الجمهورية الجزائرية الديمقراطية الشعبية

RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE

وزارة التعليم العالي و البحث العلمي

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

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

Université Frères Mentouri Constantine 1 — جامعة الإخوة منتوري قسنطينة 1

Faculté des Sciences de la Nature et de la Vie — كلية علوم الطبيعة والحياة

Département de Biologie Appliquée — قسم البيولوجيا التطبيقية

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

Domaine : Sciences de la Nature et de la Vie

Filière : Sciences biologiques

Spécialité : *Bioinformatique*

N° d'ordre :

N° de série :

Intitulé :

Deep learning pour la classification taxonomique des bactéries causant des infections nosocomiales à partir des données génomiques.

Présenté par : **BOUGUENDOURA Zakaria** et **MESSELEM Anis** Le 18/06/2023

Jury d'évaluation :

Président : HAMIDECHI Mohamed Abdelhafid (Professeur, Université Frères Mentouri Constantine 1).

Encadreur : DAAS Mohamed Skander (MCA, Université Frères Mentouri Constantine 1).

Examinatrice : DJAMAA Ouahiba (MCB, Université Frères Mentouri, Constantine 1).

Année universitaire

2022- 2023

Remerciements

Avant tout , nous remercions ALLAH : le tout Miséricordieux , l'unique , le puissant , Maitre des cieux et de la terre pour nous avoir guidé , protégé , aidé et nous a permis de mener à bien ce travail .

C'est avec un grand plaisir que nous réservons ces lignes en signe de gratitude et de reconnaissance à ceux qui ont contribué de près ou de loin à l'élaboration de ce travail :

Notre responsable monsieur Dr. DAAS Mohamed Skander encadreur de notre mémoire, merci pour son aide, la correction du manuscrit, et pour sa patience.

Ensuit nous tenons à remercier les membres du jury Pr. HAMIDECHI Mohamed Abdelhafid et Dr. DJAMAA Ouahiba pour avoir pris le temps d'évaluer ce mémoire.

Nous n'oublierons pas de remercier tous ceux qui nous ont soutenu et encouragé tout au long de la réalisation de ce travail.

Merci à tous.

Dédicace

Je tiens personnellement à remercier mon collègue et ami Anis avec qui j'ai pris beaucoup de plaisir à travailler. Nous avons formé une équipe formidable. Enfin, je tiens à remercier ma famille pour leur soutien, leur amour et leur encouragement constant. En particulier, je remercie ma mère, mon père, mes frères et ma sœur, qui ont toujours été là pour moi dans toutes les épreuves. "Zakarya"

Je dédie ce mémoire à Zakarya, mon partenaire de mémoire, mon collègue et mon ami avec qui j'ai beaucoup apprécié travailler. Je te remercie pour ton soutien et le lien spécial qui s'est créé entre nous ... et à mes chers parents, ma mère et mon père et ma grande sœur, pour leur patience, leur amour et leur soutien. "Anis"

RÉSUMÉ

Les infections nosocomiales constituent une menace importante pour la santé publique, nécessitant le développement de méthodes précises et efficaces pour leur identification et leur classification. Ce mémoire étudie la classification taxonomique des bactéries responsables d'infections nosocomiales en utilisant des techniques du Deep Learning sur des données de séquençage de l'ARNr 16S. Les méthodes traditionnelles d'identification bactérienne n'ont pas la résolution nécessaire pour classer avec précision les bactéries au niveau de l'espèce. Un modèle d'apprentissage en profondeur est développé et les performances du modèle sont évaluées sur les quatre espèces bactériennes couramment associées (*Escherichia coli, Enterococcus faecalis, Klebsiella pneumoniae et Pseudomonas aeruginosa*) aux infections nosocomiales. Les résultats démontrent la robustesse et la précision du modèle dans la classification taxonomique de ces agents pathogènes. Les résultats de cette recherche contribueront au développement de stratégies de diagnostic et de surveillance plus efficaces, permettant des interventions ciblées pour atténuer l'impact des infections nosocomiales sur les soins aux patients et les systèmes de santé.

Mots clés : Infections nosocomiales, Classification taxonomique, Données génomiques, Apprentissage automatique, Pathogènes bactériens, ARNr 16s.

ABSTRACT

Nosocomial infections pose a significant threat to public health, requiring the development of accurate and efficient methods for their identification and classification. This thesis studies the taxonomic classification of bacteria responsible for nosocomial infections using Deep Learning techniques on 16S rRNA sequencing data. Traditional bacterial identification methods lack the resolution necessary to accurately classify bacteria to the species level. A deep learning model is developed and the performance of the model is evaluated on the four bacterial species commonly associated (*Escherichia coli, Enterococcus faecalis, Klebsiella pneumoniae and Pseudomonas aeruginosa*) with nosocomial infections. The results demonstrate the robustness and accuracy of the model in the taxonomic classification of these pathogens. The results of this research will contribute to the development of more effective diagnostic and surveillance strategies, enabling targeted interventions to mitigate the impact of nosocomial infections on patient care and health systems.

Keywords : Nosocomial infections, Taxonomic classification, Genomic data, Machine learning, Bacterial pathogens, 16s rRNA.

الملخص

تشكل عدوى المستشفيات تهديدًا كبيرًا للصحة العامة ، مما يتطلب تطوير طرق دقيقة وفعالة لتحديدها وتصنيفها. تدرس هذه الأطروحة التصنيف التصنيفي للبكتيريا المسؤولة عن عدوى المستشفيات باستخدام تقنيات التعلم العميق على بيانات تسلسل 16s rRNA. تفتقر طرق تحديد البكتيريا التقليدية إلى الدقة اللازمة لتصنيف البكتيريا بدقة إلى مستوى الأنواع. تم تطوير نموذج التعلم العميق وتقييم أداء النموذج على الأنواع البكتيرية الأربعة المرتبطة بشكل شائع (*Escherichia coli* و *Enterococcus faecalis* و *Klebsiella pneumoniae* و *Pseudomonas aeruginosa*) مع عدوى المستشفيات. تظهر النتائج متانة ودقة النموذج في التصنيف التصنيفي لمسببات الأمراض. ستساهم نتائج هذا البحث في تطوير استراتيجيات تشخيص ومراقبة أكثر فعالية ، مما يتيح التدخلات المستهدفة للتخفيف من تأثير عدوى المستشفيات على رعاية المرضى والأنظمة الصحية.

الكلمات المفتاحية: عدوى المستشفيات ، التصنيف التصنيفي ، البيانات الجينومية ، التعلم الآلي ، مسببات الأمراض البكتيرية ، 16s rRNA.

LISTE DES FIGURES

LISTE DES TABLEAUX

ACRONYMES :

- ADN : Acide Désoxyribonucléique
- ANN : Artificial Neural Network
- ARN : Acide Ribonucléique
- ARNm : Acide Ribonucléique messager
- ARNr : Acide Ribonucléique ribosomique
- ARNr : 16S – ARN ribosomique 16S
- ARNt : Acide Ribonucléique de transfert
- CDC : Centers for Diseases Control Prevention (Centers for Disease Control and Prevention)
- CNN : Convolutional Neural Networks (Réseaux de Neurones Convolutifs)
- CSV : Comma-Separated Values
- DL : Deep Learning
- FNN : Feedforward Neural Networks (Réseaux de Neurones à Propagation Avant)
- IA : Intelligence Artificielle
- ML : Machine Learning
- PCR : Polymerase Chain Reaction
- ReLU : RectifiedLinear Unit (activation function)
- RNN : Recurrent Neural Networks (Réseaux de Neurones Récurrents)

TABLE DES MATIÈRES

Table des matières

CHAPITRE 2 : DEEP LEARNING (APPRENTISSAGE PROFOND)

CHAPITRE 3 : MATÉRIEL ET MÉTHODES

Introduction générale

Introduction générale

L'évolution de la recherche scientifique dans le domaine de la microbiologie permet aujourd'hui d'identifier les bactéries responsables des infections nosocomiales, c'est-à-dire des infections contractées dans les établissements de santé. Ces infections sont souvent causées par des bactéries multi-résistantes aux antibiotiques, ce qui pose un défi majeur pour la prise en charge des patients et la lutte contre ces infections. Une classification taxonomique plus précise de ces bactéries est essentielle pour comprendre leur comportement, leur épidémiologie et pour adapter les traitements antimicrobiens de manière appropriée.

L'intelligence artificielle, et en particulier le Deep Learning, se révèle être un outil puissant pour les problèmes de classification. Les algorithmes d'apprentissage profond peuvent analyser de grandes quantités de données microbiologiques, notamment les séquences génétiques des bactéries, pour identifier de manière rapide et précise les espèces bactériennes responsables des infections. Cela permettrait d'améliorer le diagnostic microbiologique des infections nosocomiales, d'optimiser les traitements antimicrobiens et de mettre en place des mesures de prévention efficaces pour contrôler la propagation de ces bactéries dans les établissements de santé.

Dans ce mémoire, notre objectif est d'apporter une contribution en utilisant l'approche du Deep Learning pour la classification taxonomique des bactéries causant des infections nosocomiales. Nous utiliserons des données microbiologiques, notamment les séquences génétiques des bactéries, pour entraîner un modèle d'apprentissage profond capable de reconnaître et de classer les espèces bactériennes responsables des infections nosocomiales avec une haute précision. Notre approche vise à faciliter et à accélérer le diagnostic microbiologique des infections nosocomiales, ce qui pourrait avoir un impact significatif sur la prise en charge des patients et sur la lutte contre ces infections dans les établissements de santé.

Ce mémoire est organisé comme suit : dans le premier chapitre nous présentons quelques concepts de base de la biologie moléculaire et la classification taxonomique des bactéries d'infections nosocomiales. Ensuite, une présentation des approches de l'apprentissage profond sont discutées dans le deuxième chapitre. La partie matériels et méthodes est présentée dans le chapitre 3. Les résultats et leur discussion sont présentés dans le dernier chapitre.

Chapitre 1 :

Classification taxonomique des bactéries d'infections nosocomiales

1. Biologie moléculaire

La biologie moléculaire est l'étude des processus de réplication, de transcription et de traduction du matériel génétique. La biologie moléculaire où le matériel génétique est transcrit en ARN, puis traduit en protéines. L'essentiel du travail en biologie moléculaire est quantitatif, et récemment beaucoup de travaux ont été faits à l'intersection de la biologie moléculaire et de l'informatique, dans la bioinformatique et dans la biologie computationnelle. Depuis les années 2000, l'étude de la structure et de la fonction des gènes, la génétique moléculaire, fait partie des sous-domaines les plus saillants de la biologie moléculaire [1].

1.1. Histoire de biologie moléculaire

La biologie moléculaire (parfois abrégée BM) est une discipline scientifique au croisement de la génétique, de la biochimie et de la physique, dont l'objet est la compréhension des mécanismes de fonctionnement de la cellule au niveau moléculaire. Le terme « biologie moléculaire », utilisé la première fois en 1938 par Warren Weaver, désigne également l'ensemble des techniques de manipulation d'acides nucléiques (ADN, ARN), appelées aussi techniques de génie génétique. La biologie moléculaire est apparue au XXe siècle, à la suite de l'élaboration des lois de la génétique, la découverte des chromosomes et l'identification de l'ADN comme support chimique de l'information génétique. après la découverte de la structure en double hélice de l'ADN en 1953 par James Watson (1928-), Francis Crick (1916-2004), Maurice Wilkins (1916-2004) et Rosalind Franklin (1920-1958) la biologie moléculaire a connu d'importants développements pour devenir un outil incontournable de la biologie moderne à partir des années 1970 [1].

1.2. ADN et sa structure

L'ADN est une chaine résultante de la polymérisation de nucléotides triphosphate porteurs des quatre types de bases. La polymérisation conduit à la perte de deux phosphates et les nucléotides sont liés par le biais du phosphate restant. Plus précisément, le phosphate en 5' d'un nucléotide est attaché par une liaison phosphodiester au groupement hydroxyle en 3du nucléotide suivant. Le polynucléotide a donc un 5 phosphate libre à une extrémité (son

extrémité 5) et un 3 OH libre à l'autre extrémité (son extrémité 3'). C'est la séquence des bases ordre de leur enchainement) qui détermine l'information génétique codée (voir Figure 1) [2].

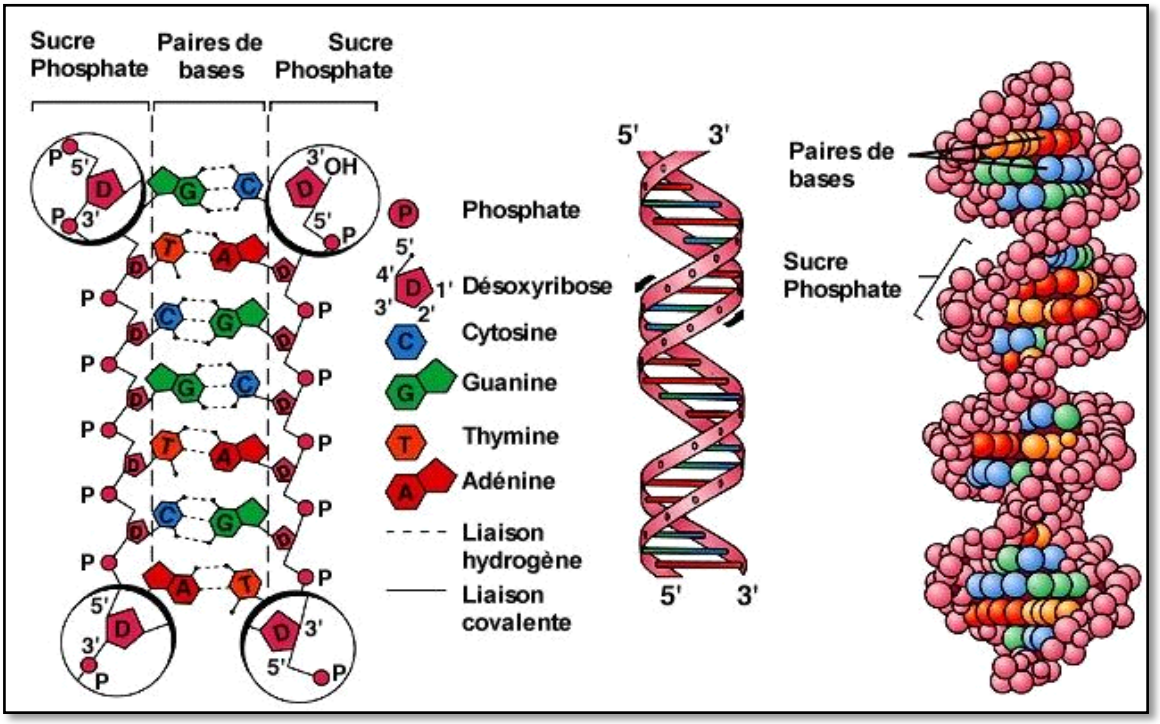

Figure 1 : Structure d'ADN [3].

1.3. ARN et sa structure :

L'ARN, ou acide ribonucléique, est une molécule présente dans les cellules vivantes qui joue un rôle essentiel dans la transmission et l'expression de l'information génétique. Il est composé de nucléotides, qui sont les unités de base de l'ARN.

La structure de l'ARN est similaire à celle de l'ADN (acide désoxyribonucléique), mais avec quelques différences clés. L'ARN est une chaîne simple plutôt qu'une double hélice, et il utilise l'uracile (U) comme base nucléique au lieu de la thymine (T) présente dans l'ADN. Les nucléotides de l'ARN sont liés entre eux par des liaisons phosphodiester pour former une chaîne linéaire. Les quatre bases nucléiques de l'ARN sont l'adénine (A), la cytosine (C), la guanine (G) et l'uracile (U) (Voir Figure 2) [4].

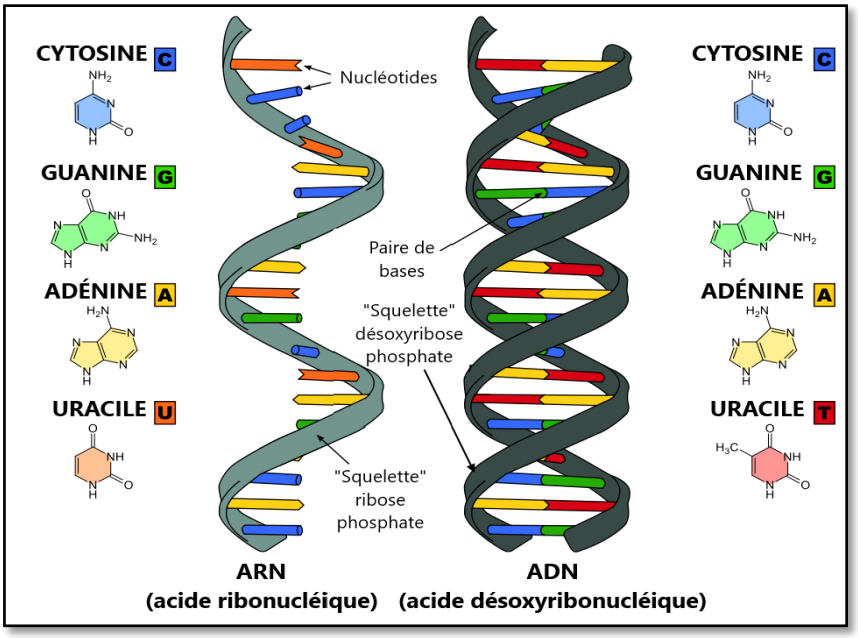

Figure 2 : Structure d'ARN [5].

1.4. Classification d'ARN

La classification des ARN est basée sur les fonctions et les caractéristiques structurelles de ces molécules. Il existe plusieurs types d'ARN qui jouent des rôles différents dans la cellule comme le montre le Tableau 1 :

Tableau 1 : Types d'ARN [6].

Type d'ARN	Description
ARN messager (ARNm)	Transcrit à partir d'un gène et traduit en protéines par les ribosomes
ARN ribosomique (ARNr)	Composant structurant des ribosomes, qui sont les usines de la synthèse protéique
ARN de transfert (ARNt)	Transporte les acides aminés vers les ribosomes pour la construction de la protéine
ARN régulateur	Régule l'expression génique en se liant à l'ADN ou aux ARN messagers

1.5. Gène

Un gène est une unité d'information et il correspond à un segment discret d'ADN qui code la séquence des acides aminés d'une protéine. Les gènes sont dispersés et séparés par des régions inter-géniques d'ADN non codant. L'information est codée sur le brin matrice qui gouverne la synthèse de l'ARN. Les deux brins de l'ADN peuvent jouer ce rôle de matrice. Les molécules d'ADN ont une capacité énorme de stockage d'information (Voir Figure 3).

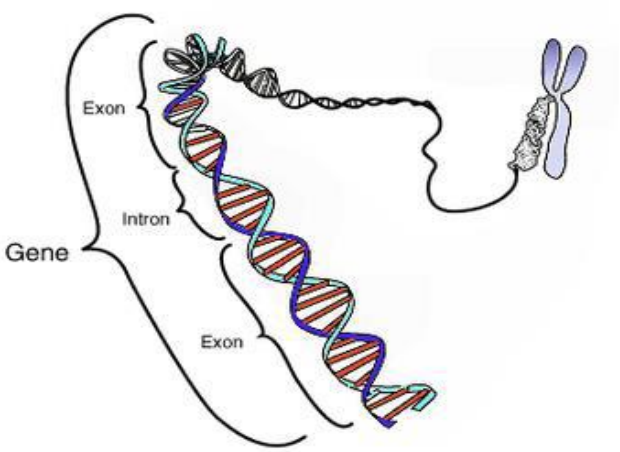

Figure 3 : structure de Gène.

1.6. Génome

C'est l'ensemble complet de l'information génétique d'un organisme, y compris tous ses gènes et les séquences d'ADN non codantes. Le génome contient toutes les instructions nécessaires pour construire et maintenir un organisme, ainsi que pour réguler ses processus biologiques. Chez les organismes à ADN, le génome est composé de molécules d'ADN double brin [6].

1.7. Protéines

Les protéines sont des macromolécules biologiques complexes constituées d'une ou plusieurs chaînes d'acides aminés enroulées en une structure tridimensionnelle spécifique. Elles remplissent de nombreuses fonctions vitales dans les cellules et les organismes, telles que la catalyse de réactions biochimiques, le transport de molécules, la communication cellulaire, la régulation de l'expression génique, la structure cellulaire, la réponse immunitaire, entre autres.

Les acides aminés qui composent les protéines sont liés par des liaisons peptidiques,

formant ainsi une chaîne polypeptidique. Cette chaîne peut ensuite se replier surelle-même pour former une structure tridimensionnelle stable, déterminée par la séquence d'acides aminés. La structure de la protéine est essentielle pour sa fonction et est influencée par de nombreux facteurs tels que les interactions chimiques entre les acides aminés et l'environnement cellulaire. les protéines sont synthétisées dans les cellules par un processus appelé traduction, qui utilise l'ARNm comme modèle pour assembler la chaîne polypeptidique à partir des acides aminés correspondants (Voir Figure 4) [6].

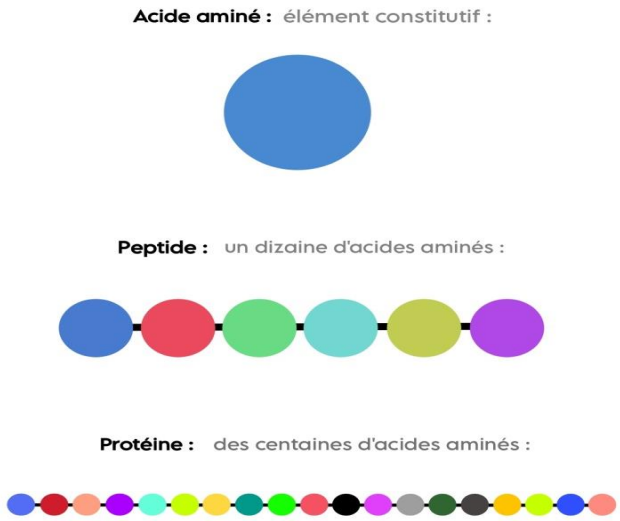

Figure 4 : structure de la protéine.

2. Infections nosocomiales

Les infections nosocomiales, également appelées infections hospitalières, sont des infections contractées pendant un séjour à l'hôpital, qui n'étaient pas présentes ni en incubation lors de l'admission du patient. Les infections qui surviennent plus de 48 heures après l'admission sont généralement considérées comme nosocomiales. Des définitions ont été établies pour identifier les infections nosocomiales de différentes localisations, telles que les infections urinaires et pulmonaires, en utilisant des critères cliniques et biologiques. Ces définitions ont été élaborées par les Centers for Diseases Control Prevention (CDC) aux États-Unis ou lors de conférences internationales pour surveiller environ 50 sites infectieux potentiels [7].

2.1. Types d'infections nosocomiales

Les infections sont très fréquemment liées à des interventions invasives : sondage urinaire ou trachéal (ventilation assistée), cathéter veineux, intervention chirurgicale, endoscopie.

Les infections urinaires sont les plus nombreuses (30%). Elles sont souvent liées à la pose de sondes urinaires mais sont rarement graves. Viennent ensuite les pneumonies (17%) souvent concomitantes à l'intubation et la ventilation assistée, les infections du site opératoire (14%) après une intervention chirurgicale, et les bactériémies/septicémies (10%) liées à l'introduction de cathéters dans les voies sanguines. Des infections de la peau et les tissus mous ou encore des voies respiratoires supérieures sont également observées (29%) (Voir Figure 5) [8].

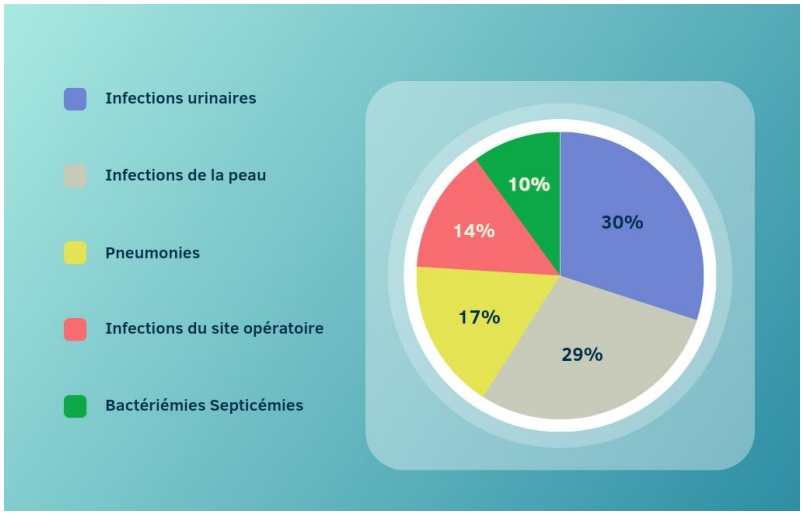

Figure 5 : Types d'infections nosocomiales contractées.

2.2. Transmission des infections nosocomiales

La transmission des infections nosocomiales peut se faire par plusieurs voies [9], notamment:

- Contact direct : lorsqu'un individu en contact direct avec un patient infecté est contaminé par le biais de la peau ou des muqueuses. Cette voie de transmission peut inclure le contact avec des fluides corporels tels que le sang, l'urine ou les selles.
- Contact indirect : lorsqu'un individu est contaminé par le biais d'objets ou de surfaces contaminées par le patient infecté. Cette voie de transmission peut inclure le contact avec des dispositifs médicaux, des vêtements de lit ou des surfaces de la chambre du patient.

- Voie aérienne : lorsque des particules infectieuses sont dispersées dans l'air et inhalées par un individu. Cette voie de transmission peut inclure la toux ou l'éternuement d'un patient infecté ou la manipulation d'objets contaminés qui libèrent des particules dans l'air.

2.3. Bactéries responsables

Il y a plusieurs bactéries qui sont responsables d'infections nosocomiales. Voici quelques exemples de bactéries (décrites dans le Tableau 2) souvent impliquées dans les infections nosocomiales [10].

Tableau 2 : Bactéries provoquant des infections nosocomiales [10].

Bactérie	**Classification taxonomique**	**Habitat naturel**	**Mode de transmission**	**Types d'infections**
Staphylococcus aureus	Genre : *Staphylococcus* Famille : *Staphylococcaceae*	Peau et muqueuses nasales	Contact direct avec la peau ou les sécrétions infectées	Infections des plaies, infections respiratoires, infections du sang
Escherichia coli	Genre : *Escherichia* Famille : *Enterobacteriaceae*	Flore intestinale	Contact direct avec les surfaces contaminées	Infections urinaires
Pseudomonas aeruginosa	Genre : *Pseudomonas* Famille : *Pseudomonadaceae*	Sol, eau, flore intestinale	Contact direct avec les surfaces contaminées ou les dispositifs médicaux	Infections respiratoires, infections des plaies, infections du sang
Klebsiella pneumoniae	Genre : *Klebsiella* Famille : *Enterobacteriaceae*	Flore intestinale	Contact direct avec les matières fécales ou les surfaces contaminées	Infections des voies urinaires, pneumonies, infections du sang

Enterococcus faecium	Genre : *Enterococcus* Famille : *Enterococcaceae*	Flore intestinale	Contact direct avec les matières fécales ou les surfaces contaminées	Infections des voies urinaires, infections du sang

3. Classification taxonomique des bactéries

La classification taxonomique des bactéries est un système hiérarchique qui permet de classer les différentes espèces de bactéries en fonction de leurs caractéristiques physiques, morphologiques, biochimiques et génétiques [11].

Le système de classification taxonomique des bactéries est basé sur la nomenclature binomiale, qui consiste en l'utilisation de deux noms latins pour désigner chaque espèce de bactérie. Le premier nom correspond au genre auquel appartient l'espèce, tandis que le second nom correspond à l'espèce elle-même. le système de classification taxonomique des bactéries a été développé par le microbiologiste Carl Woese dans les années 1970. Il est basé sur l'analyse de la séquence de l'ARN ribosomique 16S, qui permet de distinguer les différentes lignées évolutives des bactéries [11].

La classification taxonomique des bactéries est donc organisée en plusieurs niveaux hiérarchiques, du plus général au plus spécifique (voir Figure 6) :

Figure 6 : Classification classique des organismes.

L'ARNr 16S (ou ARN ribosomique 16S) est une molécule d'ARN qui fait partie des composants des ribosomes bactériens. Il s'agit d'une molécule d'ARN monocaténaire de petite taille (environ 1,5 kilo bases) qui se trouve au cœur du ribosome et qui joue un rôle important dans la traduction de l'information génétique en protéines. L'ARNr 16S est également utilisé en microbiologie pour déterminer la classification phylogénétique des bactéries. En effet, les séquences d'ARNr 16S varient d'une espèce bactérienne à l'autre, et cette variation peut être utilisée pour établir des relations évolutives entre les différentes espèces bactériennes [11].

L'utilisation de l'ARNr 16S est très courante dans la classification taxonomique des bactéries car il s'agit d'une molécule universelle et conservée dans le temps. En effet, cet ARN ribosomique est présent chez toutes les bactéries et subit une évolution relativement lente, ce qui le rend très utile pour étudier la phylogénie et la taxonomie des bactéries [12]. Les séquences d'ARNr 16S sont comparées entre différentes bactéries pour déterminer leur parenté évolutive et leur classification taxonomique.

Chapitre 2 :

Deep Learning

(Apprentissage profond)

« Comparer l'intelligence artificielle et l'intelligence humaine, c'est comme comparer une calculatrice à un mathématicien. Bien que l'IA puisse traiter les données à la vitesse de l'éclair, l'intelligence humaine englobe la créativité, les émotions et la prise de décisions éthiques, ce qui la rend vraiment unique.»

Stephen Hawking

1. Intelligence artificielle

L'objectif de la recherche dans le domaine de l'intelligence artificielle (IA) est de donner aux systèmes informatiques la capacité de penser comme les humains. Par conséquent, comprendre la pensée humaine est un défi, mais surtout, elle est difficile à modéliser et à reproduire.

L'IA est un sujet brûlant dans les médias et les revues scientifiques en raison des nombreuses réalisations qui sont le résultat des progrès de l'apprentissage automatique. De grandes entreprises telles que Google, Facebook et Microsoft, ainsi que des constructeurs automobiles tels que Toyota et Volvo, sont activement engagés dans la recherche sur l'IA et prévoient d'investir davantage dans l'avenir. La recherche sur l'IA a fait de grands progrès dans divers domaines au cours de la dernière décennie. L'avancée la plus connue est celle de l'apprentissage automatique. En particulier, grâce au développement d'architectures d'apprentissage en profondeur, des réseaux de neurones convolutifs multicouches entraînés à partir de grandes quantités de données dans des architectures à forte intensité de calcul [13].

Ce chapitre présente les bases de l'IA et ses différents domaines, en particulier l'apprentissage automatique, y compris l'apprentissage profond. Ce dernier est l'outil principal pour mener à bien notre travail.

2. Apprentissage automatique

L'apprentissage automatique, également connu sous le nom de Machine Learning en anglais, est une branche de l'intelligence artificielle (IA) qui se concentre sur le développement de techniques permettant aux ordinateurs d'apprendre à partir de données et d'améliorer leurs performances sans être explicitement programmés. L'apprentissage automatique repose sur l'idée que les ordinateurs peuvent apprendre à partir de modèles et à prendre des décisions autonomes à partir de données [14].

2.1. Phases de l’apprentissage automatique

L'apprentissage automatique se compose généralement de plusieurs phases [15]:

- Définition du problème
- Collecte de données
- Prétraitement des données [16]
- Fractionnement des données

- Sélection du modèle
- Entraînement de modèle
- Évaluation du modèle
- Réglage des Hyperparamètres
- Déploiement du modèle [17]

2.2. Types de techniques d'apprentissage automatique

Les algorithmes d'apprentissage automatique se répartissent en quatre catégories principales Comme le montre la Figure 7, il existe un apprentissage supervisé, un apprentissage non supervisé, un apprentissage semi-supervisé et un apprentissage par renforcement [18]. Vous trouverez ci-dessous une brève description de chaque type de technique d'apprentissage.

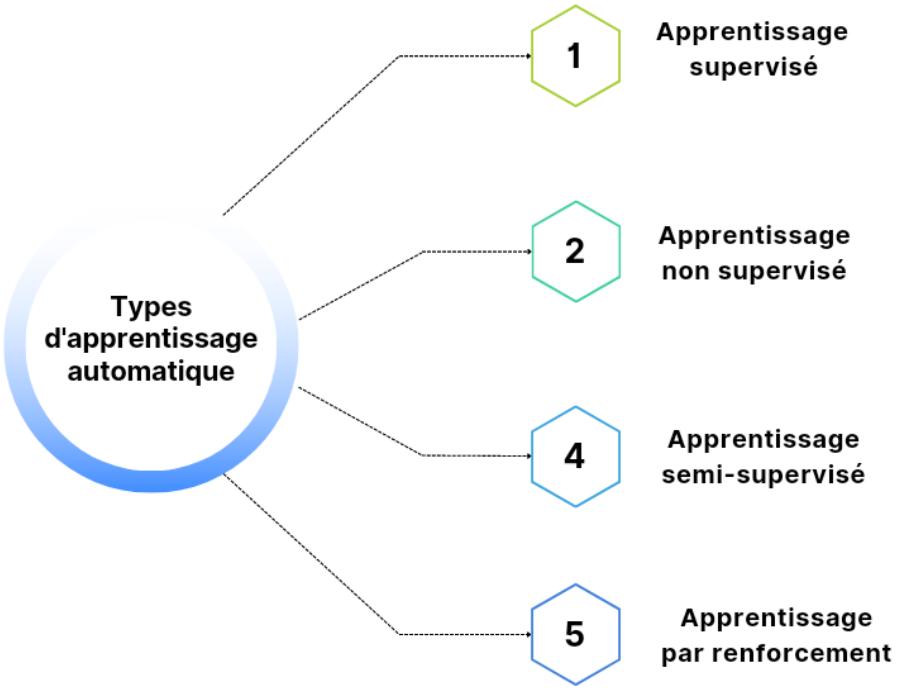

Figure 7 : Types de techniques d'apprentissage automatique.

2.2.1. Apprentissage supervisé

Dans le domaine de l'apprentissage automatique, l'apprentissage supervisé est une technique couramment utilisée pour apprendre une fonction qui associe une entrée à une sortie, en se basant sur des paires d'entrées-sorties étiquetées. Cette méthode repose sur l'utilisation de données d'entraînement annotées, composées d'une collection d'exemples d'entraînement permettant de déduire une fonction. Elle est mise en œuvre lorsque des objectifs spécifiques doivent être atteints à partir d'un ensemble de données d'entrée, c'est-à-dire une approche axée sur les tâches. Les tâches supervisées les plus couramment rencontrées sont la "classification", qui consiste à séparer les données, et la "régression", qui vise à prédire des données numériques [18].

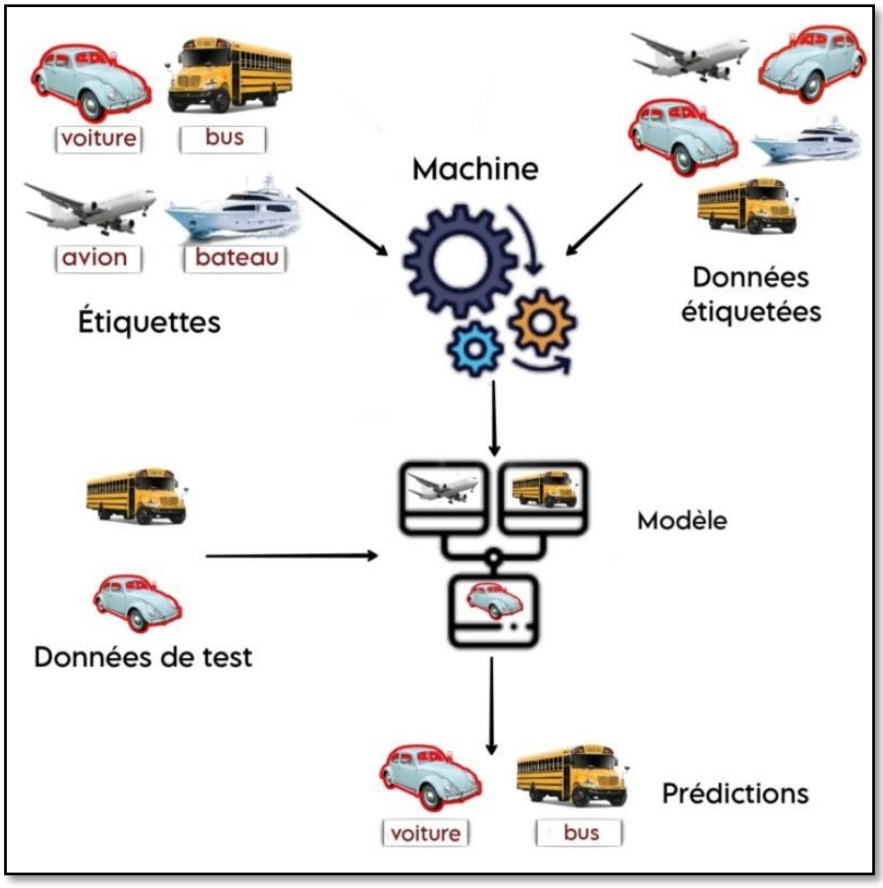

Figure 8: Apprentissage supervisé.

- Classification : Dans le domaine de l'apprentissage automatique, la classification est une méthode d'apprentissage supervisé qui vise à prédire une étiquette de classe pour un exemple donné. Cette méthode consiste à mapper mathématiquement une fonction (f) des variables d'entrée (X) aux variables de sortie (Y) en tant que cible, étiquette ou catégories. Elle peut être utilisée pour prédire la classe des points de données structurées ou non structurées. Par

exemple, la détection de pourriels, comme "spam" et "non spam", chez les fournisseurs de services de courrier électronique peut être considérée comme un problème de classification. Ci-dessous, nous présentons les problèmes de classification les plus courants [19].

- *Classification binaire* : est une méthode de classification qui implique la catégorisation de données en deux étiquettes distinctes, telles que "vrai" ou "faux" ou "oui" et "non". Dans ce type de classification, l'une des étiquettes représente l'état normal, tandis que l'autre étiquette représente l'état anormal [19]. Ceci est illustré dans la Figure 9.

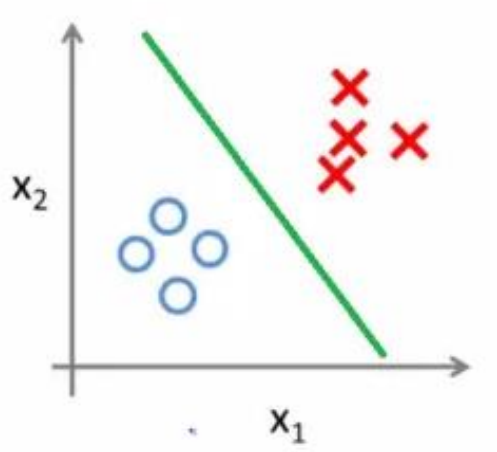

Figure 9: Classification binaire [19].

- *Classification multi-classe* :est généralement utilisée pour classer des données en plusieurs étiquettes distinctes, contrairement à la classification binaire qui se limite à deux étiquettes. la classification multi-classe ne repose pas sur le principe de résultats normaux et anormaux. Au lieu de cela, les exemples de données sont classés en fonction des classes spécifiées, en étant attribués à une étiquette spécifique [19], comme illustré dans la Figure 10.

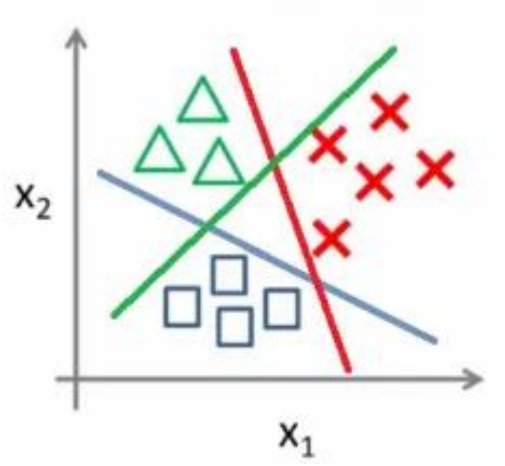

Figure 10: Classification multi-classe [20].

- Régression : La régression est un type d'apprentissage automatique qui vise à prédire une variable continue en se basant sur des données d'entrée. Contrairement à la classification qui prédit des classes ou des catégories discrètes, la régression est utilisée pour prédire des valeurs numériques [21].

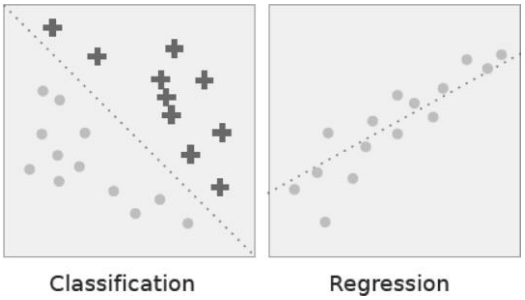

Figure 11: Différence entre classification linéaire et régression linéaire [22].

2.2.2. Apprentissage non supervisé

L'apprentissage non supervisé est une méthode d'analyse de données qui implique l'utilisation d'ensembles de données non étiquetés, sans intervention humaine, dans un processus axé sur les données. Cette méthode est couramment utilisée pour extraire des caractéristiques génératives, identifier des tendances et des structures significatives, regrouper les résultats et effectuer des explorations. Les tâches les plus courantes de l'apprentissage non supervisé sont le regroupement, l'estimation de la densité, l'apprentissage des fonctionnalités, la réduction de la dimensionnalité, la recherche de règles d'association, la détection d'anomalies [18].

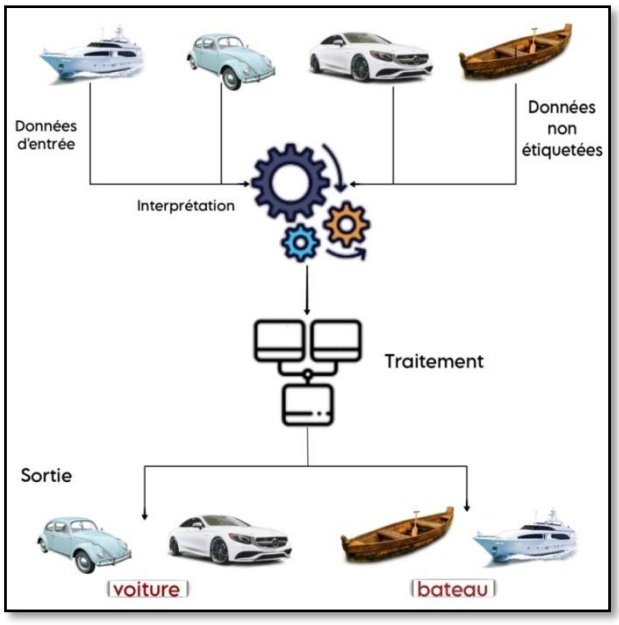

Figure 12: Apprentissage non supervisé.

2.2.3. Apprentissage Semi-supervisé

L'apprentissage semi-supervisé est une combinaison des méthodes supervisées et non supervisées mentionnées ci-dessus, qui fonctionne sur les données étiquetées et non étiquetées. Il se situe donc entre l'apprentissage "sans supervision" et l'apprentissage "avec supervision". Dans certains contextes du monde réel, il peut être rare d'avoir des données étiquetées, mais les données non étiquetées peuvent être abondantes, ce qui rend l'apprentissage semi-supervisé utile. L'objectif principal d'un modèle d'apprentissage semi-supervisé est de fournir une prédiction de meilleure qualité que celle obtenue en utilisant uniquement des données étiquetées. Certains domaines d'application courants de l'apprentissage semi-supervisé incluent la traduction automatique, la détection de la fraude, l'étiquetage des données et la classification de texte [18].

2.2.4. Apprentissage par Renforcement :

L'apprentissage par renforcement est une méthode d'apprentissage automatique permettant aux machines et aux agents logiciels d'apprendre à optimiser leur comportement dans un contexte ou un environnement donné, en se basant sur des récompenses ou des pénalités. Cette approche se concentre sur l'environnement et a pour objectif d'utiliser les connaissances acquises pour maximiser la récompense ou minimiser le risque. Elle est particulièrement utile pour améliorer l'efficacité de systèmes sophistiqués, tels que la robotique, la conduite autonome, la fabrication ou la logistique de la chaîne d'approvisionnement, en automatisant les tâches et en optimisant les processus[18].

Tableau 3 : Exemple d'utilisation des techniques d'apprentissage automatique [18].

Type d'apprentissage	**Modélisation**	**Exemples**
Supervisé	Des algorithmes ou modèles qui sont dérivés à partir de données avec des étiquettes	- Classification - Régression
Non supervisé	Des algorithmes ou des modèles dérivés de données non annotées	- Regroupement - Associations - Réduction des dimensions

Semi-supervisé	Les modèles sont construits en utilisant une combinaison de données étiquetées et non étiquetées.	- Classification - Regroupement
Renforcement	Les modèles sont fondés sur la récompense ou la pénalité (approche axée sur l'environnement)	- Classification - Contrôle

3. Apprentissage profond (Deep Learning)

Le Deep Learning, également connu sous le nom d'apprentissage profond ou DL, est un sous-domaine de l'intelligence artificielle considéré comme une évolution du Machine Learning (apprentissage automatique), voir Figure 13. Contrairement à la programmation, où la machine se contente d'exécuter des règles prédéterminées, le DL permet à la machine d'apprendre de manière autonome. Le DL repose sur des approches mathématiques pour modéliser les données. Le DL se base sur des réseaux de neurones artificiels et est conçu pour gérer de grandes quantités de données en ajoutant des couches au réseau [23].

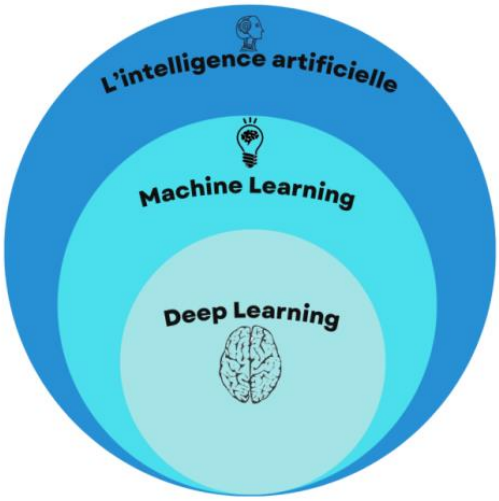

Figure 13: Positionnement du Deep Learning.

3.1. Modèle d'apprentissage profond

3.1.1. Réseaux de neurones artificiels

Les réseaux de neurones artificiels représentent une avancée majeure dans les domaines de l'informatique et de la neuro-informatique. Ces systèmes d'auto-apprentissage offrent aux ordinateurs la capacité de résoudre des problèmes à l'aide de techniques d'intelligence artificielle avancées . Les réseaux de neurones artificiels sont constitués de processeurs élémentaires fortement connectés, travaillant en parallèle pour calculer des sorties individuelles sur la base des informations qu'ils reçoivent [24]. La Figure 14 montre exemple de tels réseaux.

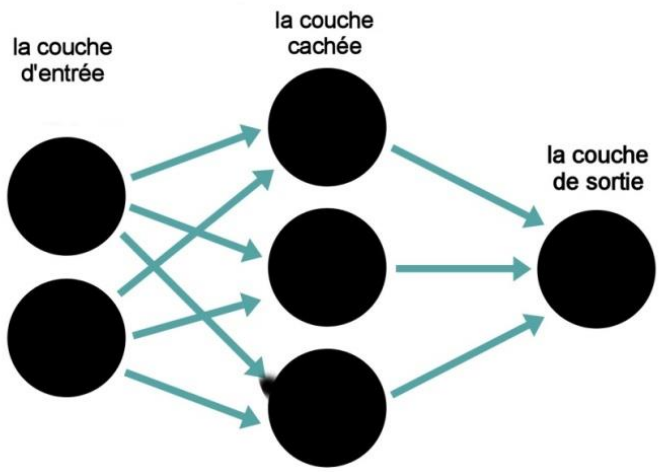

Figure 14 : Forme des réseaux de neurones artificiels.

3.1.2. Types de Réseaux de Neurones

Il existe plusieurs types de réseaux de neurones, qui ont été développés pour résoudre différents types de problèmes. Voici quelques exemples de types de réseaux de neurones [25]:

- Réseaux de Neurones Convolutifs (Convolutional Neural Networks : CNN) : ces réseaux sont souvent utilisés pour les tâches de vision par ordinateur, car ils sont capables de reconnaître des motifs dans des images.
- Réseaux de neurones à propagation avant (Feed Forward Neural Networks : FNN) : il s'agit du type de réseau de neurones le plus simple, où les signaux se propagent de l'entrée vers la sortie sans boucle de rétroaction.
- Réseaux de neurones récurrents (Recurrent Neural Networks : RNN) : ces réseaux utilisent des boucles de rétroaction pour permettre à l'information de se propager dans le temps, ce qui les rend utiles pour les tâches qui impliquent des séquences temporelles.

3.1.3. Neurone réel (Anatomie) et neurone artificiel (Formel)

Les modèles mathématiques des réseaux de neurones artificiels sont basés sur l'inspiration biologique. Leur origine remonte au désir de modéliser le fonctionnement des neurones biologiques, d'où le nom "neurone artificiel" (voir Figure 13 et Tableau 5) [26].

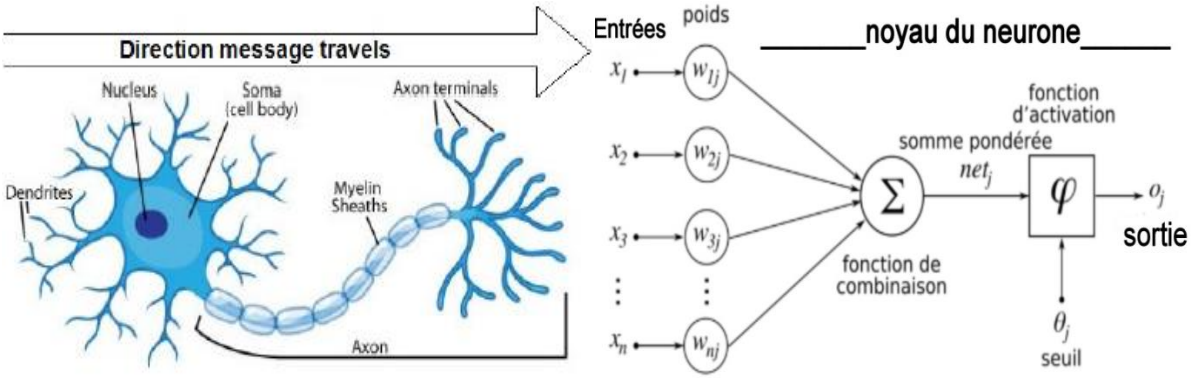

Figure 15: Neurone réel et neurone artificiel [26].

3.2. Applications de l'apprentissage automatique en biologie

Il y a plusieurs d'applications du Deep learning dans la biologie notamment :

- L'analyse de séquences d'ADN : l'apprentissage profond est utilisé pour identifier les motifs de séquences, prédire la structure des protéines et découvrir de nouveaux médicaments.
- La génomique comparative : l'apprentissage profond est utilisé pour comparer les génomes de différentes espèces et comprendre l'évolution des espèces.
- La modélisation des réseaux de régulation génétique : l'apprentissage profond est utilisé pour modéliser les interactions complexes entre les gènes et les protéines et comprendre comment les cellules régulent leur expression génique.
- La médecine personnalisée : l'apprentissage profond est utilisé pour prédire les réponses des patients aux traitements médicaux en fonction de leur génome et de leur historique médical.
- La recherche en neuroscience : l'apprentissage profond est utilisé pour analyser les signaux neuronaux et comprendre comment le cerveau fonctionne [27].

3.3. Evaluation d'un modèle d'apprentissage profond pour la classification

Il existe plusieurs performances pour vérifier le rendement du modèle d'apprentissage profond pour des problèmes de classification. Ci-après les principales mesures :

- **Matrice de confusion** : est utilisée lors de l'exécution des prédictions de classification pour évaluer les résultats obtenus. Elle comprend quatre types de résultats possibles :

Tableau 4 : Matrice de confusion.

	Réel positif	**Réel négatif**
Prédit positif	Vrai positif (VP)	Faux positif (FP)
Prédit négatif	Faux négatif (FN)	Vrai négatif (VN)

- Les vrais positifs (VP) représentent les prédictions correctes où le classificateur prédit une classe qui correspond réellement à cette classe.
- Les vrais négatifs (VN) représentent les prédictions correctes où le classificateur prédit une classe négative qui correspond réellement à cette classe négative.
- Les faux positifs (FP) se produisent lorsque le classificateur prédit une classe positive alors que la donnée en question appartient en réalité à la classe négative.
- Les faux négatifs (FN) se produisent lorsque le classificateur prédit une classe négative alors que la donnée en question appartient en réalité à la classe positive [15].

Ces quatre résultats sont représentés dans une matrice de confusion, qui est un tableau utilisé pour visualiser ces informations.

- **Accuracy** : est une métrique largement utilisée pour évaluer les performances des modèles de classification. Elle mesure la fréquence à laquelle un algorithme attribue correctement des étiquettes aux données. L'exactitude est calculée en divisant le nombre de points de données correctement prédits par le nombre total de points de données. Dans le contexte d'une matrice de confusion, l'exactitude peut être déterminée en sommant les Vrais Négatifs et les Vrais Positifs, puis en divisant cette somme par la somme des Vrais Négatifs (TN), des Vrais Positifs (TP), des Faux Négatifs (FN) et des Faux Positifs (FP) [28].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision** : est une mesure couramment utilisée en classification, en recherche d'information et dans les problèmes liés à la reconnaissance de motifs. Elle évalue la proportion d'observations pertinentes parmi les observations récupérées. Dans le cadre d'une

matrice de confusion, la précision peut être calculée en divisant le nombre de Vrais Positifs par la somme des Vrais Positifs et des Faux Négatifs [28].

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Sensibilité** : également appelée rappel (recall), est le ratio entre les événements positifs réels prédits comme positifs. Elle met l'accent sur la capacité du modèle à détecter les véritables positifs. Pour la calculer à partir de la matrice de confusion, il faut diviser la valeur des Vrais Positifs par la somme des Vrais Positifs et des Faux Négatifs [28].

$$\text{sensibilité} = \frac{TP}{TP + FN}$$

- **Spécificité** : mesure la capacité d'un modèle à prédire correctement les cas négatifs réels. Elle est définie comme le rapport entre les négatifs réels prédits comme négatifs et le total des négatifs réels et des faux positifs. Dans le contexte de la matrice de confusion, la spécificité peut être calculée en divisant les Vrais Négatifs par la somme des Vrais Négatifs et des Faux Positifs [28].

$$\text{spécificité} = \frac{TN}{TN + FP}$$

Chapitre 3 :

Matériel et méthodes

1. Matériel

1.1. Données biologiques

Le matériel utilisé dans cette étude est composé de 11700 séquences d'ARNr 16s de quatre types de bactéries (*Escherichia coli, Enterococcus faecalis, Klebsiella pneumoniae* et *Pseudomonas aeruginosa*), qui ont été obtenues à partir de la base de données de référence SILVA [29]. D'abord, nous avons construit un fichier texte à partir des quatre fichiers Fasta téléchargés. Chaque ligne représente une séquence qui contient des nucléotides (A,C,G et U). Ces données biologiques ont été extraites et traitées afin de permettre leur analyse et leur interprétation. Les séquences d'ARNr 16S sont des marqueurs génétiques largement utilisés pour étudier la diversité bactérienne dans les échantillons environnementaux. Elles ont été analysées à l'aide de méthodes bioinformatiques avancées pour obtenir des informations sur la composition taxonomique et la diversité bactérienne dans l'échantillon étudié.

1.2. Configuration de la machine

Ces caractéristiques sont détaillées dans le tableau suivant:

Tableau 5 : Caractéristiques de la machine utilisée.

Composant	Caractéristiques
Processeur	Intel i5-5300U CPU @ 2,30 GHz
RAM	8,00 Go
Stockage	500 Go
Système d'exploitation	Windows 7
Type de système	Système d'exploitation 64 bits

1.3. Outils et bibliothèques :

1.3.1. Outils :

Le présent travail est réalisé en utilisant le langage de programmation Python, via Jupyter notebook de l'outil Anaconda et Google Colab (Voir Tableau 6)

Tableau 6 : Outils utilisés.

Outil	Description
Python	Python est un langage de programmation interprété, orienté objet et de haut niveau avec une sémantique dynamique. Ses structures de données intégrées de haut niveau, combinées à un typage dynamique et une liaison dynamique, le rendent très attrayant pour le développement rapide d'applications [30].
Anaconda	Une distribution open-source de Python, conçue pour les scientifiques des données et les développeurs. Elle fournit un environnement de développement intégré (IDE) pour la programmation Python, ainsi qu'une gestion des packages et des environnements virtuels [31].
Jupyter Notebook	Un environnement interactif de développement pour la création et le partage de documents qui contiennent du code Python, des visualisations, du texte explicatif et d'autres éléments multimédias. Les notebooks Jupyter sont souvent utilisés pour l'analyse de données, la visualisation et la documentation [32].
Google Colab	Un environnement de développement pour l'exécution de notebooks Jupyter, hébergé dans le cloud par Google. Colab permet l'exécution de code Python gratuitement, avec l'accès à des ressources telles que des processeurs graphiques (GPU) et des unités de traitement tensoriel (TPU) pour l'apprentissage en profondeur et le machine learning [33].

1.3.2. Les bibliothèques :

Les fonctions python utilisées à ce travail sont indiquées dans le tableau 7.

Tableau 7 : Bibliothèques python utilisées.

Bibliothèque	**Description**
Pandas	Bibliothèque Python pour la manipulation et l'analyse de données tabulaires. Elle offre des fonctionnalités avancées pour la gestion des données, la transformation, le nettoyage et la préparation des données génomiques pour l'entraînement des modèles [34].
NumPy	Bibliothèque Python pour la manipulation de tableaux et les calculs scientifiques. Elle offre des fonctionnalités pour la manipulation efficace des données numériques, notamment pour les opérations mathématiques nécessaires dans le traitement des données génomiques [35].
Keras	Une interface de haut niveau pour la construction de modèles d'apprentissage profond en Python. Elle offre une syntaxe simple et intuitive pour la création de modèles de réseaux de neurones, ce qui la rend largement utilisée dans les projets de fin d'étude [36].
TensorFlow	Une bibliothèque open source d'apprentissage automatique et d'apprentissage profond développée par Google. Elle offre des fonctionnalités pour la création, l'entraînement et le déploiement de modèles d'apprentissage profond, notamment pour la classification de données génomiques [37].
Scikit-learn	Bibliothèque Python d'apprentissage automatique (machine learning) qui offre une large gamme d'algorithmes de classification, d'évaluation de modèles et d'outils de prétraitement des données. Elle est couramment utilisée pour l'entraînement et l'évaluation de modèles de classification dans les projets de fin d'étude [38].
Matplotlib	Bibliothèque Python pour la visualisation de données. Elle offre des fonctionnalités pour la création de graphiques et de visualisations pour l'analyse des résultats des modèles de classification, notamment dans le domaine de la bioinformatique et de la génomique [39].

2. Méthodes

Le processus de travail est divisé en deux sections principales, à savoir le nettoyage des données et le DL :

2.1. Prétraitement des données :

Dans cette étude, nous avons utilisé la bibliothèque *Biopython* pour effectuer des opérations biologiques, telles que la manipulation des séquences. Les étapes de modification du fichier ont été exécutées de manière rigoureuse en utilisant un code basé sur la bibliothèque *BioPython*. Les séquences ont été converties en chaînes de nucléotides en utilisant la fonction *SeqIO.parse*() de *BioPython*.

Une vérification systématique de la longueur de chaque séquence a été considérée, suivie d'une troncature, si nécessaire, pour limiter leur longueur à un maximum. Les positions excédantes ont été remplacées par des tirets. Chaque caractère de chaque séquence a été séparé par des espaces pour faciliter leur analyse ultérieure et améliorer la lisibilité des données. Pour marquer l'origine et l'attribution de chaque séquence, une valeur spécifique a été ajoutée à la fin de chaque séquence. Par exemple, le numéro "1" représente *Escherichia coli*, le numéro "2" représente *Enterococcus faecalis*, le numéro "3" représente *Klebsiella pneumoniae* et le numéro "4" représente *Pseudomonas aeruginosa*. Le résultat final de ces étapes de modification a été consigné dans un fichier texte, soigneusement préparé pour être utilisé dans des analyses ultérieures dans le cadre de notre projet d'étude.

Nous avons formé quatre fichiers portant les noms suivants : "Escherichia coli1", "Enterococcus faecalis2", "Klebsiella pneumoniae3" et "Pseudomonas aeruginosa4". Enfin, nous avons extrait 3000 séquence de chaque fichier et les avons combinées pour former un fichier final.

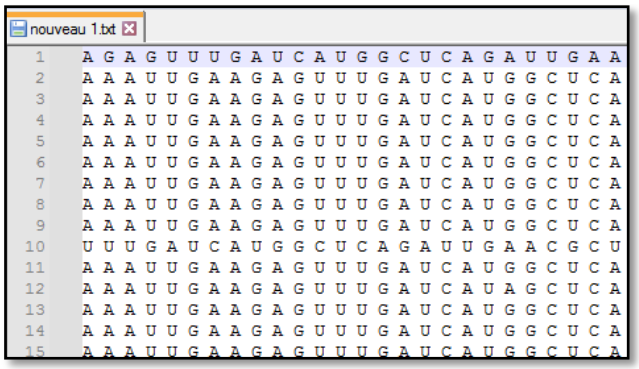

Figure 16 : Fichier final.

2.2. Apprentissage:

➢ **Importation des bibliothèques et chargement des données :** Ce code concerne l'importation des bibliothèques nécessaires à l'exécution du code, à savoir *numpy, tensorflow, sklearn, matplotlib* et *seabornet pandas*, Les données sont chargées à partir du fichier texte. Les séquences de nucléotides sont stockées dans la variable "*sequences*" et les étiquettes de classe correspondantes sont stockées dans la variable "*labels*".

```
import numpy as np
import tensorflow as tf
from tensorflow import keras
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
import matplotlib.pyplot as plt
import pandas as pd
from sklearn.metrics import confusion_matrix
import seaborn as sns

# Charger les données à partir du fichier
data = np.loadtxt("/content/nouveau 1.txt", dtype=str)
sequences = data[:, :-1]
labels = data[:, -1]
```

Figure 17 : Bibliothèques utilisées et chargement des données.

➢ **Encodage des séquences et des étiquettes :** Les séquences et les étiquettes sont encodées à l'aide de la méthode *LabelEncoder* de *sklearn*. La variable "*num_classes*" est définie comme le nombre total de classes, tandis que les séquences sont encodées en utilisant *one-hot*.

```
# Encoder les séquences et les étiquettes de classe à l'aide de LabelEncoder
encoder = LabelEncoder()
encoded_labels = encoder.fit_transform(labels)
num_classes = len(encoder.classes_)

# Encoder les séquences avec one-hot encoding
def encode_sequence(seq):
    encoded_seq = np.zeros((len(seq), 5))
    mapping = {'A': 0, 'U': 1, 'C': 2, 'G': 3}
    for i, char in enumerate(seq):
        if char in mapping:
            encoded_seq[i, mapping[char]] = 1
        else:
            encoded_seq[i, 4] = 1
    return encoded_seq
encoded_sequences = np.array([encode_sequence(seq) for seq in sequences])
```

Figure 18 : Code pour l'encodage.

➢ **Division des données en ensembles d'entraînement et de test :** Les données sont divisées en ensembles d'entraînement et de test à l'aide de la fonction *train_test_split* de *sklearn*. Les données d'entraînement représentent 80% des données totales, tandis que les

données de test représentent 20%. Les données d'entraînement et les étiquettes sont stockées dans les tableaux "*X_train*" et "*y_train*", respectivement, tandis que les données de test et les étiquettes sont stockées dans les tableaux "*X_test*" et "*y_test*", respectivement.

```
# Diviser les données en ensembles d'entraînement et de test
X_train, X_test, y_train, y_test = train_test_split(
    encoded_sequences, encoded_labels, test_size=0.2, random_state=42
)
```

Figure 19 : Code pour division des données.

- **Définition du modèle de réseau de neurones :** Le modèle de réseau de neurones est défini à l'aide de la méthode *Sequential* de *keras*. Le modèle se compose d'une couche d'aplatissement suivie d'une couche dense de 128 neurones avec une fonction d'activation relu, suivie d'une couche dense de sortie avec une fonction d'activation *softmax*. La fonction de perte utilisée est *categorical_crossentropy*.

```
# Définir le modèle de réseau de neurones
model = keras.Sequential([
    keras.layers.Flatten(input_shape=(1300, 5)),
    keras.layers.Dense(128, activation='relu'),
    keras.layers.Dense(num_classes, activation='softmax')
])
# Compiler le modèle en utilisant la fonction de perte categorical_crossentropy
model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy'])
```

Figure 20 : Modèle de réseau de neurones.

- **Entraînement du modèle** : Le modèle est entraîné sur les données d'entraînement en utilisant la méthode "*fit*". Les résultats d'entraînement sont stockés dans la variable "*history*".

```
# Former le modèle
history = model.fit(
    X_train, keras.utils.to_categorical(y_train, num_classes=num_classes),
    validation_split=0.2, epochs=10, batch_size=32
)
```

Figure 21 : Code pour l'entraînement du modèle.

- **Évaluation du modèle** : Le modèle est évalué sur l'ensemble de test à l'aide de la méthode "*evaluate*". Les résultats sont stockés dans les variables "*test_loss*" et "*test_acc*".

```
# Évaluer le modèle sur l'ensemble de test
test_loss, test_acc = model.evaluate(
    X_test, keras.utils.to_categorical(y_test, num_classes=num_classes)
)
```

Figure 22 : Code pour Évaluation du modèle

➢ **Affichage des résultats** : Le modèle est utilisé pour effectuer des prédictions sur l'ensemble de test à l'aide de la méthode *predict* de *Keras*. Les prédictions sont stockées dans la variable "*predictions*".

```
# Afficher les résultats d'entraînement (perte et exactitude)
print('Training loss:', history.history['loss'])
print('Training accuracy:', history.history['accuracy'])
# Afficher la perte et l'exactitude du modèle sur l'ensemble de test
print('Test loss:', test_loss)
print('Test accuracy:', test_acc)
# Faire des prédictions sur l'ensemble de test
predictions = model.predict(X_test)
print('Predictions:', predictions)
```

Figure 23 : Affichage des résultats d'entraînement, test et prédictions.

➢ **Affichage des graphiques** : ce code permet de visualiser la performance du modèle pendant l'entraînement à l'aide de graphiques affichant la perte et l'exactitude du modèle sur les données d'entraînement et de validation. Les valeurs de perte et d'exactitude sont obtenues à partir de l'historique de l'entraînement et sont utilisées pour créer les graphiques avec la bibliothèque *Matplotlib*. L'objectif est de détecter d'éventuels problèmes tels que l'Overfitting ou l'Underfitting.

```
# Afficher un graphique de la perte et de l'exactitude du modèle pendant l'entraînement
plt.plot(history.history['loss'])
plt.plot(history.history['accuracy'])
plt.title('Training Loss and Accuracy')
plt.xlabel('Epoch')
plt.legend(['Loss', 'Accuracy'])
plt.show()
# Obtenir les valeurs de perte et d'exactitude pour l'entraînement et la validation
train_loss = history.history['loss']
train_acc = history.history['accuracy']
val_loss = history.history['val_loss']
val_acc = history.history['val_accuracy']
# Afficher un graphique de la perte pendant l'entraînement et la validation
plt.plot(train_loss, label='Training Loss')
plt.plot(val_loss, label='Validation Loss')
plt.title('Training and Validation Loss')
plt.xlabel('Epoch')
plt.ylabel('Loss')
plt.legend()
plt.show()
# Afficher un graphique de l'exactitude pendant l'entraînement et la validation
plt.plot(train_acc, label='Training Accuracy')
plt.plot(val_acc, label='Validation Accuracy')
plt.title('Training and Validation Accuracy')
plt.xlabel('Epoch')
plt.ylabel('Accuracy')
plt.legend()
plt.show()
```

Figure 24 : Code pour afficher les graphiques.

➢ **Crée une table qui affiche les prédictions et affiche une matrice de confusion et affiche une représentation graphique de la matrice de confusion sous forme de graphique à barres** :

```
# Afficher une table avec les prédictions et les étiquettes réelles pour l'ensemble de test
table = {'Predictions': np.argmax(predictions, axis=1), 'True Labels': y_test}
print(pd.DataFrame(table))
# Afficher une matrice de confusion pour évaluer les performances du modèle
conf_matrix = confusion_matrix(y_test, np.argmax(predictions, axis=1))
plt.figure(figsize=(8, 6))
sns.heatmap(conf_matrix, annot=True, fmt='d', cmap='Blues')
plt.xlabel('Predicted Labels')
plt.ylabel('True Labels')
plt.title('Confusion Matrix')
plt.show()

# Calculer la matrice de confusion
conf_matrix = confusion_matrix(y_true, y_pred)
# Extraire les valeurs de la matrice de confusion
true_positive = conf_matrix[1, 1]
true_negative = conf_matrix[0, 0]
false_positive = conf_matrix[0, 1]
false_negative = conf_matrix[1, 0]
# Créer une liste de labels
labels = ['True Negative', 'False Positive', 'False Negative', 'True Positive']
# Créer une liste de valeurs correspondant aux labels
values = [true_negative, false_positive, false_negative, true_positive]
# Créer un graphique à barres
plt.bar(labels, values)
plt.xlabel('Labels')
plt.ylabel('Counts')
plt.title('Confusion Matrix')
plt.show()
```

Figure 25 : Code pour affichage de matrice de confusion.

Chapitre 3 :

Résultats et discussion

RÉSULTATS ET DISCUSSION

Afin d'accomplir la tâche de classification des bactéries, cette section expose en détail le matériel informatique utilisé ainsi que les méthodes spécifiques de deep learning qui ont été développées et appliquées.

La Figure 28 montre les résultats des performances de l'apprentissage profond des séquences des bactéries.

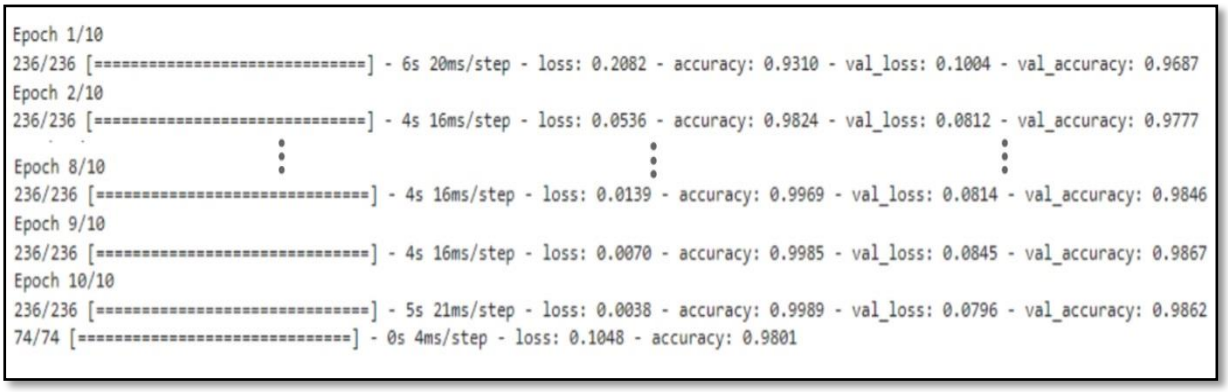

Figure 26 : Affichage de quelques résultats du modèle.

- La précision (*accuracy*) sur l'ensemble d'apprentissage (*train*) augmente progressivement au fil des époques (*epochs*), atteignant une précision de **0,9989** à l'époque 10.
- La précision sur l'ensemble de validation (*val*) est également élevée, atteignant une précision de 0,9862 à l'époque 10. Cela indique que le modèle est bien généralisé et ne sur-ajuste pas les données d'apprentissage.
- La perte (*loss*) sur l'ensemble d'apprentissage diminue progressivement au fil des époques, indiquant que le modèle s'améliore.
- La perte sur l'ensemble de validation est également faible
- Les résultats finaux sur l'ensemble de test (*test*) sont également très bons, avec une précision de **0,9801**.

Dans l'ensemble, ces résultats indiquent que le modèle est capable de reconnaître les séquences de bactéries avec une grande précision.

Les graphiques des performances jouent un rôle essentiel dans l'analyse des résultats obtenus lors de l'entraînement de modèles, permettant d'évaluer la qualité de l'apprentissage. Dans la Figure 29 nous examinerons de près les courbes de perte de validation, de perte d'entraînement et d'exactitude obtenues lors de notre étude, afin de mieux comprendre les performances de notre modèle et d'en extraire des enseignements pertinents.

Une diminution de la perte et une augmentation de la précision pour les données d'entraînement et de validation indiquent que le modèle a pu identifier les caractéristiques importantes des données d'entraînement et généraliser cette compréhension à de nouvelles données de validation (comme il est résumé dans la Figure 27).

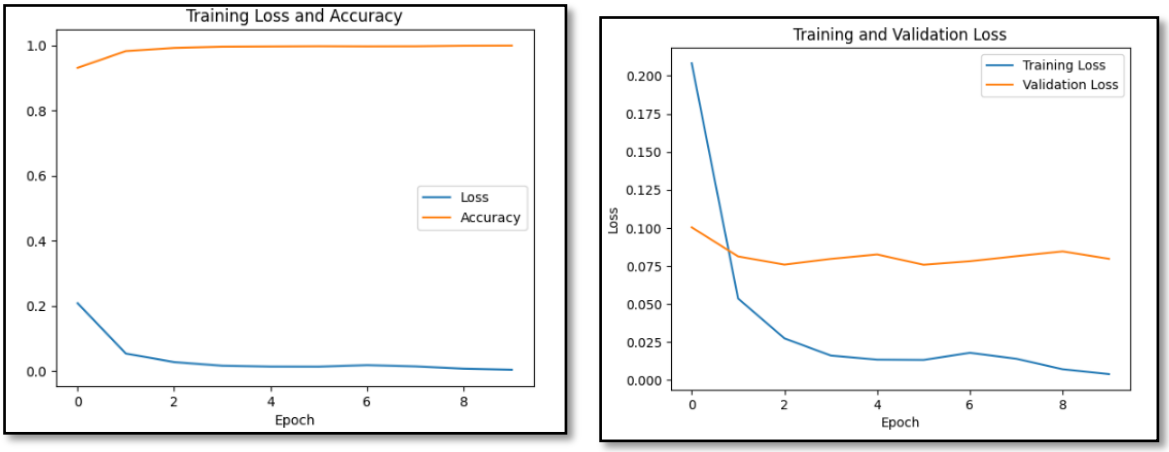

Figure 28 : Présentation graphique des résultats.

Les résultats de la Figure 30 montrent un exemple de prédictions du modèle pour les classes de bactéries. Chaque ligne représente la prédiction pour une séquence de bactéries individuelle, où les valeurs numériques sont les probabilités associées à chaque classe de bactéries.

```
Predictions: [[1.1997404e-06 1.7679308e-10 9.9999839e-01 3.3648178e-07]
 [2.5784843e-06 9.9985242e-01 1.5586842e-06 1.4355739e-04]
 [7.8561839e-05 9.9560696e-01 1.8306573e-05 4.2962255e-03]
 ...
 [9.0075184e-07 2.2055742e-10 9.9999833e-01 6.9197472e-07]
 [9.0145174e-04 7.6266837e-10 9.9909830e-01 2.1423904e-07]
 [1.1176407e-06 9.7530117e-10 1.8243051e-11 9.9999893e-01]]
```

Figure 29 : Affichage des prédictions.

En général, les prédictions semblent indiquer que le modèle est capable de classer les séquences de bactéries avec une précision raisonnablement élevée.

La matrice de confusion fournie dans la Figure 31 montre les résultats de la classification des bactéries à partir du modèle d'apprentissage profond. Chaque cellule de la matrice représente le nombre d'échantillons de chaque classe prédite par le modèle.

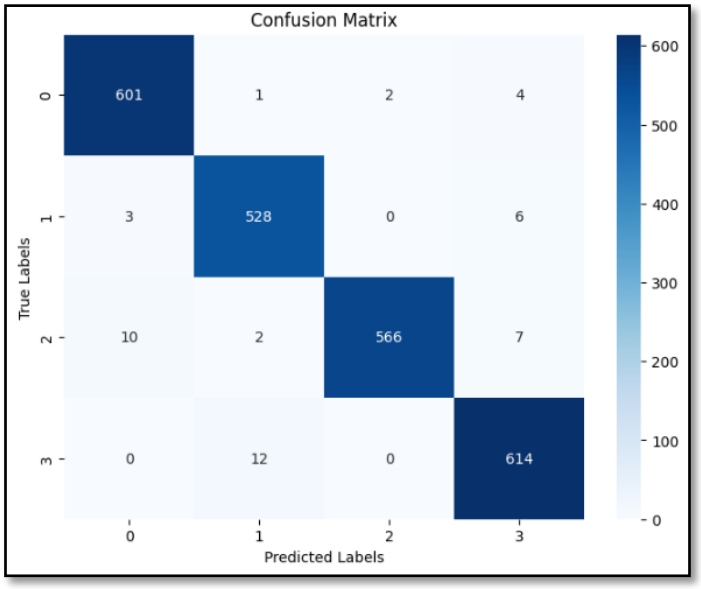

Figure 31 : Matrice de confusion.

La matrice de confusion montre les performances du modèle pour la classification de séquences de quatre bactéries différentes. Les nombres dans la matrice indiquent le nombre de séquences qui ont été classées dans chaque catégorie. Les quatre bactéries sont étiquetées comme suit :

- Le premier type de bactérie est représenté par la première colonne et la première ligne (étiqueté par 0)
- Le deuxième type de bactérie est représenté par la deuxième colonne et la deuxième ligne (étiqueté par 1)
- Le troisième type de bactérie est représenté par la troisième colonne et la troisième ligne (étiqueté par 2)
- Le quatrième type de bactérie est représenté par la quatrième colonne et la quatrième ligne (étiqueté par 3)

En analysant la matrice de confusion :

- La classe 1 (*Escherichia coli*) a été prédite correctement pour la plupart des instances (601 prédictions correctes). Cependant, il y a eu quelques erreurs de prédiction pour cette classe, où le modèle a prédit à tort les classes 2, 3 ou 4 pour certaines instances réelles de la classe 1.

- La classe 2 (*Enterococcus faecalis*) a également été prédite avec précision pour la plupart des instances (528 prédictions correctes). Cependant, il y a eu quelques erreurs de prédiction, où le modèle a prédit les classes 1, 3 ou 4 à la place de la classe 2.

- La classe 3 (*Klebsiella pneumoniae*) a également été prédite avec une précision raisonnable (566 prédictions correctes). Cependant, il y a eu quelques erreurs de prédiction, où le modèle a prédit les classes 1, 2 ou 4 à la place de la classe 3.

- La classe 4 (*Pseudomonas aeruginosa*) a été prédite correctement pour la plupart des instances (614 prédictions correctes). Il y a eu quelques erreurs de prédiction où le modèle a prédit à tort les classes 1, 2 ou 3 pour certaines instances réelles de la classe 4.

En résumé, la matrice de confusion et la table permettent d'évaluer les performances du modèle en termes de prédictions correctes et incorrectes pour chaque classe. Cela peut vous aider à comprendre les forces et les faiblesses du modèle dans la classification des différentes classes de bactéries.

Conclusion

Conclusion

Les résultats obtenus pour le modèle de DL utilisé dans la classification multi-classe taxonomique des bactéries responsables d'infections nosocomiales à partir des données génomiques sont hautement prometteurs. Les performances exceptionnelles sur les ensembles de formation, de validation et de test attestent de la capacité du modèle à généraliser efficacement à de nouvelles données. Les résultats de la matrice de confusion démontrent également que le modèle a réussi à classer de manière précise les séquences bactériennes pour chaque classe prédite. Ces résultats soutiennent l'efficacité de DL en tant qu'approche pour la classification taxonomique des bactéries à partir de données génomiques, notamment les séquences ARNr 16s. Ces constatations peuvent avoir des implications significatives dans la lutte contre les infections nosocomiales, en permettant une identification plus rapide et précise des bactéries responsables. Cela pourrait contribuer à l'élaboration de mesures préventives plus efficaces.

Références bibliographiques

RÉFÉRENCES

[1] « Biologie moléculaire - Définition et Explications ». https://www.techno-science.net/glossaire-definition/Biologie-moleculaire.html (consulté le 31 mars 2023).

[2] J. D. Watson et F. H. Crick, « Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid », *Nature*, vol. 171, nº 4356, p. 737‑738, avr. 1953, doi: 10.1038/171737a0.

[3] « La structure de l'ADN ». https://tp-svt.pagesperso-orange.fr/adn.htm (consulté le 3 avril 2023).

[4] Alberts, Johnson, Levis, et Raff, *Molecular Biology Of The Cell*, 4e éd. Garland Science, 2002. Consulté le: 29 mai 2023. [En ligne]. Disponible sur: http://gen.lib.rus.ec/book/index.php?md5=3e8cff10a833e7051dab709bbaeb05ca

[5] Peter B. Moore et Harry F. Noller, « RNA Structure: Reading the Ribosome », 2000. https://www.researchgate.net/publication/7620373_RNA_Structure_Reading_the_Ribosome (consulté le 30 mai 2023).

[6] B. Alberts, *Molecular biology of the cell*, Sixth edition. New York, NY: Garland Science, Taylor and Francis Group, 2015.

[7] « WHO_CDS_CSR_EPH_2002.12_fre.pdf ». Consulté le: 27 mars 2023. [En ligne]. Disponible sur: https://apps.who.int/iris/bitstream/handle/10665/69751/WHO_CDS_CSR_EPH_2002.12_fre.p?sequence=1

[8] « Infections nosocomiales · Inserm, La science pour la santé », *Inserm*. https://www.inserm.fr/dossier/infections-nosocomiales/ (consulté le 3 avril 2023).

[9] *Infection Prevention and Control of Epidemic- and Pandemic-Prone Acute Respiratory Infections in Health Care*. in WHO Guidelines Approved by the Guidelines Review Committee. Geneva: World Health Organization, 2014. Consulté le: 2 avril 2023. [En ligne]. Disponible sur: http://www.ncbi.nlm.nih.gov/books/NBK214359/

[10] M. Rupp et P. Fey, « Extended spectrum beta-lactamase (ESBL)-producing Enterobacteriaceae: Considerations for diagnosis, prevention and drug treatment », *Drugs*, vol. 63, p. 353‑65, févr. 2003.

[11] C. R. Woese, « Bacterial evolution. », *Microbiol Rev*, vol. 51, nº 2, p. 221‑271, juin 1987.

[12] Stackebrandt, E et Ebers, J, « Stackebrandt, E. and Ebers, J. (2006) Taxonomic Parameters Revisited Tarnished Gold Standards. Microbiology Today, 33, 152-155. - References - Scientific Research Publishing », 2006. https://www.scirp.org/(S(lz5mqp453edsnp55rrgjct55))/reference/referencespapers.aspx?referenceid=2470567 (consulté le 29 mai 2023).

[13] Inria, « Inria - Livre blanc intelligence artificielle (seconde édition 2021) », 11:48:17 UTC. Consulté le: 25 février 2023. [En ligne]. Disponible sur: https://fr.slideshare.net/INRIA/inria-livre-blanc-intelligence-artificielle-seconde-dition-2021-250202445

[14] Y. LeCun, Y. Bengio, et G. Hinton, « Deep learning », *Nature*, vol. 521, nº 7553, p. 436‑444, mai 2015, doi: 10.1038/nature14539.

[15] K. Nighania, « Various ways to evaluate a machine learning models performance », *Medium*, 30 janvier 2019. https://towardsdatascience.com/various-ways-to-evaluate-a-machine-learning-models-performance-230449055f15 (consulté le 4 avril 2023).

[16] A. A. Movassagh *et al.*, « Artificial neural networks training algorithm integrating invasive weed optimization with differential evolutionary model », *Journal of Ambient Intelligence and Humanized Computing*, mars 2021, doi: 10.1007/s12652-020-02623-6.

[17] S. M. A. Burney, T. Jilani, et C. Ardil, « A Comparison of First and Second Order Training Algorithms for Artificial Neural Networks. », janv. 2004, p. 12‑18.

[18] I. H. Sarker, « Machine Learning: Algorithms, Real-World Applications and Research Directions », *SN COMPUT. SCI.*, vol. 2, nº 3, p. 160, mars 2021, doi: 10.1007/s42979-021-00592-x.

[19] M. Mohri, A. Rostamizadeh, et A. Talwalkar, *Foundations of machine learning*. in Adaptive computation and machine learning series. Cambridge, MA: MIT Press, 2012.

[20] Mohammed Terry Jack, « Tips and Tricks for Multi-Class Classification ». https://medium.com/@b.terryjack/tips-and-tricks-for-multi-class-classification-c184ae1c8ffc (consulté le 3 avril 2023).

[21] J. Friedman, T. Hastie, et R. Tibshirani, *The Elements of Statistical Learning*. 2001. Consulté le: 30 mai 2023. [En ligne]. Disponible sur: http://gen.lib.rus.ec/book/index.php?md5=daf890ca93ba97f2a6f182cea21d9111

[22] « Initiez-vous au Machine Learning », *OpenClassrooms*. https://openclassrooms.com/fr/courses/4011851-initiez-vous-au-machine-learning (consulté le 3 avril 2023).

[23] C. Hardy, « Contribution au développement de l'apprentissage profond dans les systèmes distribués ».

[24] C. Touzet, « LES RESEAUX DE NEURONES ARTIFICIELS, INTRODUCTION AU CONNEXIONNISME ».

[25] I. Goodfellow, Y. Bengio, et A. Courville, *Deep Learning*. MIT Press, 2016.

[26] Y. Djeriri, « Les Réseaux de Neurones Artificiels », 2017.

[27] B. Alipanahi, A. Delong, M. T. Weirauch, et B. J. Frey, « Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning », *Nat Biotechnol*, vol. 33, nº 8, Art. nº 8, août 2015, doi: 10.1038/nbt.3300.

[28] I. Priyadarshini et C. Cotton, « A novel LSTM–CNN–grid search-based deep neural network for sentiment analysis », *J Supercomput*, vol. 77, nº 12, p. 13911‑13932, déc. 2021, doi: 10.1007/s11227-021-03838-w.

[29] « Search ». https://www.arb-silva.de/search/ (consulté le 12 juin 2023).

[30] « Welcome to Python.org », *Python.org*, 29 mai 2023. https://www.python.org/ (consulté le 30 mai 2023).

[31] « Anaconda | The World's Most Popular Data Science Platform ». https://www.anaconda.com/ (consulté le 30 mai 2023).

[32] « Project Jupyter ». https://jupyter.org (consulté le 30 mai 2023).

[33] « Google Colaboratory ». https://colab.research.google.com/ (consulté le 30 mai 2023).

[34] « pandas - Python Data Analysis Library ». https://pandas.pydata.org/ (consulté le 30 mai 2023).

[35] « NumPy ». https://numpy.org/ (consulté le 30 mai 2023).

[36] « Keras: Deep Learning for humans ». https://keras.io/ (consulté le 30 mai 2023).

[37] « TensorFlow », *TensorFlow*. https://www.tensorflow.org/?hl=fr (consulté le 30 mai 2023).

[38] « scikit-learn: machine learning in Python — scikit-learn 1.2.2 documentation ». https://scikit-learn.org/stable/ (consulté le 30 mai 2023).

[39] « Matplotlib — Visualization with Python ». https://matplotlib.org/ (consulté le 30 mai 2023).

Année universitaire : 2022-2023	**Présenté par :** **BOUGUENDOURA Zakaria** et **MESSELEM Anis**

Deep Learning pour la classification taxonomique des bactéries causant des infections nosocomiales à partir des données génomiques

Mémoire pour l'obtention du diplôme de Master en Bioinformatique

Les infections nosocomiales constituent une menace importante pour la santé publique, nécessitant le développement de méthodes précises et efficaces pour leur identification et leur classification. Ce mémoire étudie la classification taxonomique des bactéries responsables d'infections nosocomiales en utilisant des techniques du Deep Learning sur des données de séquençage de l'ARNr 16S. Les méthodes traditionnelles d'identification bactérienne n'ont pas la résolution nécessaire pour classer avec précision les bactéries au niveau de l'espèce. Un modèle d'apprentissage en profondeur est développé et les performances du modèle sont évaluées sur les quatre espèces bactériennes couramment associées (*Escherichia coli, Enterococcus faecalis, Klebsiella pneumoniae et Pseudomonas aeruginosa*) aux infections nosocomiales. Les résultats démontrent la robustesse et la précision du modèle dans la classification taxonomique de ces agents pathogènes. Les résultats de cette recherche contribueront au développement de stratégies de diagnostic et de surveillance plus efficaces, permettant des interventions ciblées pour atténuer l'impact des infections nosocomiales sur les soins aux patients et les systèmes de santé.

Mots clés : Infections nosocomiales, Classification taxonomique, Données génomiques, Apprentissage automatique, Pathogènes bactériens, ARNr 16s.

Président : Pr. HAMIDECHI Mohamed Abdelhafid

Encadreur : Dr**.** DAAS Mohamed Skander

Examinatrice : Dr. DJAMAA Ouahiba