الجمهورية الجزائرية الديمقراطية الشعبية

République Algérienne Démocratique et Populaire

وزارة التعليم العالي والبحث العلمي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

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

Faculté des Sciences de la Nature et de la Vie — 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 :

Méthode d'Apprentissage Automatique pour identification des plantes. Cas d'étude : les céréales

Présenté par **:** KARAALI Ouissal

MAZZI Meryem

MECIBAH Akram

Le 18/06/2023

Président : **Dr. TAMAGOULT Mahmoud** (Université Frères Mentouri, Constantine 1).

Encadreur : **D r. GHERBOUDJ Amira** (Université Frères Mentouri, Constantine 1).

Examinateur : **Dr. DJAMAA Ouahiba** (Université Frères Mentouri, Constantine 1).

Année universitaire

2022 – 2023

Remerciement

*Tout d'abord, nous remercions **Allah** le tout-puissant de nous avoir guidés tout au long de ce parcours académique et de nous avoir donné la force et la motivation nécessaires pour mener à bien ce travail.*

*Nous souhaitons exprimer notre profonde gratitude envers tous les professeurs qui nous ont accompagnés tout au long de notre cursus universitaire, en particulier notre encadrante, **Mme Dr Amira GHERBOUDJ**, pour son soutien, sa patience, sa rigueur et sa disponibilité tout au long de ce travail. Nous remercions également le docteur **Yacer BOUDERSA** pour ses conseils et ses suggestions qui nous ont permis de progresser et de mieux structurer nos idées.*

*Nous sommes profondément reconnaissants envers le **Mlle* **ABDELAZIZ Aya** *pour sa disponibilité et son soutien continu. Le **Mlle** Aya a pris le temps de répondre à nos questions, de discuter de nos idées et de nous prodiguer des conseils précieux pour améliorer notre méthodologie. Sa passion pour la bio-informatique a été une source d'inspiration pour notre équipe.*

*Nous aimerions exprimer notre reconnaissance envers les membres du jury, en particulier le Président du jury, **Dr. TAMAGOULT Mahmoud**, et l'Examinateur, **Dr. DJAMAA Ouahiba**, qui ont honoré ce travail de leur évaluation.*

Nous n'oublions pas nos amis et notre famille qui nous ont soutenus et encouragés tout au long de ce travail. Leurs encouragements et leur soutien moral ont été déterminants dans la réussite de ce projet.

Enfin, nous tenons à remercier tous ceux qui, de près ou de loin, ont contribué à la réalisation de ce mémoire.

Dédicace

Je dédie ce mémoire à :

Tout d'abord, à mes chers parents. Leur amour, leur éducation, leurs sacrifices et leur encouragement tout au long de ce parcours ont été une source d'inspiration et de motivation inestimable pour moi.

Également, à vous, mes chers frères Moncef, Nadjib et Abed El Djalil, en témoignage de mon amour et de ma gratitude éternelle. Vous êtes et resterez toujours mes plus grands soutiens dans la vie.

Je tiens à exprimer ma gratitude sincère et mon appréciation envers ma partenaire dans ce projet et mon amie Ouissal. Tu as été une source d'inspiration et de motivation pour moi. Tes perspectives éclairées et tes efforts inlassables ont grandement enrichi notre travail et ont fait ressortir le meilleur de moi-même.

Et sans oublier mes collègues Yasmina, Loubna et Nour, vous avez été d'une grande aide pour nous. J'apprécie sincèrement tout ce que vous avez fait.

Également, à mes amies Aicha, Cherifa, Hadil, Meryem et Leila. Leur amitié sincère, leur soutien moral et leurs encouragements tout au long de cette aventure ont été précieux pour moi. Leur présence m'a apporté réconfort et motivation.

Enfin, à mes professeurs de Bioinformatique, spécialement. Leur expertise, leur enseignement de qualité et leur accompagnement dans ma formation en Bioinformatique ont été d'une valeur inestimable. Leurs connaissances et leurs conseils précieux ont grandement contribué à mon apprentissage.

Avec toute ma gratitude,

Mazzi Meryem

Dédicace

Je dédie ce travail à mon père, Mohamed Cherif. Je suis honorée d'être votre fille et je vous suis infiniment reconnaissante pour les sacrifices que vous avez consentis pour mon éducation.

Je dédie également ce travail à ma mère, Nadjet. Votre amour inconditionnel, vos encouragements constants et votre soutien inébranlable ont été les piliers de ma vie. Votre présence bienveillante et votre dévouement ont été essentiels pour surmonter les défis et persévérer dans mes études.

Je souhaite exprimer ma gratitude à mes frères, Meriem, Noussair et Nour, qui ont été mes compagnons de route et mes soutiens indéfectibles. Leur présence et leur soutien ont été d'une importance cruciale pour moi, et je suis profondément reconnaissante de les avoir à mes côtés.

Je souhaite exprimer ma profonde reconnaissance envers ma collègue et amie, Mazzi Meryem, pour sa contribution inestimable à notre projet commun. Sa persévérance, sa créativité et son énergie positive ont été une source d'inspiration constante pour moi. Je suis reconnaissante d'avoir pu compter sur elle tout au long de cette expérience.

Sans oublier mes amis, et mes collègues Yasmina, Loubna et Akram pour leur amitié précieuse, leurs encouragements et leur soutien tout au long de ce parcours. Leur soutien inconditionnel, leurs conseils précieux et leur présence réconfortante ont contribué à ma motivation et à ma persévérance.

À toutes ces personnes, je vous suis profondément reconnaissante pour votre amour, votre soutien et votre présence dans ma vie. Ce travail est dédié à vous tous, et je vous remercie du fond du cœur pour votre contribution précieuse à mon parcours académique.

Karaali Ouissal

Dédicace

Ce travail a été achevé grâce à l'aide de Dieu le Tout-Puissant.

Cette mémoire est dédiée à toi, ma merveilleuse mère, qui m'a encouragé à aller de l'avant et m'a donné tout ton amour pour reprendre mes études. Je t'aime plus que les mots ne peuvent le dire.

En mémoire de mon père, qui nous a quittés il y a 16 ans, ainsi qu'à ma sœur Nedjla et mon frère Bilal, pour leur soutien.

Je souhaite également dédier ce travail à toute ma famille, en particulier ma cousine Soufiane et B. Alae Khadija. Mes amis Aymen et Yahia, vous avez été précieux dans mon parcours.

Je tiens également à remercier mes collègues, Meryem et Ouissal, avec qui j'ai pu réaliser un projet dont nous sommes fières.

Je souhaite exprimer ma gratitude envers notre encadrante, ***Mme Dr Amira GHERBOUDJ****, pour tous ses précieux conseils, son écoute active et sa disponibilité. Un grand merci également à* ***Mlle ABDELAZIZ Aya*** *et au docteur* ***Yacer BOUDERSA****.*

Enfin, je tiens à remercier chaleureusement tous ceux qui m'ont aidé de près ou de loin tout au long de mes années d'études. Votre soutien a été d'une valeur inestimable.

MECIBAH Akram

RÉSUMÉ

Cette étude présente une approche basée sur l'intelligence artificielle (IA) pour la classification des espèces de céréales en utilisant les séquences ITS (Internal Transcribed Spacer). La recherche a débuté par une analyse des caractéristiques morphologiques du blé, de l'orge, de l'avoine et du riz. Par la suite, les séquences ITS ont été alignées et ont permis de développer un modèle d'apprentissage automatique basé sur le classifieur de la forêt aléatoire (Random Forest abrégé RF). Le modèle a atteint une précision remarquable de 98% dans la prédiction du genre des céréales en se basant sur les séquences ITS. Cette étude met en évidence le potentiel considérable de l'IA dans la classification des espèces de céréales, avec des implications majeures pour les domaines de l'agriculture et une meilleure compréhension de la diversité des céréales.

Mots clés : Apprentissage Automatique, Céréales, Classification, Forêt Aléatoire, Intelligence Artificiel, ITS.

ABSTRACT

This study presents an Artificial Intelligence (AI) based approach for classifying cereal species using ITS sequences. The research began with an analysis of the morphological characteristics of wheat, barley, oats, and rice. Subsequently, the ITS sequences were aligned, and a Random Forest (RF) classifier-based machine learning model was developed. The model achieved an impressive accuracy of 98% in predicting the cereal genus based on the ITS sequences. This study highlights the significant potential of AI in cereal species classification, with major implications for the fields of agriculture and a better understanding of cereal diversity.

Key words : Machine Learning, Cereals, Classification, Random Forest, Artificial Intelligence, ITS.

ملخص

تقدم هذه الدراسة نهجًا قائمًا على الذكاء الاصطناعي لتصنيف أنواع الحبوب باستخدام تسلسلات ITS .بدأ البحث بتحليل الخصائص المورفولوجية للقمح والشعير والشوفان والأرز. تم محاذاة تسلسلات ITS وتم تطوير نموذج تصنيف قائم على الغابة العشوائية. حقق النموذج دقة مذهلة تبلغ 98٪ في توقع جنس الحبوب بناءً على تسلسلات .ITS تسلط هذه الدراسة الضوء على الإمكانات الكبيرة للذكاء الاصطناعي في تصنيف أنواع الحبوب، مع تداعيات هامة لمجالات الزراعة وفهم أفضل لتنوع الحبوب.

كلمات مفتاحية : التعلم الالي ؛الحبوب ؛التصنيف ؛الغابة العشوائية ؛ الذكاء الاصطناعي؛ ITS .

LISTES DES FIGURES

LISTE DES TABLEAUX

ACRONYMES :

- ADN : Acide désoxyribonucléique
- ARN : Acide ribonucléique
- ARNm : Acide ribonucléique messager
- ARNr : Acide ribonucléique ribosomal
- CAD : Computer-Aided Diagnosis
- CNN : Convolutional neural network
- CSS : Cascading Style Sheets
- CSV : Comma separated values
- ETS : External Transcribed Spacer
- FN : Faux négatif
- FP : Faux positif
- JS : JavaScript
- HTML : HyperText Markup Language
- IA : Intelligence artificielle
- IGS Inter-Genic Spacer
- IRM : Imagerie par résonance magnétique
- ITS : Internal Transcribed Spacer
- KNN : K-Nearest Neighbors
- ML : Machine Learning
- MUSCLE : Multiple Sequence Comparison by Log-Expectation
- NTS : Non-Transcribed Spacer
- RF : Random Forest
- RELP : Restriction Fragment Length Polymorphism
- SNP : Single Nucleotide Polymorphism
- SSP : Subspecie
- SVM : Support Vector Machines
- VN : Vrai négatif
- VP : Vrai positif
- VS : Visual Studio

TABLE DES MATIÈRES

INTRODUCTION GÉNÉRALE

INTRODUCTION GÉNÉRALE

La classification des espèces revêt une importance fondamentale au sein de la communauté scientifique, car elle permet d'identifier et de regrouper les différentes entités en fonction de leurs caractéristiques communes. Au fil du temps, diverses méthodes ont été utilisées, allant des techniques morphologiques traditionnelles aux approches moléculaires avancées. Toutefois, lorsqu'il s'agit de classer les espèces de céréales, ces méthodes traditionnelles peuvent présenter des limitations significatives en termes de subjectivité, de coût et de complexité. En particulier, ces problèmes découlent de l'interprétation des caractères morphologiques, ce qui peut entraîner des résultats subjectifs et peu fiables, créant ainsi une confusion dans l'identification des espèces de céréales. Par conséquent, afin d'améliorer l'efficacité et la précision de la classification des espèces de céréales, il est essentiel d'explorer de nouvelles approches basées sur l'intelligence artificielle, qui permettent d'obtenir des résultats plus objectifs et fiables.

Au fil du temps, l'intelligence artificielle (IA) a joué un rôle de plus en plus important dans différents domaines y compris le domaine de la biologie. Ses avancées technologiques, telles que l'augmentation de la puissance de calcul et l'accès à de grandes quantités de données agricoles, ont permis de développer des applications pratiques dans différents domaines. L'IA est utilisée pour la gestion des cultures, la prédiction des rendements, la détection des maladies des plantes et la classification des espèces végétales. Grâce à des techniques d'apprentissage automatique et d'apprentissage profond, l'IA fournit aux biologistes et agriculteurs des informations précieuses pour prendre des décisions éclairées.

Dans notre mémoire, nous avons développé un modèle d'intelligence artificielle basé sur la classification en utilisant l'algorithme Random Forest que nous avons appliqué sur les séquences ITS (Internal Transcribed Spacer) des céréales. Les séquences ITS sont des régions d'ADN non codant situées entre les gènes ribosomaux. Elles présentent une variabilité génétique spécifique à chaque espèce de céréale. Notre attention se portera principalement sur l'entraînement et l'évaluation de ce modèle afin d'améliorer la précision de la classification des céréales.

Notre manuscrit fournit une présentation détaillée des étapes de construction du modèle de classification. Il est organisé en trois chapitres :

Dans le premier chapitre, nous abordons les généralités sur les céréales, leur importance et leur classification basée sur les caractères morphologiques et les séquences ITS.

Cette compréhension préliminaire est essentielle pour passer au deuxième chapitre, qui se concentre sur l'apprentissage automatique. Ce chapitre explique les phases d'apprentissage automatique, les différents types d'apprentissage et les algorithmes utilisés, en mettant en avant le modèle Random Forest.

Finalement, le troisième chapitre présente l'étude expérimentale, reliant ainsi les deux premiers chapitres. Il décrit le matériel utilisé, la méthode appliquée pour explorer les caractéristiques morphologiques et génétiques, ainsi que la création du modèle d'apprentissage automatique. Les résultats obtenus dans ce chapitre permettent d'évaluer l'efficacité et l'applicabilité du modèle de classification des céréales basé sur l'intelligence artificielle.

Chapitre 01 :

Un regard biologique sur les Céréales

1. INTRODUCTION

Les céréales ont une longue histoire de culture et d'utilisation par les civilisations humaines, sont un aliment de base pour l'homme depuis des milliers d'années, les premières céréales à avoir été domestiquées sont le blé et l'orge, au Moyen-Orient, il y a environ 10 000 ans. Ces cultures étaient bien adaptées au climat sec et chaud de la région et constituaient une source fiable de nourriture pour les premiers agriculteurs. Au fil du temps, d'autres céréales ont été domestiquées et cultivées dans différentes parties du monde, notamment le riz en Asie, le maïs en Amérique, le sorgho et le millet en Afrique [1].

Le terme céréale il vient du mot latin ‘Ceres’ désigne un ensemble de plantes qui sont principalement cultivées pour leurs grains riches en amidon (environ 58 à 72 %) et d’autre constituions (faible quantité de protéines, environ 8 à 13 %, lipides dans une petite quantité 2 à 5%, etc.) [2].

Les céréales occupent à l’échelle mondiale, une place primordiale dans les programmes de recherche agricoles. Elles sont les principales sources de la nutrition humaine (une grande variété de produits alimentaires) et animale dans le monde. Ces dernières années, l'utilisation des céréales pour leurs propriétés fonctionnelles et bénéfiques pour la santé a suscité un intérêt croissant. Par exemple, il a été démontré que certains composants des céréales, tels que les fibres alimentaires, l'amidon résistant et les composés bioactifs, ont des effets bénéfiques sur la santé humaine, notamment en réduisant le risque de maladies chroniques telles que les maladies cardiaques, le diabète et le cancer [1].

Les céréales monocotylédones sont des plantes dont les graines contiennent un embryon avec une seule feuille de cotylédon. Les graines de ces plantes sont également caractérisées par une enveloppe protectrice appelée péricarpe. Les principales céréales monocotylédones les plus importantes cultivées aujourd'hui comprennent le blé (*Triticum.*), le riz (*Oryza sativa*), le maïs (*Zea mays*), l'orge (*Hordeum vulgare*), l'avoine (*Avena sativa*), le seigle (*Secale cereale*), le sorgho (*Sorghum bicolor*) et le millet (*Pennisetum glaucum*). Ces espèces sont toutes diploïdes (2n) ou polyploïdes(exemple, 4n « tétraploïdes » et 6n « hexaploïdies »), avec différents niveaux de ploïdie [3] [1].

2. IMPORTANCE DES CEREALES

Selon le Programme Alimentaire Mondial, les céréales sont la principale source d'alimentation pour plus de 4 milliards de personnes dans le monde, fournissant environ 30 % des calories consommées par l'ensemble de l'humanité. Elles sont relativement faciles à cultiver et à stocker, et peuvent être transformés en une variété de produits alimentaires. Les céréales jouent également un rôle important dans l'économie mondiale, de nombreux pays les considérant comme un produit d'exportation clé [1]**.**

En résumé, les céréales jouent un rôle stratégique dans l'alimentation, l'économie et la politique :

- Alimentation humaine : Elles constituent la base de nombreux régimes alimentaires dans le monde, notamment dans les pays en développement.
- Alimentation animale.
- Économie : Les céréales sont une importante culture de rente dans de nombreux pays.
- Commerce international : Les céréales sont l'une des principales marchandises échangées sur le marché mondial.
- Sécurité alimentaire : Les céréales sont une source de nourriture de base pour de nombreuses populations dans le monde. Les gouvernements et les organisations internationales ont mis en place des politiques et des programmes pour garantir la sécurité alimentaire en cas de pénuries de céréales [4], [5].

3. POSITION SYSTEMATIQUE DES PLANTES ÉTUDIÉES

Les céréales sont des plantes cultivées importantes qui appartiennent à la famille botanique des Poacées (voir figure 1**)**, ou des graminées, ce sont des plantes angiospermes (plantes à fleurs) monocotylédones, c’est-à-dire la graine possède un cotylédon unique. Cette famille comprend plus de 10 000 espèces [6].

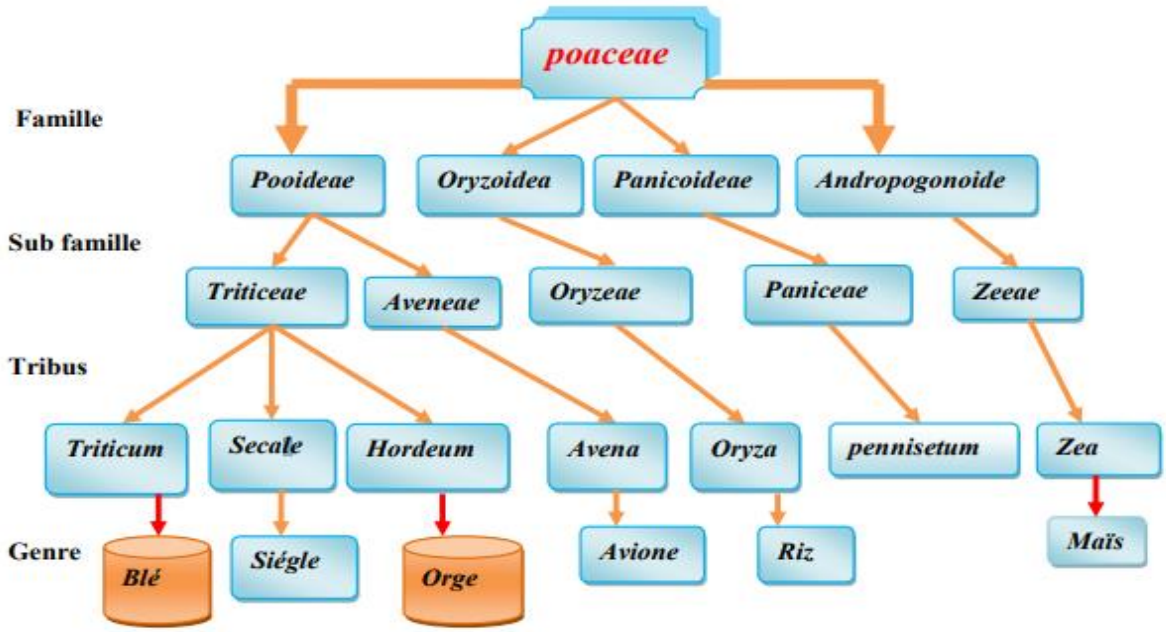

Figure 1 : Taxonomie des céréales [7].

3.1 Blé

Blé, l'une des nombreuses espèces de graminées céréalières du genre *Triticum*. Il existe plus de 20 espèces de blé, différant par leur nombre de base de chromosomes (diplo-, tétra- et hexaploïdies), et plusieurs milliers de variétés. Le blé tendre (*Triticum aestivum*) et le blé dur (*Triticum durum*) sont des variétés de blé les plus couramment utilisée dans la production alimentaire de nombreuses populations à travers le monde, en particulier en Europe, en Amérique du Nord et en Asie [7].

Tableau 1 : classification taxonomique du blé selon van Slageren [9], [10].

Règne		*Plantae*
Division		*Magnoliophyta*
Classe		*Liliopsida*
Ordre		*Poales*
Famille		*Poaceae*
Genre		*Triticum*
Espèce	Diploïde	• *Triticum Urartu* • *Triticum monococcum* (petit épeautre) • *Triticum carthlicum* (blé noir persan)
	Tétraploïde	• *Triticum dicoccoides* (blé amidonnier sauvage) • *Triticum timopheevii* (blé Sanduri) • *Triticum turgidum* (blé vernal) • *Triticum polonicum* (blé polonais)
	Hexaploïdie	• *Triticum aestivum* (tendre)
Sous-espèce		○ *Triticum aestivum ssp. Aestivum* ○ *Triticum turgidum ssp. polonicum* ○ *Triticum turgidum spp. dicoccum* ○ *Triticum turgidum ssp. Durum* ○ *Triticum turgidum ssp. Turgidum* ○ *Triticum turgidum ssp. Turanicum* ○ *Triticum turgidum ssp. carthlicum* ○ *Triticum turgidum ssp. paleocolchicum* ○ *Triticum monococcum ssp. aegilopoides* ○ *Triticum turgidum Ssp.dicoccoides* ○ *Triticum timopheevii ssp. Timopheevii*

3.2 Avoine

Le genre *Avena* comprend environ 450 espèces, dont la plupart sont des plantes sauvages et ornementales. *Avena sativa* est l'espèce la plus importante du genre *Avena* en termes de production agricole et de valeur économique. C'est une plante céréalière annuelle qui pousse dans des climats frais et humides. Elle est principalement cultivée en Europe, en Amérique du Nord et en Australie [8].

Tableau 2 : classification taxonomique du l'avoine [12], [10].

Règne		*Plantae*
Division		*Magnoliophyta*
Classe		*Liliopsida*
Ordre		*Poales*
Famille		*Poaceae*
Genre		*Avena*
Espèce	Diploïde	• *Avena longiglumis* • *Avena eriantha*
	Tétraploïde	• *Avena macrostachya* • *Avena barbata* (avoine barbelée)
	Hexaploïdie	• *Avena sativa* (avoine cultivée) • *Avena sterilis*

3.3 Orge

L'orge appartient à la famille des Poacées (ou des graminées), du genre *Hordeum*. Le nom scientifique de l'orge cultivée est *Hordeum vulgare* L .Il existe également d'autres espèces d'orge (voir tableau 3) [9].

Tableau 3 : classification taxonomique du l'orge [10].

Règne		*Plantae*
Division		*Magnoliophyta*
Classe		*Liliopsida*
Ordre		*Poales*
Famille		*Poaceae*
Genre		*Hordeum*
Espèce	Diploïde	• *Hordeum vulgare* • *Hordeum secalinum* (orge des seigles) • *Hordeum murinum* (orge des rats)
	Tétraploïde	• *Hordeum marinum* (orge marin) • *Hordeum bulbosum* (orge à bulbe) • *Hordeum jubatum* (orge jubatée)

3.4 Riz

Le riz(genre Oryza ; tableau) est une plante annuelle qui est principalement cultivée dans des zones à forte pluviométrie, où les conditions sont favorables pour sa croissance. Le riz est une céréale cultivée en Afrique et en Amérique, bien que sa production ne soit pas aussi importante que dans certaines régions d'Asie. Le riz est une culture céréalière cruciale pour la sécurité alimentaire à l'échelle mondiale. Sa production en quantité suffisante permet de nourrir des millions de personnes chaque jour et contribue ainsi à lutter contre la faim et la malnutrition dans le monde [10].

Tableau 4 : classification taxonomique de riz

Règne		*Plantae*
Division		*Magnoliophyta*
Classe		*Liliopsida*
Ordre		*Poales*
Famille		*Poaceae*
Genre		*Oryza*
Espèce	Diploïde	*Oryza glaberrima* *Oryza punctata*
	Tétraploïde	*Oryza eichingeri* *Oryza latifolia*
	Tétraploïde	*Oryza alata* *Oryza sativa*

4. CLASSIFICATION DES CÉRÉALES

Il existe plusieurs façons selon lesquelles se fait la classification des céréales, et ce travail se concentre en particulier sur les critères morphologiques et génétiques :

4.1 Caractères morphologiques étudiés

Les caractéristiques physiques externes des plantes peuvent être déterminées à partir de l'observation directe de la plante ou à partir des photos, de plus ces caractères peuvent être classés en caractères qualitatifs et quantitatifs.

4.1.1 Qualitatifs

Les caractères qualitatifs des plantes sont des propriétés visibles et discrètes, comme :

- Racine : les caractéristiques morphologiques des racines sont les suivantes [11] :
  - Racine pivotante : c'est-à-dire une racine principale qui se développe directement sous la graine.

- Racines adventives : qui se développent à partir de la base de la tige et qui contribuent à renforcer l'ancrage de la plante dans le sol.
- Système racinaire fasciculé : est composé de nombreuses racines fines et ramifiées.

- Tiges : plusieurs caractéristiques, telles que la rigidité de la tige, Certaines espèces ont des tiges plus rigides que d'autres, ce qui peut être important pour la résistance au vent ainsi les forces externes. D'autres caractéristiques telles que la couleur, la forme, la présence de nœuds, etc., peut également varier selon l'espèce et la variété.
- Feuilles :
  - La forme, la couleur et la disposition des feuilles.
  - La ligule et les oreillettes peuvent également être utilisées dans l'identification et la caractérisation de différentes espèces de plantes [12].
- L'appareil reproductif :
  - La forme et la couleur des graines.
  - La forme de l'épi : l'épi est une inflorescence caractéristique des graminées, qui se compose d'un axe central portant des fleurs unisexuées ou fleuron.
  - La forme de la panicule : Elle est composée d'un axe principal qui porte des ramifications secondaires, appelées rameaux, sur lesquels sont disposées des fleurs. Les fleurs peuvent être unisexuées ou hermaphrodites, selon les espèces de plantes. La panicule peut avoir une organisation plus ou moins dense, selon les espèces [12].

4.1.2 Quantitatifs

Les caractères quantitatifs sont des propriétés qui peuvent être mesurées.

- Tige : la longueur, nombre d'épi, etc.
- Les feuilles :
  - Nombre des feuilles (les céréales produisent généralement plusieurs feuilles avant l'apparition de la tige principale).
  - La longueur et la largeur des feuilles.

- L’appareil reproductif :
  - Nombre d'épis.
  - Nombre de grains par épi.
  - Longueur de l'épi.
  - La taille du grain.

4.2 Critères génétiques

Les critères génétiques peuvent inclure l'analyse de l'ADN pour identifier les similitudes et les différences entre les différentes espèces de céréales.

Les données les plus couramment utilisées pour les études phylogénétiques et les différences entre les espèces sont les séquences biologiques de type acides nucléiques telles que les séquences de gènes unique, des ARNm, des RFLP, des microsatellites, des SNP, des IGS (ARNr et mitochondries), des ITS (ARNr et mitochondries), des séquences de cytochrome c, des séquences alpha du facteur d'élongation et même des protéines enzymatiques ou des séquences structurelles. [13]. Dans notre étude, les données que nous avons utilisées sont les séquences ITS.

4.3 Définition d’ITS

The internal transcribed spacer en anglais, abrégé ITS, sont des régions d'ADN situées entre les gènes d'ARN ribosomal conservés présent que chez les eucaryotes, constitue l'un des marqueurs moléculaires les plus utilisés dans les études phylogénétiques et la différenciation des espèces en particulier champignons des et d’autre plantes [14]. Les ITS sont divisés en ITS1 et ITS2.

- ITS1

L'ITS1 est un morceau d’ADN situé entre les gènes de l'ARNr 18S et 5,8S chez les eucaryotes. La longueur de l'ITS1 est d'environ 200-300 pb. L’ITS1 est moins conservé que l’ITS2, et est impliqué dans le processus de maturation de l'ARN ribosomal [15], [16].

- ITS2

L'ITS2 est un morceau ADN situé entre les gènes de l'ARNr 5,8S et 28S chez les eucaryotes. La variation de longueur de la région ITS2 est plus courte que celle de la région ITS1. La longueur

de l'ITS2 est d'environ 180-240 pb. Toutefois, la région ITS2 est plus conservée que la région ITS1 dans de nombreuses unités taxonomiques. Il est impliqué dans la formation de la structure spatiale de l'ARN ribosomal [16].

Ces régions ont également des fonctions importantes dans la biologie cellulaire et sont souvent utilisées pour construire des arbres phylogénétiques qui reflètent les relations évolutives entre les espèces. En outre, l'analyse de l'ITS1 et l'ITS2 permet de mettre en évidence des phénomènes tels que l'hybridation et l'introgression génétique entre espèces.

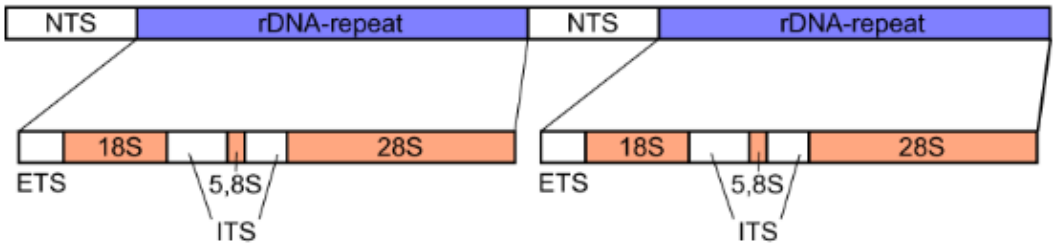

Figure 2 : Les régions d'ITS1 et ITS2 chez les eucaryotes [16].

Afin de prouver les similitudes génétiques entre les genres étudiés de céréales, l'alignement des séquences d'ADN est devenu un outil clé dans les études scientifiques modernes. En fournissant des alignements de séquences des régions ITS1 et ITS2 à travers nos études, nous avons pu mettre en évidence la similarité génétique entre ces différentes espèces de céréales.

4.3.1 Alignement des séquences

L'alignement de séquences est un outil fondamental en bio-informatique qui consiste à aligner deux ou plusieurs séquences d'ADN, d'ARN ou de protéines afin d'identifier les régions de similarité qui peuvent indiquer des relations fonctionnelles, structurelles ou évolutives. En alignant les séquences, les chercheurs peuvent comparer et analyser l'information génétique contenue dans les séquences et identifier les motifs, domaines ou éléments régulateurs conservés [17].

Dans le contexte de l'alignement de séquences multiples (figure 3), l'objectif est d'aligner plusieurs séquences pour identifier les similarités et les différences entre elles. Ce procédé est également utilisé en phylogénie pour reconstruire les histoires évolutives, les arbres

phylogénétiques ont ainsi été essentiels pour avancer notre compréhension de l'histoire évolutive et de la diversité génétique [17], [18].

```
DQ995455.1_1-218        -TCGTGACCCTGACCAAAACAGACCGAGCACGCGTTATCTATTCCTGCTGAGTGGCGGCA
DQ995458.1_1-218        -TCGTGACCCTGACCAAAACAGACCGAGCACGCGTTATCTATTCCTGCTGAGTGGCGGCA
DQ995453.1_1-218        -TCGTGACCCTGACMAAAACAGACCGAGCACGCGTTATCTATTCCTACTGAGTGGCGGCA
KP295986.1_1-216        ---GTGACCCTGACCAAAACAGACCGAGCACGCGTTATCTATTCCTGCTGAGTGGCGGCA
KU883490.1_56-274       GTCGTGACCCTGACCAAAACAGACCGAGCACGCGTTATCTATTCCTGCTGAGTGGCGGCA
                           *********** ******************************.************
```

Figure 3: Alignement multiple des cinq séquences du genre *Avena*.

4.3.2 Arber phylogénétique

Vient de la phylogénie, une des disciplines de la biologie évolutive. Elle consiste à établir les liens de parenté entre des espèces, vivantes ou disparues. Les espèces et leurs relations de parenté peuvent être représentées sous la forme d'un arbre phylogénétique [19]. Un arbre phylogénétique (figure 4) est constitué de nœuds, qui représentent des organismes ancestraux partagés, de feuilles, qui sont les organismes étudiés, et de branches reliant les nœuds entre eux et aux feuilles qui représentent les lignées évolutives d'ancêtre à descendant [20].

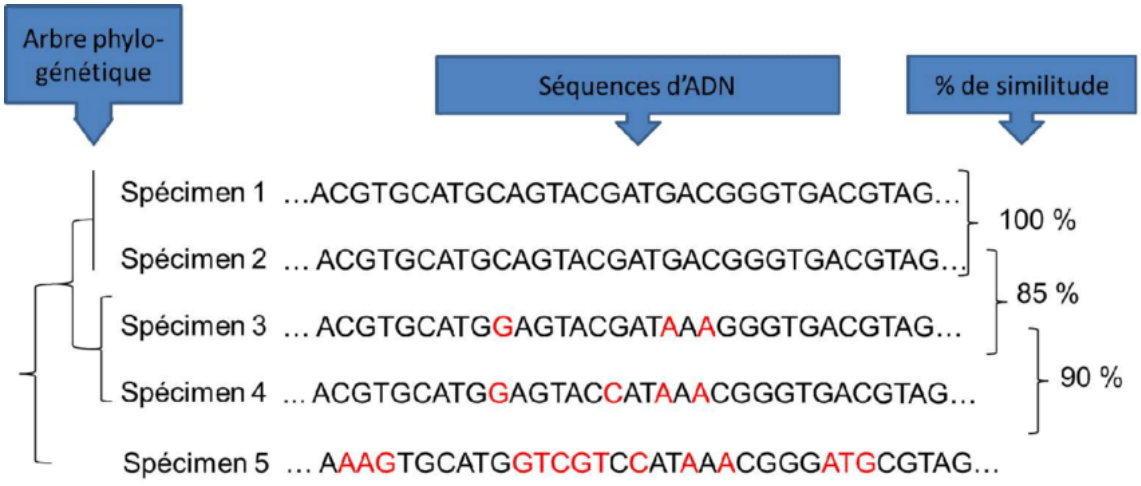

Figure 4 : Alignement de cinq séquences spécimens ADN [24].

Chapitre 02 :
L’apprentissage Automatique

1. INTRODUCTION

L'intelligence artificielle (IA) est une technologie qui existe depuis plusieurs décennies, mais elle a connu une croissance exponentielle ces dernières années grâce aux avancées de la science des données et de l'apprentissage automatique.

Les premiers travaux sur l'IA remontent aux années 1950, mais à l'époque, les capacités des ordinateurs étaient limitées, ce qui rendait difficile le développement de systèmes d'IA efficaces. Avec l'essor des ordinateurs plus rapides, des algorithmes plus sophistiqués et des données plus volumineuses, l'IA est devenue plus utile et plus accessible à un large éventail d'applications pratiques.

Aujourd'hui, l'IA est utilisée dans de nombreux domaines, tels que la médecine, la finance, l'automobile et l'industrie manufacturière, pour améliorer les processus, la qualité, accroître l'efficacité et résoudre des problèmes complexes, réduire les coûts et d'accélérer la prise de décision.

Autrement dit, L'IA ou Intelligence Artificielle, est un domaine de l'informatique qui vise à développer des machines capables de réaliser des tâches qui dépassent normalement l'intelligence humaine, telles que la reconnaissance de la parole, la prise de décisions, la résolution de problèmes, l'apprentissage, la compréhension du langage naturel, etc. [21].

L'IA utilise une variété de techniques, telles que l'apprentissage automatique, la logique symbolique, les réseaux de neurones artificiels, la vision par ordinateur, le traitement du langage naturel et bien plus encore, pour créer des machines qui sont de plus en plus intelligentes et autonomes [21].

Ce chapitre contient les notions essentielles nécessaires pour mieux comprendre les techniques d'Intelligence Artificielle, et nous avons abordé essentiellement l’apprentissage automatique dans notre travail.

2. PRESENTATION APPRENTISSAGE AUTOMATIQUE

L’apprentissage automatique est une branche de l'intelligence artificielle qui permet aux ordinateurs d'apprendre à partir de données et d'expériences, sans être livrés programmés. À partir de ces apprentissages, les ordinateurs sont en mesure de réaliser des tâches telles que la reconnaissance d'images, la traduction automatique, la détection de fraudes, etc. [22].

Cette définition met l'accent sur le fait que l'apprentissage automatique implique l'utilisation de données et d'expériences pour permettre aux machines d'apprendre et de s'améliorer dans leurs tâches, d’identifier des modèles et prendre des décisions avec un minimum d'intervention humaine, cependant l’IA ne remplace pas l’intelligence humaine, mais de l'améliorer en combinant les forces des deux.

« La Machine Learning est la science de donner à une machine la capacité d’apprendre, sans la programmer de façon explicite. »[23].

3. PHASES D’APPRENTISSAGE AUTOMATIQUE

L'objectif principal d'apprentissage automatique est d'entrainer des algorithmes pour réaliser des tâches telles que le traitement des entrés d'une manière efficace et rentable.

L'apprentissage automatique implique un processus d'apprentissage à partir de données pour faire des prévisions ou des décisions sans être explicitement programmé. Ce processus peut être divisé en deux phases : l'apprentissage et l'inférence (prédiction).

3.1 Phase d'apprentissage (ou Entrainement)

Pendant la phase d'apprentissage, un algorithme d'apprentissage automatique est formé sur un ensemble de données d'entrée. L’algorithme va modifier les paramètres et également les données utilisées, puis obtenir une sortie. De plus l’algorithme apprend à identifier des motifs et des relations entre les entrées et les sorties grâce à un processus itératif d'ajustement de ses paramètres. Au fur et à mesure que l'algorithme est formé, il devient meilleur pour faire des prévisions ou des décisions précises basées sur les données d'entrée.

3.2 Phase de prédiction (ou Inférence)

Une fois que l'algorithme a été formé, il peut être utilisé pour faire des prévisions ou des décisions sur de nouvelles données d'entrée pendant la phase d'inférence (prédiction). L'algorithme traite les données d'entrée et utilise les motifs et les relations qu'il a appris pendant la phase d'apprentissage pour générer une sortie [24].

4. CHAMPS D'APPLICATION D'APPRENTISSAGE AUTOMATIQUE

L'apprentissage automatique est devenu une partie intégrante de notre vie quotidienne, des recommandations personnalisées sur les plateformes de médias sociaux aux assistants personnels activés par la voix tels que Siri et Alexa.

Il est également utilisé dans les systèmes de détection de fraude pour protéger nos finances et dans les soins de santé pour un diagnostic précoce et des plans de traitement personnalisés. Les voitures autonomes, les assistants virtuels et la reconnaissance d'images dans les systèmes de sécurité, sans oublier que ce dernier est intervenu dans les sciences biologiques pour l'analyse de séquences génétiques, la prédiction de la structure des protéines, la conception de médicaments, la classification de tissus et la détection de maladies, cela permet aux biologistes d'obtenir des résultats plus précis et plus rapides.

À mesure que la technologie progresse et que les algorithmes s'améliorent, l'impact de l'apprentissage automatique sur la vie humaine ne fera qu'augmenter (figure 5).

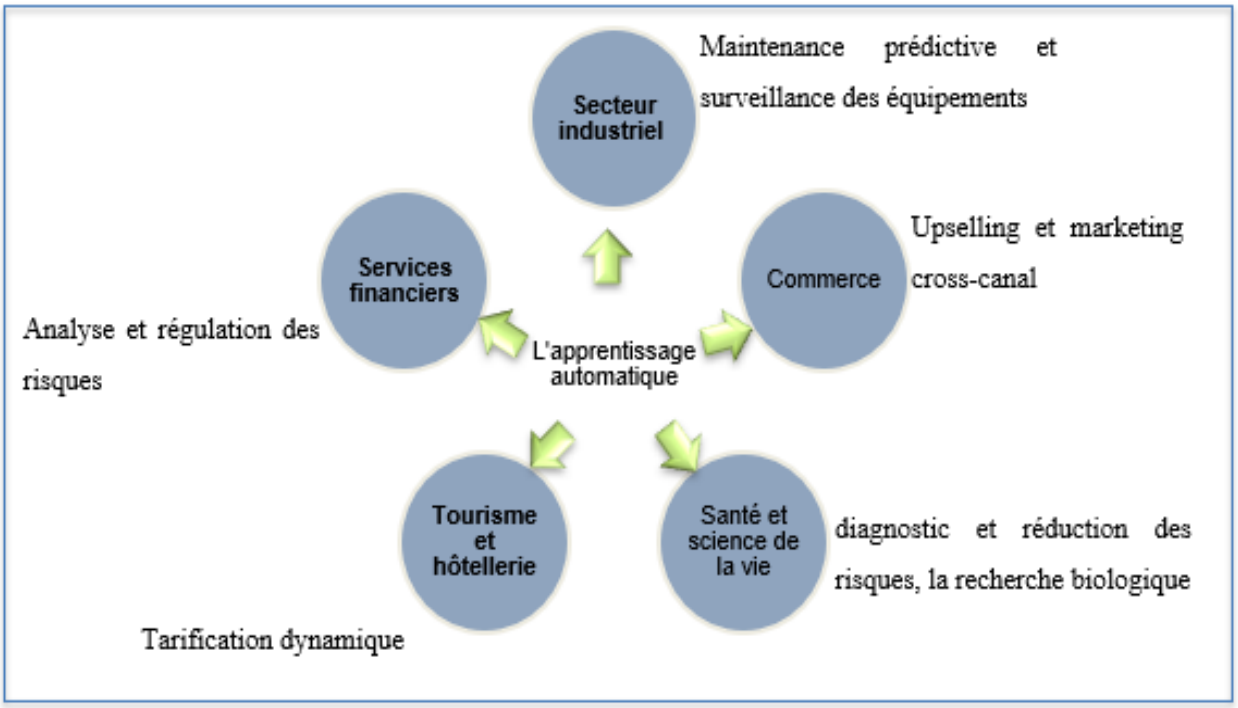

Figure 5 : Champs d'application de la machine learning.

5. EXAMPLE D'APPLICATION DU MACHINE LEARNING

Diagnostic assisté par ordinateur (ou Computer-aided diagnosis (CAD) en anglais) est une méthode qui utilise des techniques de reconnaissance des formes pour identifier les structures suspectes dans l'image. Dans ces cas l'apprentissage supervisé est utilisé pour effectuer cette tâche. Quelques milliers d'images étiquetées sont données à l'algorithme d'apprentissage automatique. Le classificateur qui en résulte est censé classer correctement les nouvelles images médicales [25] (figure 6).

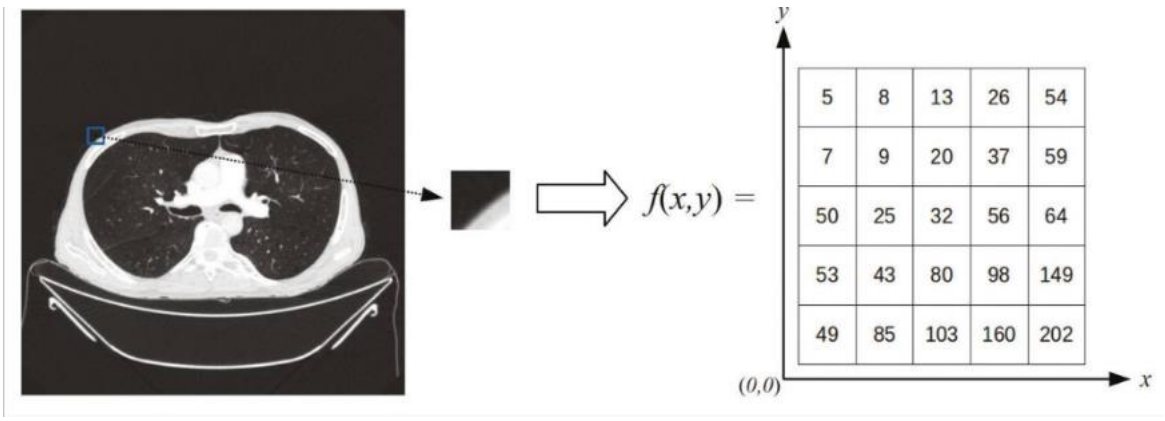

Figure 6 : Représentation de la fonction (matrice) d'une image numérique gris(coupe axiale d'un sein) [30].

6. TYPES D'APPRENTISSAGE AUTOMATIQUE

Il existe plusieurs types d'apprentissage automatique (figure 7), dont les principaux sont :

- L'apprentissage par renforcement.
- L'apprentissage non supervisé.
- L'apprentissage supervisé.

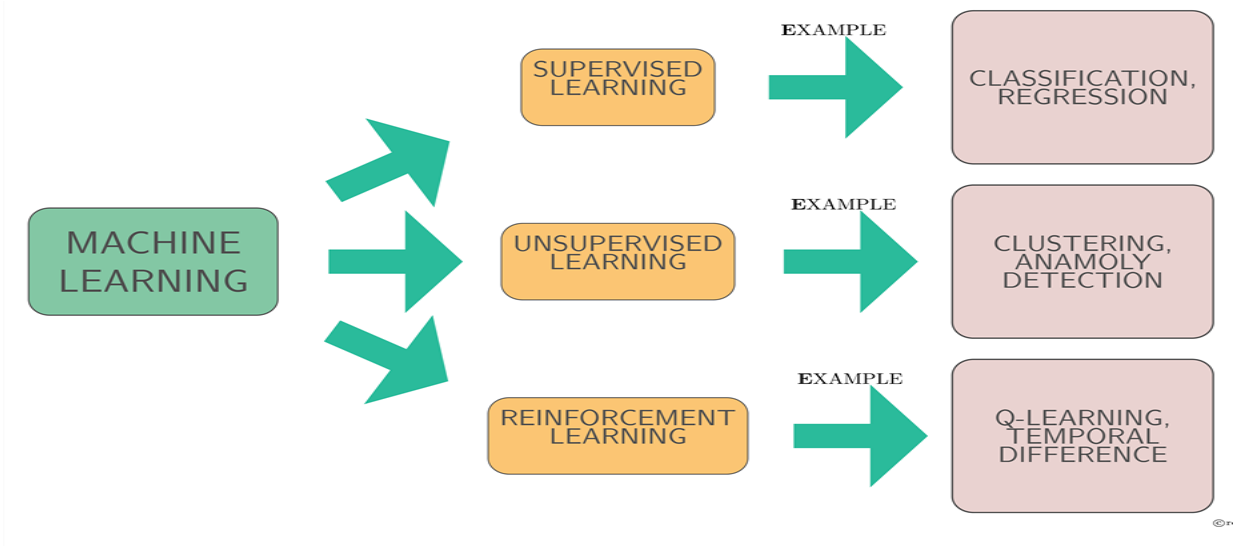

Figure 7 : Les 3 types d'apprentissage automatique avec des exemples [31].

6.1 Apprentissage par renforcement

Une méthode d'apprentissage automatique qui vise à utiliser les observations recueillies lors de l'interaction avec l'environnement pour prendre des actions qui maximisent la récompense ou minimisent le risque, dans le but de produire des programmes intelligents (également appelés agents). Cette méthode est utilisée dans différents domaines pour résoudre des problèmes d'optimisation et de prise de décision [25].

6.2 Apprentissage non supervisé

L'apprentissage non supervisé est une méthode d'apprentissage automatique qui permet d'extraire des informations utiles à partir de données non étiquetées en utilisant des algorithmes de clustering (regrouper des donnés dans des clusters), de réduction de dimension ou de détection d'anomalies.

En d'autres termes, l'apprentissage non supervise manquant de superviseurs ou de données d'entraînement. C'est-à-dire tout ce que nous avons, ce sont des données non étiquetées. L'idée est de trouver une structure cachée dans ces données [25].

6.3 Apprentissage supervisé

L'apprentissage supervisé est une technique d'apprentissage automatique qui consiste à fournir à un algorithme des données d'entrée avec des étiquettes correspondantes pour entraîner le modèle à prédire des sorties pour de nouvelles données d'entrée [26].

D'autre façon, l'apprentissage supervisé est utilisé pour développer des modèles prédictifs, c'est-à-dire des modèles capables de prédire une certaine valeur Y en fonction de variables.

Exemples :

Prédire le nom d'une personne (Y) à partir d'une photo (X), ou bien prédire le prix d'un appartement (Y) en fonction de sa surface habitable (X1) et du nombre de pièces (X2). Pour développer de tels modèles, il faut en premier lieu fournir à la machine une grande quantité de données (X, Y). On appelle cela un dataset (un jeu de données).

Ensuite, on demande à la machine de développer une fonction d'approximation qui représente au mieux la relation $X \rightarrow Y$ présente dans nos données [23].

L'apprentissage supervisé se divise en deux grandes familles de problèmes de l'apprentissage supervisé : la classification et la régression (figure 8), nous porterons une attention particulière à la classification dans ce travail.

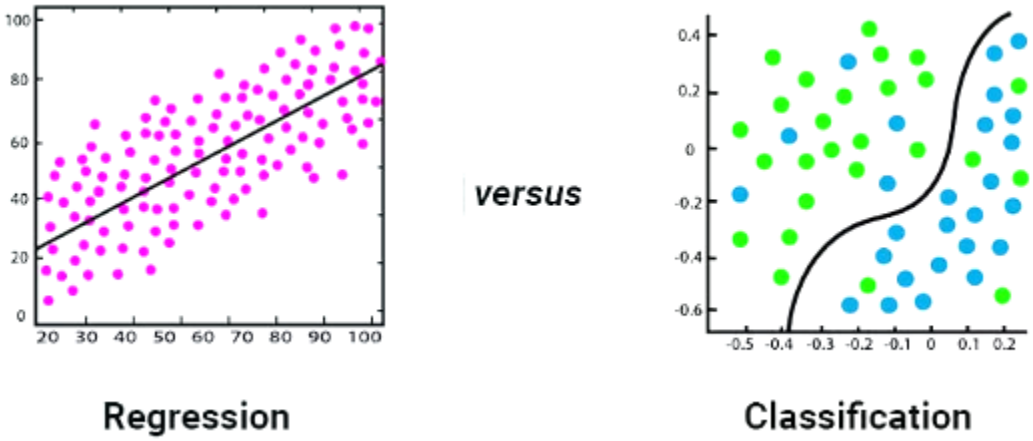

Figure 8 : Illustration de la différence entre classification linéaire et régression linéaire [33].

6.3.1 Régression

La régression est une autre méthode d'apprentissage supervisé qui utilise un algorithme pour comprendre la relation entre les variables dépendantes et indépendantes. Les modèles de régression sont utiles pour prédire des valeurs numériques basées sur plusieurs points de données. Des exemples d'algorithmes de régression sont la régression linéaire, la régression logistique et la régression polynomiale [27].

6.3.2 Classification

L'algorithme de classification fait partie des méthodes d'apprentissage supervisé, c'est-à-dire que les prédictions sont réalisées à partir de données historiques.

À l'inverse de l'apprentissage non supervisé où il n'y a pas de classes prédéfinies, il faut donc constituer les catégories en fonction des attributs communs, pour ensuite réaliser la prédiction, lorsque le résultat attendu est une classe (groupe) [28].

Par exemple :

- Classer un animal : chat, chien, vache ou autre (Multi-label classification) en se basant sur le poids, la longueur et le type de nourriture.
- Ou bien un système d'évaluation des spams : spam ou non.

7. ALGORITHMES D'APPRENTISSAGE AUTOMATIQUE

Les algorithmes les plus couramment utilisés dans le domaine de la bioinformatique sont :

- Support Vector Machines (SVM).
- Random Forest (RF).
- Naïve Bayes.
- K-Nearest Neighbors (KNN).
- Arbres de décision.

7.1 Choix de modèle

Dans ce travail, nous nous concentrerons sur Forêt d'arbres décisionnels (Random Forest) : est un modèle d'apprentissage flexible capable de résoudre les problèmes de régression et de classification. Il fonctionne en construisant plusieurs "arbres de décision" lors de la phase d'apprentissage et en produisant une prévision moyenne à partir de tous les arbres de décision impliqués [29]. Le modelé RF conserve de nombreux avantages des arbres de décision tout en obtenant de meilleurs résultats, il gère les valeurs manquées, une variété de variables (continues, binaires, catégorielles) et est bien adaptée à la modélisation de données à haute dimension. Contrairement aux arbres de décision classiques [30].

Pour mettre en œuvre cette stratégie dans notre cas spécifique de classification des céréales par ITS1 et ITS2, nous construirons un ensemble de nombreux arbres de décision. Chaque arbre devra être capable de prédire la plante de manière acceptable et devra différer des autres arbres. Les forêts aléatoires tirent leur nom de l'introduction de l'aléatoire dans la construction des arbres, garantissant ainsi que chaque arbre est différent. L'aléatoire est utilisé pour sélectionner les données utilisées pour construire chaque arbre et pour sélectionner les caractéristiques utilisées lors de chaque test de division.

7.2 Indicateurs de performance d’un classifieur

le résultat d'un système de classification peut être :

- Vrai positif (True positive) : Le modèle prédit correctement la classe positive. Cela signifie que le modèle identifie correctement qu'un message est indésirable.
- Vrai négatif (True negative) : Le modèle prédit correctement la classe négative. Cela signifie que le modèle identifie correctement qu'un message est désirable.
- Faux positif (False positive) : Le modèle prédit incorrectement la classe positive. Cela signifie que le modèle prédit à tort qu'un message est indésirable.
- Faux négatif (False negative) : Le modèle prédit incorrectement la classe négative. Cela signifie que le modèle prédit à tort qu'un message est désirable [31].

VP, VN, FP et FN respectivement, ces variables (résultats) entrent dans des formules pour mesurent : La justesse, la Précision et Le rappel.

Chapitre 03 :

Étude Expérimentale

1. INTRODUCTION

Ce chapitre combine les connaissances théoriques sur les céréales, la classification des espèces et les techniques d'apprentissage automatique. Nous débutons en présentant le matériel utilisé dans notre étude, y compris les données morphologiques et nucléiques collectées. Ensuite, nous décrivons en détail la méthode que nous avons utilisée, en mettant en avant des techniques traditionnelles telles que l'alignement des séquences ITS et l'analyse phylogénétique. Ensuite, nous détaillons la création du modèle d'apprentissage automatique basé sur l'algorithme Random Forest, ainsi que les résultats obtenus.

2. MATERIEL

2. 1 Données biologiques

Deux types de données ont été utilisés pour la réalisation de notre travail : des données sous forme des caractères morphologiques sur des parties précises des quatre céréales étudiées (blé, avoine, orge et riz) et des séquences nucléiques.

2.1.1 Données morphologiques

Les caractères ont déjà été préalablement identifiées dans d'autres études. Le tableau ci-dessous récapitule les différentes parties des plantes étudiées ainsi que leurs caractéristiques morphologiques qualitatifs et quantitatifs.

Tableau 5 : Caractères morphologiques des quatre céréales.

Organe	Caractère	Caractère state			
		Blé	Avoine	Riz	Orge
Tige	Longueur	1 jusqu'à 1.4 m [32]	0.9 à 1.5 m [33]	0.6 m à 1.3 m [34]	0.3 à 1.2 m
	Nombre d'épi	1 à 5 [35]	2 à 6	3 à 4	2-4-5 et plus [36]
	Forme des pailles (Chaume)	Creuses, pleines ou remplie de moelle [37], tubulaire [38]	Non ramifiées et creuses [39]	ronde, plate ou carrée, dressé ou étalé [40]	Creuses [41]
	Couleur	Verte, jaunâtre ou brune et Dans	Verte ou jaunâtre	blanc à brun	Verte, jaunâtre ou brune

		certains cas, une couleur rougeâtre ou pourpre.	Rougeâtre ou violacée		
	Rigidité	Tige rigide	Dressées et grêles [39]	robuste et stable	dressées et rigides formant une touffe [42]
	Nœuds	Plus lisses	Plus lisses ou poilé	cylindriques, plus épais entre-nœuds	Lisses et plus rapprochés et plus serrés
Racine	Forme	Ramifié, Fusiforme.	Adventice,	Pivotante, Ramifié.	Fibreuse
Panicule	Forme	//	Lâche, plus long et clairsemée [33]	pyramidale ou conique	//
	Nombre des épillets	1	25 à 75 épillets [33]	50 à 500 [43]	1
	Longueur	10 à 20 cm	12 à 20 cm [44]	10 à 30 cm [45]	10 à 15 cm
Epi	Longueur	4 à 15 cm [46]	1.5 à 4.8 cm	10 à 40 cm	5-10 m.
	Forme	Effilées, à bords parallèles, demi-claviformes, claviformes [47]	Grains rangés autour axe central [47]	Épis lâches : ouverts, grains espacés. Épis compacts : denses, grains regroupés	Effilé, à bords parallèles (cylindrique, non divergent, rubané) [47] [48]
	Nombre des graines	25 à 106 [49]	2 à 6	50 à 200	2 ou 6 rangs 24 a 25 [50]
Graine	Taille	3 à 9 mm [51]	0.5 à 1 cm de long [52]	2 et 3mm [45]	Presque 1cm [53]
	Forme	Ovale ou en forme de goutte [54]	Allongé et plutôt fin [46]	allongée et ovale, pointu[54]	Ovales et allongées
	Couleur	variant du roux au blanc[55]	blanches, jaunes/grises, noires [39]	blanches, noirs, verts [56],brun, jaunes, rouge ou violet	Jaune doré, Blanc, Rouge, noir [57]

Feuilles	Largeur	1.5 à 2 cm [58]	5 à 16 mm [59]	0.5 à 1.45 cm [45]	5 mm [60]
	Longueur	15 à 20 cm [58]	5 à 40 cm [59]	12-28 cm [45]	5 à 15 mm [61]
	Nombre	7 ou 8 [62]	3 à 5	5 à 15	Presque 14 feuilles [50]
	Présence des oreillettes	Des oreillettes velues [58]	Sans oreillettes	Sans oreillettes	Très embrassant
	Forme	rubanées, à nervures parallèles, et terminées en pointe [38]	glabres, longues et effilées et engainent les tiges [63]	lancéolées, étroites, parallèles [45]	tordues avec dommages dus au gel [64]
	Couleur	vert on glauque[65]	Vert plus pâle [66]	vert ou vert noirâtre [45]	vert plus clair [64]
Fleure	Forme	Très petite et sans éclat visible [67]	En panicules d'épillets tombants, protégés par deux glumes [59]	hermaphrodites, panicule regroupée [68]	Les fleurs sont arrangées en épillets
	Nombre	2 à 4 par épillet [58]	2 à 3 fleurs dans la panicule [69]	6 dans la panicule [68]	1 seule fleur par épillet [58]
	Longueur	//	20 à 25 mm [70]	//	90 à 200 mm [61]

2.1.2 Données nucléiques

Le dataset utilisé dans la deuxième partie de ce travail est un ensemble de séquences d'ITS1 et d'ITS2 des quatre genres des céréales. Les séquences correspondantes ont été téléchargées depuis la banque de données Gene Bank au format fasta, qui est couramment utilisé pour stocker des séquences nucléiques ou protéiques.

- Les dataset téléchargé est composé de huit fichiers (tableau 6) :

Tableau 6 : Description du contenu du 8 fichiers utilisées.

Nom du fichier	Taille	Description
avena-ITS1	103 KB	Une collection de huit fichiers, chacun contenant environ 200 séquences. Ces séquences appartiennent à quatre types de céréales différents : le blé, le riz, l'avoine et l'orge. Les séquences sont réparties entre les fichiers, et chaque fichier contenir un nombre variable de séquences pour chaque type de céréales. Les séquences représentent des régions spécifiques du génome des céréales étudiées.
avena-ITS2	55.0 KB	
hordeum-ITS1	95.8 KB	
hordeum-ITS1	97.8 KB	
oryza-ITS1	44.0 KB	
oryza-ITS2	46.5 KB	
triticum-ITS1	95.9 KB	
triticum-ITS2	66.2 KB	

- Organisation du fichier FASTA

- En-tête de séquence : Chaque séquence est précédée d'une ligne d'en-tête qui contient l'identifiant ou une description de la séquence. La ligne d'en-tête commence par le symbole '>'.

- Ligne de séquence : La ligne suivante contient la séquence elle-même, qui peut être de l'ADN, de l'ARN ou des protéines, voici un exemple (figure 9).

>AJ242402.1 Triticum Urartu internal transcribed spacer 1 (ITS1)

TCGTGACCCTGACCAAAACAGACCGTGCACGCGTCATCCAATCCGTCGGCGGTGG
CATTGTCCGTCGCTCGGCCAATGCCTCGACCACCTCCCCTCCTTGGAGTGGGTGGG
GGCTCGGGGTAAAAGAACCCACGGCGCCGAAGGCGTCAAGGAACACTGTGCCTAA
CCCGGGGGCATGGCTAGCTTGCTAGTCATCCCTTGTGTTGCAAAGCTATTTAATC

Figure 9 : Exemple d'un fichier format FASTA.

2. 2 Configuration de la machine

Le tableau suivant détaille les caractéristiques de l'ordinateur utilisé, qui est une machine simple.

Tableau 7: Les caractéristiques de l'ordinateur utilisé pour l'apprentissage.

Ordinateur	Caractéristiques
Processeur	Intel(R) Core (TM) i3-3110M CPU @ 2.40 GHz 2.40 GHz
Mémoire installée RAM	4.00 Go DDR3
Stockage	500 Go
Système d'exploitation	Windows 10 professionnel
Type de système	Système d'exploitation 64 bits

2. 3 Logiciels

2.3.1 Environnement de travail

Le travail est réalisé en utilisant le langage de programmation Python, via Jupyter notebook de l'anaconda, et également le logiciel Geneious prime :

- Python 3.8.7 : est un langage de programmation open source créé en 1990 par Guido van Rossum et est développé par la Python Software Foundation. Python se distingue par sa lisibilité et son esthétique, privilégiant l'utilisation de mots anglais plutôt que de symboles de

ponctuation et Il dispose également d'une syntaxe plus concise par rapport à d'autres langages structurés [71], [72].

- Anaconda 22.9.0 : Anaconda est une distribution open-source qui regroupe les langages de programmation Python et R dans le domaine de la science des données. Son objectif principal est de simplifier la gestion et le déploiement des paquets. En plus de cela, Anaconda offre une interface utilisateur graphique appelée Anaconda Navigator, qui propose une alternative visuelle à l'interface en ligne de commande. Anaconda Navigator est inclus dans la distribution d'Anaconda et permet aux utilisateurs de rechercher, installer, exécuter et mettre à jour des paquets de manière conviviale [73].
- Jupyter Notebook 6.4.12 : Jupyter Notebook est une application web open-source qui permet aux utilisateurs de créer et de partager des documents contenant du code en temps réel, des équations, des visualisations et du texte explicatif. Il s'agit d'une application client-serveur qui permet d'éditer et d'exécuter des documents de notebook directement à partir d'un navigateur web. Jupyter Notebook est polyvalent et peut être utilisé dans diverses tâches telles que le nettoyage de données, la simulation, la modélisation, la visualisation, l'apprentissage automatique, et bien d'autres. Il facilite la collaboration et le partage de notebooks avec d'autres utilisateurs, et il prend en charge plusieurs langages de programmation [74], [75].
- Geneious Prime 2023.1 : Geneious Prime est une application logicielle développée par Biomatters, un leader technologique qui fournit des logiciels intuitifs mais puissants pour l'analyse des données d'ADN. Fondée en 2003, l'objectif principal est de créer des solutions de bio-informatique pour l'analyse, l'interprétation et le traitement des séquences moléculaires. Il vise à simplifier la gestion des données et la complexité computationnelle, permettant ainsi aux scientifiques de se concentrer sur leurs recherches [76].
- Visual studio code 1.78.0 : est un éditeur de code open-source développé par Microsoft en 2015. Il est hautement personnalisable, permettant aux utilisateurs d'ajouter des langages, des débogueurs et des outils à leur installation pour répondre à leurs besoins spécifiques en développement. VS Code propose une vaste bibliothèque d'extensions et est également utilisé dans divers projets.

- HTML : est un langage de balisage standard utilisé pour créer et structurer des pages web. Il est considéré comme relativement simple à apprendre et constitue souvent le point de départ pour ceux qui souhaitent se lancer dans le développement web [77].
- CSS : est un langage de feuille de style utilisé pour décrire et séparer la présentation d'un document HTML de son contenu. Il permet aux développeurs de créer un aspect cohérent sur plusieurs pages ou sites et de créer des designs réactifs qui s'adaptent à différentes tailles d'écran et appareils [78].
- JS : JavaScript est un langage de programmation utilisé principalement pour créer des pages web interactives. Il permet aux développeurs d'ajouter des fonctions dynamiques aux pages web, telles que des animations, des effets visuels et des interactions avec l'utilisateur. JavaScript est souvent utilisé en conjonction avec HTML et CSS pour créer des sites web modernes et interactifs. Il s'agit d'un langage côté client, ce qui signifie qu'il s'exécute sur le navigateur de l'utilisateur plutôt que sur le serveur [79].

2.3.2 Bibliothèques

Le tableau ci-dessous présente une liste des différentes bibliothèques utilisées dans notre travail, accompagnées de leur description et de leur version correspondante.

Tableau 8 : Les bibliothèques python utilisées.

Bibliothèque	Description
Matplotlib 3.5.2	Une bibliothèque de visualisation de données en Python qui permet de créer des graphiques de qualité professionnelle. Elle offre une flexibilité pour créer une variété de graphiques en 2D et en 3D [80].
NumPy 1.23.5	Un package fondamental pour le calcul scientifique avec Python qui fournit un objet de tableau multidimensionnel, ainsi qu'un ensemble de routines pour des opérations rapides sur des tableaux, y compris les opérations mathématiques, logiques, la manipulation de forme, le tri, la sélection, l'algèbre linéaire de base, les opérations statistiques de base, la simulation aléatoire et bien plus encore [81].
Pandas 1.4.4	Est une bibliothèque open-source de traitement de données pour le langage Python. Elle offre des structures de données et des fonctions pour la manipulation et l'analyse de données tabulaires et des séries temporelles [82].
Scikit-learn 1.0.2	Également connu sous le nom de sklearn, est une bibliothèque de machine learning libre et open-source pour le langage de programmation Python. Il est efficace pour la construction de modèles d'apprentissage automatique [83].
Joblib 1.2.0	Est une bibliothèque Python populaire utilisée pour le traitement parallèle, la sérialisation et la persistance des objets Python. Elle est principalement utilisée pour le caching des résultats de calcul coûteux et la parallélisation des tâches [84], [85].

TensorFlow 2.12.0	Est une bibliothèque open-source de calcul numérique et de machine learning, développée par Google. Elle permet de créer et d'exécuter des graphes de calcul pour les tâches de traitement de données et d'apprentissage automatique. Il est également compatible avec plusieurs langages de programmation, notamment Python, C++ et Java [86].
Seaborn 0.12.2	Est une bibliothèque permettant de créer des graphiques statistiques en Python. Elle est basée sur Matplotlib, et s'intègre avec les structures Pandas [87].
Regex 2023.6.3	En Python, une expression régulière (ou "regex") est une séquence de caractères qui forme un motif de recherche. Elle peut être utilisée pour vérifier si une chaîne de caractères contient un motif spécifique ou pour extraire des informations spécifiques d'une chaîne de caractères[88], [89].

3. MÉTHODES

L'ensemble du processus du travail est divisé en deux sections principales, la section d'analyse des caractéristiques morphologiques des céréales et la section de l'alignement et de la réalisation du model ML.

3.1 Explorations des caractéristiques morphologiques

Dans cette partie, nous avons développé une application web qui permet d'identifier les céréales étudiées en utilisant les caractères morphologiques collectés au préalable. Cette application est conçue sous la forme d'une interface utilisateur interactive. La structure du code de cette interface est la suivante :

- Fichier HTML : Code de structure de la page (Figure 10), en incluant des éléments tels que le titre, la description, le conteneur principal et les emplacements pour les choix et les résultats. Il lie également une feuille de style externe (style.css) pour gérer l'apparence de la page.

```
<!DOCTYPE html>
<html lang="en">
<head>
    <meta charset="UTF-8">
    <meta http-equiv="X-UA-Compatible" content="IE=edge">
    <meta name="viewport" content="width=device-width, initial-scale=1.0">
    <title>Quiz</title>
    <link rel="stylesheet" href="style.css">
</head>
<body>
```

Figure 10 : Structure HTML.

- Fichier CSS : Le code CSS fournit des styles pour les éléments de la page, tels que l'arrière-plan, les dimensions du conteneur, la mise en page des choix, les styles des résultats, et l'affichage des images.
- Fichier JS : Il s'agit d'un code JavaScript qui utilise un objet et des fonctions pour permettre aux utilisateurs de classifier les céréales en fonction de leurs caractéristiques morphologiques. Les caractéristiques morphologiques sont définies dans un objet appelé "plantes" qui contient des informations sur différentes parties des plantes (Figure 11).

```
var start = plantes;

function CreateQuiz(obj){
    var keys = Object.keys(obj)
    var t = ""
    keys.forEach((key)=>{
        t+=`
            <div class="choice" onclick="NextClass('${key}')">
                ${key}
            </div>
        `
    })
    document.querySelector('#choices').innerHTML = t;
}
```

Figure 11 : Structure de code JS.

3.2 Exploration génétique

Cette étape est elle-même divisée en deux approches :

3.2. 1 Alignement des séquences

Nous avons utilisé le logiciel Geneious Prime pour réaliser un alignement global de nos séquences en utilisant l'algorithme MUSCLE qui est réputé pour sa capacité à traiter des ensembles de données de taille modérée et des séquences présentant un niveau modéré de divergence. Il utilise une approche itérative pour aligner les séquences.

L'alignement a été réalisé à trois niveaux : intraspécifique, interspécifique et entre genres.

- Au premier niveau : nous avons aligné les séquences d'ITS provenant de différentes variants d'une même espèce.
- Au deuxième niveau : dans le deuxième niveau, nous avons comparé et alignée les séquences consensus (une séquence composite résultant de l'alignement de multiples séquences similaires) obtenus pour chaque espèce afin d'analyser les similarités et les différences entre les espèces étudiées.
- Au troisième niveau : fait un alignement entre les quatre genres étudiés, en utilisant les séquences consensus obtenus de l'étape précédente.

3.2. 2 Création du model ML

Cette partie décrit les différentes méthodes utilisées pour atteindre l'objectif de notre approche. Pour commencer notre travail, nous importons les bibliothèques Python nécessaires en utilisant les commandes présentées dans la figure 12.

```
import re
import joblib
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.metrics import accuracy_score,precision_score,recall_score
from sklearn.preprocessing import LabelEncoder
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences
from sklearn.metrics import confusion_matrix
```

Figure 12 : Code python pour l'importation des bibliothèques nécessaires.

i Prétraitement des données

Nous utilisons différentes fonctions pour filtrer et nettoyer les données, les rendant ainsi plus appropriées. Voici les principales étapes :

- Collecte et Nettoyage des données
  - Les données utilisées proviennent de la base de données GenBank sous la forme de fichiers FASTA et ont été regroupés en un seul fichier.
  - Par la suite, les séquences redondantes, les séquences de moins de 100 paire de bases et celles contenant des caractères ambigus autres que A, T, G et C ont été éliminées de l'ensemble de données (figure 13).

```
    for line in contenu:
        all_ += 1
        if len(line) > 100:
            seq = "\n".join(line.split("\n")[1:])
            _1 = line.split("\n")[0]
            if seq not in sequences and "N" not in seq and len(seq) >= 100:
                fichier.append(">" + _1 + "\n" + seq)
                p += 1
                sequences.append(seq)
                labels.append(fasta_file)

fichier = ">".join(fichier)  # Ajouter ">" au début de toutes les séquences

open("dt/traite/sequences_traite.fasta", "w").write(fichier)
print("terminé ", all_, p)
```

```
terminé  1694 796
```

Figure 13 : Code de nettoyage et fusion des séquences dans fichier FASTA.

- Les séquences obtenues sont stockées consécutivement dans un fichier csv (un format de fichier couramment utilisé pour stocker des données tabulaires, telles que des tableaux ou des bases de données simples) pour faciliter leur utilisation ultérieure (Figure 14). Ensuite, nous utilisons la méthode ‘str.extract’ sur la colonne "Header" d'un DataFrame pour extraire les noms de céréales en utilisant des expressions régulières. Les noms extraits sont ensuite stockés dans une nouvelle colonne appelée "Cereal" (Figure15).

```
fasta_file = 'dt/traite/sequences_traite.fasta'
csv_file = 'unique_sequences1.csv'
fasta_to_csv(fasta_file, csv_file)
# Read the dataset containing ITS sequences
data = pd.read_csv('unique_sequences1.csv')
data
```

Figure 14 : Code de conversion du ficher FASTA en CSV.

```
# Extract cereal names from the Header column using regular expressions
data['Cereal'] = data['Header'].str.extract(r'(Avena|Oryza|Triticum|Hordeum)', flags=re.IGNORECASE)

# Filter out rows where the cereal name is missing
data = data.dropna(subset=['Cereal'])

# Drop the Header column
data = data.drop('Header', axis=1)

# Reset the index of the DataFrame
data = data.reset_index(drop=True)

# Print the updated DataFrame
data
```

Figure 15: Code d'ajout de la colonne 'Cereal' contenant les noms extraits.

- Diviser les données

Les colonnes 'Sequence' et 'Cereal' sont conserves en deux listes (figure 16), séquences d'entrée (X) et étiquettes correspondantes (y). Cette préparation des données est importante pour faciliter le traitement ultérieur.

```
data['Sequence'] = data['Sequence'].astype(str)
data['Cereal'] = data['Cereal'].astype(str)

# Split the data into input sequences (X) and labels (y)
X = data['Sequence'].values.tolist()  # Convert X to a list
y = data['Cereal'].values
```

Figure 16 : Code de séparation du Dataset en deux listes.

ii Apprentissage :

- Représentation numérique

Les séquences d'ADN peuvent être converties en une forme numérique pour facilite leur utilisation dans les modèles d'apprentissage automatique (figure 17). Dans ce travail, les séquences d'entrée sont encodées en entiers en utilisant la classe Tokenizer de TensorFlow au niveau des caractères. Les étiquettes sont également encodées en entiers à l'aide de la classe 'LabelEncoder' de 'scikit-learn'.

```
# Tokenize the input sequences
tokenizer = Tokenizer(char_level=True)
tokenizer.fit_on_texts(X)
X = tokenizer.texts_to_sequences(X)

# Perform label encoding on the label column
label_encoder = LabelEncoder()
y = label_encoder.fit_transform(y)
```

Figure 17 : Encodage des séquences et les étiquètes.

- Remplissage des séquences

Le remplissage des séquences a été réalisé en utilisant une fonction ‘pad_sequences’. Cette fonction vise à remplir les séquences numériques qui ont une longueur inférieure à la longueur cible en ajoutant des zéros à la fin, afin d'obtenir la même longueur cible pour toutes les séquences.

- Création du modèle

Tout d'abord, Les données sont divisées en ensembles d'apprentissage et de test. L'ensemble d'apprentissage comprend 80% des données (636 séquences), tandis que l'ensemble de test comprend 20% des données (160 séquences). Cette division est réalisée de manière aléatoire en utilisant la fonction `train_test_split`.

Ensuite, un modèle de classification Random Forest est créé en instanciant la classe `RandomForestClassifier`. Le modèle est ensuite ajusté aux données d'apprentissage en utilisant la méthode `fit`. Les ensembles d'apprentissage sont utilisés pour entraîner le modèle, permettant ainsi au modèle d'apprendre à partir des données et de trouver les relations entre les séquences d'entrée et les étiquettes correspondantes (figure 18).

```
# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=100)
print("Training set size:", len(y_train))
print("Testing set size:", len(y_test))
# Create an instance of the Decision Tree classifier and fit it to the training data
model = RandomForestClassifier()
model.fit(X_train, y_train)
```

Figure 18: Code et modèle de classification

Ensuite, le modèle est utilisé pour faire des prédictions sur les données de test en utilisant la méthode ‘predict’ (voir figure 19).

```
# Make predictions on the testing data
predictions = model.predict(X_test)
```

Figure 19 : Code pour faire des prédictions sur les données de test.

- Evaluation

Nous utilisons les étiquettes réelles `y_test` et les étiquettes prédites `predictions` pour calculer les mesures d'évaluation du modèle (figure 20). Ces mesures nous aident à évaluer les performances du modèle en termes de précision, de rappel et d'exactitude.

```
accuracy = accuracy_score(y_test, predictions)
precision = precision_score(y_test, predictions, average='weighted')
recall = recall_score(y_test, predictions, average='weighted')
```

Figure 20 : Code de calcule des performances du modèle.

- Enregistrement

Cela nous permet de reprendre l'utilisation du modèle sans avoir à le réentraîner à chaque fois.

```
joblib.dump(model, 'ML_Model.joblib')
```

Figure 21: Code pour la sauvegarde du modèle.

iii Prédiction

En utilisant le modèle, nous faisons des prédictions sur la nouvelle séquence en utilisant la méthode `predict` (figure 22). Le résultat de cette prédiction est le type de céréale qui correspond à la séquence donnée.

```
# Define the new sequence to predict
new_sequence = "TCGTGACCCTGACCAAAACAGACCGCGCACGCGTCATCCAATCCGTCGGCGACGGCACCGTCCGTCGCTCGGCCAATGCCTCGACCACCTCCCCTCCTCGGAGCG(

# Preprocess the new sequence
new_sequence = tokenizer.texts_to_sequences([new_sequence])
new_sequence = pad_sequences(new_sequence, maxlen=max_sequence_length)

# Make predictions on the new sequence
new_prediction = model.predict(new_sequence)
# Convert the predicted label back to its original value
new_predicted_label = label_encoder.inverse_transform(new_prediction)
print("Predicted Cereal:", new_predicted_label)
```

Figure 22 : Code de prédictions sur la nouvelle séquence.

4. RESULTATS

Le résultat de ce travail est divisé en trois parties : la première partie consiste à la classification des quatre céréales étudiées à partir d'un ensemble de caractères morphologiques de ces plantes. Et également à l'implémentation de l'alignement multiple des séquences ITS dans le but de définir une autre facette de différence entre les céréales avant d'appliquer des méthodes modernes.

La seconde partie consiste à la création du modèle ML pour prédire la classification des séquences en utilisant Random Forest Classifier.

4.1 Résultat de la phase 1

4.1.1 Classification morphologique

Dans cette partie, nous présentons l'interface utilisateur permettant l'identification des céréales en utilisant les caractères morphologiques comme base (figure 24). L'objectif de cette fonctionnalité était de permettre aux utilisateurs de tester leurs connaissances et de leur fournir des résultats personnalisés en fonction de leurs réponses. Le code affiche des choix de classification et met à jour les résultats en fonction des sélections de l'utilisateur, en affichant une image correspondante aux plantes classées (figure 25).

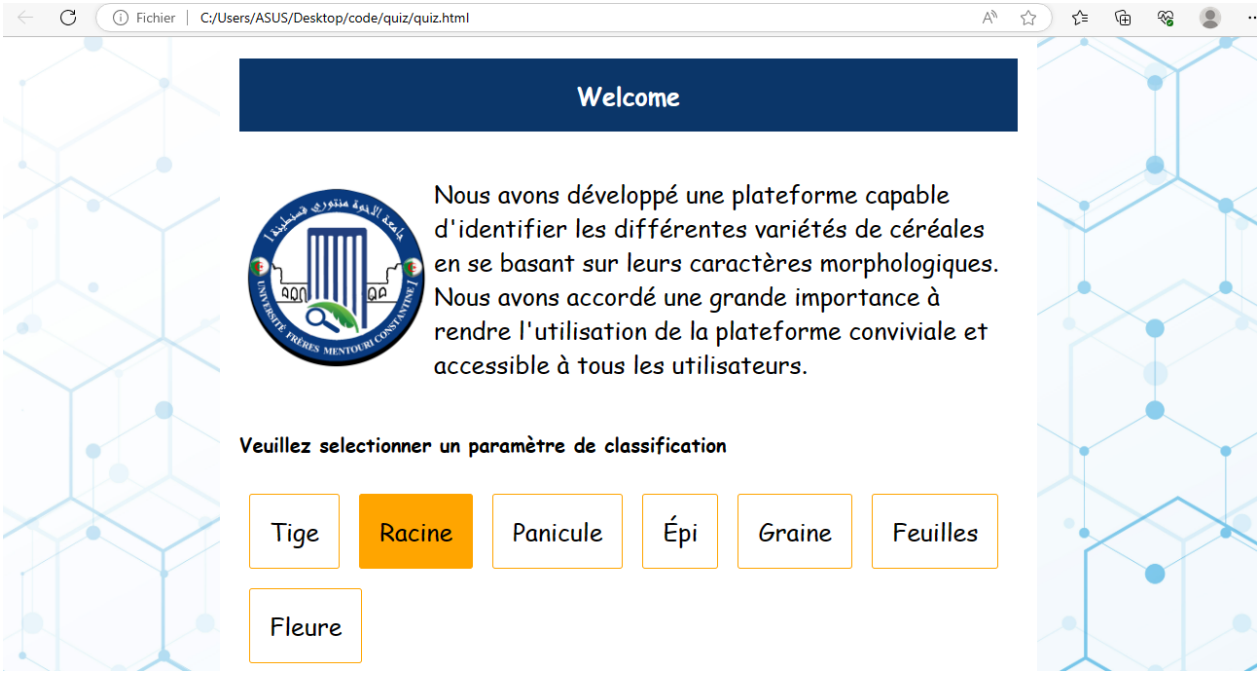

Figure 23 : Interface utilisateur

Figure 24 : Exemple du résultat.

4.1.2 Résultat d'alignement des séquences

L'Alignement multiple des séquences nucléiques d'ITS1 et d'ITS2 chez les quatre plantes *Triticum*, *Oryza*, *Avena* et *Hordeum* sur trois niveaux, les résultats obtenus sont comme le suit :

i. Au premier niveau

Nous avons effectué un alignement multiple des séquences de chaque espèce en regroupant un ensemble de séquences provenant de différentes variantes. Cette procédure a été répétée pour toutes les espèces étudiées, les résultats sont affichés dans les figures suivantes :

```
Consensus          TAAAAGAACCCACGGCGCCGAAGGCGTCAAGGAACACTGTGCCTAACCCGGGGGCATGGC  180
AJ242402.1         TAAAAGAACCCACGGCGCCGAAGGCGTCAAGGAACACTGTGCCTAACCCGGGGGCATGGC  180
AY450265.1:1-222   TAAAAGAACCCACGGCGCCGAAGGCGTCAAGGAACACTGTGCCTAACCCGGGGGCATGGC  180
KF031132.1:43-263  TAAAAGAACCCACGGCGCCGAAGGCGTCAAGGAACACTGTGCCTAACCCGGGGGCATGGC  180
```

Figure 25 : Alignement multiples des séquences d'espèces *Triticum urartu.*

```
Consensus   AGAACCCACGGCGCCGAACGCGTCAAGGAACACTGTGCCTAACCCGGGGGGAATGGCTAGC  180
AJ288102.1 AGAACCCACGGCGCCGAACGCGTCAAGGAACACTGTGCCTAACCCGGGGGGAATGGCTAGC  179
AJ288101.1 AGAACCCACGGCGCCGAACGCGTCAAGGAACACTGTGCCTAACCCGGGGGGAATGGCTAGC  179
AY255038.1 AGAACCCACGGCGCCGAACGCGTCAAGGAACACTGTGCCTAACCCGGGGGGAATGGCTAGC  180
```

Figure 26 : Alignement multiple des séquences de l'espèce *Oryza alta.*

```
Consensus          CCGTCGTCGCTCAGCCAAATCCTCGATAACCTCCTCTCCTTGGAGTGGGGGCTTGGGGTA  120
KP295986.1:1-216   CCGTCGTCGCTCAGCCAAATCCTCGATAACCTCCTCTCCTTGGAGTGGGGGCTTGGGGTA  117
KU883490.1:56-274  CCGTCGTCGCTCAGCCAAATCCTCGATAACCTCCTCTCCTTGGAGTGGGGGCTTGGGGTA  120
DQ995453.1:1-218   CCGTCGTCGCTCAGCCAAATCCTCGATAACCTCCTCTCCTCGGAGTGGGGGCTTGGGGTA  119
```

Figure 27 : Alignement multiple des séquences de l'espèce *Avena sterilis.*

```
Consensus  AGAACCCACGGCGCCGAACGCGTCAAGGAACACTGTGCCTAACCCGGGGGGAATGGCTAGC  180
AJ288102.1 AGAACCCACGGCGCCGAACGCGTCAAGGAACACTGTGCCTAACCCGGGGGGAATGGCTAGC  179
AJ288101.1 AGAACCCACGGCGCCGAACGCGTCAAGGAACACTGTGCCTAACCCGGGGGGAATGGCTAGC  179
AY255038.1 AGAACCCACGGCGCCGAACGCGTCAAGGAACACTGTGCCTAACCCGGGGGGAATGGCTAGC  180
```

Figure 28 : Alignement multiple des séquences de l'espèce *Hordeum bulbosum.*

ii. Au deuxième niveau

À cette étape, le résultat est un alignement entre les séquences consensus de chaque espèce d'une plante, obtenues à partir de l'étape précédente.

```
Consensus                                                 GGTAAAAGAACCCACGGCGCCGAAGGCGTCAAGGAACACTGTGCCTAACCCGGGGGCATG
Triticum aestivum - realigned consensus sequence          GGTAAAASAACCCACGGCGCCGAAGGCGTCAAGGAACACTGTGCCTAACCCGGGGGCATG
Triticum carthlicum - realigned consensus sequenceGGTAAAAGAACCCACGGCGCCGAAGGCGTCAAGGAACACTGTGCCTAACCCGGGGGCATG
Triticum dicoccoides - realigned consensus sequencGGTAAAAGAACCCACGGCGCCGAAGGCGTCAAGGAACACTGTGCCTAACCCGGGGGCATG
Triticum monococcum - realigned consensus sequenceGGTAAAAGAACCCACGGCGCCGAAGGCGTCAAGGAACACTGTGCCTAACCCGGGGGCATG
Triticum polonicum alignment consensus sequence   GGTAAAAGAACCCACGGCGCCGAAGGCGTCAAGGAACACTGTGCCTAACCCGGGGGCATG
Triticum timopheevii - realigned consensus sequencGGTAAAAGAACCCACGGCGCCGAAGGCGTCAAGGAACACTGTGCCTAACCCGRGGGCATG
Triticum turgidum alignment consensus sequence    GGTAAAAGAACCCACGGCGCCGAAGGCGTCAAGGAACACTGTGCCTAACCCGGGGGCATG
Triticum urartu alignment consensus sequence      GGTAAAAGAACCCACGGCGCCGAAGGCGTCAAGGAACACTGTGCCTAACCCGGGGGCATG
```

Figure 29 : Partie d'alignement multiple entre les séquences ITS des espèces du Blé.

iii. Au troisième niveau

Le résultat sera un alignement entre quatre séquences consensus, chacune provenant d'un alignement entre les espèces de chaque plante.

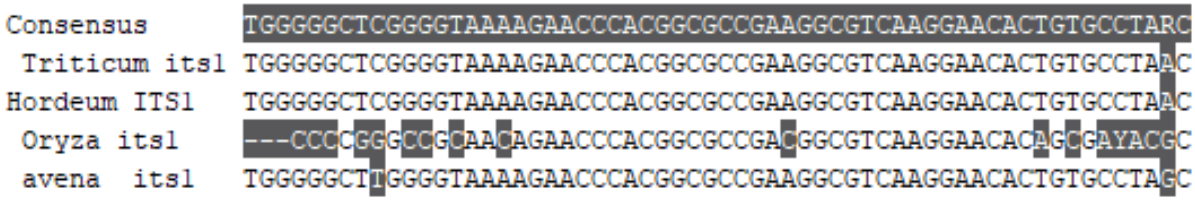

Figure 30 : Partie d'alignement multiple des quatre genres étudiés.

iv. Création d'arbre phylogénétique

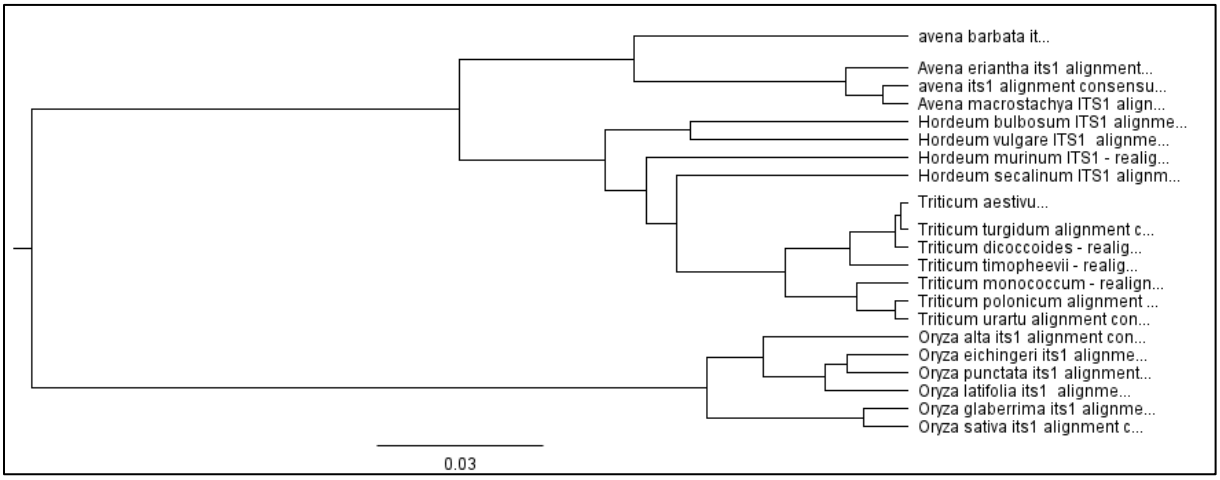

Figure 31 : Arbre phylogénétique des espèces étudie

4.2 Résultat de la phase 2

4.2.1 Résultat de prétraitement des données

Après avoir effectué le processus de nettoyage de notre ensemble de données et filtré les informations utilisables pour l'apprentissage, nous avons obtenu un fichier CSV contenant les données finales (voir Figure 24). Ce fichier est composé de 796 séquences ITS1 et ITS2 associées à un type de céréale spécifique.

	Sequence	Cereal
0	GTGACCCTGACCAAAACAGACCGAGCACGCGTTATCTATTCCTGCT...	avena-its1
1	GTCGTGACCCTGACCAAAACAGACCGAGCACGCGTTATCTATTCCT...	avena-its1
2	TCGTGACCCTGACMAAAACAGACCGAGCACGCGTTATCTATTCCTA...	avena-its1
3	TCGTGACCCTGACCAAAACAGACCGAGCACGCGTTATCTATTCCTG...	avena-its1
4	TCGTGACCCTGACCAAAACAGACCGAGCACGCGTTATCTATTCCTG...	avena-its1
5	TCGTGACCCTGACCAAAACAGACCGAGCACGCGTTATCTATTCCTG...	avena-its1
6	TCGTGACCCTGACCAAAACAGACCGAGCACGCGTTATCTATTCCTA...	avena-its1

Figure 32 : Fichier CSV du résultat du prétraitement.

4.2.2 Résultat de la représentation numériques et remplissage des séquences

Toutes les séquences auront la même longueur fixe (343) et peuvent être traitées en parallèle par notre modèle de ML.

	0	1	2	3	4	5	6	7	8	9	...	334	335	336	337	338	339	340	341	342	Label
0	2	4	2	3	1	1	1	4	2	3	...	0	0	0	0	0	0	0	0	0	0
1	2	4	1	2	4	2	3	1	1	1	...	0	0	0	0	0	0	0	0	0	0
2	4	1	2	4	2	3	1	1	1	4	...	0	0	0	0	0	0	0	0	0	0
3	4	1	2	4	2	3	1	1	1	4	...	0	0	0	0	0	0	0	0	0	0
4	4	1	2	4	2	3	1	1	1	4	...	0	0	0	0	0	0	0	0	0	0

Figure 33: Encodages et remplissages.

4.2.3 Résultat de l'apprentissage :

Après avoir achevé l'apprentissage, le modèle a été soumis à des tests pour évaluer son efficacité et sa précision. Le test a été fait sur 20% des données utilisées dans ce travail. voici les résultats obtenus :

	Real Cereal	Predicted Cereal	Correct Prediction
0	hordeum-ITS1	hordeum-ITS1	True
1	avena-its2	avena-its2	True
2	oryza-its1	oryza-its1	True
3	avena-its2	avena-its2	True
4	hordeum-ITS2	hordeum-ITS2	True
5	hordeum-ITS2	hordeum-ITS2	True
6	avena-its2	avena-its2	True
7	avena-its2	avena-its2	True
8	triticum-ITS1	triticum-ITS1	True

Figure 34: Les résultats avec une colonne indiquant si la prédiction est correcte ou non.

La figure 27 montrer la sortie affichera les étiquettes d'origine, les étiquettes prédites et les valeurs True si la prédiction est correcte, et False si elle est incorrecte. Après avoir les résultats du test on a calculé les valeurs d'évaluation suivantes :

Accuracy = 0.9748427672955975 = 98%

Précision = 0.976870299511809 = 98%

Rappel = 0.9748427672955975 = 98%

Par la suite nous fournissons la matrice de confusion suivante :

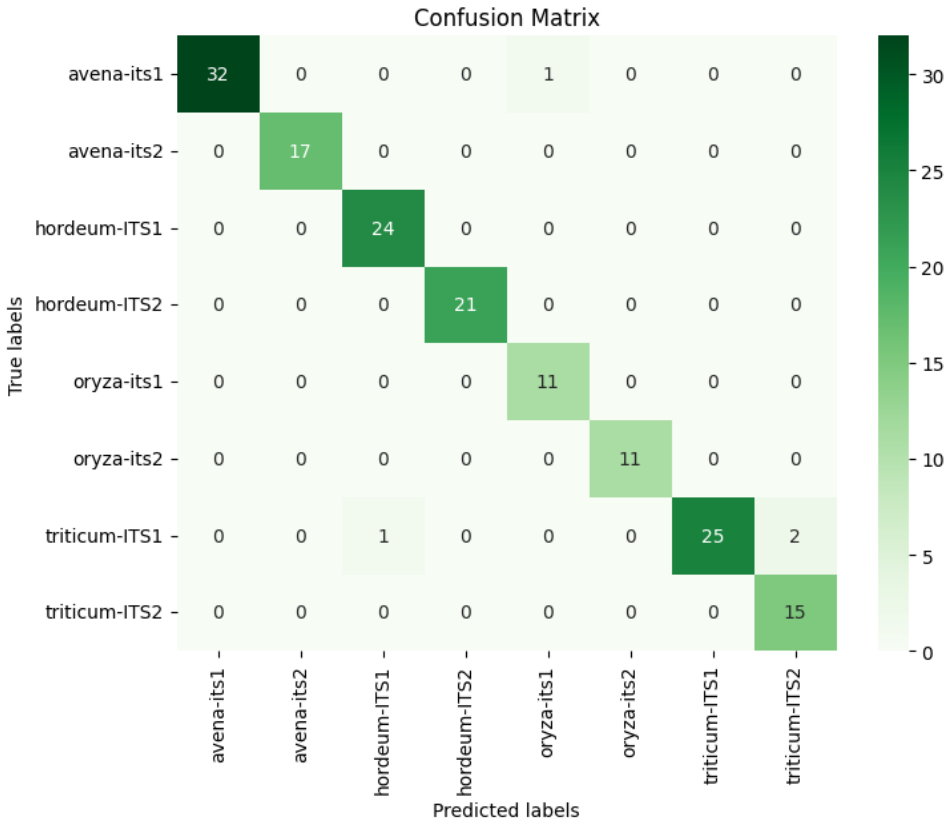

Figure 35 : Matrice de confusion du test du modèle.

L'objectif de la matrice est d'évaluer la performance du notre modèle en comparant les prédictions du modèle avec les valeurs réelles des données de test, et de fournir une vue détaillée des erreurs de classification effectuée. La matrice de confusion est organisée sous forme de tableau à deux dimensions, où les lignes représentent les classes réelles et les colonnes représentent les classes prédites par le modèle. Chaque cellule de la matrice contient le nombre d'instances appartenant à une classe réelle et prédite spécifique, la ligne diagonale principale représente les séquences correctement prédites (Vrais Positifs : VP), et les éléments en dehors de la diagonale indiquent les erreurs effectuées (Faux Négatifs : FN).

5. DISCUSSION

Les méthodes de classification modernes ont largement supplanté les méthodes traditionnelles dans les études et les recherches scientifiques en raison de leur efficacité, de leur rapidité et de leur précision dans le traitement et la manipulation de vastes ensembles de données.

Dans le cadre de notre travail, nous avons cherché à éclairer ce point en procédant de la manière suivante : premièrement, nous avons étudié les caractéristiques morphologiques des quatre genres de céréales afin d'identifier les principales différences observables. Ces différences ont été présentées sous la forme d'une interface utilisateur interactive développée pour évaluer les connaissances des utilisateurs et leur fournir des résultats personnalisés en fonction de leurs réponses.

Deuxièmement, nous avons effectué un alignement global et multiple des séquences nucléiques d'ITS1 et d'ITS2 (qui sont couramment utilisées car elles présentent des variations génétiques informatives pour l'analyse phylogénétique) pour chaque genre étudié (Triticum, Oryza, Avena et Hordeum) à trois niveaux : intraspécifique, entre espèces d'un même genre et entre les genres. Cette démarche a été accompagnée de la construction d'un arbre phylogénétique en utilisant le logiciel Geneious Prime.

L'objectif initial de cette étude était d'identifier les similitudes, les différences et de mettre en évidence les positions ou les divergences génétiques qui ont conduit, au fil du temps, à l'apparition et à l'évolution de ces espèces. Cet arbre représenterait les relations entre les espèces de céréales étudiées, où les espèces les plus proches dans l'arbre seraient considérées comme étant génétiquement plus similaires et ayant un ancêtre commun plus récent, tandis que les espèces plus éloignées seraient génétiquement plus distinctes et auraient un ancêtre commun plus lointain.

L'analyse de l'arbre phylogénétique révélerait des informations sur les relations évolutives entre ces céréales, montrant, par exemple, que l'orge et l'avoine sont plus étroitement liées du point de vue de la parenté génétique, tandis que le riz et le blé sont plus éloignés. Cela suggérerait que l'orge et l'avoine partagent un ancêtre commun plus récent que le riz et le blé.

En passant à l'utilisation de méthodes d'intelligence artificielle, plus précisément l'apprentissage automatique (ML), pour classer et prédire les céréales à partir de leurs séquences ITS1 et ITS2. Nous avons utilisé le classifieur Random Forest pour construire notre modèle. Les résultats obtenus démontrent une précision et une exactitude de 98% dans cette tâche.

Cette performance élevée constitue un indicateur positif de la qualité et de l'efficacité du modèle dans sa capacité à classifier avec précision les différentes espèces de céréales. Malgré le nombre réduit de séquences disponibles pour l'analyse des ITS, notre modèle a réussi à obtenir des résultats fiables, mettant en évidence sa robustesse et son potentiel dans la classification précise des céréales.

À notre connaissance, notre travail se distingue en étant le premier à utiliser un modèle basé sur l'IA pour classer les genres de céréales en utilisant les séquences ITS. Toutefois, il convient de noter qu'il est possible de comparer notre travail sur les céréales avec des études antérieures qui ont appliqué l'intelligence artificielle à la classification des champignons. Cette comparaison nous permet d'avoir une vision plus complète de l'efficacité de notre méthode et de son potentiel d'application dans différents domaines biologiques.

En comparant notre travail à l'étude CNN_FunBar [90], nous pouvons prendre en compte différents critères tels que le modèle utilisé et la précision (Accuracy). Cependant, étant donné que notre étude porte sur la classification à partir des séquences ITS, il est important de se concentrer sur les références qui se penchent également sur la classification des espèces en utilisant les séquences ITS.

Tableau 9 : Comparaison du notre travail avec un autre approche.

	Description des données utilisées	Méthode utilisée	Accuracy
Notre tavail	796 séquences ITS des céréales.	Random Forest	98%
CNN_FunBar	500 séquences ITS des champignons.	Les réseaux neuronaux convolutionels (CNN)	97%

Conclusion et Perspectives

CONCLUSION ET PERSPECTIVES

Le travail a commencé par une collection des caractéristiques morphologiques des quatre genres de céréales étudiés. Cette analyse a permis d'identifier les principales différences observables entre ces genres, ce qui a fourni une base solide pour notre approche de classification. En utilisant les séquences ITS, nous avons procédé à un alignement multiple des séquences ITS1 et ITS2. Ces séquences présentent des variations génétiques informatives pour l'analyse phylogénétique. Cet alignement a grandement facilité la classification ultérieure des céréales.

Les résultats obtenus ont été très prometteurs, avec une précision de classification très élevés. Ces résultats démontrent clairement l'efficacité et la capacité de notre modèle. La méthode ML utilisée dans cette étude présente des avantages significatifs par rapport à la méthode traditionnelle de classification des céréales. Elle permet une analyse plus précise et efficace des séquences de céréales, offrant ainsi des résultats plus fiables et des perspectives prometteuses pour l'identification et la classification des espèces de céréales.

Ces avancées pourraient avoir des implications significatives dans les domaines de la Bioinformatique et de la conservation des ressources génétiques, permettant ainsi une meilleure compréhension de la diversité des céréales et de leur évolution.

Cependant, il convient de poursuivre les recherches pour améliorer et affiner les modèles de machine learning utilisés, ainsi que pour explorer de nouvelles caractéristiques et approches pour une meilleure classification et prédiction des séquences de céréales. Comme perspectives, nous envisageons l’étude d’un plus grand nombre de genres de céréales ce qui ouvrent la voie à des recherches approfondies visant à rendre la classification des espèces et des sous-espèces de céréales plus précise et détaillée. Une autre perspective intéressante pour améliorer cette étude seraient d'explorer l'utilisation de modèles de deep learning pour l'identification des céréales à partir d'images. En outre, une application pratique de notre projet consisterait à l'appliquer à des échantillons réels provenant de différents sites agricoles ou de collections de ressources génétiques. Cela permettrait de caractériser et de conserver les variétés de céréales de manière plus concrète et applicable dans un contexte réel.

Références

Liste des références :

[1] Sergio O- Serna-Saldivar, «Cereal Grains _ Properties, Processing, and Nutritional Attributes (2010, CRC Press)»

[2] C. Jean-Paul, "CÉRÉALES - Encyclopædia Universalis," Encyclopædia Universalis, Apr. 2023. Available: https://www.universalis.fr/encyclopedie/cereales/. Accessed: Apr. 23, 2023.

[3] Britannica, The Editors of Encyclopaedia. "cereal". Encyclopedia Britannica, 3 Mar. 2023, https://www.britannica.com/topic/cereal. Accessed 12 May 2023.

[4] W. Laskowski, H. Górska-Warsewicz, K. Rejman, M. Czeczotko, et J. Zwolińska, « How Important are Cereals and Cereal Products in the Average Polish Diet? », Nutrients, vol. 11, no 3, p. 679, mars 2019, doi: 10.3390/nu11030679.

[5] K. Martin et M. Matthew, « The cereals imperative of future food systems », International Rice Research Institute, oct. 2019, Consulté le: 28 avril 2023. URL https://www.irri.org/news-and-events/news/cereals-imperative-future-food-systems

[6] C. A. E. Strömberg, « Evolution of Grasses and Grassland Ecosystems », Annu. Rev. Earth Planet. Sci., vol. 39, no 1, p. 517-544, mai 2011, doi: 10.1146/annurev-earth-040809-152402.

[7] B. F. Carver, Wheat Science and Trade. 2009, p. 569. doi: 10.1002/9780813818832.

[8] I. G. Loskutov et H. W. Rines, « Avena », in Wild Crop Relatives: Genomic and Breeding Resources: Cereals, C. Kole, Éd., Berlin, Heidelberg: Springer, 2011, p. 109-183. doi: 10.1007/978-3-642-14228-4_3.

[9] A. Goyal et M. Ahmed, « Barley: Production, Improvement, and Uses », Crop Science, vol. 6, p. 2852-2853, janv. 2012, doi: 10.2135/cropsci2012.12.0003br.

[10] FAO, Éd., Livestock in the balance. in The state of food and agriculture, no. 2009. Rome: FAO, 2009.

[11] L. Taiz et E. Zeiger, Plant physiology, Fifth edition. Sunderland, Massachusetts: Sinauer Associates, Inc., Publishers, 2010.

[12] P. H. Raven, R. F. Evert, et S. E. Eichhorn, Biology of plants, 7th ed. New York: W.H. Freeman, 2005.

[13] A. DJEKOUN et M. A. HAMIDECHI, « cours de phylogénie moléculaire », UNIVERSITE CONSTANTINE 1, p. 49.

[14] R. Freire, A. Arias, J. Méndez, et A. Insua, « Sequence variation of the internal transcribed spacer (ITS) region of ribosomal DNA in Cerastoderma species (Bivalvia: Cardiidae) », Journal of Molluscan Studies, vol. 76, no 1, p. 77-86, févr. 2010, doi: 10.1093/mollus/eyp047.

[15] C. L. Schoch et al., « Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi », Proceedings of the National Academy of Sciences, vol. 109, no 16, p. 6241-6246, avr. 2012, doi: 10.1073/pnas.1117018109.

[16] Samanthi, « What is the Difference Between ITS1 and ITS2 », Compare the Difference Between Similar Terms, juill. 2021, Consulté le: 27 avril 2023. URL https://www.differencebetween.com/what-is-the-difference-between-its1-and-its2/

[17] V. Simossis, J. Kleinjung, et J. Heringa, « An overview of multiple sequence alignment », Current protocols in bioinformatics, vol. 3, no 1, p. 3-7, 2003.

[18] I. Ovcharenko et al., « Mulan: multiple-sequence local alignment and visualization for studying function and evolution », Genome Res, vol. 15, no 1, p. 184-194, janv. 2005, doi: 10.1101/gr.3007205.

[19] T. Bigot, « Recherche automatisée de motifs dans les arbres phylogénétiques », juill. 2014.

[20] G. Deléage, M. Gouy, et A. de Brevern, « Chapitre 6. Bases théoriques de la phylogénie moléculaire », Sciences Sup, vol. 3, p. 59-77, 2021.

[21] P. Domingos, The master algorithm: how the quest for the ultimate learning machine will remake our world. New York: Basic Books, a member of the Perseus Books Group, 2015.

[22] T. M. Mitchell, Machine Learning. in McGraw-Hill series in computer science. New York: McGraw-Hill, 1997.

[23] G. Saint-Cirgue, « Introduction au Machine Learning », Machine Learnia, 21 juin 2019. https://machinelearnia.com/machine-learning-introduction/ (consulté le 13 avril 2023).

[24] T. H. Davenport, « Artificial Intelligence For The Real World: Don't Start With Moon Shots. » in Harvard Business Review,. 2018.

[25] M. Mohammed, M. B. Khan, et E. B. M. Bashier, Machine learning: algorithms and applications. Boca Raton: CRC Press, Taylor & Francis Group, 2017.

[26] E. Alpaydin, Introduction to machine learning, 2nd ed. in Adaptive computation and machine learning. Cambridge, Mass: MIT Press, 2010.

[27] mobiskill, « Apprentissage supervisé vs apprentissage non supervisé », Mobiskill, 22 mars 2021. https://mobiskill.fr/blog/conseils-emploi-tech/apprentissage-supervise-vs-apprentissage-non-supervise/?fbclid=IwAR0B397uUNw5ox5B3pt34tZvqyJ-cFd0hI-kjIK4_ne-cA5GxPxZl7gkqsk (consulté le 14 avril 2023).

[28] alexandre, « Algorithme De Classification: Définition Et Principaux Modèles », Formation Data Science | DataScientest.com, 22 novembre 2022. https://datascientest.com/algorithme-de-classification-definition-et-principaux-modeles (consulté le 14 avril 2023).

[29] S. Borah, V. Emilia Balas, et Z. Polkowski, Éd., Advances in Data Science and Management: Proceedings of ICDSM 2019, vol. 37. in Lecture Notes on Data Engineering and Communications Technologies, vol. 37. Singapore: Springer Singapore, 2020. doi: 10.1007/978-981-15-0978-0.

[30] Y. Qi, « Random Forest for Bioinformatics », Machine Learning Department, NEC Labs, p. 18.

[31] A. Abdelkrime, « Introduction à l'apprentissage automatique », mars 2019, Consulté le: 14 avril 2023. URL : https://github.com/projeduc/intro_apprentissage_automatique/blob/41010fa18afe493d7a33fe691cad9bab7282930e/introduction.md

[32] D. Soltner, Les bases de la production vegetale : le sol, le climat, la plante meteorologie, pedologie, bioclimatologie. Tome II, Le climat / Dominique Soltner,... - Universite de Lorraine, 7e edition. 1995. Consulté le: 29 mai 2023. URL : https://ulysse.univ-lorraine.fr/discovery/fulldisplay/alma991001952739705596/33UDL_INST:UDL

[33] P. Salgado, L. H. Bình, V. C. Cương, et T. V. Thủ, « Production et utilisation de l'avoine fourragère (Avena strigosa et Avena sativa) au nord du Vietnam : une solution pour résoudre le déficit fourrager en hiver ».

[34] West Africa Rice Development Association « WARDA Annual Report 1996. Mbé, Côte d'Ivoire ». Consulté le: 6 juin 2023. URL ://pdf.usaid.gov/pdf_docs/Pnach999.pdf

[35] Soltner, D. (1955) Les Grandes productions végétales: Phytotechnie "spéciale: Les Céréales, Les Plantes DITEs 'sarclées', Prairies. 18e éd. Bressuire: Collection sciences et techniques agricoles.

[36] V. Thirulogachandar et T. Schnurbusch, « 'Spikelet stop' determines the maximum yield potential stage in barley », Journal of Experimental Botany, vol. 72, no 22, p. 7743-7753, déc. 2021, doi: 10.1093/jxb/erab342.

[37] C. Moule, Céréales. Phytotechnie spéciale : bases scientifiques et techniques de la production des principales espèces de grande culture en France. Paris: La Maison Rustique, 1980.

[38] M. CLÉMENT-GRANDCOURT, J.PATRS, "les_cereales," COLLECTION D'ENSEIGNEMENT AGRICOLE. Disponible sur: https://excerpts.numilog.com/book /9782307635826.pdf?fbclid=IwAR0zeG3B2pUDbikaK_BfN-Ki8EACcUy UPQeW2nKmPkEZ2yzcLmteLTx_k (consulté le 6 avril 2023).

[39] H. BOULAL, O. ZOGHOUANE, M. EL MOURID, and S. REZGUI, "Guide pratique du conduit des céréales d'automne (blés,orge et l'avoine) dans le Maghreb (Algérie, Maroc, Tunisie)," 2007, p. 20.

[40] M. T. Adagbe, « Utilisation de la terre renforcée par des tiges de paille de riz comme matériau des éléments porteurs des bâtiments armés avec le rônier », Université Paul Sabatier - Toulouse III, p. 165, juill. 2021, doi: https://theses.hal.science/tel-03551654/preview/2021TOU30084b.pdf.

[41] S. B. Alaoui, Référentiel pour la Conduite de l'orge (Hordeum vulgare). 2005.

[42] « Orge (Hordeum vulgare), une très ancienne céréale », Binette & Jardin. https://jardinage.lemonde.fr/dossier-1413-orge-hordeum-vulgare-ancienne-cereale.html (consulté le 6 avril 2023).

[43] D. Toon, C. S. W. Marco, D. Salif, et I. Philip, "Curriculum D'apprentissage Participatif Et Recherche Action (APRA) Pour La Gestion Intégrée de La Culture de Riz de Bas-Fonds (GIR) À Madagascar - Manuel Du Facilitateur". Madagascar, 2008. Consulté le: 6 juin 2023. URL : https://fr.scribd.com/document/83943525/Curriculum-d-apprentissage-participatif-et-recherche-action-APRA-pour-la-gestion-integree-de-la-culture-de-riz-de-bas-fonds-GIR-a-Madagascar-Ma

[44] R. Valluru, M. P. Reynolds, et T. Lafarge, « Food security through translational biology between wheat and rice », Food Energy Secur, vol. 4, no 3, p. 203-218, oct. 2015, doi: 10.1002/fes3.71.

[45] I.Wopereis, S. Brand-Gruwel, I.Vermetten, 2008.« Connaissance de la plante de ri». p 25-35

[46] D. Soltner, « Phytotchnie spéciale : les grandes productions végétales : céréales - plantes sarclées – prairies », Sciences et techniques agricoles, 20edition, 2005. 472p.

[47] A. canadienne d'inspection des aliments Gouvernement du Canada, « Instructions particulières (IP 142.1.2-2) : Procédures d'inspection des cultures de semences de céréales et de céréales à petits grains », 9 septembre 2012. https://inspection.canada.ca/protection-des-vegetaux/semences/methodes-d-inspection/cultures-de-semences-genealogiques-de-cereales/fra/1347203083351/1347203347397 (consulté le 7 avril 2023).

[48] « Barley grass | Agriculture and Food ». https://www.agric.wa.gov.au/grains-research-development/barley-grass (consulté le 7 avril 2023).

[49] D. Soltner, « Phytotchnie spéciale : les grandes productions végétales : céréales - plantes sarclées – prairies », Sciences et techniques agricoles, 20edition, 2005. 472p.

[50] Agriculture and Horticulture Development Board (AHDB) . https://ahdb.org.uk/ (consulté le 6 avril 2023).

[51] E.Firatligil-Durmuş, et al. (2010) 'Geometric parameters of wheat grain determined by image analysis and FEM approach', Cereal Research Communications, 38(1), pp. 122–133. doi:10.1556/crc.38.2010.1.13.

[52] D.HOUASSINE , « Etude de l'effet de l'association féverole (Vicia faba L.) – Avoine (Avena sativa L.) sur la biodisponibilité du phosphore dans la rhizosphère», ECOLE NATIONALE SUPERIEURE AGRONOMIQUE, 24 july 2019, P156

[53] Le Monde " Orge (Hordeum vulgare), une très ancienne céréale", Binette & Jardin. URL : https://jardinage.lemonde.fr/dossier-1413-orge-hordeum-vulgare-ancienne-cereale.html#:~:text=L'orge%20(Hordeum%20vulgare),herbe%20%C3%A0%20chat%2C%20par%20erreur. (Accessed: 08 April 2023).

[54] R. N. Spengler, B. Cerasetti, M. Tengberg, M. Cattani, et L. M. Rouse, « Agriculturalists and pastoralists: Bronze Age economy of the Murghab alluvial fan, southern Central Asia », Veget Hist Archaeobot, vol. 23, no 6, p. 805-820, nov. 2014, doi: 10.1007/s00334-014-0448-0.

[55] Cbprfad, C. and patissiers, C. boulangers (no date) Les analyses des protéines du blé. " CREBESC CBPRFAD, CREBESC. Available at: https://levainbio.com/cb/crebesc/les-proteines-du-ble/ (Accessed: 08 April 2023).

[56] M.-H. Dabat, B. Pons, et S. Razafimandimby, « Des consommateurs malgaches sensibles à la qualité du riz », Économie rurale. Agricultures, alimentations, territoires, no 308, Art. no 308, déc. 2008, doi: 10.4000/economierurale.330.

[57] A. Dyhia "contribution à l'étude de l'influence d'un herbicide, le Glyphosate et d'un fongicide, le Mancozèbe, sur la germination, la croissances et la physiologie de deux céréales : Hordeum vulgare L. et Avena sativa L." p.83, 2019

[58] C.MOUL « CÉRÉALES II PHYTOTECHNIE SPÉCIALE». LA MAISON RUSTIQUE-PARIS, p95, 1971.

[59] Klorane Botanical Foundation, "L'Avoine Plantes herbacées, DESCRIPTION DE L'AVOINE Avena sativa L. (Famille des Poaceae)", Les Cauquillous 81506 LAVAUR Cedex - France, (Accessed: 01 April 2023) URL https://www.kloranebotanical.foundation/en/node/399

[60] C. Katy ,(2007) Hordeum jubatum (foxtail barley), Minnesota Wildflowers. URL : https://www.minnesotawildflowers.info/grass-sedge-rush/foxtail-barley (Accessed: 20 March 2023).

[61] Botarela, “Hordeum vulgare L. | Orge cultivé ”, 2012, (Accessed : 03 march 2023) URL http://botarela.fr/Poaceae/Taxons/Hordeum-vulgare.html.

[62] M.Clement-Grandcourt, J.Parts, “LES CERELES”, COLLECTION D'ENSEIGNEMENT AGRICOLE, vol. 3, 1966. p.36

[63] A. Dyhia “contribution à l'étude de l'influence d'un herbicide, le Glyphosate et d'un fongicide, le Mancozèbe, sur la germination, la croissances et la physiologie de deux céréales : Hordeum vulgare L. et Avena sativa L.” , 2019. p.83

[64] Sally , P. and Alex , D. "Barley grass, Agriculture and Food”. (2019) URL: https://www.agric.wa.gov.au/grains-research-development/barley-grass (Accessed: 15 February 2023).

[65] G. Heuzé, Les plantes céréales. Librairie agricole de la maison rustique, 1896.

[66] Le Monde « Avoine (Avena sativa), céréale alimentaire et engrais vert », Binette & Jardin.URL https://jardinage.lemonde.fr/dossier-1331-avoine-avena-sativa-cereale-alimentaire-engrais-vert.html (consulté le 7 avril 2023).

[67] L. Zohra, « Biologie des cereale, Impact des pratiques culturales sur le rendement des céréales ». Université de Bisk,2022,p.50, URL http://archives.univ-biskra.dz/bitstream/123456789/22568/1/ZOHRA_LORBI.pdf

[68] M. K. A. Imrana, « LA DIVERSITE PATHOGENIQUE DU RICE YELLOW MOTTLE VIRUS (RYMV) DANS LES ZONES RIZICOLES DE MAROVOAY- REGION DE BOENY », 2009.

[69] R. Lásztity, The chemistry of cereal proteins, 2nd ed. Boca Raton: CRC Press, 1996.

[70] A. Dyhia “contribution à l'étude de l'influence d'un herbicide, le Glyphosate et d'un fongicide, le Mancozèbe, sur la germination, la croissances et la physiologie de deux céréales : Hordeum vulgare L. et Avena sativa L.” , 2019. p.83

[71] « What is Python? Executive Summary », Python.org. https://www.python.org/doc/essays/blurb/ (consulté le 14 mai 2023).

[72] Techno-Science.net. “Python (langage) - Définition et Explications.“URL : https://www.techno-science.net/glossaire-definition/Python-langage.html. [Accessed: May 14, 2023].__Please note that the access date should be written in English and the format is day month year.

[73] « What is Anaconda? | Domino Data Science Dictionary ». https://www.dominodatalab.com/data-science-dictionary/anaconda (consulté le 14 mai 2023).

[74] L, B.”Jupyter Notebook : Tout savoir sur le notebook préféré des data scientists”, LEBIGDATA.FR,2021. URL://www.lebigdata.fr/jupyter-notebook (Accessed: 27 March 2023).

[75] Aubert, A.”Qu’est-ce que jupyter et comment faire plus avec vos notebooks ?, Saagie”.2022,URL: https://www.saagie.com/fr/blog/quest-ce-que-jupyter-et-pourquoi-est-il-un-outil-incontournable/ (Accessed: 09 June 2023).

[76] Bioinformatics software for Sequence Data Analysis (2023) Geneious.URL: https://www.geneious.com/ (Accessed: 11 April 2023).

[77] A. C. Schwickert, « HTML - Hypertext Markup Language », Informatik-Spektrum, vol. 20, no 3, p. 168-169, juin 1997, doi: 10.1007/s002870050065.

[78] L. Pan et J. S. Ma, « HTML+CSS Implementation based on Image Intelligent Scene Recognition Algorithm », 2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), p. 532-536, nov. 2022, doi: 10.1109/ICAISS55157.2022.10011034.

[79] E. D. C. Lima, J. de P. S. Ferreira, et A. Lobato, « AMBIENTE VIRTUAL DE APRENDIZAGEM DE JAVASCRIPT (LAB JS) », avr. 2018. Consulté le: 30 mai 2023. [En ligne]. Disponible sur: https://www.semanticscholar.org/paper/AMBIENTE-VIRTUAL-DE-APRENDIZAGEM-DE-JAVASCRIPT-(LAB-Lima-Ferreira/667e131cbc5c39d5cc55fa47f64a8b1f7aa9894a

[80] Visualization with python (no date) Matplotlib. Available at: https://matplotlib.org/ (Accessed: 14 June 2023).

[81] NumPy. URL: https://numpy.org/ (Accessed: 1 June 2023).

[82] Pandas. URL : https://pandas.pydata.org/ (Accessed: 05 June 2023).

[83] P. Bach, V. Chernozhukov, M. S. Kurz, et M. Spindler, « DoubleML -- An Object-Oriented Implementation of Double Machine Learning in Python ». arXiv, 20 décembre 2021. doi: 10.48550/arXiv.2104.03220.

[84] J. Faouzi et H. Janati, « pyts: A Python Package for Time Series Classification », J. Mach. Learn. Res., 2020, Consulté le: 3 juin 2023. [En ligne]. Disponible sur: https://www.semanticscholar.org/paper/pyts%3A-A-Python-Package-for-Time-Series-Faouzi-Janati/acd8ae94f9102fdfc809a9eeaaeb999cdb892c49

[85] P. Bach, V. Chernozhukov, M. S. Kurz, et M. Spindler, « DoubleML -- An Object-Oriented Implementation of Double Machine Learning in Python ». arXiv, 20 décembre 2021. doi: 10.48550/arXiv.2104.03220.

[86] Learn scikit. URL https://scikit-learn.org/stable/ (Accessed: 5 June 2023).

[87] Seaborn,” Tout Savoir Sur l’outil de data visualization en python, Formation Data Science | DataScientest.com”,2023. URL https://datascientest.com/seaborn-tout-savoir (Accessed: 09 June 2023).

[88] A.M.Kuchling, "Regular expression howto, Python documentation”,2023. URL: https://docs.python.org/3/howto/regex.html (Accessed: 10 June 2023).

[89] G.Lino ,”Maîtriser les expressions régulières, Python pour la data-science”.2020. URL https://pythonds.linogaliana.fr/regex/ (Accessed: 26 April 2023).

[90] R. Das, A. Rai, et D. C. Mishra, « CNN_FunBar: Advanced Learning Technique for Fungi ITS Region Classification », Genes, vol. 14, no 3, Art. no 3, mars 2023, doi: 10.3390/genes14030634.

Année universitaire : 2022-2023	Présenté par : MAZZI Meryem KARAALI Ouissal MECIBAH Akram

Thème :

Méthode d'Apprentissage Machine pour identification des plantes.

Cas d'étude : les céréales

Mémoire pour l'obtention du diplôme de Master en :

Bioinformatique

Domaine : Science de la nature et la vie

Département de Biologie Appliquée

Cette étude présente une approche basée sur l'intelligence artificielle (IA) pour la classification des espèces de céréales en utilisant les séquences ITS. La recherche a débuté par une analyse des caractéristiques morphologiques du blé, de l'orge, de l'avoine et du riz. Par la suite, les séquences ITS ont été alignées et ont permis de développer un modèle d'apprentissage automatique basé sur le classifieur de la forêt aléatoire (RF). Le modèle a atteint une précision remarquable de 98% dans la prédiction du genre des céréales en se basant sur les séquences ITS. Cette étude met en évidence le potentiel considérable de l'IA dans la classification des espèces de céréales, avec des implications majeures pour les domaines de l'agriculture et une meilleure compréhension de la diversité des céréales.

Mots-clefs : Apprentissage Automatique, Céréales, Classification, Forêt Aléatoire, Intelligence Artificiel, ITS.

Encadreur : D r. GHERBOUDJ Amira (Université Frères Mentouri, Constantine 1).

Examinateur 1 : Dr. TAMAGOULT Mahmoud (Université Frères Mentouri, Constantine 1).

Examinateur 2 : Dr. DJAMAA Ouahib (Université Frères Mentouri, Constantine 1).