.11575	196.07355	298.16813
Num Peaks	8	4	6
Peaks	[91.0542, 245.0, 125.0233, 999.0, 154.0499, 80.0, 155.0577, 355.0, 185.0961, 349.0, 200.107, 45.0, 229.0859, 142.0, 246.1125, 734.0]	[85.0284, 54.0, 137.0597, 54.0, 155.0703, 999.0, 197.0808, 323.0]	[144.0808, 37.0, 168.0808, 32.0, 196.1121, 82.0, 208.1121, 54.0, 239.1543, 877.0, 299.1754, 999.0]
EncodedName	19645	11637	19834
EncodedFormula	2515	92	3259
ScaledEncodedName	1.136009337	-0.234208837	1.168348403
ScaledEncodedFormula	-0.629818142	-1.563845754	-0.343018083
ProcessedPeaksNumerical	91.0542,245.0,125.0233,999.0,154.0499,80.0,155.0577,355.0,185.0961,349.0,200.107,45.0,229.0859,142.0,246.1125,734.0	85.0284,54.0,137.0597,54.0,155.0703,999.0,197.0808,323.0	144.0808,37.0,168.0808,32.0,196.1121,82.0,208.1121,54.0,239.1543,877.0,299.1754,999.0

IV.4. Choix du Modèle

Les analyses exploratoires réalisées sur le jeu de données transformé nous ont permis de mieux comprendre la distribution et les relations entre les différentes caractéristiques. Les résultats obtenus à partir des histogrammes et de la matrice de corrélation fournissent des indications précieuses pour le choix du modèle d'apprentissage automatique (Fig. VI.1).

IV.4.1. Distributions des Caractéristiques

Les histogrammes des caractéristiques *MW*, *ExactMass*, *Num Peaks*, et *MeanIntensity* montrent des distributions asymétriques avec une concentration de valeurs dans les plages inférieures. Cette asymétrie suggère que les modèles robustes aux distributions biaisées, tels que les forêts aléatoires (Random Forests) ou les modèles de gradient boosting, pourraient être appropriés.

Les caractéristiques encodées et normalisées (*EncodedName*, *EncodedFormula*, *ScaledEncodedName*, *ScaledEncodedFormula*) montrent une distribution centrée autour de la moyenne, indiquant que la normalisation a été efficace pour mettre les différentes caractéristiques sur une échelle comparable.

IV.4.2. Relations entre les Caractéristiques

La matrice de corrélation, visualisée via une carte thermique, révèle des corrélations significatives entre certaines caractéristiques. Les fortes corrélations entre *MW* et *ExactMass*, par exemple, indiquent une redondance potentielle qui pourrait être traitée par des techniques de réduction de dimensionnalité telles que l'analyse en composantes principales (PCA) avant l'entraînement du modèle.

Les caractéristiques montrant de faibles corrélations les unes avec les autres suggèrent que chaque caractéristique apporte une information unique, ce qui est favorable pour les modèles qui bénéficient de la diversité des données, tels que les réseaux de neurones.

IV.4.3. Recommandations de Modèles

Les résultats des analyses exploratoires indiquent que des modèles tels que :

- Forêts Aléatoires (Random Forests) : Ces modèles sont robustes aux distributions biaisées et peuvent gérer des ensembles de données avec de nombreuses caractéristiques. Ils sont également efficaces pour gérer les données avec des relations non linéaires et peuvent offrir une bonne performance de classification.
- LightGBM : Une version optimisée du gradient boosting, LightGBM utilise des techniques innovantes pour améliorer la vitesse et l'efficacité, ce qui le rend particulièrement adapté

aux grandes quantités de données et aux tâches avec de nombreuses caractéristiques, offrant à la fois rapidité et haute précision.

- Réseaux de Neurones : Pour des données avec des caractéristiques diversifiées et de grandes dimensions, les réseaux de neurones peuvent capturer des relations complexes et non linéaires entre les caractéristiques, offrant une flexibilité et une puissance de modélisation élevées.
- Gradient Boosting : Les modèles de gradient boosting, comme XGBoost, peuvent également gérer des données biaisées et offrir une performance élevée en se concentrant sur les erreurs de prédiction des modèles précédents.
- Ces modèles devraient être testés et évalués pour déterminer lequel offre la meilleure performance en termes de précision et de robustesse pour nos objectifs spécifiques d'apprentissage automatique.

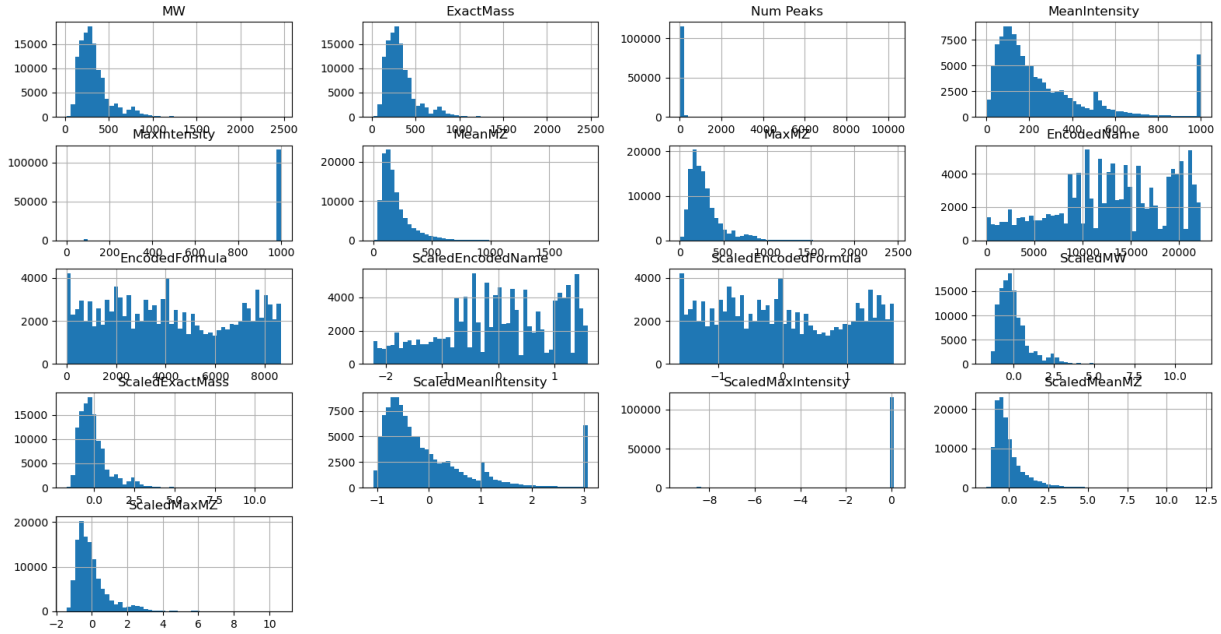

Figure IV.1. Distributions des caractéristiques des données métabolomiques transformées et normalisées pour l'analyse exploratoire et la sélection de modèles d'apprentissage automatique.

IV.5. Entraînement du Modèle

Lors des premières étapes de notre expérimentation, nous avons tenté d'utiliser des modèles de *Forêt Aléatoire* et *LightGBM* pour entraîner notre modèle d'identification de métabolites. Malheureusement, ces tentatives ont été entravées par des problèmes de surcharge du système et de saturation de la RAM. Les caractéristiques complexes et volumineuses de notre jeu de données ont exercé une pression considérable sur les ressources du système, entraînant des arrêts prématurés du processus d'entraînement. Ces résultats soulignent l'importance de l'optimisation des ressources informatiques lors de la manipulation de grands ensembles de données et de modèles d'apprentissage automatique sophistiqués.

En réponse aux défis rencontrés avec les modèles de *Forêt Aléatoire* et *LightGBM*, nous avons dirigé notre attention vers l'utilisation d'un *réseau de neurones* multi-sorties. Contrairement aux tentatives précédentes, le modèle de réseau de neurones a été entraîné avec succès sans rencontrer de problèmes de surcharge du système.

IV.6. Évaluation et Optimisation du Modèle

Lors de l'évaluation du modèle, il est apparu que toutes les prédictions générées par le modèle étaient incorrectes. En conséquence, il n'a pas été possible de calculer la précision en valeur exact, qui est définie comme le ratio des prédictions correctes par rapport au nombre total de prédictions. Plusieurs facteurs peuvent expliquer cette incapacité à obtenir des résultats précis, Les pics étaient écrits sous forme de liste, ce qui a rendu leur conversion en format numérique très difficile, La méthode utilisée pour le prétraitement et la conversion des pics en format numérique peut avoir introduit des erreurs importantes. Si les caractéristiques essentielles des pics n'ont pas été préservées correctement, le modèle ne dispose pas des informations nécessaires pour faire des prédictions exactes.

IV.7. Déploiement du Modèle

L'interface graphique utilisateur (GUI) développée en utilisant *Tkinter* a démontré son efficacité et sa convivialité lors des tests. Nous avons pu charger des fichiers CSV, exécuter des prédictions et visualiser les résultats sans difficulté majeure. La simplicité et l'intuitivité de la GUI ont été largement appréciées, permettant même aux utilisateurs sans expertise technique de naviguer facilement à travers les différentes fonctionnalités (Fig. IV.2).

Le chargement des fichiers nécessaires, incluant le modèle, le scaler, et les encodeurs, s'est déroulé sans accroc. La fonction de prétraitement des pics a converti avec succès les données des pics en listes de valeurs numériques, garantissant une longueur uniforme grâce au *padding*. Ce

prétraitement a été crucial pour assurer que les données d'entrée soient dans le format approprié pour les prédictions du modèle.

La fonction de téléchargement de fichier a permis de traiter efficacement des fichiers CSV contenant des données de plusieurs molécules. Le modèle a pu prédire les noms et formules de chaque ligne du fichier, rendant le processus de prédiction massivement parallèle et rapide. Nous avons pu observer les résultats en temps réel, ce qui a considérablement amélioré l'efficacité des analyses.

L'ajout de la fonction d'explication des caractéristiques a été un atout précieux, fournissant aux utilisateurs des informations détaillées sur chaque caractéristique utilisée par le modèle. Cette transparence a permis de mieux comprendre le processus de prédiction et les facteurs influençant les résultats, renforçant ainsi la confiance dans les prédictions du modèle.

La configuration de la GUI avec des champs d'entrée et des boutons pour les différentes fonctionnalités a permis une navigation fluide et intuitive. Les tests utilisateurs ont montré que la disposition claire et les instructions explicites ont permis de réduire les erreurs d'utilisation et d'augmenter la satisfaction des utilisateurs.

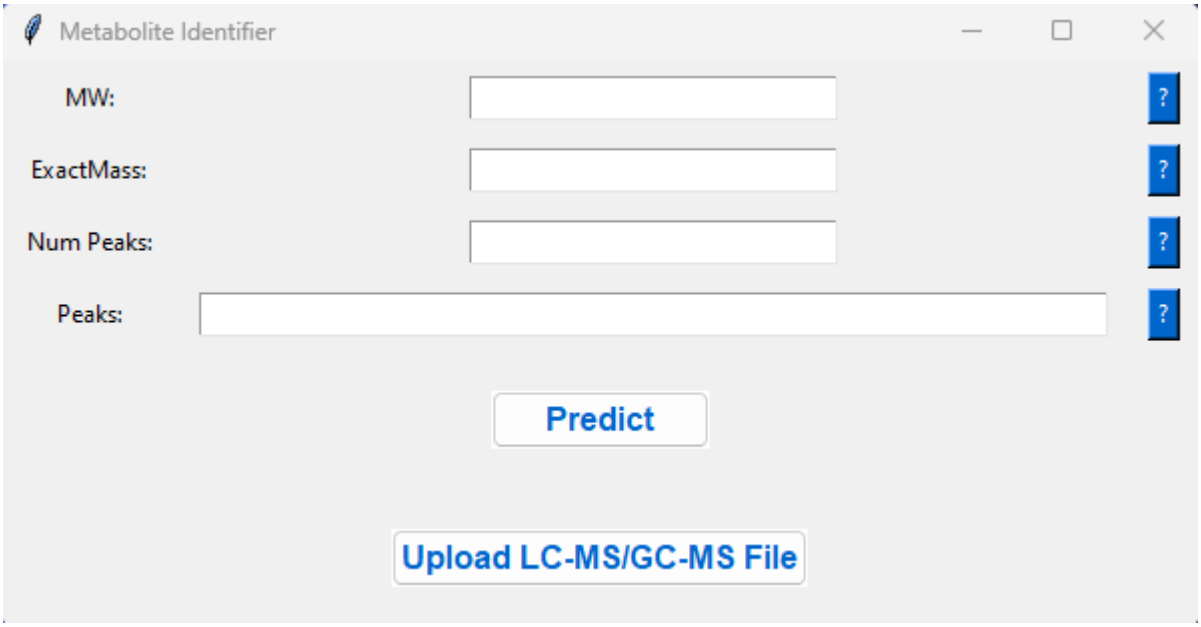

Figure IV.2. Interface graphique utilisateur (GUI) pour l'identification des métabolites par prédiction des noms et formules à partir des données de spectrométrie de masse.

Une revue par Pomyen et al. (2020) met en lumière les applications croissantes de l'apprentissage profond en métabolomique, notant que les techniques d'apprentissage profond, en particulier les réseaux neuronaux convolutifs (CNNs), ont montré un potentiel significatif dans le prétraitement des données et l'identification des structures. Pomyen et al. (2020) identifient

également plusieurs défis qui doivent être relevés pour que l'apprentissage profond soit appliqué plus efficacement à la métabolomique, y compris la nécessité d'architectures spécifiques au métabolome, la gestion des problèmes de dimensionnalité et l'établissement de régimes d'évaluation de modèles robustes.

De plus, Hall et al. (2015) ont évalué l'optimisation des paramètres de modélisation des réseaux de neurones artificiels (ANN) en utilisant des données d'indice de rétention (RI) pour 390 composés. Ils ont évalué quatre paramètres de modélisation des ANN : l'annulation du taux d'apprentissage, les critères d'arrêt, la méthode de division des données et l'architecture du réseau. L'étude a révélé que le meilleur modèle, construit en utilisant une division des données par clustering de Ward et une architecture de réseau minimement non linéaire.

Enfin, une étude par Bonetta et al. (2023) discute de l'utilisation de l'apprentissage automatique pour l'analyse des voies métaboliques et explique comment construire et entraîner un modèle de réseau de neurones profond pour effectuer une prédiction multi-étiquettes des composés (métabolites) vers leurs classes respectives de voies métaboliques. Ils détaillent le processus étape par étape, y compris la collecte de données, l'ingénierie des caractéristiques, la sélection du modèle, l'entraînement et l'évaluation. Ce processus de construction et d'évaluation de modèles peut facilement être transféré à d'autres domaines d'intérêt.

Les futures itérations de notre modèle pourraient bénéficier de l'incorporation de telles techniques d'optimisation et de ces perspectives, améliorant potentiellement les performances en capturant mieux la variabilité biologique sous-jacente.

Conclusion et Perspectives

les leçons tirées de cette recherche ouvrent la voie à des améliorations significatives et des avancées futures dans le domaine de l'identification des métabolites par l'intelligence artificielle. Cette étude a permis de mettre en évidence les défis et les opportunités dans l'application des réseaux de neurones à la métabolomique, offrant une base solide pour des recherches futures. Avec des ajustements et des optimisations appropriés, ce travail a le potentiel de fournir des outils précieux pour la recherche en métabolomique et au-delà, contribuant ainsi à une meilleure compréhension des processus biologiques et à des avancées dans le diagnostic et le traitement des maladies.

Pour les travaux futurs, plusieurs axes d'amélioration et de recherche sont envisageables :

- Optimisation des Algorithmes de Prétraitement : Il est crucial de développer et tester des algorithmes de prétraitement plus sophistiqués pour convertir les données des pics de

manière à préserver leurs caractéristiques essentielles. Cela inclut l'exploration de nouvelles méthodes de remplissage (padding) et de mise à l'échelle des pics. L'objectif est de s'assurer que les données d'entrée du modèle conservent les informations pertinentes nécessaires pour une identification précise.

- Optimisation des Hyperparamètres : Un ajustement plus fin des hyperparamètres du modèle de réseau de neurones pourrait améliorer ses performances. Des techniques telles que la recherche en grille (*grid search*) ou l'optimisation bayésienne pourraient être employées pour trouver les combinaisons d'hyperparamètres optimales qui maximisent la précision du modèle.
- Exploration de Nouvelles Architectures de Modèles : L'évaluation d'autres architectures de réseaux de neurones, telles que les réseaux neuronaux convolutifs (CNN) ou les réseaux de neurones récurrents (RNN), pourrait offrir des avantages en termes de capture des caractéristiques complexes des données de métabolomique. Ces architectures avancées sont connues pour leur capacité à extraire des caractéristiques pertinentes à partir de données complexes et peuvent potentiellement améliorer les performances du modèle.
- Collaboration Interdisciplinaire : La collaboration avec des experts en biologie, chimie analytique et informatique pourrait apporter des perspectives précieuses et des solutions innovantes pour les défis rencontrés. Travailler avec des spécialistes de différents domaines peut permettre de combler les lacunes dans les connaissances et d'apporter des solutions intégrées aux problèmes complexes.

Références Bibliographiques

IV.8. Limites de l'Étude

Bien que cette recherche ait exploré avec succès le développement et le déploiement d'un modèle d'intelligence artificielle pour l'identification des métabolites, plusieurs limites importantes ont été identifiées. Premièrement, la méthode de prétraitement et de conversion des pics en format numérique n'a pas réussi à capturer toutes les caractéristiques essentielles nécessaires pour une identification précise, ce qui a entraîné une faible précision des prédictions. Deuxièmement, le temps et les ressources nécessaires pour affiner les algorithmes de prétraitement et optimiser les hyperparamètres du modèle étaient considérables, ce qui a restreint l'étendue des expériences réalisées dans cette étude. Enfin, bien que la GUI ait facilité l'interaction utilisateur, elle pourrait bénéficier de techniques de visualisation plus avancées pour améliorer l'interprétation des résultats.

Conclusion

Références bibliographiques

Al-Bukhaiti, W., & Al-Farga, A. (2017). Gas Chromatography: Principles, Advantages and Applications in Food Analysis. *International Journal of Science Innovations and Discoveries* .

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

Beale, D. J., Jones, O. A., Karpe, A. V., Dayalan, S., Oh, D. Y., Kouremenos, K. A., Ahmed, W., & Palombo, E. A. (2016). A Review of Analytical Techniques and Their Application in Disease Diagnosis in Breathomics and Salivaomics Research. *Int J Mol Sci*, *18*(1). https://doi.org/10.3390/ijms18010024

Buttazzo, G. (2023). Rise of artificial general intelligence: risks and opportunities. *Front Artif Intell*, *6*, 1226990. https://doi.org/10.3389/frai.2023.1226990

Cardoso Rial, R. (2024). AI in analytical chemistry: Advancements, challenges, and future directions. *Talanta*, *274*, 125949. https://doi.org/https://doi.org/10.1016/j.talanta.2024.125949

Celeghin, A., Borriero, A., Orsenigo, D., Diano, M., Méndez Guerrero, C. A., Perotti, A., Petri, G., & Tamietto, M. (2023). Convolutional neural networks for vision neuroscience: significance, developments, and outstanding issues. *Front Comput Neurosci* ,17 , .1153572https://doi.org/10.3389/fncom.2023.1153572

Das, L., Sivaram, A., & Venkatasubramanian, V. (2020). Hidden Representations in Deep Neural Networks: Part 2. Regression Problems. *Computers & Chemical Engineering*, *139*, 106895. https://doi.org/10/1016.j.compchemeng.2020.106895

Das, S., Tariq, A., Santos, T., Kantareddy, S. S., & Banerjee, I. (2023). Recurrent Neural Networks (RNNs): Architectures, Training Tricks, and Introduction to Influential Research. In O. Colliot (Ed.), *Machine Learning for Brain Disorders* (pp. 117-138). Humana

Copyright 2023, The Author(s). https://doi.org/10.1007/978-1-0716-3195-9_4

Degano, I. (2019). Liquid chromatography: Current applications in Heritage Science and recent developments. *4*(5). https://doi.org/doi:10.151/5psr-2018-0009 (Physical Sciences Reviews)

Donatti, A., Canto, A. M., Godoi, A. B., da Rosa, D. C., & Lopes-Cendes, I. (2020). Circulating Metabolites as Potential Biomarkers for Neurological Disorders-Metabolites in Neurological Disorders. *Metabolites* .(10)10 ,https://doi.org/10.3390/metabo10100389

Emwas, A. H., Al-Talla, Z. A., Yang, Y., & Kharbatia, N. M. (2015). Gas chromatography-mass spectrometry of biofluids and extracts. *Methods Mol Biol*, *1277*, 91-112. https://doi.org/10.1007/978-1-4939-2377-9_8

García-Ordás, M., Benítez-Andrades, J., García, I., Benavides, C., & Alaiz Moreton, H. (2020). Detecting Respiratory Pathologies Using Convolutional Neural Networks and Variational Autoencoders for Unbalancing Data. *Sensors*, *20*. https://doi.org/10.3390/s20041214

Gathungu, R. M., Kautz, R., Kristal, B. S., Bird, S. S., & Vouros, P. (2020). The integration of LC-MS and NMR for the analysis of low molecular weight trace analytes in complex matrices. *Mass Spectrom Rev*, *39*(1-2), 35-54. https://doi.org/10.10/02mas.21575

Gbadago, D. Q., Moon, J., Kim, M., & Hwang, S. (2021). A unified framework for the mathematical modelling, predictive analysis, and optimization of reaction systems using computational fluid dynamics, deep neural network and genetic algorithm: A case of butadiene synthesis. *Chemical Engineering Journal*, *409*, 128163. https://doi.org/https://doi.org/10.1016/j.cej.2020.128163

Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2022). A guide to machine learning for biologists. *Nat Rev Mol Cell Biol*, *23*(1), 40-55. https://doi.org/10.1038/s41580-021-00407-0

Hanson, L. (2008). Is Quantum Mechanics Necessary for Understanding Magnetic. *Concepts Magn. Reson. Part A*, *32*. https://doi.org/10.1002/cmr.a.20123

Jakhar, D., & Kaur, I .(2020) .Artificial intelligence, machine learning and deep learning: definitions and differences. *Clinical and Experimental Dermatology*, *45*(1), 131-132. https://doi.org/10.1111/ced.14029

Jiang, Y., Luo, J., Huang, D., Liu, Y., & Li, D. D. (2022). Machine Learning Advances in Microbiology: A Review of Methods and Applications. *Front Microbiol*, *13*, 925454. https://doi.org/10.3389/fmicb.2022.925454

Karabagias, I. K. (2020). Advances of Spectrometric Techniques in Food Analysis and Food Authentication Implemented with Chemometrics. *Foods*, *9*(11). https://doi.org/10.3390/foods9111550

Kaur, G., & Sharma, S. (2018). Gas Chromatography -A Brief Review .

Korfmacher, W. A. (2005). Principles and applications of LC-MS in new drug discovery. *Drug Discov Today*, *10*(20), 1357 .1367-https://doi.org/10.1016/s1359-6446(05)03620-2

Kumar, R., Bohra, A., Pandey, A. K., Pandey, M. K., & Kumar, A. (2017). Metabolomics for Plant Improvement: Status and Prospects. *Front Plant Sci*, *8*, 1302. https://doi.org/10.3389/fpls.2017.01302

Lindley, S. E., Lu, Y., & Shukla, D. (2024). The Experimentalist's Guide to Machine Learning for Small Molecule Design. *ACS Appl Bio Mater*, *7*(2), 657-684. https://doi.org/10.1021/acsabm.3c00054

Mahum, R., Irtaza, A., Nawaz, M., Nazir, T., Masood, M., Shaikh ,S., & Abouel Nasr, E. (2022). A robust framework to generate surveillance video summaries using combination of zernike moments and r-transform and deep neural network. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-022-13773-4

Markley ,J. L., Brüschweiler, R., Edison, A. S., Eghbalnia, H. R., Powers, R., Raftery, D., & Wishart, D. S. (2017). The future of NMR-based metabolomics. *Curr Opin Biotechnol*, *43*, 34-40. https://doi.org/10.1016/j.copbio.2016.08.001

Medina, S., Perestrelo, R ,.Silva, P., Pereira, J., & Câmara, J. (2019). Current trends and recent advances on food authenticity technologies and chemometric approaches. *Trends in Food Science & Technology*, *85*. https://doi.org/10.1016/j.tifs.2019.01.017

Mintz, Y., & Brodie, R. (201 .(9Introduction to artificial intelligence in medicine. *Minim Invasive Ther Allied Technol*, *28*(2), 73-81. https://doi.org/10.1080/13645706.2019.1575882

Molavian, R., Fatahi, A., Abbasi, H., & Khezri, D. (2023). Artificial Intelligence Approach in Biomechanics of Gait and Sport: A Systematic Literature Review. *J Biomed Phys Eng*, *13*(5), 383-402. https://doi.org/10.31661/jbpe.v0i0.2305-1621

Nichols, J. A., Herbert Chan, H. W., & Baker, M. A. B. (2019). Machine learning: applications of artificial intelligence to imaging and diagnosis. *Biophys Rev*, *11*(1), 111-118. https://doi.org/10.1007/s12551-018-0449-9

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

Patrizi, B., Cumis, S., Viciani, & D'Amato. (2019). Dioxin and Related Compound Detection: Perspectives for Optical Monitoring. *International Journal of Molecular Sciences*, *20*, 2671. https://doi.org/10.3390/ijms20112671

Rajawat, J., & Jhingan, G. (2019). Mass spectroscopy. In *Data processing handbook for complex biological data sources* (pp. 1-20). Elsevier .

Rodrigues, J. A., Krois, J., & Schwendicke, F. (2021). Demystifying artificial intelligence and deep learning in dentistry. *Braz Oral Res*, *35*, e094. https://doi.org/10.1590/1807-3107bor-2021.vol35.0094

Rouessac, F., Rouessac, A., & Cruch??, D. (2004). *Analyse chimique*. Dunod .

Strathmann, F. G., & Hoofnagle, A. N. (2011). Current and future applications of mass spectrometry to the clinical laboratory. *Am J Clin Pathol*, *136*(4), 609-616. https://doi.org/10.1309/ajcpw0ta8obbngck

Theodosiou, A. A., & Read, R. C. (2023). Artificial intelligence, machine learning and deep learning: Potential resources for the infection clinician. *J Infect*, *87*(4), 287-294. https://doi.org/10.1016/j.jinf.2023.07.006

Thompson, J. M. (2018). *Infrared spectroscopy*. Jenny Stanford Publishing .

Xu, M., Tang, Z., Duan, Y., & Liu, Y. (2016). GC-Based Techniques for Breath Analysis: Current Status, Challenges, and Prospects. *Crit Rev Anal Chem*, *46*(4), 291-304. https://doi.org/10.1080/10408347.2015.1055550

Zagorchev, L., Seal, C. E., Kranner, I., & Odjakova, M. (2013 .(A central role for thiols in plant tolerance to abiotic stress. *Int J Mol Sci*, *14*(4), 7405-7432. https://doi.org/10.3390/ijms14047405

Zhang, H., & Moon, S. K. (2021). Reviews on Machine Learning Approaches for Process Optimization in Noncontact Direct Ink Writing. *ACS Appl Mater Interfaces*, *13*(45), 53323-53345. https://doi.org/10.1021/acsami.1c04544

Zhu, L.-T., Chen, X., Ouyang, B., Yan, W.-C., Lei, H., Chen, Z., & Zheng-hong, L. (2022). Review of Machine Learning for Hydrodynamics, Transport, and Reactions in Multiphase Flows and Reactors. *Industrial & Engineering Chemistry Research*, *61*. https://doi.org/10.1021/acs.iecr.2c01036