REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE Dr. MOULAY TAHAR – SAIDA

FACULTE DE TECHNOLOGIE

DEPARTEMENT D'INFORMATIQUE

N° D'ORDRE : .....

THESE

Présentée par

MOKRI Miloud Aboubakeur El Sadek

Pour l'obtention du diplôme de

DOCTORAT «L. M. D» en INFORMATIQUE

Spécialité: Informatique

Option: Informatique

Le Biomémétisme et le data mining dans la sécurité informatique dans le web et le big data

Défendu publiquement, en ../../2021

Devant le jury composé de:

AMINE Abdelmalek	Professeur	Université Tahar Moulay de Saida	Président
BENSLIMANE Sidi Mohammed	Professeur	ESI SBA	Examinateur
YAHLALI Mebarka	MCA	Université Tahar Moulay de Saida	Examinateur
MEKKAOUI Kheireddine	MCA	Université Tahar Moulay de Saida	Examinateur
HAMOU Reda Mohamed	Professeur	Université Tahar Moulay de Saida	Directeur de thèse

Année Universitaire 2020-2021

Laboratoire GeCoDe, Université de Saida

Remerciements

Au nom d' **ALLAH** le tout Miséricordieux et que la prière et la paix soient sur notre prophète **Mohamed** aalayh Elssalet wa Elssalem.

Je remercie ALLAH, Le Tout Puissant, pour son aide et sa protection, et de m'avoir donné la patience, la capacité, la volonté et le courage pour accomplir ce travail.

Cette thèse est le résultat d'un effort. Cet effort n'aurait pas pu aboutir sans la contribution d'un nombre de personnes. Ainsi se présente l'occasion des les remercier.

Je tiens à exprimer ma profonde gratitude et mes sincères remerciements à mon professeur **HAMOU Reda Mohamed** le meilleur directeur de thèse que je pouvais imaginer ainsi qu'une personne extraordinaire. Je le remercie pour m'avoir fait travailler sur un sujet que j'ai l'adoré, d'avoir partagé avec moi tout son savoir-faire et d'être disponible, je le remercie pour toutes ces heures durant lesquelles il a tenu à corriger mon travail et avoir me guider dans cette thèse.

Je suis également très reconnaissant à Monsieur **AMINE Abdelmalek** professeur et chef de laboratoire **GECODE** pour son soutient et ses encouragements durant ma période dans cette formation doctorale.

Je tiens à remercier aussi mes collègues **Shwan Khaled** et **Bensaid Tayeb** pour être toujours la pour me donner de l'aide et les conseilles.

Je tiens a remercier aussi les membres de jury d'avoir examiné mon travail représenté par :

- Monsieur BENSLIMANE Sidi Mohammed, Professeur à ESI SBA - Ecole Supèrieure en Informatique de Sidi Bel Abbès ;
- Madame YAHLALI Mebarka, Maître de Conférences à l'université Tahar Moulay de Saida ;
- Monsieur MEKKAOUI Kheireddine, Maître de Conférences à l'université Tahar Moulay de Saida.

Je suis ravi et honoré que messieurs BENSLIMANE Sidi Mohammed et MEKKAOUI Kheireddine, et Madame YAHLALI Mebarka aient accepté d'être les rapporteurs de cette thèse. Je tiens à remercier aussi Monsieur AMINE Abdelmalek, Professeur à l'université Tahar Moulay de Saida pour m'avoir fait l'honneur de présider le jury.

Dédicace

Je dédie ce modeste travail

Á mes chers parents, que dieu les protège Pour leur amour, leurs encouragements et leurs sacrifices.

Á mes chères sœurs et mes chers frères

Á ma future épouse

Ainsi qu'à toute ma famille et amies.

- ملخص -

في الآونة الأخيرة مع التطور التكنولوجي، ومع حوسبة الشركات ومشاركة البيانات، ومع تبادل المعلومات العديدة على مختلف المنصات، شهدت البيانات انفجار كبير، بحيث أنها تصل من مصادر مختلفة، بأنواع مختلفة، هذه الكتلة من البيانات لا تتوقف عن النمو بسرعة عالية. من بين الشركات الكبيرة التي لديها كمية هائلة من البيانات نجد قوقل ، فايسبوك وأمازون. المعلومات التي يتم تداولها على الويب تستوجب التحليل والتحكم فيها من خلال تقنيات قوية ومتطورة، لأنها تتعرض لمخاطر تجعلها مهددة أو معرضة للإتلاف، وهذا هو دور منصة البيانات الكبيرة.

حفز النمو الهائل للبيانات الباحثين على إنشاء توجهات جديدة في تخزين البيانات والتحكم فيها، مثل منصات البيانات الكبيرة والتخزين السحابي .

تلعب أنظمة المعلومات اليوم دورًا استراتيجيًا في الشركات، حيث يرتبط المستخدمون ببعضهم البعض، بحيث أنه مع الاتجاهات الجديدة يمكن للمستخدم مشاركة موارده، أو جزء من النظام على السحابة، أو يمكنه أيضًا تبادل معلومات سرية على البريد الإلكتروني، مما يجعل البيانات الحساسة عرضة للاختراق أو التجسس من خلال محاولات المتسللين الذين يريدون أخذ المعلومات الغير الشرعية.

بعد كثرة محاولات المتسللين المختلفة للتجسس على المعلومات السرية، يعد أمن أنظمة المعلومات والبيانات المتداولة على الويب ضروريًا ومهمًا للغاية، فمن الضروري تحديد المحيطات الحساسة للبيانات المراد حمايتها، و تحديد سياسة أمن الموارد من أجل توفير عملية خاضعة للرقابة.

قد صاحب الاستخدام الواسع لمنصات تبادل البيانات المختلفة على الويب محاولات لانتهاك السياسات الأمنية عن طريق التسللات الخبيثة والهجمات، بحيث أن هنالك الكثير من الأبحاث التي أجريت والتي تهدف إلى محاربة هذه الأعمال غير المشروعة. الهدف من هذه الأطروحة هو أن تندرج في إطار ضمان أمن منصات تبادل البيانات، وقد اخترنا مشكلتين أساسيتين، حماية البريد الإلكتروني من جهة، وأنظمة الأندرويد من ناحية أخرى. وبهذه الروح، اقترحنا نهجًا جديدًا يمثل نموذجًا استرشاديًا مستوحى من الطبيعة لمكافحة الرسائل الإلكترونية الخبيثة مبني على عمل الأخطبوطات البحرية لتحسين نتائج التقنيات الموجودة، وتعتبر النتائج التي تم الحصول عليها مرضية و مقبولة بعد المقارنة مع

البحوث الأخرى التي تم إجراؤها بالفعل، بحيث أظهرت خوارزميتنا قدرة تصفية جيدة. استخدمنا أيضًا خوارزميات التصنيف والتعلم العميق لمكافحة تطبيقات الأندرويد الضارة استنادًا على تراخيص الوصول التي تطلبها هاته التطبيقات.

الكلمات المفتاحية البيانات، الويب، البيانات الكبيرة، السحابة، المتسللين، أمن، الهجمات، البريد الإلكتروني، نظام أندرويد، استرشاديا، مستوحى من الطبيعة، الأخطبوطات، التصنيف، التعلم العميق، تراخيص.

- Abstract -

Recently with the technological development and the computerization of the companies and the sharing of data, and following to the numerous exchanges of information on the various platforms an explosion of data has seen the light, these data arrive from different sources with varied type on several formats, this mass of data is growing at high speed. Among the big companies which have an exponential capacity of data, we find Google, Facebook and Amazon. The information that circulates on the web needs to be analyzed and controlled by powerful and robust techniques because it poses risks which makes it threatened or lost, that's the role of the BIG Data platform.
The gigantic growth of data has motivated researchers to create new trends in data storage and control, such as Big Data platforms and cloud storage.
Information systems today play a strategic role in companies, the users are interconnected with each other. With the new trends, a user can share his resources or part of the system on the Cloud, or he can also exchange confidential information on electronic mail which made the sensitive data are vulnerable to be hacked or spied by intruders who want to take illegitimate information.
Following the various attempts of hackers to spy on confidential information, The security of information systems and data circulating on the web is very important, it is necessary to define sensitive perimeters of the data to be protected and a resource security policy in order to provide the control access to data.
The wide use of different data exchange platforms on the web has been accompanied by attempts to violate security policies by malicious access and attacks, a lot of research's done aims to fight these illegitimate acts. Our objective of this thesis is to be part of the framework that contribute to guaranteeing the security of data exchange platforms, in this spirit, we have opted for two issues, protect the electronic mail on the one hand, and the Android mobile systems on the other hand. In Our work we had proposed a new approach which is a bio-inspired model to fight against malicious emails based on the functioning of the marine octopod to improve the results of the techniques already existing, our results obtained by this method were compared with other existing research and showed a good filtering capacity. We had also opted to make a comparative study of multiple detectors to fight the android malware apps, we had used a multiple of classification and deep learning algorithms to detect Android malicious apps based on the access permissions requested by those apps.

Keywords : Data, Web, BIG Data, Cloud, intruder, Security, attack, email, Android system, Heuristic, Bio-inspired, octopod, classification, deep learning, permissions.

- Résumé -

Récemment avec le développement technologique et l'informatisation des entreprises et le partage des données, et suite aux nombreux échanges d'information sur les différentes plateformes, une explosion de données a vu le jour, ces dernières arrivent des différentes sources avec des types variés sur plusieurs formats, cette masse de données ne cesse de croître à grande vitesse. Parmi les grandes entreprises qui ont vu une augmentation exponentielle des données on trouve Google, Facebook, Amazone. L'information qui circule dans le web nécessite d'être analysée et contrôlée par des techniques performantes et robustes parce qu'elle se pose à des risques qui la rendent menacée ou perdue, c'est le rôle de la plate de forme BIG Data.
La croissance gigantesque des données a motivé les chercheurs de créer des nouvelles tendances de stockage et contrôle de données tel que les plateformes Big Data et de stockage dans le nuage " in the Cloud'.
Les systèmes d'information aujourd'hui jouent un rôle stratégique dans les entreprises, les utilisateurs sont interconnectés entre eux. Avec les nouvelles tendances, un utilisateur peut partager ses ressources ou une partie du système sur le Cloud, ou il peut aussi échanger de l'information confidentielle sur les messageries électroniques ce qui rend les données sensibles vulnérables d'être piratées ou espionnées par des tentatives des intrus qui veulent prendre ces données.
Suite aux différentes tentatives des pirates qui cherchent à attaquer et voler les données des victimes ce qui rendait la sécurité des systèmes d'informations et des données qui circulent dans le web en échec, il est nécessaire et très important de protéger les utilisateurs en tout préservant la sécurité de leurs données, il faut définir des périmètres sensibles des données à protéger et les politiques de sécurité des ressources afin de fournir un fonctionnement maîtrisé.
La grande utilisation des différentes plateformes d'échanges de données dans le web était accompagnée par des tentatives de violations des politiques de sécurité par des accès malveillants, beaucoup des recherches ont été faites pour lutter contre ces actes, notre objectif dans cette thèse est de s'inscrire dans le cadre de garantir la sécurité des plateformes d'échanges de données. Dans cet esprit nous avons opté pour deux problématiques, protéger les messageries électroniques d'une part et les systèmes Android mobile d'autre part. Dans notre travail nous avons proposé une nouvelle approche basée sur une technique bio-inspirée pour lutter contre les emails malveillants. Cette technique est inspirée du fonctionnement de l'octopode marin pour améliorer les résultats des techniques existantes, notre résultat obtenu par cette méthode a été comparé avec d'autres travaux, notre modèle a montré une bonne capacité de filtrage. Dans notre deuxième travail nous avons utilisé des algorithmes de classification et d'apprentissage profond pour lutter contre les applications Android malveillantes en se basant sur les autorisations d'accès demandées par ces applications.

Mots clés : Données, Web, BIG Data, Cloud, intrus, Sécurité, attaque, mes-

sagerie électronique, Système Android, Heuristique, Bio-inspiré, octopode, classification, apprentissage profond, las autorisations.

Table des matières

TABLE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ACRONYMES

SE	Système d'Exploitation
KDD	Knowledge Data Discovery
HTTP	HyperText Transfer Protocol
SGBD	Système gestion base de données
Hadoop	High-Availability Distributed Object-Oriented Platform
HDFS	Hadoop Distributed File system
SQL	Structured Query Language
SMS	Short Message Service
DOS	Denial Of Service
IDS	Intrusion detection system
HIDS	Host Intrusion Detection System
NIDS	Network Intrusion Detection System
ECB	Electronic Code Book
CBC	Cipher Block Chaining
CFB	Cipher FeedBack
AA	Apprentissage Automatique
DM	Data Mining
IA	Intelligence Artificiel
VP	Vrai Positif
FP	Faux Positif
VN	Vrai Négatif
FN	Faux Négatif
ROC	Receiver Operating Characteristic
TD-Learning	Temporal Difference Learning
Q-Learning	Quality Learning
MLP	Multilayer Perceptron

NP-Complet	Non déterministe Polynomial
P	Polynomial
AG	Genetic Algorithm
RCGA	Real Coded Genetic Algorithm

Introduction générale

Suite à l'avancement rencontré dans la technologie numérique, et le grand partage et échange de données rencontrées dans le web ce qui s'est manifesté par une explosion de données. Une multitude des services performants de traitement et de stockage de données ont été apparues pour contrôler et gérer ce gros volume, ces services sont garanti par l'utilisation de la technologie Big Data.
Dernièrement avec l'ouverture des entreprises à l'utilisation d'internet et la grande utilisation des services de transmission des données telle que le courrier électronique, et suite au partage des parties des systèmes d'entreprise avec les fournisseurs et l'utilisation et la synchronisation des données avec les plateformes de stockages de données telles que dans le nuage " in the Cloud ", les données des utilisateurs sont devenues la cible des différentes tentatives de vols par des intrus pour des différents buts.
L'augmentation d'utilisation des différents services électroniques a obligé les utilisateurs d'échanger les informations sensibles qui veulent le partager avec toute confidentialité, mais malheureusement ce partage a été visé par des intrus qui cherchent à voler cette information. Les pirates veulent prendre les données des utilisateurs de façon illégitime chacun à un but spécifique, certains pirates veulent prendre des mots de passe, certains d'autres sont payés par des entreprises pour voler les futurs plans des autres entreprises compétitives...etc.
Les entreprises et les utilisateurs doivent maîtriser les ressources à partager pour qu'elles soient exploitables de façon intelligente pour lutter contre les tentatives malveillantes, prendront un exemple qui se trouve dans les entreprises, il faut limiter le contrôle d'accès aux données pour chaque utilisateur pour qu'il manipule les taches autorisées de les utiliser, et aussi protéger les ressources partagées dans le système d'information par des pare-feux et des antivirus.
Le courrier électronique est l'un des services d'échanges de données qui est beaucoup utilisé par les entreprises et les utilisateurs, ils partagent leurs données confidentielles et sensibles quotidiennement dans cette plateforme telle que les photos, les vidéos, des instructions...etc. La grande importance de courrier électronique aujourd'hui le fait la cible des différentes attaques spams, ces dernières sont des

courriels non sollicités envoyés vers les utilisateurs pour les pirater.
Les systèmes d'exploitation Android mobile sont aussi la cible des différentes attaques des pirates, ces systèmes ont connu une immense évolution, ils sont utilisés sur les différents appareils importants tels que le mobile. Aujourd'hui les applications Android sont utiles et praticables dans des différents domaines, elles sont connectées avec les sites web, les systèmes d'information et les services de stockage de données dans les nuages. Cette utilisation importante des SE Android l'ont rendu vulnérable aux attaques des pirates qui cherchent à voler les données par le développement des applications Android malveillantes, l'opération de pirate s'achever avec succès lorsque l'utilisateur installe ces applications malveillantes et donne des autorisations d'accès à ces applications de façon indirecte, le pirate puisse voler ce qu'il veut du mobile et ce qui peut se propager vers d'autres appareille connectées au mobile piraté.
Dans nos recherches on a touché deux grands problèmes, le premier est le problème des messageries électroniques menacées par les spams e-mails, et le deuxième est le problème des applications android malveillantes. Le courrier électronique est un outil très important et nécessaire dans le fonctionnement des entreprises et pour les différents utilisateurs dans des différents domaines, ces boîtes de messagerie électronique sont menacées par des e-mails malveillants sous le nom des spams, ces derniers sont générés généralement de la publicité. les spammeurs injectent des menaces dans les courriels et exploitent les faiblesses des victimes, le spam est un grand problème pour les utilisateurs des messageries, beaucoup de recherche existantes cherche à détecter et lutter contre ces actes des intrus. Dans notre étude, on a proposé un détecteur qui filtre les spams des messages normaux, ce détecteur est un algorithme heuristique basé sur le fonctionnement naturel des octopodes, cette technique est une méthode bio-inspirée d'apprentissage automatique basé sur le système de défense des octopodes adaptés au problème des spams afin de défendre et lutter contre ces derniers, notre algorithme est détaillé dans le quatrième chapitre.
Le deuxième problème tangible est le problème des applications Android malveillantes. Récemment le SE Android est le SE le plus populaire et le plus utilisé dans de nombreux appareils, cette utilisation importante de ce système le rendait vulnérable à de nombreux actes malveillants représentés par des applications malveillantes. Depuis le lancement de la version "6.0 Marshmallow", Android a lancé le système des autorisations d'accès des applications lors de leur installation. Dans notre étude, on a utilisé des différents détecteurs pour lutter contre les applications malveillantes en se basant sur les autorisations d'accès (des détecteurs de la fouille de données et un autre détecteur d'apprentissage profond), dans cette étude on a essayé de tester le réseau de neurones récurrent le LSTM pour la détection des applications intruses, et faire une étude comparative des résultats obtenus avec les résultats des méthodes de la fouille de données.

Plan de la thèse

Notre thèse est Divisé à **4 chapitres**, **le premier** parle des technologies du Web et du Big Data, **le deuxième** touche la sécurité informatique en général, **le troisième** chapitre comprend le data mining, les méta-heuristiques et la bio-

inspiration, dans **le quatrième** chapitre nous présentons nos travaux réalisés, et nous illustrons et discutons les résultats obtenus par nos travaux.
Dans le premier chapitre "Technologies du WEB et BIG DATA Analytics", On a présenté un aperçu général sur le Web et le Big Data. Dans la première partie de ce chapitre on a commencé par introduire le Web et voir l'architecture du Web, ensuite on a abordé le World Wide Web et le W3C et présenté aussi la transition du Web statique vers le Web dynamique, après on a touché les protocoles d'échanges sur le web, et à la fin on a mentionné quelques avantages et inconvénients du Web. La deuxième partie de ce chapitre couvrait le Big data, on a commencé par une introduction de Big data et abordé son fonctionnement et ses caractéristiques, et quels sont les outils de big data, ensuite on a mentionné quelques domaines touchés par cette technologie.
Dans le deuxième chapitre "La sécurité Informatique" on a présenté les différents risques humains et matériels qui compromettent la sécurité des données et défini aussi les attaques, ensuite on a mentionné quelques types et classifications de ces derniers et on a parler de la sécurité et ses services, et à la fin on a présenté quelques mécanismes de sécurité.
Le troisième chapitre comprend "Data Mining, Meta heuristique et Bio-inspiration", ce chapitre est riche par les concepts. Dans la première partie de ce chapitre on a défini la fouille de données et vu ses différentes taches, ensuite on a présenté le processus d'extraction de connaissance KDD et défini l'apprentissage automatique avec ses différents types. La deuxième partie de ce chapitre parle des méta-heuristiques, on a commencé par définir les heuristiques et les méta-heuristiques et ait les types de ces derniers. La troisième partie de ce chapitre contient la bio-inspiration, cette partie parle d'historique de la bio-inspiration en transitons du terme bionique au terme bio-mimétisme jusqu'au terme bio-inspiration, ensuite on a vu les différentes sources d'inspiration et aborder les étapes d'inspiration à partir de la nature.
Le quatrième chapitre "Approches, Résultats et Expérimentation" contient les approches et les algorithmes utiliser dans notre étude, on a abordé deux grands problèmes, le premier est le problème des attaques spams sur la messagerie électronique, et le deuxième est le problème des attaques des applications Android malveillantes. Dans la première partie de ce chapitre, on a présenté notre nouvelle technique bio-inspiré basé sur les octopodes pour le filtrage des spams, on a commencé par définir notre problématique et mentionner quelques études existantes, ensuite on a abordé l'approche naturelle de l'octopode et présenter la transition vers notre proposition artificielle, on a testé notre algorithme avec un ensemble de données appeler "SMS Spam collection" et illustrer et discuter les résultats obtenus par notre travail, on a expérimenté notre algorithme en utilisant les différentes représentations de texte (N-gram et sac de mot) avec et sans nettoyer les mots vides, les meilleurs résultats obtenus sont mis pour être comparé avec des études existant pour évaluer notre travail. La deuxième partie de ce quatrième chapitre contient notre travail manifesté dans la détection des applications Android malveillantes en se basant sur les autorisations, on a commencé cette partie par une présentation de la problématique, ensuite on a mentionné quelques études existantes et définir les SE Android, en outre on a vu les différentes versions d'Android et définir les

autorisations des applications, à la fin on a présenté notre travail qui est une étude comparative d'un ensemble des détecteurs de la fouille de données et un détecteur d'apprentissage profond(réseau de neurones récurent LSTM). Dans cette étude on a utilisé le corpus de données appeler "Android Malware/Benign permissions " pour détecter et filtrer les applications Android, ensuite on a illustré et discuté les résultats obtenus par cette étude comparative des différents algorithmes pour la détection des applications Android malveillante.
à la fin on a conclu avec une brève synthèse de ce qu'on a fait dans notre thèse et identifié quelques perspectives et travaux à aborder dans le futur.

Problématique

L'informatisation des systèmes et le développement du matériel informatique performant ont affecté sur le rendement des entreprises dont lequel les services présentés sont améliorés. Les utilisateurs et les entreprises ont suivi ce progrès technologique avec la multitude des plateformes et des services dans le web tels que les plateformes de paiements électroniques et de stockage dans les nuages (Cloud), les services des Big data, les messageries électroniques. . . etc. Ce progrès a touché aussi les smartphones avec la naissance des SE mobile tel qu'Android et IOS.

L'avancement technologique a obligé les utilisateurs et les entreprises de s'adapter et s'orienter vers les différentes nouvelles plateformes pour améliorer les services aux clients, certaines entreprises aujourd'hui sont obligées de partager une partie du système avec les clients. Donc l'échange de données a été grandi de façon exponentielle ce qui a obligé la naissance des plateformes big data pour gérer ce grand volume.

Ce partage et échange de données dans le monde ont été suivis par des tentatives de vols par des pirates qui veulent menacer les données des gens, chaque pirate a un but spécifique qui veut l'atteindre et une méthode de vol. Les pirates informatiques cherchent à voler soit de l'argent, prendre des photos, prendre des idées sur des projets des entreprises, espionner la vie privée des gens . . . etc.

Le partage de données nécessite une bonne maîtrise des outils et des systèmes d'information pour contrôler l'accès aux données, chaque utilisateur doit accéder aux données qu'il est autorisé de les manipuler donc il faut limiter les accès. Les pirates cherchent à nuire la sécurité des systèmes d'information et accéder aux données de façon illégitime de plusieurs façons, certains injectent des virus et des programmes malveillants dans les systèmes et les applications, certains d'autres appelés spammeurs envoient des emails non sollicités qui contiennent des virus pour les utilisateurs de la messagerie électronique, et plusieurs d'autres types d'attaques qui ne se comptent pas.

Il faut protéger les données et les utilisateurs par offrir des mécanismes de sécurité qui luttent contre les actes malveillants des pirates. Les administrateurs

des systèmes peuvent protéger les réseaux et les accès internet par des par feux et des antivirus, ils peuvent aussi protéger les sites web par des systèmes de détection d'intrusion. Les fournisseurs des messageries électroniques protègent les utilisateurs et ses données par offrir des filtres qui détectent les emails spams et les bloquer. Les utilisateurs des systèmes d'exploitation mobiles peurs des applications malveillantes, ils cherchent toujours à trouver des détecteurs qui peuvent protéger leurs données. Un utilisateur de Smartphone peut recevoir des menaces lors de l'utilisation du mobile, et le pirate peut rendre l'utilisateur comme une machine zombie qui fait tout ce qu'il veut, ce qu'il permet de voler les données et l'identité d'utilisateur, et même il peut prendre des conversations secrètes avec d'autres personnes... etc.
Les détecteurs classiques sont des détecteurs très limités et lents et ne s'adaptent pas avec les nouveaux techniques de traitements en parallèle et de stockage dans le nuage, les chercheurs utilisent des techniques et des algorithmes de data mining, et aussi ils utilisent des techniques heuristiques qui sont adaptables aux matériels disponibles, et qui donnent des résultats approchés pour détecter et filtrer les actes malveillants. On a concentré dans notre travail sur deux grands problèmes, le premier des messageries électroniques et le deuxième des systèmes d'exploitation android.

1

TECHNOLOGIES DU WEB ET BIG DATA ANALYTICS

Table des matières

1.1 WEB

1.1.1 Introduction

Internet peut être défini comme un ensemble des ordinateurs du monde entier relié entre eux afin d'échanger des données, les ordinateurs sont reliés par des câbles qui traversent les pays et les océans, ou bien par des satellites [88]. Le mot Internet n'est utilisé qu'à partir les années 1982 avec la définition du protocole TCP/IP [88], Internet est le réseau informatique mondial qui rend accessibles plusieurs services

au public comme le courrier électronique et le World Wide Web ...etc [89]. Le web n'est pas l'Internet, cette dernière est le grand ensemble des réseaux interconnecté entre eux, mais le web peut être défini comme une application associé au protocole http qui permet de prendre et lire les données à partir du serveur et les affiches dans le navigateur client (demandeur).

1.1.2 Bref historique d'internet et du WEB

L'histoire d'Internet a commencé à la fin des années 60 avec un projet appelé ARPANET (Advanced Research Project Agency), en 1969 une connexion de 4 premiers ordinateurs d'ARPANET qui ont été créé un système de transmission de données militaires aux États-Unis, et qui ont ensuite été utilisés par des universités, et plusieurs tentatives de développement ont été suivi après. En 1982 le protocole TCP/IP a été apparu et devenu un standard protocole de communications pour le réseau informatique. Ensuite en 1991 le World Wide Web est né, et en 1993 le premier navigateur Mosaic a été apparu, ensuite en 1994 Yahoo et W3C font leur apparition [116].

1.1.3 Architecture du web

Elle est définie en architecture client/serveur, le principe est de s'appuyer sur un poste central qui est le serveur, et ce dernier envoi des données aux machines clientes, ce fonctionnement est illustré dans la figure 1.1 au dessous [89].

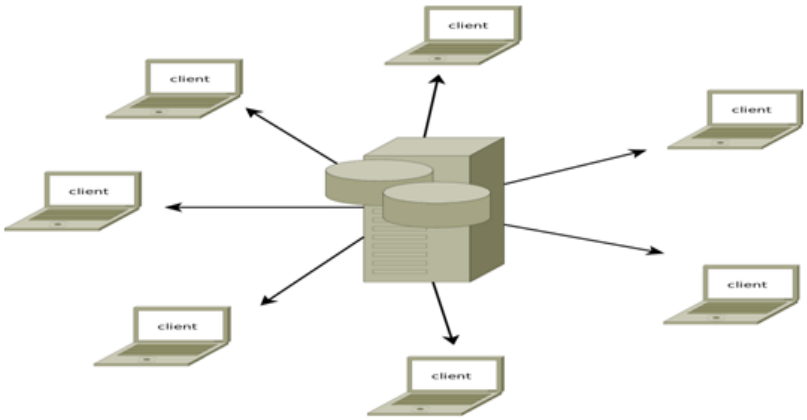


FIGURE 1.1: Architecture générale Client/Serveur.

Le Serveur

Le serveur est un programme qui offre un service sur le réseau, il accepte des requêtes, les traite et renvoie le résultat au demandeur (client), généralement le terme serveur s'applique sur une machine puissante en capacité d'entrée-sortie et qui fournit des services et donne des réponses au demandeur [89].

Le client

Le logiciel client est un programme qui utilise le service offert par un serveur, le client envoie une requête et reçoit la réponse, les programmes qui exploitent les

services de serveur sont appelés programmes clients (client mail, navigateur), le client et le serveur doivent utiliser le même protocole, un serveur peut répondre à plusieurs clients en simultané [89].

1.1.4 Exemple

On peut prendre l'exemple de la consultation d'une page sur un site web, l'internaute connecté via son navigateur web est le client, le serveur est composé des ordinateurs contenant les programmes qui servent les pages demandées et Le protocole de communication HTTP est utilisé [89].

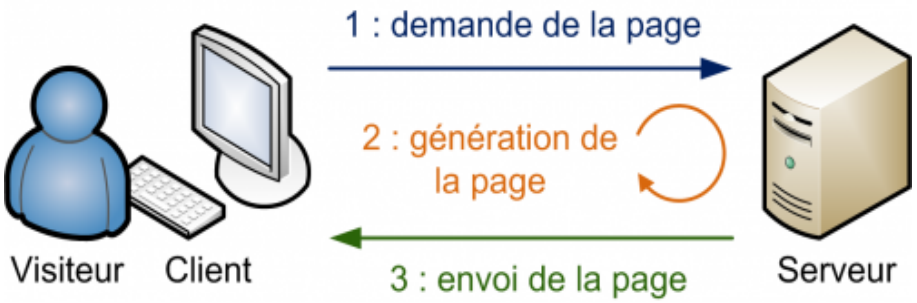


FIGURE 1.2: Consultation d'une page sur un site web.

1.1.5 World Wide Web

La toile mondiale en français, abrégé WWW est un système hypertexte qui permet de consulter des pages sur des sites à l'aide d'un navigateur via internet comme le courrier électronique(email), il est développé par Tim-Berners Lee et Robert Cailliau à la fin des années 1980 [89]. Le Web n'est qu'une des applications d'internet. On appelle le Web tout ensemble des applications mises en œuvre sur internet qui rend accessibles des pages Web via des données stockées dans des sites ou des serveurs qui on les regardent grâce à un logiciel appelé navigateur (par exemple :Mozilla Firefox, Microsoft Internet Explorer) [89]. L'image de la toile vient des hyperliens qui lisent les pages web entre elles [88].

1.1.6 Le W3C

Le W3C est un organisme de normalisation fondé en octobre 1994 comme un consortium chargé de promouvoir la compatibilité des technologies du World Wide Web. Le W3C n'émet pas des normes au sens européennes mais des recommandations à valeur de standards industriels [116].

1.1.7 Web 1.0, web 2.0, web 3.0 et le web 4.0

Le web 1.0 a vu le jour dans le début des années 1990 en utilisant des pages HTML, les pages ont été statiques, les webmasters partagent leurs informations dans des sites et les internautes lisent le contenu de ces sites seulement, ils ont été des récepteurs des informations seulement et leurs comportements étaient passifs, ce web est appelé le web statique (aussi le web traditionnel) [117, 18, 20].
Le web 2.0, certain l'appelle le web social, ce type a été apparu dans les années

2000, les réseaux sociaux font leurs apparition et des communautés sont émergées et des nouvelles plateformes media ont vu le jour comme Facebook et Youtube, le contenu est produit par les internautes et l'entreprise garantit l'hébergement, le web 2.0 a donné la possibilité de partager et échanger des informations, les internautes peuvent interagir et partager quelques ressources ou informations, beaucoup de blogs personnels sont écrits par les internautes, ils ne sont pas justes des récepteurs, ils sont devenus actifs et ils peuvent changer des informations avec des entreprises, ou bien donner des avis pour des produits, ajouter des commentaires... etc, de façon général ils sont devenus des acteurs actifs [117, 18, 20].
Le web 3.0 appelé aussi le web sémantique, ce type est né au cours des années 2010, il ajoute la notion du contexte sur les informations disponibles. Le web 3.0 donne un sens aux données pour faciliter et optimiser le fonctionnement aux utilisateurs, ce type de web est intelligent et il se base sur la manipulation des métadonnées et des données sous forme des ontologies (OWL) des triplets RDF, RDFS, langage XML ... etc [117, 18, 20].
Les nouveaux besoins des utilisateurs dans le WEB et la nécessité d'optimiser les fonctionnements afin de fournir une bonne qualité de service ont donné naissance aux web 4.0, appelé aussi le WEB intelligent (smart), ce web est une extension du web 3.0 mais il se concentre sur le contrôle des données et la protection de la vie privée des gens, il a commencé dans les années 2020, ce type de web est une novelle technologie qui regroupe plusieurs disciplines telles que l'apprentissage automatique, IOT (internet of things), la robotique, le traitement sur le Cloud... etc. Ce type de web est rapide et intelligent, et il est pour but que l'humain interagit avec la machine de façon similaire à la communication des humaines entre eux [117, 18, 20].

1.1.8 Web statique Vers le web dynamique

Au début du World Wide Web (1991) le contenu des pages était fixé, ils ont été des pages statiques [116] mais depuis l'apparition et l'utilisation des langages des scripts exécutables sur le serveur dans différents langages (PHP, Python...etc.) a rendu possible de faire varier le contenu des pages et donner le début du web dynamique [116].

Le script

Le script est un programme informatique chargé d'exécuter une fonction bien précise lorsqu'un utilisateur réalise une action ou lorsqu'une page web est en cours d'affichage sur un écran [69].

1.1.9 Développement web d'applications

Aujourd'hui avec la naissance des langages de développement du web comme : HTML/CSS, JavaScript, PHP... etc. Le développement du web se repose sur l'utilisation de ces langages par les développeurs web pour créer des applications du WWW, ou des applications exécutées dans le serveur sur un navigateur via un protocole http. Chaque développeur rédige les lignes de code et développe les applications web en fonction des exigences présentées dans un cahier des charges bien précis, il analyse les besoins et donne une solution technique désirée [116, 92].

Le développement web n'est pas assez facile, il faut maîtriser plusieurs langages de différentes technologies par exemple pour les structures des documents il faut maîtriser le XML , XHTML ...etc. Pour le style des feuilles il faut connaître le CSS, pour les interactions quelque soit côté client ou bien serveur il faut connaître les langages comme : Javascript , PHP , Python , Java, apache2...etc [116].

1.1.10 Protocoles d'échange sur le Web

L'échange de données sur le Web se fait avec le protocole HTTP (Hyper Text Transfert Protocol) qui permet au client de demandé via une requête, et le serveur répond (requête/réponse). Dans le web beaucoup de données circulent quelque soient images(GIF, PNG...etc), textes, vidéos(mp4...etc), audios(mp3). Le web nous permet de faire des formations en ligne, partager des données, voir les actualités du monde, acheter des produits en ligne...etc. On peut faire beaucoup de choses, mais ces services fournis par le web sont exposés à des risques par des intrus qui cherche à voler les données [116].

1.1.11 Avantages et Inconvénients du Web

Le web a réduit les distances, grâce au web on peut partager des données à des distances très éloignées et faire des connaissances avec des personnes de différents pays. On peut aussi acheter des produits, naviguer, chatter...etc. Le web nous a fourni beaucoup de services. Comme le web a des avantages, il a aussi des inconvénients, parmi ces derniers on pris par exemple les sites qui pourront être retirés (indisponibles), et d'autres sites qui contiennent des virus injectés par des pirates qui veulent espionner la vie privée des gens (contenu malicieux). L'augmentation rapide des données a causé une difficulté d'utilisation des informations et un risque de contrôle, et des connexions occupées et lentes [89].

1.1.12 Conclusion

Le web a fourni beaucoup de services aux utilisateurs, il a donné la possibilité de faciliter les échanges de données entre les personnes de différents lieux, il a permis aussi de suivre l'actualité au monde et plusieurs choses avantageuses, mais ce partage et échange de données dans le web peuvent s'expose à des risques des intrus qui peuvent espionner l'information confidentielle de la vie privée des gens par exemple : certains sites obligent les utilisateurs à fournir des informations sensibles pour qu'ils puissent accéder à leurs contenus. Avec le développement des sites web et des réseaux sociaux, et avec la liaison des gens dans le web et la contribution forte du web dans la vie des utilisateurs, les données sont devenues en danger, donc la sécurité des sites web et des données confidentielles des utilisateurs reste un problème majeur et essentiel à résoudre.

1.2 BIG DATA

1.2.1 Introduction

Durant les dernières années, les données numériques de différents formats ont été augmentés de façon exponentielle, cette augmentation importante a obligé les développeurs de trouver des nouvelles techniques pour gérer et analyser ce grand volume de données. Ces quantités de données stockées à l'échelle mondiale ne cessent de croître à grande vitesse. Pierre Nerzic a mentionné dans ces études l'énorme augmentation des données, par exemple : en 2015 Google a déclaré une augmentation avec une quantité de 10 Eo (10 milliards de Go) [97], en 2018 Facebook a déclaré une augmentation avec une quantité de 1 Eo (avec un moyen de 7 Po de nouvelles donnée par jour) [97], Amazon a déclaré une augmentation avec une quantité de 1 Eo [97]. Ces données gigantesques sont difficiles de les stocker et les traiter, donc cela nécessite de trouver des nouvelles manières pour stocker et contrôler ces données, ce qui a conduit à la naissance du ***big data***, ou bien ***méga-données***, ou encore ***données massives***.

1.2.2 Le Big Data

Le big data ou méga donnée, grosses données ou encore données massives. Selon le rapport IDC (International Data Corporation) la masse de données en ligne a été gonflée en 2011 avec 1,8 zettaoctets. Actuellement la masse de données produites est estimée à près de 3 trillions (3.10^{18}) d'octets de données [14], ces données s'accroissent de manière colossale et rapide et de différents formats, elles peuvent être des messages textuels, des vidéos, des images, des sons vocaux, des signaux GPS. . . etc. Gartner a défini le Big Data comme un concept qui regroupe les trois « V » (Variété, Volume, Vélocité), il voit que les données se présentent avec différents formats, de différents sources avec un volume croissant, et une vitesse de collecte et de partage très rapide [102, 14]. Cette définition de Gartner est abondante dans l'arène du big data, elle peut être comme un standard de définition de big data. Beaucoup de chercheurs ont basé sur cette définition pour ajouter des nouvelles sur cette technologie, certaines entreprises ajoutent deux autres « V » à cette définition qui sont la valeur et la véracité [102], sur ces deux derniers "V" , le premier se concentre aux données et qu'ils doivent ont une valeur mystique (les données collectées signifient quelque chose), et le deuxième "V" proposé par les entreprises s'intéresse sur la fiabilité et la crédibilité des sources de données avant de les utiliser. Cependant il n'existe pas une définition bien précise pour le big data, sa définition se varie d'une communauté à une autre, d'un utilisateur à un autre, d'un fournisseur de services à un autre.

Fonctionnement du Big Data

Intégrer le big data permettent de rassembler les données de différents formats arrivants des sources diverses. Les méthodes d'intégration de données traditionnelles sont limitées et ne terminent pas la tâche facilement de manière parfaite avec une bonne qualité. Donc pour faire du big data, il est nécessaire de donner la naissance à des nouveaux techniques qui permettent d'intégrer des données de grands

volumes (on parle des pétaoctets et des zettaoctets) avec une haute performance de réalisation de la tache [102].

Gérer pour faire du big data il faut que vous garantissiez l'opération de stockage des données. Le cloud peut la fournir, il prend en considération vos besoins de traitement de données et laisse la possibilité d'ajouter des données en fonction de ce que vous voulez [102].

Analyser faire exploiter vos données, cette étape est très importante dans le big data, grâce aux gros volumes de données on peut mieux extraire de l'information, on cherche à trouver des décisions et des nouvelles solutions pour des divers problèmes en exploitant ces données en utilisant des techniques d'apprentissage automatique, d'apprentissage profond et d'intelligence artificielle [102].

Les caractéristiques du big data

volume L'une des caractéristiques les plus importantes dans le big data est bien le volume, il peut se définir par la masse de données produite de différentes sources, ou d'une autre façon c'est la taille de données [102, 58, 10]. Cette masse s'augmente de façon exponentielle chaque jour. Un ensemble de chercheurs présentent le volume de façon que ce dernier n'a pas de limite, d'autres chercheurs disent que si vous consulteriez des grands volumes de données vous allez toucher le big data, pour d'autres chercheurs si vous consulteriez la plus grande masse de données que vous avez utilisée dans vos recherches donc vous allez être dans une plateforme big data. Certaines entreprises voient que pour parler de volume, il faut que la masse de données puisse s'agir à 10 téraoctets, pour d'autres à 100 pétaoctets [105, 58, 10].

vélocité En d'autre termes la vitesse, manifesté dans la réception rapide des nouvelles données générées des différents sources, ces données doivent être traitées d'une façon rapide par analyser la masse traitée, et essayer de trouver des réponses dans un temps réel ou quasi réel par exemple : une entreprise qui traite ses données en plusieurs jours, elle peut réaliser les mêmes tâches en quelques minutes grâce aux nouvelles technologies de big data telle que Hadoop et map-reduce, cette rapidité de traitements de données à un impact économique évident parce qu'un traitement lent de l'information peut menacer l'avenir d'une entreprise [102, 29, 58, 10].

variété La variété peut être définie dans les différents types et formats de données disponibles quelque soit structuré, semi-structuré ou non structuré. Auparavant l'information traité été clairement structuré dans des bases de données, mais aujourd'hui avec l'augmentation des données plusieurs types de données ont été mise en œuvre tels que les données semi-structuré et non structuré (texte, audio, vidéo, image...etc). Ces nouveaux types de données nécessitent un prétraitement pour qu'elles soient utilisables, la technologie de big data offrent des outils tels que NoSQL qui peut traite tous les types de données à sa forme originelle [102, 29, 58, 10].

véracité La véracité représentée par un contrôle préalable se fait aux données collectées afin de voir si ces derniers sont crédibles et fiables pour donner un sens d'utilisation et qu'elles ne sont pas collectées sans aucun intérêt. La véracité cherche à faire un contrôle aux données avant qu'elles soient collectées pour qu'elles

soient validées avec un intérêt d'utilisation pour un but bien spécifié, et ne pas ajouter des données sans aucun sens, cela permet de perdre de l'espace de stockage et de temps sans aucun résultat [29, 17, 58, 10].

valeur La valeur de big data signifie que les données ajoutées de différentes sources ont un sens et signifient quelque chose, et n'est pas qu'un gaspillage de ressources, et que ces données ont un objectif à atteindre [29, 17, 58, 10].

Variabilité Cette caractéristique se concentre sur l'incohérence des flux de données, il existe plusieurs nombres dans les données et cette propriété est devenue un défi lors d'utilisations des données médias numériques. Les échelles de données se varient aussi à cause des différentes sources et types de données [58, 10].

Validité Plusieurs chercheurs scientifiques font nettoyer les données afin de faire une analyse sur ces derniers, la validité vise à voir l'exactitude et à corriger les données pour être utilisé dans leurs objectifs destinés [58, 10].

Volatilité Ce "V" cherche à voir la durée de pertinence et d'utilité des données, il s'intéresse au calcul du temps quand les données sont importantes, ce v cherche à gagner de l'espace de stockage et ne pas stocker les données jusqu'à ordre indéfini. Un stockage de données non utiles peut causer des dépenses élevées si pour ça les chercheurs doivent trouver des méthodes de mise à jour pou garantir la disponibilité des données en cas de nécessité de récupérer ces derniers [58, 10].

Visualisation L'une des caractéristiques des big data est la visualisation des données, visualiser un grand nombre de données n'est pas facile, car il existe plusieurs défis en fonction du temps et techniques. Il faut représenter les données pour qu'elles soient graphiquement lisibles et significatives par exemple : par des diagrammes de réseaux, des formes arborescentes, des rayons, des cartes ... etc [58, 10].

Viralité Cette caractéristique définit la vitesse de propagation des données, elle mesure la vitesse de diffusion des données à partir d'un utilisateur vers d'autres différents utilisateurs afin d'utiliser ces données [58, 10].

Viscosité Cette caractéristique définit le décalage des événements, elle mesure la différence de temps entre l'événement produit et l'événement décrit [58, 10].

Venue Cette caractéristique se concentre sur l'arrivage de données à partir de différentes plateformes, elle s'intéresse sur les différents types de données reçues à partir des différentes sources via des différentes plateformes [58, 10].

Vocabulaire Cette caractéristique s'intéresse sur la terminologie des données (les modèles de données, les structures de données... etc.) [58, 10].

Vagueness Cette caractéristique se concentre sur l'indistinction existée dans une donnée, elle concerne la réalité des informations qui suggéraient peu ou pas de réflexion sur ce que chacune pouvait transmettre [58, 10].

1.2.3 Les Outils Du Big Data

Les SGBD sont incapables de traiter des grandes masses de données, mais actuellement avec la naissance du terme big data, des systèmes de stockage et de traitement de grosses volumes de données comme **Hadoop**, **MapReduce**,

NoSQL ont vue le jour, ces systèmes sont capables de traités un grand volume de données.

Map-Reduce

MapReduce est un modèle d'architecture de développement informatique, il est pour objectif d'effectuer des traitements de façon parallèle souvent sur des données trop volumineuses. Ce modèle contient deux tâches essentielles map et Reduce, la fonction Map découpe un problème du traitement de données en sous-problèmes et la fonction Reduce aide à traiter les données de chaque sous-problème en parallèle, ensuite les résultats des sous-problèmes sont remontés pour résoudre le problème réel posé et obtenir des résultats de façon rapide [14]. Le modèle MapReduce distribue les données dans un cluster pour les traiter, ce modèle a été propagé très rapidement par des sociétés qui possèdent un traitement de données important telles que Facebook et Amazon, il est utilisé aussi dans le cloud computing [14].

Hadoop (High-Availability Distributed Object-Oriented Platform)

Hadoop est un framework open source basé sur le principe de Map-Reduce, il a été créé par Doug Cutting, puis il a été développé par la fondation Apache en 2009, son principe repose sur le traitement distribué des grandes masses de données pour augmenter la puissance de calcul et de stockage [113]. Hadoop fractionne le traitement des données sur les machines disponibles, en cas de défaillance le traitement est reporté à une autre machine, cela permettra le traitement en parallèle, chaque machine assure une partie des calculs et du stockage. Hadoop se compose du MapReduce, du système de fichiers de données distribuées HDFS(Hadoop Distributed File system) et d'un certain nombre de projets associés, notamment Apache Hive, HBase...etc [113].
HDFS est un système de fichiers distribués permettant d'accéder rapidement aux données, Il fournit des mécanismes de résistance aux pannes [113].
Hadoop MapReduce regroupe les algorithmes de traitement en parallèle des données [113].
Un certain nombre de modules ont été ajoutés tel que Hive qui est un langage de requête proche de SQL qui permet d'interroger les données stockées dans Hadoop à travers HiveQL, encore Pig qui est un langage de script pour interagir avec de larges ensembles de données, ou encore Spark et Tez, et d'autres systèmes de traitement des données [113].
Hadoop est adapté au stockage de données non structurées (les données textuelles ou les documents multimédias ...etc), il ne nécessite pas de connaître la structure des données pour accéder aux données (comme les bases de données relationnelles) [113].

NoSQL

Des nouveaux modèles de stockage adaptés aux gros volumes de données ont vu le jour, et donnent la naissance aux bases de données NoSQL, ils ont été proposé par Carl Strozzi, ces modèles ont été apparus comme une alternative aux bases

de données relationnelles, ils permettent de stocker et accéder aux données sans l'utilisation du langage SQL(l'unité logique n'y est plus la table), et les données ne sont pas manipulées avec SQL. Ces systèmes "NoSQL" ne nécessitent pas de connaître la structure des données, ils peuvent accéder à n'importe quel type de données quelque soit structuré ou semi-structuré [14]. Ces systèmes "NoSQL" ont une grande performance dans le contexte des applications Web, ils utilisent généralement des algorithmes du type MapReduce pour paralléliser l'ensemble de données pour effectuer la tâche. Plusieurs grands acteurs dans le web ont adopté les bases NoSQL comme Google, Facebook...etc [14].

Apache spark

Apache spark est un framework open source qui permet de traiter et analyser les données de façon rapide, il est dédié au big data, il traite les données volumineuses de façon distribuée (traitement en clusters), il est rapide et facile à utiliser, et il permet aussi de traiter les fichiers HDFS et les bases NoSQL... etc. Il prend en charge le traitement dans la mémoire (in-Memory), et il permet d'utiliser aussi le traitement sur disque (conventionnel) si le volume de données est trop large pour la mémoire [49, 126].
Il se caractérise par la vitesse, et il peut exécuter des programmes 100 fois plus rapide que Hadoop MapReduce dans la mémoire et 10 fois plus rapides sur le disque [49, 126].
Apache spark est facile à utiliser pour programmer des applications pour les traitements en parallèle en langages java, python, Scala ou R. Il peut être exécuté aussi sur hadoop ou sur le Cloud, il peut aussi traiter les différents sources de données tels que HDFS, Hbase... etc [49, 126].
Il se caractérise aussi par sa généralité parce qu'il contient de plusieurs bibliothèques, il contient les bibliothèques SQL, les dataframes, les spark streaming et les traitements par graphes (GraphX), il contient aussi une bibliothèque Mlib pour l'utilisation des algorithmes de machine Learning et plusieurs d'autres bibliothèques. Ces bibliothèques peuvent être utilisées dans un seul programme [49, 126].
Des grandes entreprises utilisent apache spark, par exemple Netflix et pinterest l'utilisent en streaming pour offrir une diffusion rapide et meilleur des films et vidéos, sa dernière version 2.2 était lancée le 11 juillet 2017 [49, 126].

1.2.4 Les domaines du big data

Le big data peut toucher presque la majorité des domaines qu'ils existent, on peut mentionner quelques-uns :

La recherche médicale

Avec l'augmentation des méga-données et grâce à la plateforme big data qui nous facilite l'évaluation des données massives, actuellement les médecins peuvent trouver de meilleures solutions de traitement pour remédier à leurs patients[45].

L'industrie

Grâce aux méga-données, et en raison de la diversité et la multiplicité des données, les entreprises peuvent améliorer et augmenter leurs productions et travailler

d'une manière plus durable et renouvelable[45].

L'économie

Le big data aident les entreprises à mieux comprendre, connaître et prédire les besoins de leurs clients, et de leur proposer des offres toujours plus adaptées qui satisfont leurs besoins[45].

L'énergie

Les données peuvent aider les spécialistes à trouver des prédictions sur la consommation énergétique à long terme pour trouver une consommation durable et renouvelable [45].

Le marketing

Les données sont utilisées par des algorithmes pour l'amélioration des méthodes de vente et d'achat, par la fourniture des moyens nécessaires pour faciliter le marketing, et améliorer les relations avec les consommateurs [45].

Le secteur bancaire

Le Big Data et le web permettant aux banques d'améliorer ses services et de proposer des fonctionnalités adaptées au profil de ses clients [45].

L'enseignement

La diversité et la multiplicité des données générées dans le web aident les enseignants de différents secteurs d'améliorer ses connaissances et de mieux les former par l'apprentissage.

1.2.5 Conclusion

Nous avons présenté dans ce chapitre un bref aperçu sur le web et Le Big Data, nous avons vu que le big data est appliqué dans tous les domaines liés au Web, l'augmentation des données dans le web peut causer des difficultés sur les systèmes classiques, cependant avec le développement de la technologie big data et la naissance des outils de cette dernière, l'analyse de données et l'extraction de la connaissance sont devenues faciles et rapide qu'auparavant. Les outils tels que hadoop, map-reduce et NoSQL ont aidé les gens à mieux générer, stocker et manipuler le grand volume de données (on parle des mégas, pétas et même des zettas de données), cette augmentation de données a touché plusieurs domaines s'ils ne sont pas tous, presque tous les domaines ont bénéficié de la naissance du big data, ce gros volume est difficile de le contrôler ce qui a causé les problèmes des intrus qui veulent espionner les données sensibles et confidentielles des gens. Les fraudes ont été augmentées de façon exponentielle, donc il est nécessaire de lutter contre ces menaces et de garantir la sécurité des données par suivre des précautions et augmenter d'utiliser les mécanismes de défense.

2 LA SÉCURITÉ INFORMATIQUE

Table des matières

2.1 Introduction

Le grand échange et partage de données dans le web actuellement ont donné la naissance au terme big data, cette augmentation de volume s'expose à des difficultés de protection surtout pour le cas des données sensibles, un grand nombre des intrus ont été apparus dans le but d'espionner la vie privée des gens. Ce phénomène pose des grands risques si pour cela les chercheurs ont essayé de trouver des techniques qui visent à protéger les ensembles de données et de ressources matérielles et logicielles.
La pratique d'aujourd'hui associée aux systèmes d'information des entreprises qui deviennent ouverts et accessibles par les utilisateurs et les fournisseurs est essentielle, ce partage nécessite de bien connaître leurs ressources afin de définir la partie sensible dans le fonctionnement des entreprises pour la protéger et garantir une exploitation contrôlable avec un accès maîtrisé. **Par exemple** les nouvelles tendances informatiques de stockage dans les nuages« in the Cloud » ont permis

aux utilisateurs d'accéder aux ressources et de partager une partie de système d'information d'une société ou d'une entreprise, donc cela nécessite d'évaluer les risques pour trouver une meilleure protection.

2.2 Les risques

Les systèmes d'information sont vulnérables à de nombreux risques, donc il faut mesurer ces risques en mesurant leurs effets possibles, si les conséquences sont négligeables ou catastrophiques. Nous pourrons énumérer certains des risques ci-dessous [15].

2.2.1 Les risques humains

Les risques humains comprennent tous les risques causés par l'être humain, soit par les utilisateurs soit par les informaticiens eux-mêmes [15].

2.2.1.1 La maladresse

Par exemple : exécuter un traitement non souhaitable ou effacer involontairement des données ou des programmes...etc [15].

2.2.1.2 L'inconscience et l'ignorance

Introduire des programmes malveillants sans le savoir (par exemple lors de la réception de courrier) [15].

2.2.1.3 La malveillance

Comme les virus et les mauvaises manipulations par exemple : certains utilisateurs peuvent mettre en péril le système d'information par des modifications en introduisant volontairement des mauvaises informations dans la base de données ou par injection des logiciels malveillants, prendrons un exemple réel : un informaticien peut ajouter des fonctions cachées dans un système qui lui permet de voler de l'argent ou des informations sensibles [15].

2.2.1.4 L'ingénierie sociale (social engineering)

Cette technique vous permet d'obtenir d'une personne des informations confidentielles afin de les exploiter pour des fins spécifiques [15].

2.2.1.5 L'espionnage

Rassemble l'ensemble des techniques qui visent à obtenir des informations sensibles surtout dans le secteur industriel, pour prendre des informations sur les activités des entreprises concurrentes par exemple : les futurs produits, la politique de fabrication des produits, les projets en cours de réalisation...etc [15].

FIGURE 2.1: Exemple d'un pirate qui fait l'espionnage

2.2.2 Les risques matériels

2.2.2.1 Incidents liés au matériel

Les incidents liés au matériel peuvent se définir sur les composants électroniques qui ont des défauts et qui tombent en panne [15].

2.2.2.2 Incidents liés au logiciel

Les programmes sont de nature complexe car ils font plusieurs taches, ils nécessitent des efforts de plusieurs programmeurs, ces derniers peuvent faire des erreurs dans des instructions quelque soit de manière individuelle ou collective qui affecte passivement sur tout le système [15].

2.2.2.3 Incidents liés à l'environnement

Parfois les conditions climatiques inhabituelles affectent sur le fonctionnement des machines électroniques, il est possible qu'un ordinateur tombe en panne, même pour le cas des réseaux de communication qui sont sensibles aux changements de température par exemple : en cas d'incendie, ou pour un cas de champ magnétique [15].

2.3 Les Attaques

N'importe quelle action (tentative) qui cherche à nuire la sécurité des informations [35].
Chaque attaque a un but spécifique, il existe des attaques qui cherchent à interrompre les informations et visent à nuire la disponibilité de données, d'autres attaques ciblent la confidentialité des informations, un autre type d'attaque cherche à faire des modifications qui visent l'intégrité des informations, un dernier type d'attaque vise l'authenticité des informations [35].

2.3.1 La Première classification

La première classification des attaques comprend les attaques passives et actives [110].

2.3.1.1 Les attaques passives

Ce type d'attaque concerne les attaques qui vise à prendre ou à lire une information confidentielle transmise (la capture du contenu d'un message qui a été transmise), ou à utiliser les données sensibles communiquer par les utilisateurs pour un but spécifique, mais ce type d'attaque ne sert pas à modifier l'information ciblée **par exemple** : la lecture des messages, ou l'analyse du trafic [110]. Un courrier électronique peut contenir des informations sensibles.

2.3.1.2 Les attaques actives

Ces attaques tentent à modifier les ressources partagées dans une communication, elles visent à créer des ressources de données escroqueries, on trouve quatre catégories d'attaque de ce type qui sont : la modification de messages, le déni de service, la mascarade et le rejeu [110].

2.3.2 La deuxième classification

La deuxième classification des attaques comprend les attaques internes et externes [47].

2.3.2.1 Les attaques internes

Ce type d'attaque est causé par les utilisateurs qui sont autorisés pour accéder au système, généralement par les employés d'une entreprise, et pour lutter contre ces menaces internes, les entreprises utilisent souvent des pare-feu où des antivirus [47].

2.3.2.2 Les attaques externes

Ces types d'attaque sont générés par des utilisateurs externes qui ne sont pas autorisés d'accéder au système, et qui essayent d'accéder à des informations ou des ressources d'une manière illégitime et non autorisée pour lire, modifier ou utiliser ces données sensibles, ce sont **les pirates** les plus connus par ce type d'attaque [47].

2.3.3 La troisième classification

Selon cette classification, les attaques de cette catégorie peuvent porter atteinte à un service de sécurité ou plusieurs services au même temps, il existe des attaques qui cherchent à nuire la confidentialité des informations en brisant les règles privées, d'autres pour but de toucher l'intégrité en altérant les données, un autre type d'attaque vise la disponibilité en rendant un système ou un réseau informatique indisponible, un dernier type cible l'authenticité des informations [15].

2.3.4 Les différents types d'attaques

Il existe plusieurs types d'attaques, on peut mentionner quelques-uns.

2.3.4.1 Les Attaques sur un système informatique

Ce type d'attaque cherche à nuire un système informatique, et il essaye de prendre des informations sensibles par exemple : les virus, l'enregistreur de frappe (keylogger)...etc [15].

Les programmes malveillants un logiciel malveillant (malware en anglais) est un logiciel développé qui cherche à attaquer un système informatique pour arrêter ses fonctionnalités ou profiter de ses privilèges [15], on peut mentionner quelques-uns ci-dessous :

Virus Le virus est un programme écrit qui cherche à perturber le bon fonctionnement d'un système informatique, il peut se propager à l'intermédiaire des messages électroniques ou par publication en ligne sur internet ou par les clés USB [15].

Le ver (worm) Le ver est un programme malveillant qui se déplace dans un réseau et qui peut être considéré comme un virus [15].

Le cheval de Troie (trojan) Un cheval de Troie (Trojan Horse en anglais) est un programme malveillant qui se présente comme un logiciel légitime dans un micro ordinateur mais il contient des fonctionnalités malveillantes, une fois qu'il se propage dans un système, il peut faire des graves dégâts, on peut mentionner quelques-uns célèbres : Vundo, FlashBack, Netbus... etc [15, 147].

La porte dérobée (backdoor) est une fonctionnalité dans un logiciel malveillant qui permet d'ouvrir un accès sur un système informatique pour prendre le contrôle, l'attaqueur peut créer des réseaux de botnet(un groupe d'ordinateurs zombies contrôlé) [15, 131].

Le logiciel espion (spyware) est un programme malveillant, lorsqu'il est installé dans un ordinateur, il cherche à collecter les informations confidentielles qu'ils contiennent sans l'autorisation ou la connaissance d'utilisateur, ensuite il transfère ces données vers une autre source ou un autre ordinateur pour exploiter ces données illégitimes [15].

l'enregistreur de frappe (keylogger) est un programme malveillant qui s'installe dans un ordinateur et qui ne se montre pas aux utilisateurs (illisible), il est pour but d'enregistrer toutes les frappes de clavier réalisé par l'utilisateur [15].

le rootkit est un logiciel malveillant qui cherche à obtenir les droits d'administrateur sur un ordinateur ou une appareille mobile [15].

2.3.4.2 Les attaques d'application

Ce type d'attaque regroupe les programmes malveillants qui cherchent les failles de sécurité sur les applications et les logiciels afin de les pénétrer [15].

L'exploit peut-être défini comme un programme malveillant qui cherche aux failles de sécurité d'un logiciel afin de l'exploiter, ensuite il les utilise pour profiter de privilèges illégitime [15].

2.3.4.3 Les attaques sur les sites web

Ces types d'attaques sont faisables par des pirates qui cherchent à bloquer l'accès au site, ou à prendre le contrôle d'un site web (ajouter ou supprimer des contenus), ou encore à récupérer les données transmises par des utilisateurs [83]. Il existe plusieurs d'attaques de ce type, on peut mentionner quelques-uns les plus connues : **les injections SQL** qui permettent à un hacker d'injecter des requêtes SQL dans une base de données pour perturber ou stopper le fonctionnement des services du site web, ou de voler des informations confidentielles stockées dans les bases de données des sites [83], on a aussi l'attaque de **cross-site scripting (XSS)** qui vise à injecter un contenu malveillant en langage script dans la page web pour déclencher des actions indésirables [83].

2.3.4.4 Les attaques sur la messagerie électronique

Le courrier électronique est un service important d'échange d'informations pour les internautes, cette utilisation importante du courrier électronique en a fait la cible de diverses perturbations telles qu'elles des attaques de spam, la plupart du temps sont générées de la publicité. Le service des messages courts (SMS) est également la cible d'attaques de spam comme les emails, il est donc nécessaire de protéger les e-mails et les messages courts contre les différentes attaques des spammeurs [15].

Le pourriel (spam) de nombreuses informations sont échangées par les e-mails, ils sont devenues des outils essentiels pour les opérations commerciales, et même pour les personnes dans leur vie quotidienne, elles sont désormais utilisées dans tous les secteurs professionnels. Cette utilisation importante des e-mails peut être exposée à un risque d'attaques pour espionner les données, généralement dans un but précis notamment industriel pour obtenir des informations sur les activités concurrentes, et trouver tous les détails sur les activités des autres sociétés : (projets en cours, futurs produits, politique de prix). Ces attaques sont effectuées par un non sollicité e-mail appelé spam [15]. Le spam peut être défini comme un e-mail ou un SMS de copies identiques, envoyées automatiquement en nombre, indésirable, non sollicité, reçu sans le plein consentement du destinataire [15]. Les annonceurs sont les premiers spammeurs mais certains web-masters n'hésitent pas à promouvoir leur site à travers ce [15].

L'hameçonnage (phishing) est un courrier électronique d'où le pirate demande aux victimes de lui fournir des informations confidentielles [15].

Le canular informatique (hoax) est un courrier électronique qui force l'utilisateur à faire des opérations dangereuses sur son poste, par exemple : suppression d'un fichier, arrêt système non autorisé..etc [15].

2.3.4.5 Les attaques sur le réseau

il existe des différents types d'attaques réseau qui cherche à prendre un accès non autorisé aux ressources, ou de lire le contenu secret afin de l'exploiter à d'autres fins [15].

Les écoutes (sniffing) est une technique utilisé par les pirates informatiques pour capturer le trafic du réseau, les pirates détectent tous les messages qui cir-

culent dans le réseau en utilisant un logiciel sniffer [15].

L'usurpation d'identité (spoofing) le spoofing est une technique qui consiste à prendre une autre identité électronique d'une personne ou d'une autre machine pour masquer sa propre identité, on peut trouver trois types : l'email spoofing, l'IP spoofing et le smart-spoofing IP [15, 104].

Le déni de service (denial of service "DoS") DoS est une technique qui vise à perturber ou paralyser un serveur informatique en provoquant des interruptions de service, par exemple en menaçant l'activité d'une entreprise par saturer l'espace de stockage, ou saturer la capacité de traitement d'une base de données ce qui rendre une application informatique incapable de répondre aux requêtes de ses utilisateurs [15, 132].

2.4 La sécurité

Il est toujours plus facile d'attaquer que de défendre, donc pour fournir une bonne politique de sécurité il faut d'abord connaître les risques et les attaques qui menacent la vie privée des gens. Les problèmes techniques de la sécurité informatique peuvent être classés en deux grandes catégories, la première se concentre sur la sécurité des appareilles comme les ordinateurs, serveurs, smartphone...etc. La deuxième catégorie s'intéresse sur la sécurité des sites web et des réseaux. Donc il faut protéger les données contre les actes de malveillance et cela peuvent être garantis en élaborant une bonne politique de sécurité, par exemple pour garantir une prévention contre des intrus dans une entreprise il faut définir qu'elles sont les ressources à protéger pour assurer la fourniture de ses activités par offrir un accès limité [79]. Aujourd'hui avec le grande nombres des ordinateurs et des smartphones, et la multitude des activités qui ont fourni l'accès au donnée à partir des différentes appareilles, et avec la naissance de la technologie d'informatique en nuage (Cloud) les utilisateurs ne sait pas où sont les données stockées, ou comment ils sont sauvegardées et quels sont les traitements appliqués sur ces données ?, esque ces données sont-elles conservées en toute sécurité ?.
Donc pour offrir une bonne politique de sécurité il faut définir les risques possibles et les objets sensibles à protéger, il faut prendre des précautions en garantissant la sûreté des systèmes informatiques contre les accidents matériels qui peuvent être causés par le climat, et lutter contre les actes malveillants des intrus par proposer des mécanismes de sécurité qui cherchent à garantir les services de sécurité [79]. Dans la littérature il existe plusieurs techniques et mécanismes de sécurité des données par exemple : les techniques de chiffrement dans le cas de circulation des données dans le web ou dans un réseau informatique, les pare-feu (firewalls) pour lutter contre les intrusions dans un réseau informatique ...etc [79]. Ces mécanismes de sécurité cherchent à assurer la protection des systèmes et des réseaux informatiques pour défendre les données des utilisateurs. [79].

2.4.1 Service de Sécurité

Un service de sécurité cherche à augmenter la sécurité des échanges de données dans un système informatique, un mécanisme de sécurité peut se baser sur un ou plusieurs services de sécurité. On peut citer les services de sécurité les plus connus

au-dessous [35].

2.4.1.1 Confidentialité

Ce service assure que seules les personnes autorisées ont accès aux données échangées de la communication, la confidentialité rend l'information incompréhensible à d'autres personnes qui ne sont pas autorisées et que les seuls acteurs de la transaction ont l'accès à l'information [35].

2.4.1.2 authentification

L'identification des acteurs de la communication est vérifiée quand l'information n'est pas accessible que par les acteurs autorisés [35].

2.4.1.3 Intégrité

Ce service garantit que les données de la communication ne sont pas modifiées et que ces données échangées sont les données réelles envoyées et reçues par les acteurs de la communication [35].

2.4.1.4 Non-répudiation

Ce service garantit que les acteurs de la communication participent à l'échange de données et ne nie pas la participation dans une communication[35].

2.4.1.5 Disponibilité

L'objectif de ce service est de garantir l'accès aux données à tout moment dans des bonnes conditions pendant un échange de données, l'information n'est accessible que par les acteurs de la communication[35].

2.4.2 Mécanismes de Sécurité

Un ensemble des techniques et des stratégies qui visent à détecter et lutter contre les actes malveillants et cherchent à stopper tout type d'attaque qui veut compromettre la sécurité. On peut mentionner quelques mécanismes de sécurité dans la suite [35].

2.4.2.1 Antivirus

Un utilitaire capable de protéger un système informatique contre les virus et les malwares qui cherchent à nuire l'ordinateur ciblé et faire des dégâts sur les systèmes (effacer des données, l'hameçonnage...etc) pour voler les données sensibles [35, 74].
Aujourd'hui avec le développement des systèmes informatiques, les pirates ont profité et accompagné cette progression, beaucoup de virus et des actes malveillants ont vu le jour visant à supprimer et menacer les données ciblées, prenons l'exemple des attaques spams sur les messageries électroniques. Le but des antivirus est de protéger le système contre tous types d'intrusion. La meilleure protection des gens contre ces virus est d'être vigilant et ne croyez pas aux tromperies des publications qui disent vous avez gagné quelque chose, le même cas pour les mails qui disent la même chose car ce sont des spams. L'antivirus vous aide toujours à lutter contre les menaces et cherche à garantir la sécurité de vos données [35, 74].

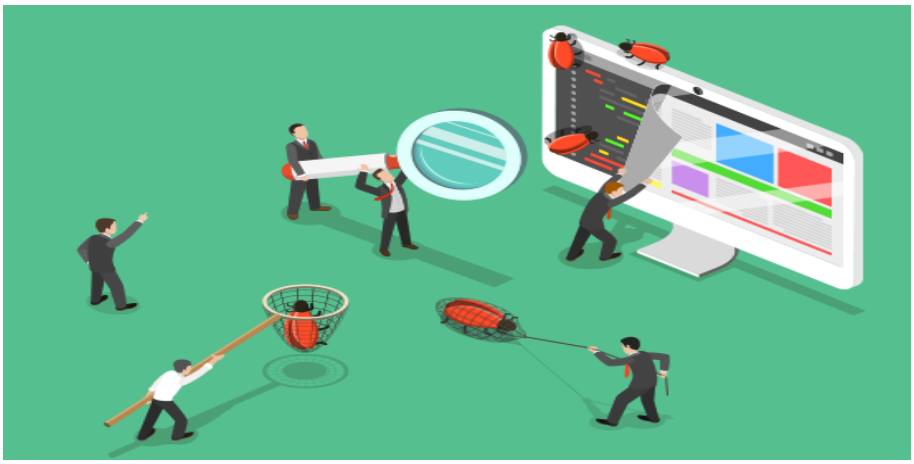


FIGURE 2.2: Anti-virus

2.4.2.2 Le pare-feu

Peut être logiciel ou matériel (ou une combinaison des deux), le par feu est pour but de contrôler le trafic de réseau quelque soit intérieur/extérieur selon une politique de sécurité [35].
Les pare-feu (en anglais firewall) sont les meilleurs mécanismes de défense des réseaux, ils font un mur entre les réseaux internes et les réseaux externes (tels qu'internet) pour détecter les attaques pendant la communication, ils sont situés à l'entrée du réseau et ils enregistrent les trafics précédents et construisent des journaux de sécurité. les pare-feu contrôlent tous type de communications entrantes ou sortantes pour lutter contre tous types de fraude, ils autorisent les communications légitimes et bloquent tout trafic identifié comme malveillant [35].
Les programmes malveillants et les intrusions sont des voleurs de données et d'identité, les utilisateurs s'exposent à ces dangers lorsqu'ils sont connectés à un réseau ou à internet. Les pare-feu protègent les réseaux et les ordinateurs et ils permettent d'assurer la sécurité des données [35].

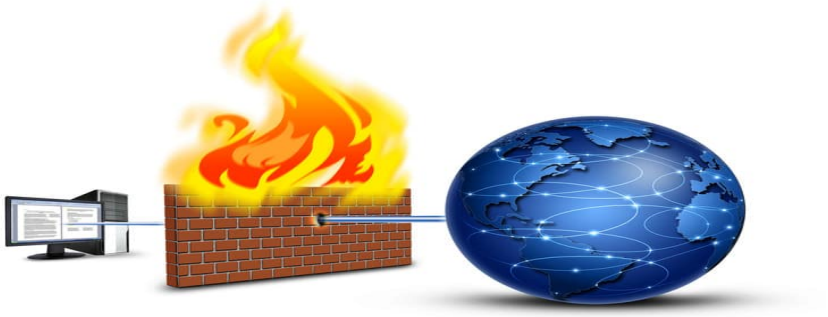


FIGURE 2.3: Pare-feu

2.4.2.3 La journalisation (Logs)

La journalisation (en anglais logging) est faire enregistrer chaque événement exécuté pendant le fonctionnement d'un système dans un fichier log, les événements exécutés sont sauvegardés de façon chronologique ce qu'ils permettent de faire des analyses statistiques sur le fonctionnement d'un système. Il existe deux types de journalisation, journalisation applicative et journalisation système [68].
Journalisation applicative enregistre tous les opérations et les événements pendant le fonctionnement d'une application, les événements sont enregistrés de façon chronologique par date et heure d'exécution [68].
Journalisation système enregistre tous les opérations et les événements pendant le fonctionnement du système, elle filtre et catégorise chaque tache par sa catégorie par exemple : information, avertissement...etc [68].

2.4.2.4 Système de détection d'intrusion (IDS)

IDS(Intrusion detection system) sont des techniques qui bloquent et arrêtent toute tentative de fraude dans un système informatique ou dans un réseau. Il existe les HIDS (host intrusion detection system) et les NIDS (network intrusion detection system) et les IDS Hybride constitués des HIDS et NIDS. Les HIDS sont des détecteurs qui analysent le bon fonctionnement des systèmes par le blocage des menaces qui cherchent à nuire la sécurité des données. Les NIDS sont des techniques qui analysent le trafic réseau par bloquer toute tentative d'intrusion. Les IDS hybride sont des mélanges des HIDS et les NIDS [106].
Avec le développement actuel des systèmes informatiques, les IDS classiques et les anciens modèles ne donnent pas un bon filtrage des actes malveillants, suite aux faiblesses des systèmes classiques, des nouvelles techniques ont été apparues représentées par des techniques du data mining et d'autres techniques heuristiques basées sur l'apprentissage automatique. Des nouvelles tendances dans les heuristiques ont donné la naissance à des techniques bio-inspiré qui sont des algorithmes inspirés de la nature, ces techniques se concentrent sur l'observation et l'analyse de la nature pour obtenir des modèles artificiels qui cherchent à trouver des solutions aux problèmes humains, l'analyse de la nature se fait par observer les formes et les procédés, les matériaux, les comportements et les interactions des êtres vivants. Toutes ces techniques ont été apparues pour lutter contre les menaces et augmenter le taux de précision pour une bonne détection des intrusions.
Dans notre travail, on a utilisé et adapté le comportement intelligent des abeilles social pour détecter les intrusions, les abeilles forment un système intelligent qui est difficile à casser, ce dernier peut être un bon système de défense contre les intrusions.

2.4.2.5 Le contrôle d'accès

Le contrôle d'accès est une technique qui limite les accès, on peut trouver deux types de contrôle d'accès : physique et logique. Le premier limite les accès aux salles qui contiennent un matériel sensible par exemple : l'accès aux salles informatiques, et le deuxième empêche et limite les accès aux réseaux, ou aux donnés dans un système informatique [35, 31]. Le contrôle d'accès bloque et gère les accès des utilisateurs dans un système informatique pour le protéger et garantir

la sécurité des données, il se fait selon des différentes identifications pour accéder aux ressources, chaque système à son propre technique de sécurité pour contrôler les accès aux données, chaque utilisateur doit s'authentifie afin d'approuver l'accès par un mot de passe, un code pin, un fichier physique, une empreinte, une détection faciale, un code à barres ...etc [35, 31]. Chaque accès se fait selon les tâches des utilisateurs et les services qui l'on est besoin. Selon la norme ISO27001, il est nécessaire de mettre une politique de sécurité pour contrôler les accès à un réseau informatique [35, 31].

FIGURE 2.4: L'accès a une salle via empreinte

2.4.2.6 Le bourrage de trafic

Un mécanisme de sécurité qui ajoute des données aux ressources du trafic, il envoie des données inutiles pour perturber les tentatives d'écoute et d'analyse sur le trafic transmis, de façon générale ce mécanisme de sécurité augmente le taux de confidentialité [35].

2.4.2.7 La notarisation

Une technique qui se base sur un ensemble des acteurs de confiance, lors d'un échange de données, les éléments de la communication sont enregistrés et envoyés vers les acteurs de confiance (la date, le contenu du message...etc), cette technique assure la non-répudiation [35].

2.4.2.8 L'horodatage

En anglais " Timestamping ", cette technique enregistre les traces d'échange de données, l'horodatage garde des enregistrements des étapes faites pendant les échanges par détails de la date et l'heure [35, 100].

2.4.2.9 Détection des SPAMS

Le courrier électronique est un service important d'échange de données pour les internautes, il est utilisé beaucoup dans li divers domaines, cette utilisation importante des courriers en a fait la cible de diverses perturbations telles qu'elles des attaques spams qui sont générées la plupart du temps de la publicité.

Le spam cible les victimes qui utilisent la messagerie électronique pour prendre des données sensibles, les spammeurs exploitent les faiblesses des victimes cibles via des mails qui contiennent des menaces. Il existe plusieurs types de spams, chacun à un but spécifique, par exemple un type de spam vise à alourdir le fonctionnement de l'appareil cible (ordinateur ou mobile de la victime) et le faire perdre le temps, d'autres types injectent des virus pour prendre une idée ou information, un autre type cible les entreprises pour voler de l'argent ou prendre une idée de production...etc.
Pour lutter contre les spams, il faut augmenter la sûreté, un utilisateur peut éviter les spams en négligeons de répondre aux messages douteux et d'éviter d'exécuter les pièces jointes dans ces messages. Un autre astuce de protéger et lutter contre les spams est d'éviter d'ajouter les e-mails personnels dans les sites sur internet qui nous demandent des informations confidentielles, un utilisateur peut créer des e-mails poubelles pour les utiliser dans les sites, et éviter de recevoir des spams aux boîtes officielles.
Le spam est un grand problème pour les utilisateurs de la messagerie électronique, beaucoup de détecteurs de spam ont été mises en service, et beaucoup de recherches ont été réalisées dans ce domaine et d'autres sont en cours de réalisation. Les détecteurs classiques sont limités et ne donnent pas des résultats satisfaisants, les chercheurs ont été orientés vers les méthodes heuristiques on se basent sur l'apprentissage automatique.
Des nouvelles heuristiques ont été apparues basées sur la nature, ces techniques ont prouvé qu'elles sont efficaces et performantes pour résoudre les problèmes humains.
Dans notre travail on a basé sur le système de défense de l'octopode pour lutter contre les attaques des prédateurs, le fonctionnement naturel de l'octopode a été adapté pour la détection des mails spams, l'approche est détaillée dans le quatrième chapitre.

FIGURE 2.5: Filtrage des spams

2.4.2.10 De-identification

Un concept de sécurité qui peut être défini comme tout processus qui permet de détecter, supprimer ou cacher les informations d'identification (les informations sensibles) dans une grande échelle de données pour garantir la confidentialité des données [57].
La fonction inverse de de-identification est appelée la re-identification, cette technique peut être définie comme un processus qui consiste à revenir à l'état original des données après l'étape de de-identification pour permettre au propriétaire de revenir à l'état initial en cas de besoin [57].

2.4.2.11 Cryptographie

Introduction
L'échange de données entre les gens est un processus important depuis le temps antique, le besoin de diffuser l'information de façon sécuritaire est aussi important pour garantir que l'information n'a pas été modifiée, supprimée ou visualisée par d'autres gens intrus.
L'histoire a commencé quand Jules César à envoyer des messages à ses généraux, il a essayé de sécuriser ses messages pour qu'ils soient illisibles au moment d'envoi, il a remplacé les lettres A par des D par un décalage de trois lettres par ordre alphabétique, les B par des E... etc. Pour toutes les lettres du message envoyé a fait le même décalage par un ordre de trois pour les déchiffrer [98].
Au moment de la deuxième guerre mondiale, Alain Turing a utilisé des techniques de substitutions et de permutations en utilisant des machines mécaniques (le chiffrement se fait automatiquement) comme la machine Enigma [54, 50].
Après aux années 70, les techniques de cryptographie étaient symétriques (à clé secrète), ensuite En 1976 W. Diffie et M. Hellman ont donné naissance au terme de chiffrement asymétrique (à clé public). Après en 1978, R. Rivest, A. Shamir et L. Adleman ont proposé le premier algorithme de chiffrement à clé publique qui est le RSA [50].

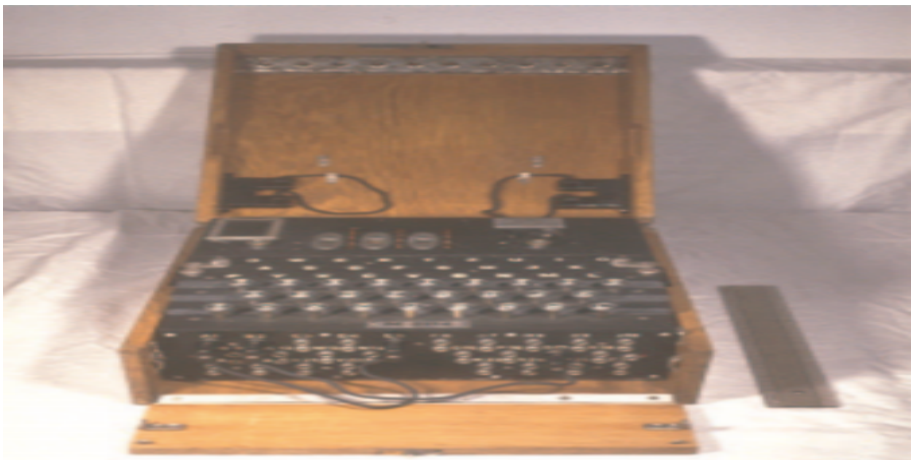


FIGURE 2.6: Machine enigma

La cryptographie cherche à chiffrer les messages pour qu'elles ne soient pas lisibles et compréhensibles que par l'émetteur et le récepteur, alors que deux parties font l'échange des messages dans un canal, une autre personne ne peut pas comprendre, modifier, supprimer ou lire les messages transmis dans ce canal [38]. Les données lisibles sont appelées **texte en clair**, la technique qui vise à rendre le texte en clair un message chiffré illisible est appelé **le cryptage, chiffrement ou codage**, le texte résultant du cryptage est appelé **texte chiffré ou cryptogramme**, la fonction inverse qui vise à remettre le texte à son état original est appelée **le décryptage** [98].

La cryptographie est le processus qui consiste à chiffrer un message pour le rendre incompréhensible pour d'autres intrus qui cherchent à espionner les données envoyées sur un canal de transmission pendant l'échange de données, sauf le destinataire peut lire le message transmis, ce processus de chiffrement utilise des fonctions mathématiques pour coder et crypter les messages transmis, le chiffrement est pour but de lutter contre les intrus et sécuriser les données [50, 38].

La cryptanalyse est un processus appliqué sur le texte chiffré pour trouver le secret de chiffrement, et déchiffrer les messages pour les rendre lisibles et compréhensibles, ce sont les pirates qui cryptanalyse les messages chiffrés et cassent les algorithmes de chiffrement [50, 38].

La cryptologie englobe et mélange les deux techniques la cryptographie et la cryptanalyse [50, 38].

Un cryptosystème peut être défini par toutes les notions de processus de chiffrement (texte clair, texte chiffré, l'algorithme qui fait le chiffrement, les clés ...etc) [43].

La cryptographie classique dans le début de la cryptographie des méthodes classiques ont vu le jour pour chiffrer les messages, on peut mentionner quelques-uns.

Substitution monoalphabétique le principe de ce type de chiffrement est que chaque lettre est remplacée par une autre lettre ou symbole, on peut mentionner quelques méthodes connues comme le chiffrement de césar et le chiffrement affine [43].
- Chiffrement de César : parmi les méthodes les plus connues et simples, son concept de chiffrement est basé sur un décalage se fait aux lettres des messages en se basant sur l'alphabet, par exemple on note "p" l'indice de la lettre de l'alphabet, "k" est le décalage a faire, la fonction de chiffrement est :

$$C = E(p) = (p + k) mod 26 \tag{2.1}$$

Pour le déchiffrement :

$$p = D(C) = (C - k) mod 26 \tag{2.2}$$

- La méthode de chiffrement de césar est facile à casser si la méthode de chiffrement est connue au pirate, elle a une faiblesse contre l'attaque d'analyse de fréquence

[43].
- Analyse de fréquence : cette technique se base sur le nombre d'occurrences d'une lettre dans le message avec le savoir de la langue utilisée, elle fait une analyse de fréquence des lettres pour déchiffrer les cryptogrammes, elle est faible dans le cas des messages courts [43].
- Pour lutter contre ce type d'attaque sur un cryptogramme, on peut chiffrer les messages par des diagrammes, trigrammes...etc. On peut aussi utiliser des homophones par remplacer une lettre par un symbole choisi au hasard et ne pas un symbole unique, on peut aussi coder chaque lettre par choisir un nombre des symboles égal à la fréquence d'apparition de la lettre dans le message [43].

Lettre	%	Chiffre
A	9	09 12 33 47 48 53 67 78 92
B	1	81
C	3	13 41 62
D	3	01 03 45
E	16	06 10 14 16 23 24 44 46 54 55 57 74 79 82 87 98
F	1	31
G	1	25
H	1	39
I	8	32 50 56 70 73 83 88 93
J	1	15
K	0	04
L	5	26 37 51 65 84
M	3	22 27 68

Lettre	%	Chiffre
N	7	18 58 59 66 71 91 99
O	5	00 05 07 54 72
P	3	38 90 95
Q	1	94
R	6	29 35 40 42 77 80
S	8	11 19 21 36 76 86 96 97
T	7	17 20 30 43 49 69 75
U	6	02 08 61 63 85 90
V	2	34 52
W	0	60
X	0	28
Y	0	24
Z	0	01

FIGURE 2.7: Exemple d'un table pour le chiffrement par homophones

- Chiffrement affine : une fonction est affine lorsqu'elle a une fonction linéaire (un polynôme de degré 1) de la forme x → a * x + b, le chiffrement affine utilise une fonction affine pour le cryptage [43].

$$y = (ax + b) mod 26 \tag{2.3}$$

a et b sont des constantes, x et y sont des nombres correspondant aux lettres de l'alphabet par exemple (A=0,B=1,...), si a = 1, alors on est dans le chiffrement de César où b est le décalage.
Si b=0 donc "a" est chiffré par "A" aucun décalage est détecté, donc aucune modification peut être faite aux messages, le message chiffré est lui même le message en clair.
-Chiffrement polygraphique (polygamique) : ce type de chiffrement substitue un groupe de lettres par un groupe de symboles comme le chiffrement de Playfair et le chiffrement de Hill [43].

Substitutions polyalphabétiques : on peut mentionner le chiffrement de Vigenère.
-Chiffrement de Vigenère : ce type de chiffrement est une amélioration de chiffrement de césar d'où il se base sur une clef de chiffrement pour faire le décalage

en utilisant le carré de Vigenère par exemple chiffrer le texte "**CHIFFRE DE VIGENERE**" avec la clef "**BACHELIER**" [43].

Clair	C	H	I	F	F	R	E	D	E	V	I	G	E	N	E	R	E
Clef	B	A	C	H	E	L	I	E	R	B	A	C	H	E	L	I	E
Décalage	1	0	2	7	4	11	8	4	17	1	0	2	7	4	11	8	4
Chiffré	D	H	K	M	J	C	M	H	V	W	I	I	L	R	P	Z	I

FIGURE 2.8: le carré de Vigenère

-ce type de chiffrement a été cryptanalyse par Kasiski et Friedman [43].
-Chiffrement de Vernam : ce type de chiffrement a le même principe que Vigenère mais dans ce type la clef de chiffrement a la même longueur que le message clair [43].

Chiffrement par transpositions ce type de chiffrement se base sur des permutations sur l'ordre des lettres, lorsque le message est plus grand, nous tombons sur un problème des nombres des permutations plus grands, dans ce type de chiffrement on peut trouver les transpositions rectangulaires en se basant sur un mot-clé et un tableau et des règles pour chiffrer les messages [43, 99].

Les concepts cryptographiques

Le chiffrement symétrique ou avec un autre sens le chiffrement à clé secrète, actuellement les chercheurs le nomment chiffrement conventionnel, ce type de chiffrement peut être considéré comme un processus de cryptage qui utilise la même clé secrète (privée) pour le cryptage et le décryptage, il est rapide, utile et praticable par plusieurs chercheurs pour chiffrer les gros volumes de données, dans ce type de chiffrement on trouve deux techniques de chiffrement, selon le mode d'opération : le chiffrement par bloc, et le chiffrement par flot [50, 38, 43].
- Le chiffrement par bloc : ce type de chiffrement divise le message en clair en blocs, il chiffre chaque bloc pour obtenir le cryptogramme final, on peut mentionner quatre modes de chiffrement en bloc selon le mode d'opération [38].
1- Le mode ECB (Electronic Code Book) : ce mode est le plus simple des modes, le chiffrement se fait par un découpage de message clair en bloc, chaque bloc résultant est codé de façon séparée, ce mode est rarement praticable, et s'expose à un risque de modification de bloc à l'intérieur sans la connaissance de l'émetteur et le destinataire [38].

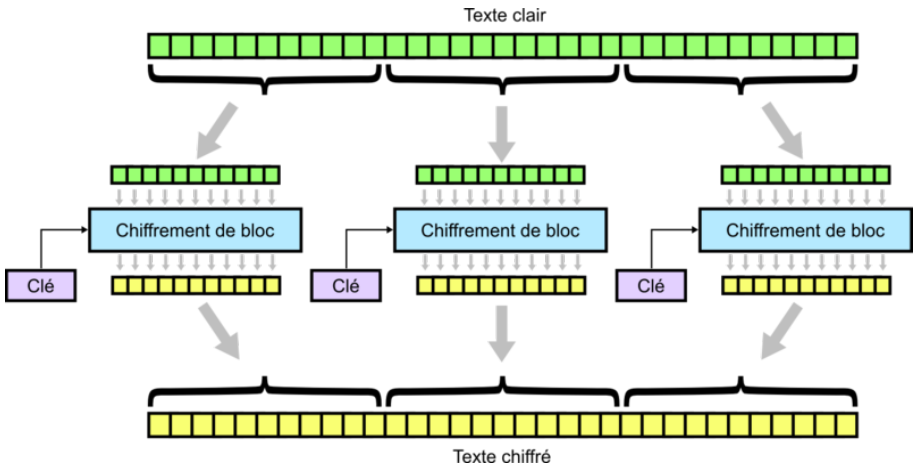


FIGURE 2.9: Le mode ECB

2- Le mode CBC (Cipher Block Chaining) : ce mode découpe le message clair en bloc, pour le cas de premier bloc, le mode CBC sélectionne un bloc initial (appelé aussi vecteur d'initialisation) de façon aléatoire, il applique un XOR avec le premier bloc de texte en clair à crypter, ensuite pour chaque bloc un XOR est appliqué entre le bloc de texte en clair et le bloc chiffré précédent. Ce mode est comme inconvénient que le chiffrement se fait de façon séquentielle, donc ça prend du temps pour le chiffrement d'un gros volume de données et pour voir des cryptogrammes dans un temps réel [38].

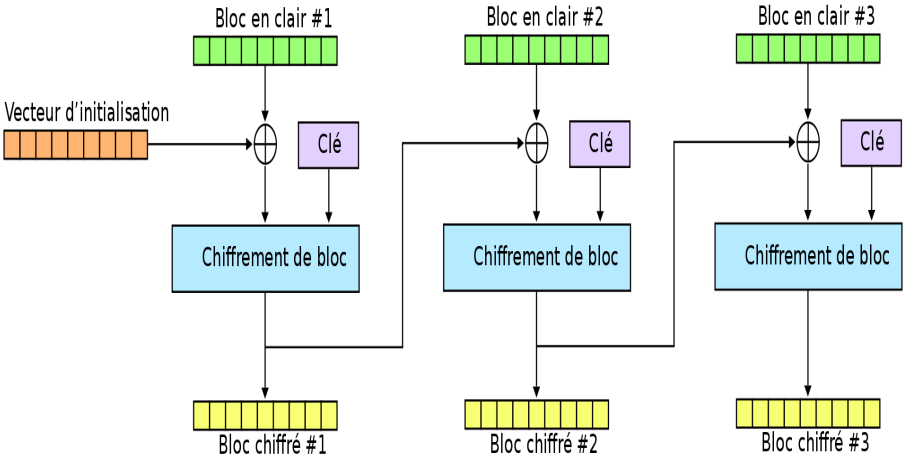


FIGURE 2.10: Le mode CBC

3- Le mode CFB (Cipher FeedBack) : ce mode a une orientation de chiffrement par flux, un ensemble de clés est généré, ils sont utilisés pour le chiffrement sur les blocs du message en clair, l'algorithme AES utilise ce mode pour le chiffrement [94].

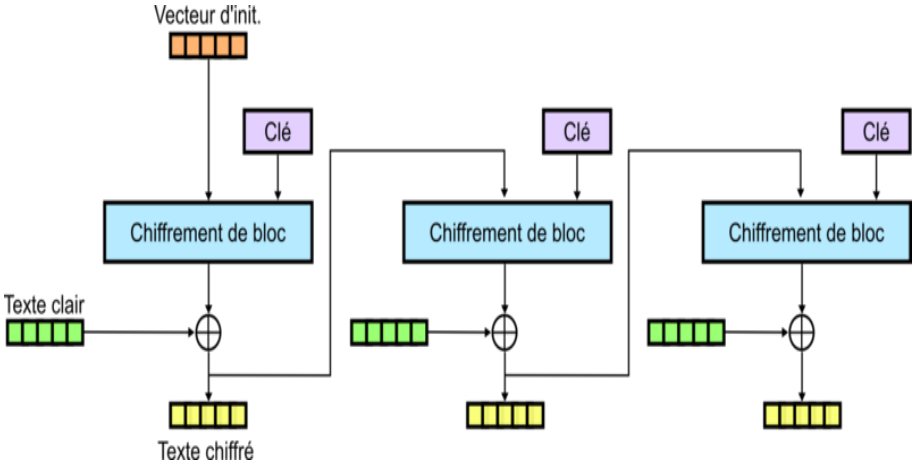


FIGURE 2.11: Le mode CFB

4- Le mode OFB (Output FeedBack) : ce mode a presque le même principe de chiffrement que le CFB, le seul changement dans ce mode, le flux de clé est obtenu en chiffrant le précédent flux de clé, ce mode a beaucoup de failles de sécurité [94].

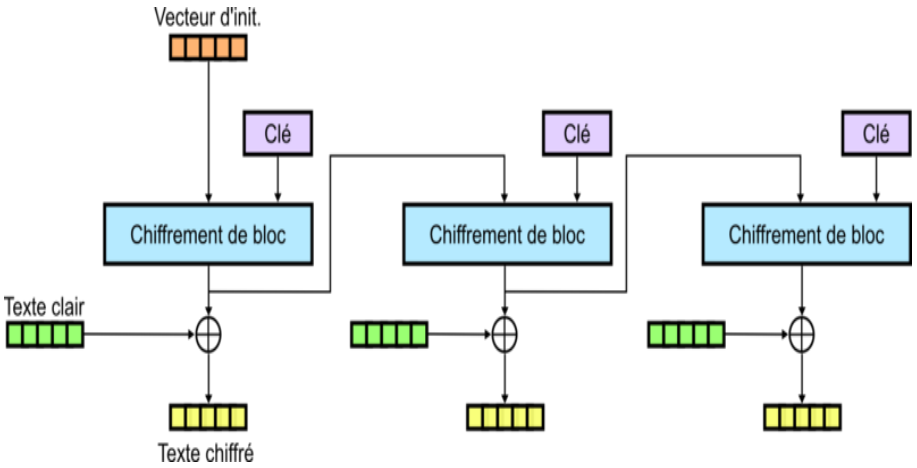


FIGURE 2.12: Le mode OFB

-Le chiffrement par flot (stream cipher) : Ce mode de chiffrement est un chiffrement symétrique selon le mode d'utilisation, il chiffre le message en clair bit par bit (ou octet par octet) par flux, il utilise des clés différentes pour chiffrer chaque bit, certains algorithmes utilisent des générateurs de clés, ce type de chiffrement effectue une opération " XOR " entre la clé et le message en clair pour obtenir le cryptogramme, l'algorithme RC4 utilise ce mode de chiffrement [140, 13].

Le chiffrement asymétrique en 1976 dans l'université de stanford, Whitfield Diffie et Martin Hellman ont proposé un nouveau paradigme de chiffrement en se basant sur des clés publiques ou chiffrement asymétrique, ce type de chiffrement utilise deux clés pour crypter et décrypter les messages, une clé partagée appelée clé publique et une clé secrète appelée clé privée [36].

L'expéditeur utilise la clé publique pour chiffrer les messages, et le destinataire déchiffre les cryptogrammes en utilisant la clé privée, la clé privée est une clé secrète qui ne doit pas être communiquée ou partagée, ce type de chiffrement asymétrique se base sur les fonctions à sens unique qui sont difficiles à inverser sauf si on a une information particulière nommée la clé secrète [38, 36].
Ce type de chiffrement s'expose aux risques des attaques de l'homme au milieu qui permet aux intrus de changer les clés publiques échangées entre l'émetteur et le récepteur du message, parmi les algorithmes de ce type de chiffrement connu on trouve le RSA proposé par Ronald Rivest, Adi Shamir et Leonard Adleman en 1978, les algorithmes de ce type de chiffrement asymétrique sont souvent appliquées dans la signature numérique [38, 36].

Fonction de hachage est une fonction qui prend en entrée des données de tailles quelconque et produit une chaîne de taille fixe et réduite, ces fonctions sont à sens unique, le résultat d'application d'une fonction de hachage est appelée **haché** de données, **empreinte** ou encore **un condensé** de ces données [38, 120].
Ces fonctions de hachage sont utilisées beaucoup dans l'authentification et la signature numérique, ils sont rapides à calculer et donnent des condensés de taille réduite et fixe, ils résistent aux collisions. Il est impossible de trouver le même haché sur des données différentes, ces fonctions disposent qu'elles sont à sens unique, donc il est difficile à inverser la fonction de hachage pour retourner à l'état original [38, 120].
On peut trouver des fonctions de hachage sans clé et d'autres avec clé, dans le premier cas la fonction de hachage se comporte de façon aléatoire, et dans le deuxième cas la fonction de hachage se base sur la clé pour hacher le message en clair, on peut mentionner quelques fonctions de hachage utilisées comme MD4 et MD5, SHA-1 et SHA-2 ...etc [38, 120, 112].

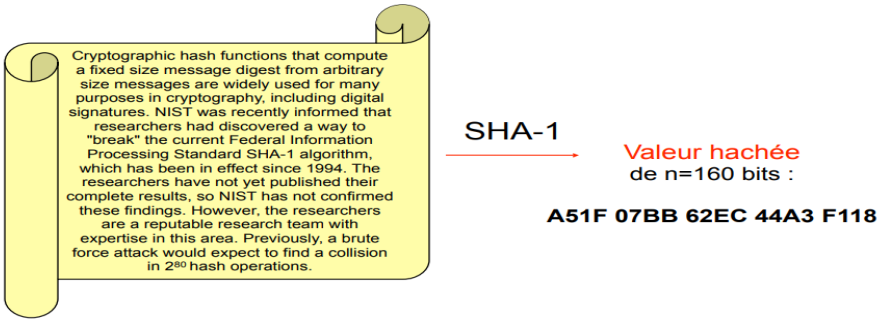


FIGURE 2.13: Exemple d"une fonction de hachage "SHA-1"

Signature numérique est valide après le hachage des messages, l'émetteur signe le condensé, il existe plusieurs mécanismes et techniques de signature numérique, généralement le condensé est signé par une clé privée, et le récepteur vérifie le message signé avec une clé publique partagée par l'émetteur (ou le signataire), les clés privées et publiques sont générées par le signataire du message [53].
La signature numérique n'a pas le même sens que la signature manuscrite, la der-

nière valide l'identité d'une personne, mais la première est ajoutée au contenu de chaque message, et elle est différente pour chaque un, la signature numérique vérifie l'authenticité et l'intégrité des données [53].

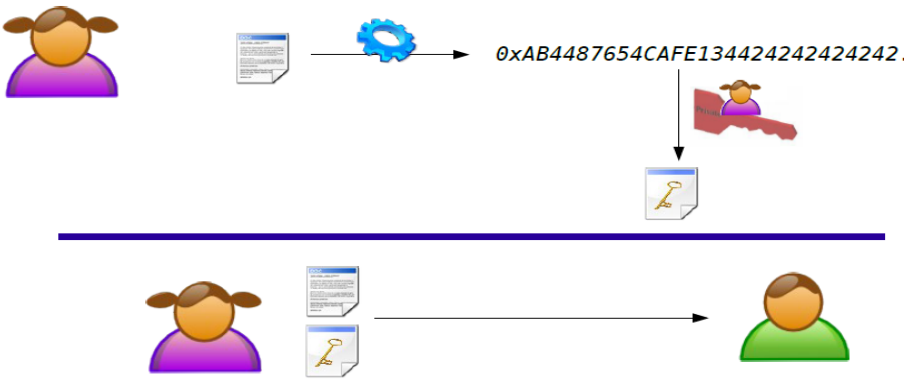


FIGURE 2.14: Exemple d'une signature numérique

2.5 Conclusion

Suite au développement rapide des services informatisés et sa nécessité dans les entreprises et les communications, et avec l'utilisation d'internet et la rapidité d'accès aux données via des différentes distances quelque soit près ou loin, la sécurité des données dans la communication pendant l'échange de données est un objectif important à atteindre.
La naissance des nouveaux services et la multitude de ces derniers et sa contribution forte dans la vie des gens ont été accompagnées par l'émergence des programmes malveillants qui cherchent à voler les données, nous avons présenté dans ce chapitre qu'elles sont les risques et les différents types d'attaques qui cherchent à nuire la sécurité des données, ensuite on a élaboré les problèmes techniques de la sécurité des données, on a vu que pour garantir une bonne protection de données et pour lutter contre les actes malveillants et tous type d'attaque qui veulent compromettre la sécurité il faut une précaution suivie par une prévention, cela est fourni par une bonne politique de sécurité qui cherche à garantir le plus grand nombre de services de sécurité possible, via des mécanismes de sécurité, dans la fin de ce chapitre on a cité quelques mécanismes de sécurité informatique connue.

3

DATA MINING, META HEURISTIQUE ET BIO-INSPIRATION

Table des matières

3.1 Introduction

Actuellement les données explosent dans le web, et ne cessent pas à croître jour après jour de façon exponentielle et rapide, et avec des différents types (quelque soit texte, image ou vidéo...etc). Le contrôle de ce grand volume de données est important et essentiel.
Le contrôle de ces données d'une part est essentiel mais faire extraire de la connaissance automatiquement à partir de ce grand volume est aussi important d'autre part. Les chercheurs veulent résoudre des problèmes tels qu'elles de la recherche

d'informations, ils veulent prendre des décisions de façon automatique, ceci peut-être résolu en utilisant l'apprentissage machine par faire apprendre à la machine pour qu'elle soit prête pour prendre des décisions. De façon générale les experts cherchent à analyser la masse de donnés gigantesque de façon automatique pour extraire de la connaissance, et prendre des décisions qui aident à trouver des solutions aux différents problèmes.
Avec les technologies adoptés qui ont apparu dans le monde tel que le commerce électronique, le stockage dans les nuages...etc. Les chercheurs ont été basé sur l'apprentissage machine en utilisant des algorithmes de fouille de données, des heuristiques et même des nouvelles techniques appelées des techniques bio-inspirées basées sur le comportement naturel pour résoudraient leurs problèmes et trouver des décisions, et même pour acquérir de la connaissance.
Le fouille de données ou **data mining** rassemble l'ensemble des techniques qui visent à rechercher et extraire de la connaissance à partir d'une masse de données qui se trouve dans des entrepôts de données, le développement des méthodes et des techniques de fouille de données est né suite à un ensemble de critères, parmi les sont : l'explosion gigantesque des données et l'évolution du matériel telle que les ordinateurs robustes de haute performance, le stockage dans les nuages, l'amélioration des protocoles de communications réseaux, la facilité d'échanges de données, le commerce électronique...etc.
Actuellement, la fouille de données (data Mining) a un impact énorme sur l'économie grâce aux techniques développées, beaucoup des prévisions ont été faites et des décisions ont été prises, et des améliorations bénéfiques aux propriétaires ont vu le jour. Le processus de la fouille de données (data Mining) se fait via un ensemble des étapes qu'on va les citer dans la suite, ce processus est nommé le processus d'extraction de la connaissance (en anglais : Knowledge data discovery).

3.2 Data Mining

Fouilles de données ou **forage de données**, en anglais **Data Mining** ont été apparu au milieu des années 1990 aux États-Unis, il rassemble l'ensemble de méthodes et des techniques d'analyse de grands volumes de données stockées dans des entrepôts de données afin de rechercher et extraire de la connaissance qui peut être utilisée par des entreprises pour faire des améliorations en réduction des coûts [111].
En 1996, fayyad et al. ont défini le data mining comme un processus non-trivial d'identification de structures inconnues, valides et potentiellement exploitables dans les bases de données [7]. Les techniques de data mining sont utilisées dans plusieurs domaines, par exemple : dans la détection des fraudes, dans la prévision des consommations énergétique, dans le traitement automatique du langage, dans la reconnaissance vocale et faciale, dans les traitements médicaux...etc [66].
Les méthodes de fouille de données sont des techniques analytiques de données, ces méthodes ont vu le jour suite à l'explosion quantitative et qualitative de données d'une part, et le développement du matériel (les ordinateurs) et la rapidité de calcul et de traitement de l'information d'autre part. Les techniques de data mining permettent de classifier les données et aident à prendre des décisions de

façon automatique, elles permettent aussi de chercher à trouver des relations entre les données et catégoriser les données...etc. Ces techniques ont facilité beaucoup des tâches aux gens et ont optimisé la gestion des ressources [111].

3.2.1 Les taches de data-mining

Les méthodes de data mining ont de nombreuses tâches :

3.2.1.1 Classification

Les techniques de data mining cherchent à organiser les données en utilisant l'AA d'où il existe une base d'apprentissage faite par un superviseur qui aide la machine à classifier les données [111, 66].

3.2.1.2 Segmentation

Parmi les taches du data mining est la segmentation, d'où les méthodes du data mining regroupent les données dans des groupes (ou clusters), ce type est considéré comme un type d'apprentissage non-supervisé [111, 66].

3.2.1.3 Association

L'objectif de cette tâche est de découvrir les éléments (ou les événements) liés, elle est utilisée beaucoup dans le marketing par trouver les associations entre les produits achetés ensemble [111, 66].

3.2.1.4 Prédiction

Il s'agit de découvrir et trouver des prédictions raisonnables dans le futur, c'est a dire à partir d'un ensemble de données on peut prédire et trouver des conclusions [111, 66].

3.2.2 Processus d'extraction de connaissance (Knowledge data discovery)

L'extraction de la connaissance avec les techniques du data minig se fait via un processus appelé **le processus d'extraction de connaissance, en anglais KDD (Knowledge data discovery)**. Le data minig est considéré comme une étape importante dans ce processus, ce dernier est réalisé en passant par plusieurs étapes, tous d'abord il faut connaître le domaine d'application et déterminer les objectifs souhaités, ensuite une étape de sélection d'un ensemble de données d'où on peut utiliser une méthode d'échantillonnage de données, une étape suivante qui fait un pré-traitement sur ces données (un nettoyage) pour le cas des données qui ne sont pas structurées comme les données textuelles, les images...etc. Le pré-traitement se fait soit par une conversion, ou par une suppression des données manquantes, ou par une réduction de dimension, ou par l'utilisation d'une méthode de sélection des attributs, selon le cas des données que nous traitons, ensuite une transformation de données vers un langage structuré compréhensible par l'algorithme du data minig est faite, puis une étape qui applique une technique du data mining, dans cette étape il faut bien choisir l'algorithme selon la tache à atteindre, si on veut une classification ou une segmentation, une régression ou

une association. Lorsqu'on termine d'appliquer la méthode du data minig nous passons à l'étape de visualisation des résultats par des histogrammes, ensuite une étape qui cherche à évaluer et interpréter les résultats obtenus par des mesures de validation, enfin l'étape dernière d'extraction de la connaissance, et on peut la utiliser et la mette en disposition [111, 128].

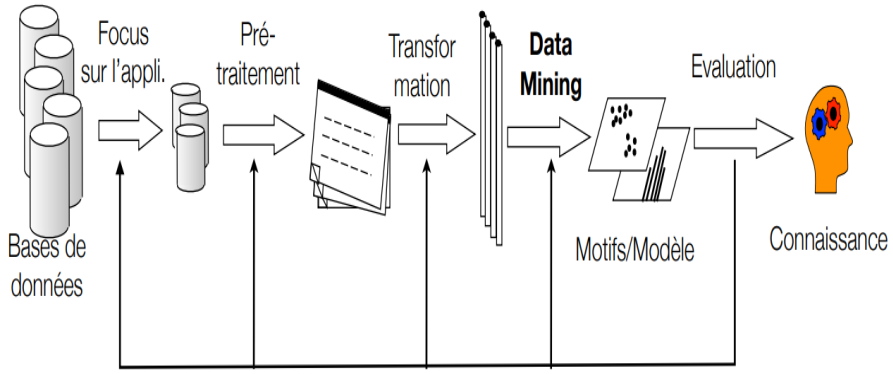


FIGURE 3.1: Processus d'extraction de connaissance

3.3 Apprentissage automatique

En anglais ***Machine Learning***, l'apprentissage automatique est une petite partie du domaine de l'intelligence artificielle, l'IA rassemble l'ensemble des systèmes et des machines qui cherchent à imiter l'intelligence humaine. L'AA rassemble l'ensemble des méthodes et des algorithmes qui enseignent automatiquement la machine pour réaliser une certaine tâche, ou pour résoudre un certain problème que nous ne connaissions pas de solution exacte [108, 103].
L'AA touche d'autres disciplines tel que la statistique, la probabilité...etc. Plusieurs domaines utilisent l'apprentissage pour réaliser un ou plusieures tâches de façon automatique, par exemple pour optimiser des modèles et des résultats, ou pour une détection d'intrusion, ou pour un traitement automatique de la langue...etc [108, 103].
Plusieurs grandes entreprises utilisent l'AA, par exemple ***Facebook*** utilise des algorithmes de reconnaissance faciale pour faciliter l'étiquetage d'un ami dans une photo que vous partagiez. L'entreprise de paiement électronique ***Paypal*** aussi utilise l'apprentissage automatique pour détecter les fraudes et protéger ses données. Le système de reconnaissance vocale de l'Iphone appelé ***"Siri"*** est l'un des meilleurs systèmes vocaux, ce système d'assistant virtuel utilise des algorithmes d'AA, c'est le même cas pour l'assistante virtuelle de Windows ***"Cortana"*** .
La grande entreprise ***"Google"*** utilise l'AA dans tous ses services tels que la recherche, la navigation dans internet, la traduction automatique, la navigation via Google maps...etc. Et plusieurs d'autres grandes entreprises qu'on ne peut pas les citer toutes qui utilisent l'AA dans ses services [125, 119].
Il existe plusieurs types d'AA, on peut trouver l'apprentissage supervisé, l'ap-

prentissage non-supervisé, l'apprentissage semi-supervisé et l'apprentissage par renforcement [108].

3.3.1 Apprentissage supervisé

L'apprentissage supervisé est un type d'AA d'où les algorithmes de ce type basent sur un ensemble des exemples étiquetés par un superviseur pour prédire la classe de chaque échantillon teste, et ensuite les algorithmes catégorisent les nouveaux exemples non-étiquetés. L'algorithme apprend de chaque exemple de la base d'apprentissage pour obtenir une bonne décision de façon automatique. Ce type d'apprentissage est le plus utilisé, et il existe plusieurs algorithmes de ce type comme la régression logistique et linéaire, Naïve bayes, SVM, KNN ...etc [108, 103].
Ce type d'apprentissage est utilisé pour résoudre beaucoup de problèmes, comme la détection des spams, la détection d'intrusion, la reconnaissance vocale, la détection des fraudes, la détection de plagiat...etc [108, 103].
Le choix d'un algorithme d'apprentissage se fait selon des critères de sortie, si la sortie de l'algorithme que l'on veut soit une valeur du type continu (nombre), on utilise un algorithme de régression, sinon si on veut une sortie de valeur du type discret (une catégorie) on utilise un algorithme de classification [125].

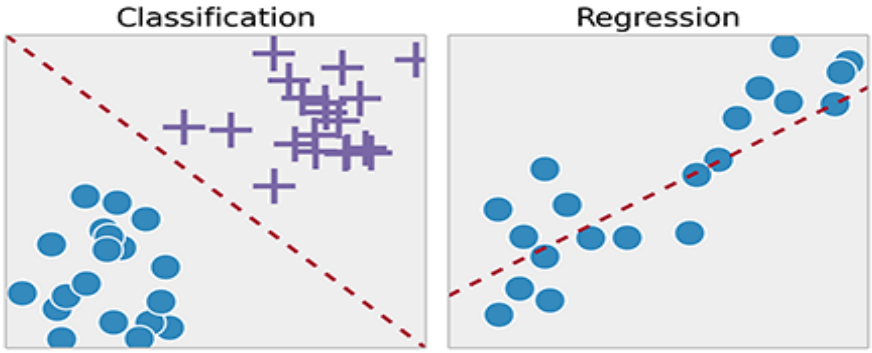


FIGURE 3.2: Exemple d'une classification et une régression

3.3.1.1 Mesures d'évaluation

Il existe plusieurs mesures qui permettent d'évaluer la qualité des résultats prédites par un algorithme qui se basent sur un processus d'apprentissage supervisé, on peut trouver l'erreur de classeur, le taux de succès, la précision, le rappel, la f-mesure, l'entropie...etc [86, 11].
Dans le cas de la classification binaire d'où la classe de sortie d'algorithme prend deux valeurs, le classeur peut construire une matrice de confusion en se basant sur les valeurs des VP, VN, FP, FN [86, 11].
VP : Le nombre de vrais positifs. Cette valeur incrémente quand les exemples de la classe sont positifs affectés par un superviseur, et dont la classe prédite par le classeur est positive [86, 11].
VN : Le nombre de vrais négatifs. Cette valeur incrémente quand les exemples de la classe sont négatifs affectés par un superviseur, et dont la classe prédite par

		Classeur	
		+	-
Superviseur	+	VP	FN
	-	FP	VN

TABLE 3.1: Matrice de confusion

le classeur est négative [86, 11].
FP : Cette valeur incrémente quand les exemples de la classe sont négatifs affectés par un superviseur, et dont la classe est prédite par le classeur est positive [86, 11].
FN : Cette valeur incrémente quand les exemples de la classe sont positifs affectés par un superviseur, et dont la classe prédite par le classeur est négative [86, 11].

La précision Cette mesure peut être calculée pour la classe des positifs, et même pour la classe des négatifs, elle permet de calculer les prédictions justes de l'algorithme. D'une autre façon, elle permet de calculer combien d'instances classées correctes par l'algorithme, sa formule pour les cas des positifs est [86, 11] :

$$Précision_{(P)} = VP/(VP + FP) \tag{3.1}$$

- Sa formule pour les cas des négatifs est [86, 11] :

$$Précision_{(N)} = VN/(VN + FN) \tag{3.2}$$

Le rappel Cette mesure aussi peut être calculée pour le cas des positifs, et même pour les négatifs, pour le cas de la classe positive, le rappel permet de calculer le taux des vrais positifs à partir de la base d'exemples des positifs de ceux qui existent, combien l'algorithme a trouvé de classes positives, d'une autre façon il mesure la capacité d'un classeur de détecter les instances correctement classées, sa formule peut être définie comme [86, 11] :

$$rappel_{(P)} = VP/(VP + FN) \tag{3.3}$$

- Pour le cas de la classe négative, le rappel permet de calculer le taux des vrais négatifs, c'est t'a dire à partir de la base d'exemples des négatifs de ceux qui existent, combien l'algorithme a trouvé des classes négatives, sa formule peut être définie comme [86, 11] :

$$rappel_{(N)} = VN/(VN + FP) \tag{3.4}$$

FN-rate En anglais False negative rate, sa formule peut être définie comme :

$$FN - rate = FN/(FN + VP) \tag{3.5}$$

FP-rate En anglais False positive rate, sa formule peut être définie comme :

$$FP - rate = FP/(FP + VN) \tag{3.6}$$

VP-rate En anglais True positive rate, sa formule peut être définie comme :

$$VP - rate = VP/(VP + FN) \tag{3.7}$$

F-mesure

$$F - mesure = 2 * rappel * précision/(rappel + précision) \tag{3.8}$$

Entropie c'est la perte d'information, elle se calcule par la formule au-dessous.

$$E = -log(précision) \tag{3.9}$$

La courbe ROC En anglais Receiver Operating Characteristic, elle est une courbe qui dessine l'évolution du taux des vrais positifs (True positive rate) en fonction du taux des faux positifs (False positive rate) [11].

3.3.2 Apprentissage non-supervisé

L'apprentissage non supervisé est un type d'apprentissage d'où la machine apprend sur un ensemble de données sans connaître la classe de sortie de ces dernières, avec un autre sens sans la présence d'un superviseur, l'algorithme fonctionne de façon autonome sur la base de ses règles et prend une décision et prédit des résultats en sortie [52].
Parmi les taches les plus connues de ce type d'apprentissage sont la segmentation (le regroupement en anglais clusternig) et les règles d'association et la réduction de dimensionnalité [52].
La segmentation consiste à regrouper un ensemble de données hétérogènes en des groupes homogènes qui ont des caractéristiques communes en fonction de similarité, on peut pratiquer ce type d'apprentissage comme un système de recommandation ou comme un détecteur des intrus dans la cybersécurité ou dans un robot [52].
Les règles associations permettent d'analyser les relations et les liens entre les variables, et trouver les associations entre les données [52].
La réduction de dimension cherche à diminuer le nombre d'attributs dans une base de données structurée, en supprimant les attributs qui peuvent être négligeables et ne contiennent pas des valeurs influentes [52].

3.3.3 Apprentissage semi-supervisé

L'apprentissage semi-supervisé est un mélange de deux types d'apprentissage supervisé et non-supervisé, ce type utilise un petit ensemble de données étiquetées par un superviseur avec un grand ensemble de données non-étiquetées, il est praticable dans le cas où l'étiquetage de données est coûteux, généralement ce type d'apprentissage est utilisé pour résoudre les problèmes de reconnaissance des formes [59, 115].

3.3.4 Apprentissage par renforcement

L'apprentissage par renforcement permet à une machine, robot ou à un algorithme de choisir une action de façon autonome sans besoin de faire l'exploitation

de tous les cas, l'algorithme dans un environnement à un état T cherche à produire une action pour aller à un état successif T+1, la production d'une action à un état quelconque est appelée une politique, à chaque action produite à un état donné une récompense est générée, cette récompense peut être positive ou négative, les récompenses sont cumulées à long terme, un algorithme de ce type d'apprentissage utilise une fonction d'évaluation (appelée aussi fonction d'utilité), cette fonction cherche à optimiser la décision à prendre dans l'état T en résumant une espérance de gain, cette fonction d'utilité apprend par les récompenses précédentes. Il existe deux types des fonctions d'utilités, la première inclut la valeur d'un état sous une politique, et la deuxième inclut la valeur d'une action dans un état sous une politique. L'algorithme ou le robot cherche à prendre une décision optimale sur la base des récompenses des actions itérées précédemment en maximisant la somme des récompenses, par exemple en peut citer quelques algorithmes de ce type d'apprentissage : le TD-learning (temporal difference) et le Q-learning (quality) [33].

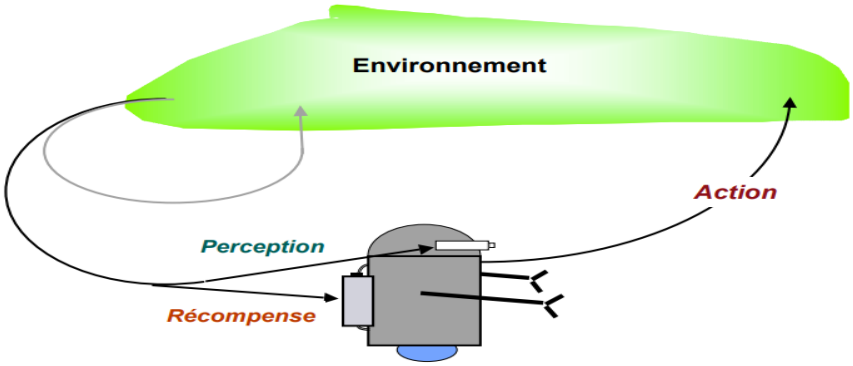


FIGURE 3.3: Schéma générale d'apprentissage par renforcement

Ce type d'apprentissage est inspiré de la psychologie comportementale naturelle, il est utilisé dans plusieurs domaines tels que la recherche opérationnelle, l'optimisation combinatoire, les systèmes multi-agents, la robotique, les algorithmes génétiques ...etc. L'agent qui utilise l'apprentissage par renforcement cherche à obtenir un comportement décisionnel et améliore les actions à prendre dans le futur à la base des récompenses générées à chaque état visité [82].

3.3.5 Apprentissage profond

En anglais deep learning, l'apprentissage profond est un type où la machine (le robot) s'apprend à prendre une décision en utilisant l'apprentissage automatique (machine Learning) [34].
L'apprentissage profond est une discipline d'intelligence artificielle qui se base sur le fonctionnement des réseaux de neurones biologiques [34].
Le premier neurone formel est apparu en 1943 par les travaux de Warren McCulloch et Walter Pitts, après, le perceptron est inventé par F. Rosenblatt en 1957, en 1986 les perceptrons multicouches (MLP) ont vu le jour, récemment avec le développement des matériels qui ne cesse d'augmenter, et grâce à la haute perfor-

mance de traitement des données en matière de calcul et de stockage. L'apprentissage profond a vu une grande révolution en matière d'utilisation, ces algorithmes d'apprentissage profond sont utilisées dans plusieurs domaines tels que la détection des fraudes, la reconnaissance vocale (Siri, Cortana et Assistant Google), la reconnaissance faciale comme l'utilise Facebook...etc [34].
Un neurone biologique reçoit des signaux transmis par d'autres neurones comme entrées, au niveau du corps cellulaire le neurone traite ces signaux, si les résultats obtenus sont supérieurs à un seuil d'activation, le neurone renvoie un potentiel d'action par son axone vers d'autres neurones [65].
Les réseaux de neurones artificiels sont inspirés des neurones biologiques naturels, les synapses biologiques sont modélisées par des poids, le corps cellulaire est représenté par la fonction de transfert (fonction d'activation), l'axone est représenté par la sortie, pour mieux comprendre voir la figure 3.4 [65].

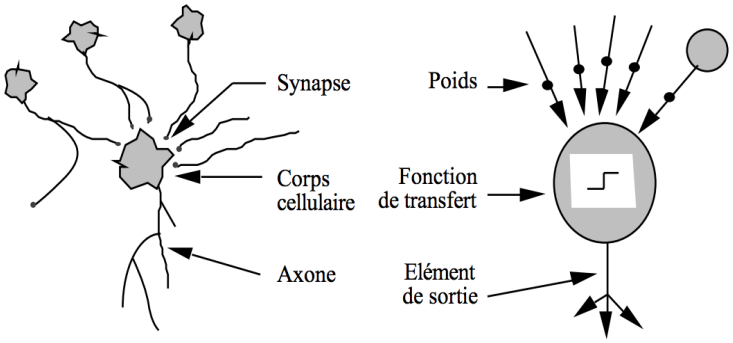


FIGURE 3.4: Neurone biologique et neurone artificiel

Chaque neurone artificiel possède des données comme entrées, les entrées possèdent des poids, une fonction d'activation est appliquée pour générer la sortie. Avec la naissance des perceptrons multicouches le principe des couches a vu le jour, une couche d'entrée qui contient les entrées, un ou plusieures couches cachées qui sont au milieu du perceptron, et une couche de sortie qui contient les sorties du perceptron. Le MLP utilise une rétro-propagation du gradient de l'erreur, cette technique modifie les poids pour diminuer l'erreur et optimiser la sortie, pour mieux comprendre voir la figure 3.5 [65].

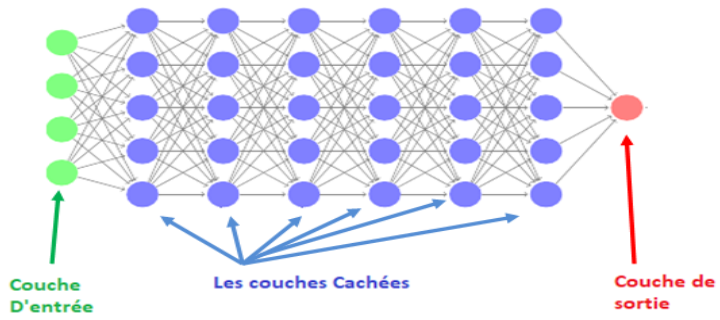


FIGURE 3.5: Perceptron Multicouche (MLP)

Les experts et les chercheurs ont vu que l'ajout des couches cachées des neurones artificiels améliore les résultats obtenus, et que les réseaux de neurones artificiels qui contiennent un grand nombre des couches cachées donnent des résultats meilleurs [34]. Les réseaux de neurones peuvent utiliser l'apprentissage superviser, non-superviser ou par renforcement. Dans le premier cas d'apprentissage superviser, l'algorithme neuronal s'apprend à partir d'un ensemble de données étiquetées par un expert (superviseur) pour obtenir les résultats finaux. Dans le deuxième cas d'apprentissage non-superviser, l'algorithme s'apprend de façon autonome des données qui ne sont pas étiquetées, et génère les résultats. Dans le troisième cas d'apprentissage par renforcement, le réseau de neurones renforce ses connaissances par les récompenses positives, et ignore les mauvais résultats, l'algorithme s'apprend par le temps en termes des récompenses cumulées [34].
Il existe plusieurs types de réseaux de neurones, on peut mentionner quelques-uns : **les réseaux de neurones feed-forward**, **les réseaux de neurones récurrents** et **les réseaux de neurones convolutifs**.

3.4 Les heuristiques et les Méta-heuristiques

3.4.1 introduction

Le progrès technologique et l'avancement en matière de calcule et de stockage de données n'ont pas résolus tous les problèmes humains, il existe toujours des problèmes difficiles à résoudre qui sont définis par des problèmes NP-complets (Non déterministe Polynomial). Le temps nécessaire pour résoudre ces problèmes en utilisant n'importe quel algorithme couramment connu augmente rapidement avec la taille du problème [141, 122].
Il existe des problèmes d'optimisation combinatoire peuvent être résolus par des méthodes exactes, ces méthodes exploitent l'espace de recherche complète pour obtient la meilleure solution, ces problèmes sont appelés des problèmes déterministes de la classe P (Polynomial), mais ces méthodes exactes pour certains problèmes prennent beaucoup de temps pour obtenir une réponse[141, 122].
D'autres problèmes sont résolus par des heuristiques et des méthodes approchées en explorant qu'une partie de l'espace de recherche pour trouver une solution approchée, cette dernière peut ne pas être la solution optimale, on peut trouver d'autres solutions optimales dans un délai de temps réduit et raisonnable [141, 122].
Les spécialistes d'optimisation combinatoire ont orienté leurs études vers les méthodes heuristiques et les algorithmes d'approximation pour traiter les problèmes np-complet [141, 122].
Les heuristiques cherchent à trouver une meilleure solution dans un ensemble des solutions réalisables d'un problème, elles vous ferez résoudre un problème d'optimisation combinatoire mais pas de façon exacte. La solution désirée comme meilleure est celle qui minimise ou maximise la fonction objective introduite dans le problème, dans un problème donné on peut trouver plusieurs solutions optimales [141, 122].

3.4.2 Les heuristiques

Le terme heuristique vient du mot grec **heuriskein** qui veut dire **"trouver"**, une heuristique est une technique qui cherche à résoudre un problème en exploitant ce dernier pour trouver une solution acceptable dans un temps de calcul réduit. Les heuristiques sont faciles à combiner avec d'autres méthodes ce qui les rend praticables beaucoup. Une heuristique peut trouver une solution acceptable, optimale ou approchée à la solution optimale mais elle ne garantit pas la solution obtenue. La solution optimale dans une heuristique cherche à maximiser ou minimiser la fonction objective d'algorithmes utilisés selon le problème reconnu [141, 122, 41]. Les heuristiques peuvent être classées en deux grands groupes, le premier contient **les techniques constructives** et le deuxième type contient **les techniques de fouilles locales**. les techniques de la première classe se basent sur une solution initiale et cherchent à trouver une solution finale complète en ajoutant des éléments dans chaque étape, ce type de technique est rapide et simple. Les techniques de la deuxième classe prennent une solution initiale complète et essayent de l'améliorer en explorant ses voisinages [122].

3.4.3 les Méta-heuristiques

Le terme méta-heuristique est composé de deux mots grecs **méta** veut dire **au delà** et **heuriskein** veut dire **trouver**, une méta-heuristique peut être définie comme un modèle générique qui cherche à obtenir une solution pour plusieurs problèmes différents, les heuristiques sont appliquées à des problèmes spécifiques mais les méta-heuristiques peuvent être adaptées et appliquées à divers problèmes [141, 122, 41].
Il existe plusieurs méthodes pour classifier et catégoriser les méta-heuristiques, la classification la plus connue est de catégoriser les méta-heuristiques dans quatre classes, la première classe pour les techniques constructives, la deuxième contient les techniques de recherche locale, la troisième pour les approches évolutives et la dernière comprend les méthodes hybrides [141].

3.4.3.1 les méthodes constructives

Ces méthodes se basent sur une solution initiale vide et insèrent des éléments à chaque étape pour arriver à une solution finale complète, donnant un exemple du problème du voyageur de commerce, ce dernier peut être résolu et illustré avec une méthode heuristique constructive, un voyageur doit visiter plusieurs clients, il fait la visite client par client, en se partant vers le client le plus proche dans chaque étape, jusqu'à qu'il visite tous les clients, après, il rentre chez lui [141].

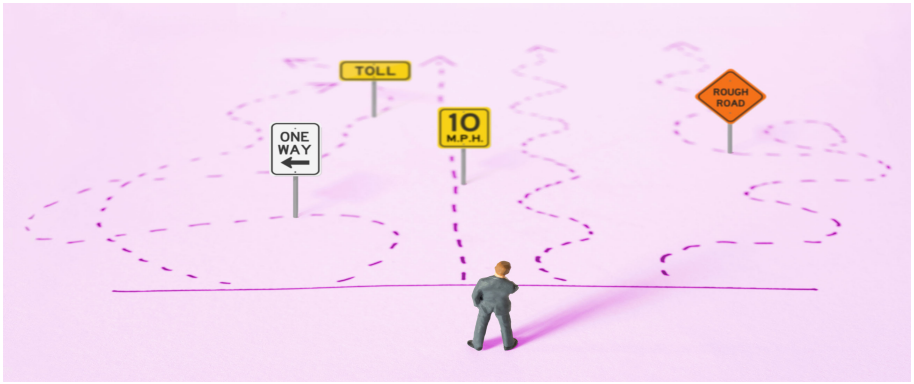


FIGURE 3.6: Voyageur de commerce

3.4.3.2 les méthodes de recherches locales

Ces méthodes sont représentées par des algorithmes itératifs, ces derniers exploitent un espace de recherche et se déplacent d'une solution à une autre, à partir d'une solution on peut trouver une autre solution, la notion de voisinage est nécessaire dans ce type d'algorithme, il faut définir les voisins et les relations entre eux, les voisins peuvent être définis par un ensemble de solutions pour un problème, ces méthodes de recherche locales exploitent les relations entre les voisins et se différencient sur le type d'exploitation de chaque méthode [141, 84]. Il existe plusieurs méthodes de recherche locales, on peut mentionner : la méthode de **recuit simulé**, la méthode **tabou**, la méthode de **descente** ...etc [141].

la méthode tabou

La méthode tabou est une heuristique parmi les méthodes de recherche locale, cette technique a un principe que la solution optimale soit la solution qui minimise la fonction fitness, elle se base sur deux listes, une liste pour les solutions interdites et l'autre liste pour les mouvements interdits qui ont déjà visité pour éviter les mouvements cycliques [90].

L'algorithme parcourt l'espace de recherche en cherchant des nouvelles solutions avec les voisins les moins distants, l'algorithme peut arrêter par un critère d'arrêt soit par un nombre d'itérations, ou un temps de calcul, ou une solution optimale ou si la recherche est stagnée lorsque la solution trouvée par l'algorithme ne s'améliore pas sur plusieurs itérations [90].

La méthode tabou utilise deux essentiels principes l'intensification et la diversification, le premier consiste à enregistrer les meilleures solutions déjà visitées, et le deuxième vise à explorer de nouveaux endroits sur l'espace recherche [90].

La méthode tabou est une méthode simple à comprendre, elle donne des bons résultats, mais elle a comme inconvénient des listes qui demande des ressources importantes pour afin de contenir les problèmes à résoudre. Elle est praticable dans plusieurs domaines tels que l'intelligence artificielle, les problèmes d'optimisation...etc [90].

3.4.3.3 les méthodes évolutives

Ce type des méthodes à populations de solutions contrairement aux méthodes constructives et de recherches locales qui se basent à une solution unique, les mé-

thodes évolutives sont basés sur le fonctionnement naturel, elles sont issues de la théorie de l'évolution, ce type de méthodes commence par une population initiale, l'algorithme fait des itérations d'améliorations et trouve une population de solutions [141, 16].
Généralement la population initiale est générée de façon aléatoire, l'algorithme reproduit à chaque étape des individus en utilisant des fonctions soit d'adaptation, soit d'évaluation...etc, jusqu'à ce qu'il arrive à la population finale qui est la solution, l'algorithme peut interrompre son fonctionnement avec un critère d'arrêt comme un nombre d'itérations défini ...etc [141, 16].
Parmi les algorithmes évolutifs on trouve les algorithmes génétiques, la stratégie d'évolution, la programmation génétique...etc. Un algorithme évolutif dépend dans son travail sur des opérateurs tels que la mutation, la sélection et le croisement. La sélection est une opération qui cherche à garder les meilleurs individus, la mutation cherche à faire des modifications dans les individus, le croisement cherche à faire des combinaisons entre les caractéristiques des individus parents pour obtenir une nouvelle génération des individus enfants avec des nouvelles caractéristiques [141, 16].

Les algorithmes génétiques
En Anglais Genetic algorithms, ces AGs ont été développés par John Holland au début des années 1970, et le livre de Goldberg a contribué à sa diffusion et généralisation en 1989, ce type d'algorithme est l'une des méthodes évolutives les plus connues et populaires, ces algorithmes sont du type stochastique pour optimisation en utilisant des opérateurs et des techniques basées sur la génétique naturelle représentée par la sélection, mutation, le croisement, ces AGs sont appliqués dans plusieurs domaines tel que la recherche opérationnelle, la recherche d'information, la sécurité informatique...etc [16, 70, 44].
Un **individu** est représenté par un ensemble de **chromosomes**, chaque **chromosome** est représenté par une suite de **gènes**, un ensemble des individus forme une **population**, **l'environnement** est représenté par l'espace de recherche [67], voir la figure 3.7.

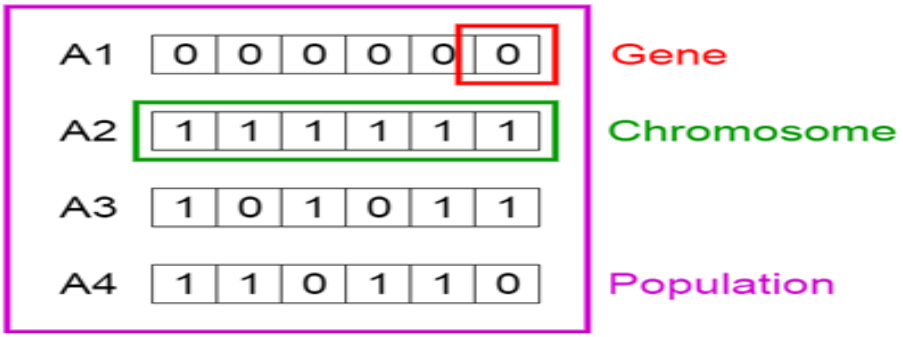


FIGURE 3.7: Gène, Chromosome, Population

Codage
Avant de parcourir les étapes et les opérateurs d'algorithme génétique, il existe

une opération trop importante appelée **le codage** [44].
Les individus qui représentent la population du problème à traiter sont codés, le choix de type de codage est essentiel et important, un bon codage guide l'AG à trouver un optimum global dans un temps court et acceptable, généralement le codage binaire est le plus utilisé et praticable. Pour les grands problèmes le codage binaire est limité, chaque individu de la structure du problème est représenté par une suite de bits [44].
Récemment les chercheurs utilisent un codage réel qui peut contenir les grands problèmes surtout dans les problèmes d'optimisation à variables continues, certains chercheurs les appellent **RCGA** (Real Coded Genetic Algorithms) [44].
Il existe d'autres types de codage comme le codage **gray** [44].

Population initiale

Une population initiale est générée, il faut produire un groupe des individus hétérogène en faisant une génération d'une population qui garantit la diversité des caractéristiques de ces individus et qu'ils sont les plus aptes, pour parcourir le plus largement possible l'espace de recherche, la phase de génération d'une population initiale est très importante car un bon choix nous amenons vers une convergence rapide et nous trouverons une solution satisfaisante, et aidons l'algorithme à fonctionner de façon rapide. Dans les cas des problèmes lesquels nous ne savons rien, la population peut être générée en essayant de toucher tous les points du problème à résoudre, ou elle peut être générée aléatoirement [44].

Sélection

Dans cette étape, l'algorithme cherche à choisir un groupe de chromosomes les plus aptes et qu'ils ont une bonne condition physique pour obtenir des combinaisons des enfants avec des meilleures caractéristiques dans les phases suivantes, il existe plusieurs types de sélection, sélection par roulette, sélection par rang et sélection par tournoi [73].

- Sélection par roulette

Appelée aussi sélection par proportionnalité ou sélection sur le fitness, ce type utilise le principe de la roue de la fortune, chaque individu est sélectionné en fonction de sa probabilité d'adaptation au problème [73].

- Sélection par rang

Les individus sont rangés par ordre selon les meilleurs scores d'adaptation au problème, les individus peuvent être rangés de façon croissante ou décroissante en fonction de la qualité du problème quelque soit de minimisation ou de maximisation, une sélection des individus se fait suite à leurs rangs [73].

- Sélection par tournoi

La sélection par tournoi consiste à choisir aléatoirement deux individus et sélectionner le meilleur parmi eux, l'individu gagne le tournoi et sera sélectionné via une probabilité de la fonction d'évaluation [73].

Croisement

En anglais crossover, cette technique permet de combiner les caractéristiques des individus parents (un ou plusieurs), et obtenir une nouvelle génération enfants (un ou plusieurs). Cette technique fait le brassage génétique des individus de la population. Il existe plusieurs types de croisement, on peut mentionner le premier

type qui est **le croisement à un point aléatoire** qui fait la coupure dans un point aléatoire et fait l'échange de la deuxième partie entre les deux chromosomes voir figure 3.8.

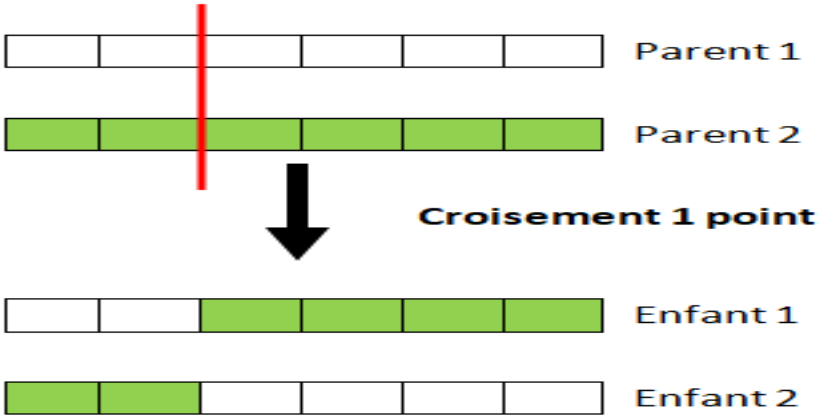


FIGURE 3.8: Croisement à une seul point

Le deuxième type est **le croisement à n points**, ce type fait la coupure dans n points aléatoires et l'échange de la partie des chromosomes se fait à n fois. Le troisième type est **le croisement uniforme** qui pour chaque bit il fait l'échange avec une probabilité fixée donnée, par exemple en donnant la probabilité de changer pour chaque bit qui reste fixé, on fait le croisement avec deux bit 1/2, voire la figure 3.9 [16, 73].

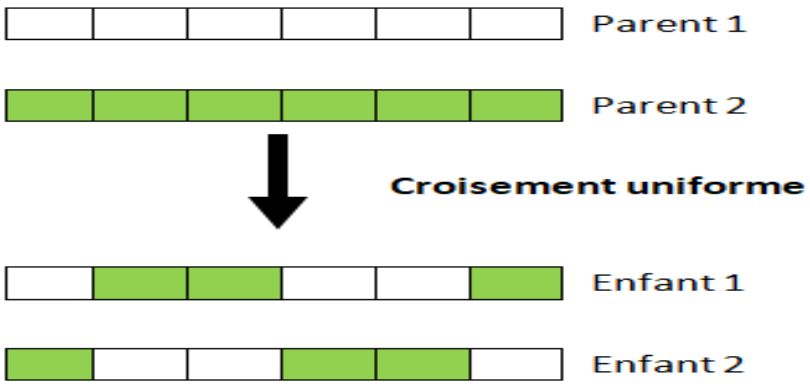


FIGURE 3.9: Croisement uniforme

Mutation

Cette technique permet de substituer aléatoirement un ou plusieurs gènes d'un chromosome selon un taux donné appelé le taux de mutation, généralement ce taux est faible pour ne pas tomber sur la recherche aléatoire [73].

Remplacement
Dans cette étape l'AG cherche à remplacer certains individus parents par des individus enfants, ce remplacement peut être total ou partiel, par prendre les meilleurs individus qui sont présents en fonction de leur performance et les remplacer à la place des individus faibles et mauvais jusqu'à la formule de la nouvelle population de la même taille de la population arrivée au début de cette étape [73].

3.4.3.4 les méthodes hybrides

Récemment les chercheurs font mixer les méthodes de différents types déjà vus au dessus, par exemple ils ont injecté une méthode de type de recherche locale dans une autre méthode évolutive ou vis versa, ce type d'opération est appelé les méthodes hybrides. Ces méthodes hybrides ont donné des résultats de bonne qualité, beaucoup de méthodes hybridées ont été proposées par les chercheurs qui ont été appliquées à plusieurs domaines. Le choix de la technique à utiliser dépend du problème en expérimentant le problème posé et en essayant de gagner et réduire le temps, et de trouver des solutions aux problèmes de grande taille [141].

3.4.4 Bio-inspiration

3.4.4.1 Introduction

La nature est grande et inspirante, elle contient de nombreux échanges des actes entre les êtres vivants qui ne se comptent pas, s'inspirer et s'imiter ces actes échangés dans la nature pour résoudre les problèmes humains peut être définie par le terme de biomimétisme ou bio-inspiration [2, 1].
Dans le monde terrestre, il existe des nombreux organismes qui échangent les actes pour sauver leurs vies et pour préserver la stabilité naturelle, le monde marin aussi contient de nombreux actes échangés entre les êtres vivants, de multiples idées intéressantes peuvent être inspirées par ces actes pour innover, ces interactions entre organismes forment un écosystème durable [2].
Faire innover a toujours été fondamental pour les ingénieurs et l'ingénierie de connaissance, le développement durable et renouvelable a une grande importance dans la vie des gens. Dans la nature les espèces animales et végétales changent ces actes et forment un écosystème durable. Depuis long temps l'homme a observé la nature, il a trouvé beaucoup de réponses et solutions à ces problèmes en vue d'analyser ce monde naturel [2].
Grâce à l'inspiration de la nature et des écosystèmes, l'homme a amélioré sa vie, les technologies ont été développées en plusieurs domaines. Faire imiter les comportements vivants à fournir un immense reflet sur les différents secteurs et il a aidé à créer des nouveaux modèles et matériaux qui sont utiles et très importants dans la vie des êtres humains. L'inspiration de la nature peut être divisée en trois classes, la première est d'inspirer à partir des formes des êtres vivants, ce type d'inspiration est utilisé beaucoup dans l'architecture, prenons l'exemple de l'esplanade Théâtre, il a une couverture inspirée de la peau des fruits du Durian, cette couverture a réduit la consommation énergétique avec 30 % et de 55 % en matière d'utilisation d'éclairage, voir la figure 3.10 [138, 96].

FIGURE 3.10: L'esplanade Théâtre inspiré par la peau des fruits du Durian

La deuxième classe vise à inspirer à partir des procédés et des matériaux présents dans la nature, par exemple les chercheurs ont vu que la peau de requin a une capacité de lutter contre les bactéries, ils ont proposé un pansement antibactérien pour remédier aux blessés de façon rapide, voir la figure 3.11 [138, 137].

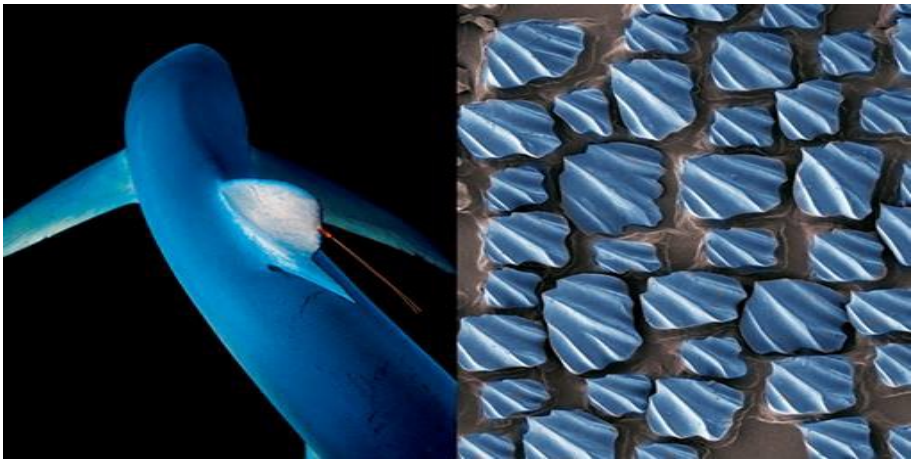


FIGURE 3.11: La peau de requin

La troisième classe cherche à imiter à partir des interactions dans les espèces et les écosystèmes et les échanges entre les êtres vivants dans la nature, prenons un exemple de la chasse dans la nature, il existe beaucoup de tentatives de la chasse, le prédateur attaque la proie pour manger et survivre, la proie peut faire des tentatives pour s'échapper, si non elle est attaquée, ce type peut être représenté artificiellement par des modèles appelés des modèles de prédateur-proie, ils sont utilisés beaucoup dans le cas de la protection des données informatique, les chercheurs qui proposent les modèles inspirés de la nature essayent de modéliser les actes naturels et les imitent par proposer des modèles artificiels qui cherchent à résoudre des problèmes humains de façon d'augmenter le taux de protection de données et assurer la sécurité [138].

3.4.4.2 Historique

Depuis l'antiquité, les gens s'inspirent de la nature pour améliorer leurs outils et offrent des solutions qui avancent et développent leurs vies, l'idée de Léonardo de Vinci (1452-1519) pour inventer les machines volantes a été inspirée du comportement des oiseaux et des chauves-souris, cette étude est pour but de permettre aux humains de voler, au cours de sa vie Léonardo a produit plus de 500 croquis de la mécanique de vols dans l'air, voir figure 3.12 [26]. Léonardo de Vinci a fait beaucoup de tentatives pour voler mais il n'a pas réussi, il a été juste un observateur. Après en 1903 les frères Wright ont réussi à proposer le premier avion qui vole, ils ont inspiré l'idée de leur avion à partir de pigeons [46].
L'inventeur américain Otto Schmitt a utilisé le terme biomimétique pour la première fois dans une conférence internationale à Boston, Schmitt était un doctorant biophysicien, en 1934 il a proposé un circuit électrique basé sur les systèmes d'impulsion neuronaux des calmars, il a proposé une idée que la biologie peut être utilisée dans la technologie, ensuite le terme est entré dans le dictionnaire Webster [46].

FIGURE 3.12: Modèle d'une machine volante inventé par Léonardo de Vinci

Le terme bionique a été inventé par l'ingénieur et psychiatre Jack Steele en 1960, Jack a proposé le mot bionique en représentant toute science des systèmes qui copient de la nature, le terme est entré dans le dictionnaire Webster en tant que science qui applique les données du fonctionnement biologiques pour résoudre les problèmes de l'ingénierie, ensuite Martin Caidin a donné un sens différent au terme bionique qui cherche à utiliser les parties artificielles du corps et augmenter la force humaine grâce à ces appareils, suite à ce changement les chercheurs après ont hésité à utiliser ce terme dans leurs recherches [46].
En 1982, le terme biomimétisme était apparu, le terme a été généralisé par le scientifique Janine Benyus dans son livre Biomimicry : Innovation Inspired by Nature en 1997, il a défini le biomimétisme en tant que science qui observe et étudie la nature, et puis inspiré de ces processus et observations naturelles pour obtenir des

modèles pour résoudre les problèmes humains, il a vu que les interactions dans la nature sont durables et elles peuvent aider les humains à développer leur vies [46]. Le terme Bio-inspiration a été beaucoup utilisé depuis les années 2000, les chercheurs ont dit qu'on ne copie pas de la nature on s'inspire, l'utilisation du mot biomimétisme a été réduit et le mot bio-inspiration a fait une grande tendance. Le biomimétisme est parmi les sciences qui peuvent donner une révolution industrielle et un développement en plusieurs disciplines de la recherche pour améliorer la vie des humains. Faire inspirer de la nature a donné des bonnes solutions, mais inspirer du monde naturel ne garantit pas toujours la conception des problèmes et la production d'un développement durable pour tous les problèmes. Les chercheurs essayent de pratiquer les comportements naturels en bénéficiant de résoudre les problèmes artificiels mais avec des conditions en voyant l'impact sur l'environnement et sur la vie des humains, les chercheurs de cette discipline veulent inspirer des productions durables et importantes pour offrir des services meilleurs et confortables pour les êtres humains.

3.4.4.3 La Bionique

Le mot bionique est rétréci de deux mots « biologie et électronique », ce terme est créé par le médecin de l'armée américaine Jack E Steele en 1960, ce paradigme étudie les systèmes et les sciences biologiques et extrait des modèles et des techniques qui cherchent à résoudre des problèmes humains artificiels. Ce terme est adopté dans le dictionnaire de Webster en tant que discipline qui cherche à appliquer le fonctionnement des systèmes biologiques dans les problèmes de l'ingénierie pour les résoudre, le chercheur Martin caidin a donné un sens différent au terme et il a référencé Jack Steele dans ses recherches, suite à ces actions les chercheurs ont hésité à utiliser ce terme, il est moins utilisé dans la recherche. Récemment la naissance des termes bio-inspiration et biomimétisme ont éliminé l'utilisation du terme bionique, ces nouveaux termes ont englobé toutes les recherches faites précédemment au terme d'inspiration du monde vivant [1, 26, 46, 55].

3.4.4.4 Biomimétisme ou Bio-inspiration

Le terme biomimétisme vient des mots grecs "**bios**" qui veut dire "'**vie**" , et "**mimêsis**" qui veut dire "**imitation**", le chercheur américain Otto Schmitt est le premier qui a utilisé le terme biomimétique dans un congrès à boston en 1934, ce mot n'était pas généralisé, ensuite en 1982 le mot biomimétisme a été apparu et généralisé par le chercheur Janine Benyus, Janine a défini le biomimétisme comme une discipline qui étudie la nature et puis imite les processus naturels pour résoudre les problèmes humains [46].
Le biomimétisme peut être défini comme un paradigme ou une science qui examine la nature et étudie les interactions qu'elle contienne, puis essayer d'imiter ou s'inspirer les systèmes naturels analysés pour obtenir des modèles et des algorithmes artificiels qui cherchent à résoudre les problèmes humains [1, 26, 46, 55].
Récemment à partir des années 2000, les chercheurs et les inventeurs utilisent le mot bio-inspiration de façon générale, le terme bio-inspiration a le même sens que le terme biomimétisme, les chercheurs ont vu qu'ils ne copient pas de la nature, ils inspirent, c'est pour ça ils se sont orientés vers le terme bio-inspiration [1, 46].

En terme de la durabilité et la diversité disponible dans la nature, les inventeurs cherchent à proposer des modèles à partir de la nature en visant à résoudre plusieurs domaines artificiels. L'approche biomimétisme cherche à améliorer les solutions déjà existées et résoudre les problèmes non-résolus en gagnant le temps avec des pertes minimes en optimisant le maximum, les modèles inventés sont caractérisés par la stabilité, l'efficacité et la robustesse [1, 46].
inspirer à partir du monde vivant ne garantit pas toujours la conception pour un développement durable, les inventeurs imitent des ressources naturelles pour trouver des innovations efficaces et rentables en condition de protéger l'environnement sans aucun impact, les chercheurs de biomimétisme voient que ce dernier peut contenir trois niveaux cités au-dessous.

Inspiration à partir des formes et structures
Imiter les formes du monde naturelles est l'un des niveaux qui ont une grande importance, ce type a été utilisé beaucoup pour résoudre les problèmes artificiels et trouver des solutions, les architectes utilisent ce type d'inspiration beaucoup pour formuler des maquettes qui cherchent à économiser les consommations énergétiques, ou pour un bon design [138, 96].
Par exemple les ingénieurs japonais de la compagnie Shinkansen ont vu que l'oiseau martin-pêcheur peut entrer avec une grande vitesse dans l'eau sans le secouer, cette compagnie des chemins de fers a inspirés l'avant de son train à partir du bec d'oiseau, elle a gagné une réduction de 15% de la consommation électrique et une augmentation de vitesse avec 10% ,et un abaissement de bruit au cours de son mouvement, voir la figure 3.13 [138, 2].

FIGURE 3.13: Inspiration du l'avant du train à partir du bec de oiseau

Un autre exemple de "Whale Power", une entreprise canadienne qui travaille dans le domaine de l'énergie renouvelable, elle utilise une éolienne inspirée des nageoires des baleines à bosses, cette éolienne a réduit le bang sonore et augmente le taux de production énergétique avec 20% voir figure 3.14 [2].

FIGURE 3.14: Une éolienne inspirée des nageoires des baleines à bosses

Inspiration à partir des procédés et des matériaux présents dans la nature

Les chercheurs dans ce niveau s'intéressent sur les matériaux disponibles dans la nature, le monde vivant contient des différentes qualités de matières, par exemple Goodyear a produit des pneus auto-régénérant en basant sur la peau et les muscles humains, Michelin a proposé un pneu biodégradable et l'imprimé en 3D.

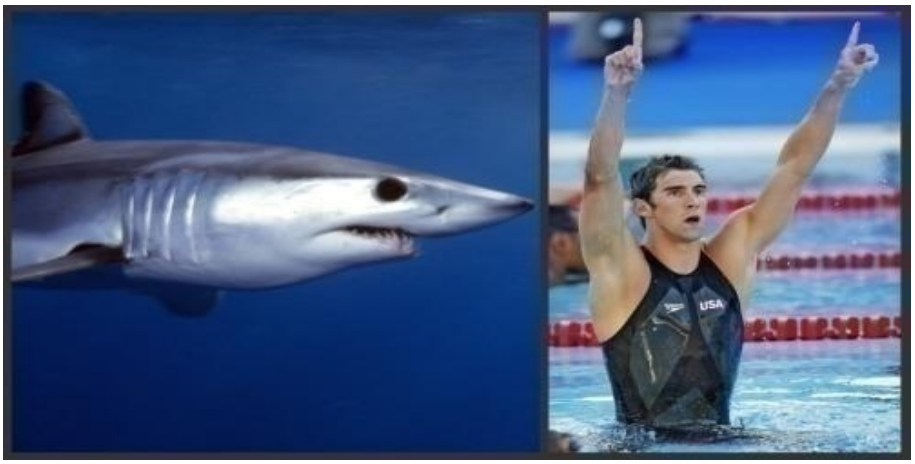


FIGURE 3.15: Fastskin

Un autre exemple d'inspiration appelé "**fastskin**", la société Speedo produit des maillots pour nager, ces tenus sont super performants, ils aident les nageurs d'être plus rapides, la société s'est basée sur la fabrication de ces tenus sur la peau du requin, voir figure 3.15 [138, 137, 107].

La peau de requin est une source d'inspiration, par exemple dans le domaine médical Sharklet a proposé **Sharkskin** qui est un pansement antibactérien qui permet aux blessés de se guérir de façon rapide, voir la figure 3.11 [137].

Inspiration à partir des interactions dans les espèces et les écosystèmes

Si tu observes la nature, tu trouves beaucoup de relations et interactions entre les

êtres vivants, ces échanges forment les espèces et les écosystèmes. Les inventeurs inspirent à partir des interactions dans les écosystèmes, ils voient que la diversité des échanges existants dans la nature peut donner des solutions durables aux problèmes humains et améliorent les technologies [138]. le biomimétisme a touché et optimisé plusieurs domaines remarquables et intéressants précédemment, par exemple les sciences aéronautiques, l'architecture...etc. Les interactions échangées dans la nature ne sont pas dénombrables, il existe un grand ensemble des systèmes intelligents dans la nature, l'humain peut se compter sur ces systèmes pour résoudre ses besoins fondamentaux, par exemple en observant le comportement des abeilles social, les abeilles sont des insectes qui vivent en colonies très organisées formant un système d'intelligence en essaim, ce système est difficile à casser, les abeilles cherchent de la nourriture ou le pollen, les butineuses sortent de la cellule pour trouver de la nourriture, elles communiquent entre eux par la danse ou par la production d'une phéromone, elles dansent frétillante si la nourriture se trouve dans des grandes distances, sinon elles dansent en rond dans des petites distances, le système de fonctionnement des abeilles social a été inspiré et utilisé par les chercheurs dans plusieurs domaines tels que la segmentation et la classification des données, l'apprentissage des réseaux de neurones, contrôle des robot...etc [138, 129].
Les colonies de fourmis comme un autre exemple sont des insectes qui vivent en colonie, c'est parmi l'un des meilleurs systèmes d'intelligence en essaim disponible.

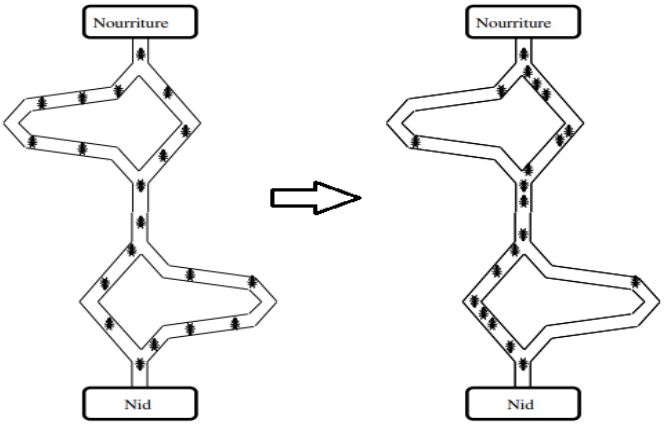


FIGURE 3.16: Les fourmis prend le plus court chemin à la nourriture

Les fourmis sortent de la fourmilière pour chercher de la nourriture aléatoirement, elles communiquent l'information entre eux par une substance chimique appelée **phéromone**, elles perçoivent ce dernier par des récepteurs qui se trouvent dans leurs antennes, après qu'elles trouvent la nourriture, les fourmis laissent des traces de phéromones pour marquer leur trajet entre la fourmilière et la source de nourriture, le chemin le plus court contient une quantité de phéromone importante, il est renforcé au cour de le retour de fourmis au nid pour qu'elles attirent les autres fourmis, la quantité de phéromone produite aide les fourmis à comprendre la quantité de nourriture trouvée et combien de nombres de fourmis peuvent suivre le chemin, à la fin les fourmis font un retour à la fourmilière avec la nourriture

trouvée pour la stocker, Deneubourg et al sont les premiers qui ont inspiré le comportement des fourmis en 1983.

3.4.4.5 Classification des algorithmes bio-inspirés

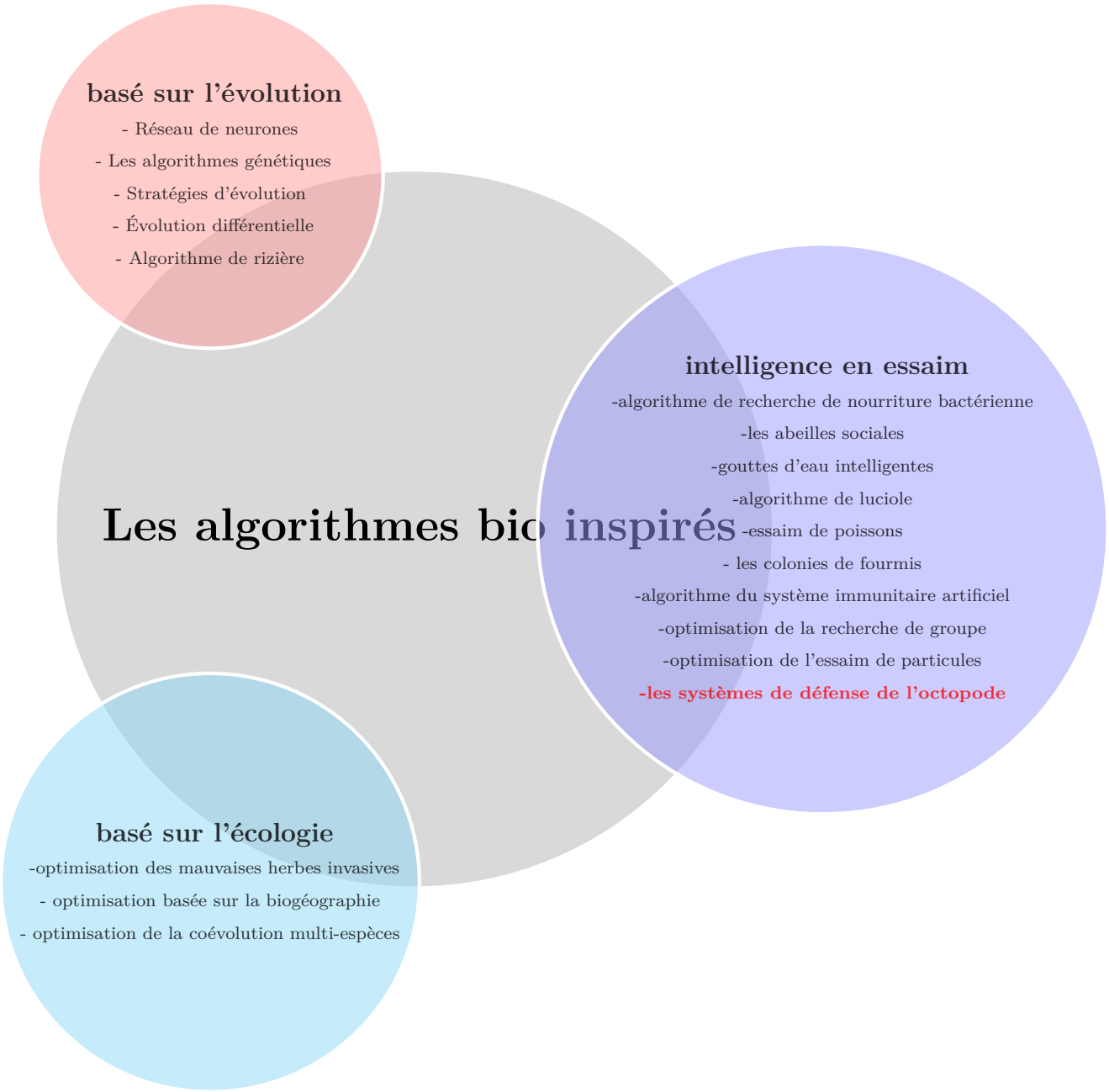


FIGURE 3.17: Classification des algorithmes bio-inspirés [144]

Les algorithmes basé sur l'évolution
Parmi les algorithmes bio-inspirés qu'ils existent en trouvent les algorithmes évolutionnaires ou bien les algorithmes basés sur l'évolution, ces algorithmes sont basés

sur la théorie de l'évolution. Le principe de ces algorithmes est qu'ils se basent sur une population des individus, et ils cherchent à assurer une évolution de cette population jusqu'au trouver une population de solutions meilleures et adaptées au problème à résoudre. Ces algorithmes se basent sur une fonction objective et ils ont un comportement stochastique. ces algorithmes utilisent plusieurs techniques telles que l'encodage des individus, la sélection, la mutation, remplacement...etc. Ils cherchent à fournir une évolution des populations vers une population mieux adaptés au problème touché. Parmi les algorithmes qui se basent sur ce fonctionnement on trouve les algorithmes génétiques, stratégie d'évolution...etc [42].

L'intelligence en essaim

Intelligence en essaim ou bien intelligence distribuée, en Anglais "Swarm Intelligence", cette discipline est inspirée du fonctionnement collectif des êtres vivants, par exemple le comportement collectif des insectes. ce comportement collectif est le résultat d'un acte soit de recherche de la nourriture, ou bien se défendre contre une attaque de prédateur pour se protéger, ou bien une attaque pour chasser une proie...etc. Ces algorithmes se basent sur un travail collectif pour réaliser une certaine tâche et résoudre le problème. Les insectes comme les abeilles et les fourmis utilisent un acte collectif pour chercher de la nourriture et survivre, chaque individu a un comportement en se communicant ensemble pour générer le comportement global qui permet de résoudre la tache. Ces algorithmes d'intelligence distribuée sont beaucoup utilisés pour réaliser les problèmes artificiels et d'optimisation combinatoire. Ils existent beaucoup d'algorithmes qui utilisent l'intelligence distribuée, dans notre cas on a utilisé le fonctionnement de l'octopode pour résoudre le problème des attaques spams sur les messageries électroniques. Les octopodes sont des animaux marins intelligents, ils peuvent se protéger contre une attaque de prédateurs en réagissant par plusieurs actes, ce fonctionnement d'attaque de prédateur/proie est un modèle collectif basé sur l'intelligence distribuée, notre modèle a prouvé qu'il peut être un bon détecteur contre les spams et assurer une meilleure protection [42].

Les algorithmes basé sur l'écologie

Dans la bio inspiration, en trouve une autre catégorie des algorithmes basés sur l'écologie, ces algorithmes écologiques sont inspirés des fonctionnements qui se présentent dans la nature et les écosystèmes, ils se basent sur les interactions complexes des organismes dans l'environnement afin d'obtenir des algorithmes d'optimisation plus complexes que les deux autres catégories déjà cité. Ces interactions ne sont pas dénombrables, ils peuvent être le résultat d'une relation entre un agent et une espèce, ou un agent et plusieurs espèces, ou bien par un agent et son environnement. Les interactions dans la nature peuvent être coopérative, ou bien compétitives, selon le fonctionnement des organismes traités [42].

3.4.4.6 comment s'inspirer de la nature ?

Les inventeurs explorent les différents sources disponibles dans la nature pour inspirer et trouver des nouvelles conceptions, ils évaluent les conceptions existées comme phase de départ, ils peuvent trouver des nouvelles conceptions en se basant sur certains principes des conceptions déjà existées, les concepteurs explorent les produits et les techniques existés, et voient leurs avantages et inconvénients,

si ces derniers ont besoin d'une amélioration ou une nouvelle solution ou ils sont satisfaisantes [134, 123].
La nature est riche, elle est considérée comme une source d'inspiration pour les concepteurs, elle contient beaucoup des interactions entre les espèces et les écosystèmes, ces mécanismes naturels contiennent un grand nombre de solutions intelligentes et originales qui peuvent être des solutions intéressantes pour les besoins humains [134, 123].
Selon les chercheurs tous les objets, les échanges ou les événements présentés dans la nature peuvent être inspirés. Un concept inspiré peut être un résultat de plusieurs idées et il peut être une solution à plusieurs problèmes, donc il faut suivre une stratégie pour trouver une solution à un problème quelconque. Une stratégie biomimétique est une stratégie puissante et performante d'où le concepteur propose des nouvelles conceptions à la base des principes des systèmes naturels [134, 123]. Les experts et les chercheurs voient que l'inspiration de la nature passe par 6 étapes selon un processus de conception, on peut les déclarer au-dessous :

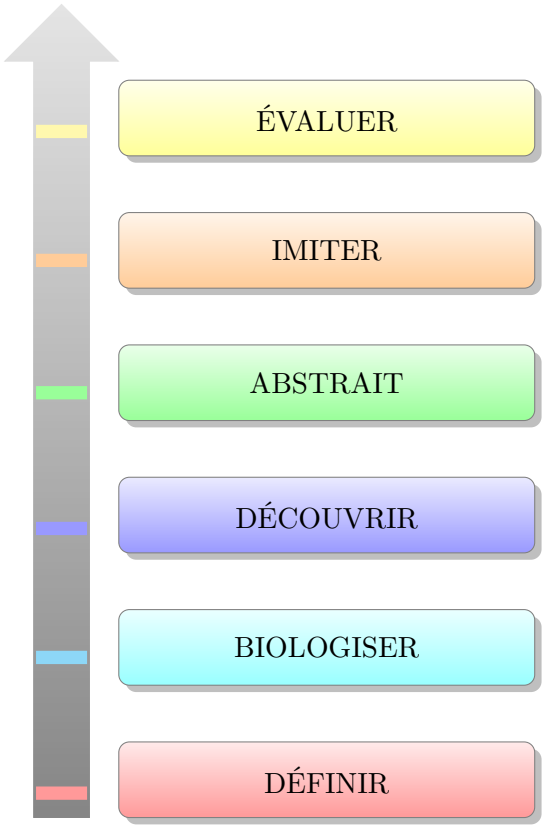


FIGURE 3.18: Le processus de biomimétisme

Définir
John Dewey dit que « Un problème bien défini est à moitié résolu. », la première étape dans le processus de biomimétisme est de définir le défi à résoudre, avec un autre terme le cadrage du travail de conception, une étape d'exploration et de définition des objectifs et de poser des questions, le but de cette étape est de comprendre ce que votre conception doit faire [134].

Quand vous avez une idée sur quoi vous allez travailler, essayer de définir le défi dans une phrase, et poser des questions comme comment pourrions-nous... ?, es qu'on peut... ?...etc [134].
Définir le contexte et les spécificités dans lesquelles vous allez travailler pour cadrer votre travail, et ne restreignez pas le champ jusqu'au vous limitez de trouver des variétés de solutions, il faut toujours bien poser les questions, une bonne question vous guidera à des meilleures solutions innovantes et impressionnantes [134].
Une autre astuce peut vous le faire pour vous faciliter de définir le problème et de voir d'autres problèmes entourant le votre, quelles sont les limites des autres systèmes et quelles sont les relations des autres systèmes avec votre proposition, ces points peuvent vous aider à schématiser votre système et de définir les contraintes pour un départ mieux [134].

Biologiser
Cette étape vient après que vous définissiez votre défi, cette étape est spéciale pour biologiser le défi, avec un autre sens est de chercher dans la nature les stratégies qui résout les problèmes de conceptions, dans cette étape il faut reformuler les questions du défi en question pour trouver des réponses dans la nature [134].
Il ne faut pas formuler des questions bêtes qui n'aident pas dans la recherche, il faut bien poser les questions par trouver des questions qui doivent clarifier vos recherches, essayer de décrivez les fonctions et les contextes en fonction des termes naturels biologiques qui ont une pertinence, donc la façon d'énoncer les fonctions et les contextes doit être changée en mode naturel biologique [134].
Cette étape est très importante dans le processus de biomimétisme, il ne faut pas se précipiter, il faut prendre tous le temps nécessaire pour répondre aux questions biologiquement, cela conduire à gagner le temps et découvrir des nouvelles perspectives [134].

Découvrir
Dans cette étape il faut chercher des fonctionnements naturels qui répondent à nos besoins de conception, par exemple un fonctionnement d'organisme, des interactions dans les espèces et les écosystèmes...etc. Cette étape se concentre sur la recherche en parcourant plusieurs espèces et écosystèmes et en voyant les interactions et les échanges entre les êtres vivants en essayant d'adapter ces fonctionnements naturels à votre défi [134].
Les experts se connectent avec la nature en sortant vers elle, ils découvrent la nature en observant et analysant les comportements. Récemment les livres et les ressources en ligne contiennent aussi beaucoup d'informations nécessaires qui décrivent la nature et les interactions existées de façon claire et détaillées, ces études ont fait par des naturalistes qui étudient la nature, ces gens ont une grande capacité de connaissance des fonctionnements naturels et ils ont une bonne vision de voir les interactions entre les organismes et les écosystèmes [134].
En observant les comportements naturels, vous pouvez poser des questions quels organismes, ou quel comportement peut résoudre mes besoins, ou quel comportement a la même fonction que votre défi...etc, ensuite vous commencer à obtenir des idées [134].

Abstrait
Après le choix des comportements naturels qui ont la même fonction que votre défi, dans cette étape vous devez traduire vos modèles naturels en des modèles conceptuels non-biologiques, il faut définir les stratégies conceptuelles par la description du fonctionnement des stratégies biologiques sans se baser sur les termes naturels. L'extraction des stratégies de conception est l'étape la plus difficile dans le processus du biomimétisme [134].
Plusieurs tentatives peuvent vous aider à trouver une stratégie conceptuelle, vous pouvez essayer de synthétiser les comportements et les modèles naturels, ou aussi vous pouvez se baser sur d'autres recherches dans des revus scientifiques qui peuvent vous aider à trouver des solutions à vos besoins, un autre astuce est d'essayer de dessiner les comportements naturels par des diagrammes simples et claires, essayer d'identifier les termes pertinents biologiques en les remplaçant par des termes synonymes non-biologiques, ensuite vous tenter de trouver la stratégie de conception, vous devez réécrire la stratégie sans utiliser les mots biologiques et elle doit contenir la fonction dans le même sens de défi, ensuite vous pouvez dessiner la stratégie conceptuelle qui contient toutes les informations comme un diagramme de processus sans la partie biologique en visant la partie artificielle et vous pouvez critiquer votre stratégie conceptuelle à la fin, est-ce qu'elle contient toutes les informations utiles par rapport au défi ? ...etc [134].

Imiter
Cette étape est la partie créative du processus est de rendre la stratégie conceptuelle simple et claire, elle doit être lisible en essayant de modéliser les stratégies conceptuelles par la recherche des relations ou en développant des concepts, vous devez proposer des contraintes en se basant sur des fonctions ou utiliser des techniques pour trouver des nouvelles idées, afin de trouver la traduction finale du modèle naturel vers le modèle artificiel [134].

Évaluer
Cette étape est l'étape finale, elle permet de tester votre modèle artificiel et évaluer votre conception de manière s'il répond aux besoins de défi de conception et aux conditions déjà faites, ce modèle doit être adapté dans la vie réelle et voir si ce modèle est fiable ou non, le modèle peut être bon ou non, il est toujours possible de faire des améliorations sur les modèles pour trouver des bons résultats, c'est rare que vous réussissiez du premier coup, il faut toujours expérimenter et tester votre modèle pour l'évaluer, vous pouvez baser sur les limites de votre modèle pour l'améliorer (un problème du coût, de matériels...etc.), jusqu'à que vous arriviez à votre modèle typique final [134].

3.4.5 Conclusion

Dans ce chapitre on a abordé plusieurs notions essentielles dans notre travail, on a commencé par le data mining et ses différentes tâches, ensuite on a défini le processus d'extraction de connaissance(KDD) et l'apprentissage automatique et les différents types de ce dernier, on a aussi présenté les différentes mesures dévaluation d'apprentissage superviser.
On a défini aussi les heuristiques et le méta-heuristiques et présenté les différents

types de ces derniers, on a aussi vu que les heuristiques et les méta-heuristiques ont pris une place importante dans la recherche et ils ont donné dés résultats satisfaisante et impressionnants. Dans la suite on a présenté aussi le nouveau paradigme d'inspiration de la nature appelé bio-inspiration (ou biomimétisme), ce dernier cherche à observer et analyser les interactions du monde naturel afin de trouver des inspirations vers des modèles artificiels pour résoudre les besoins et les problèmes humains, dans la partie de biomimétisme on a vu les différents niveaux d'ispiration et expliquer brièvement le processus d'inspiration de la nature (le processus de biomimétisme).

4

APPROCHES, RÉSULTATS ET EXPÉRIMENTATION

Table des matières

4.1 Une nouvelle technique bio-inspirée basée sur les octopodes pour le filtrage des spams

4.1.1 Introduction et problématique

De nombreuses informations sont échangées par les e-mails et les sms, elles sont devenues des outils essentiels pour les opérations commerciales et même pour les personnes dans leurs vies quotidiennes, elles sont désormais utilisées dans tous les

secteurs professionnels. Cette utilisation importante des e-mails et des sms peut être exposée à des attaques qui cherchent à espionner les données pour atteindre un but bien précis, notamment industriel pour obtenir des informations sur les activités concurrentes en cherchant à trouver tous les détails : (projets en cours, futurs produits, politique de prix...etc.), ces attaques sont effectuées par un non sollicité mail appelé spam [15]. Le spam peut être défini comme un e-mail (ou un SMS) de copies identiques, envoyées automatiquement en nombre, indésirables, non sollicités, vers un contenu indésirable, reçu sans le plein consentement du destinataire.
Les annonceurs sont les premiers spammeurs, mais certains webmasters n'hésitent pas à promouvoir leur site à travers ce [15].
Au cours des dernières années, les chercheurs ont utilisé l'intelligence informatique pour prendre des décisions et résoudre des problèmes réels et même pour acquérir des connaissances pour les grands problèmes et les systèmes en temps réel. L'intelligence informatique utilise des techniques heuristiques qui donnent de bonnes solutions approximatives en temps opportun, l'importance d'utiliser les heuristiques ne consiste pas à trouver un algorithme informatique mais à acquérir des connaissances appropriées. Dans les heuristiques nous trouvons un nouveau paradigme représenté par l'inspiration de la nature (ou bio-inspiration). La bio-inspiration peut être considérée comme une source potentiellement attrayante de connaissances en matière de conception [8]. Aujourd'hui, lorsque les activités humaines ont un impact négatif sur la planète, certains redécouvrent que la nature est une source d'inspiration, une bibliothèque de solutions pour innover tout en respectant l'environnement, la bio inspiration n'est pas simplement l'invention de gadgets ou d'objets imités d'animaux ou plantes [127, 51], elle s'inspire également du fonctionnement des milieux et des relations entre les êtres vivants aux fins du développement durable des sociétés. La nature est vaste, elle est une source d'idées, tous les êtres vivants sont soumis à des mouvements dont les causes et les modalités sont très variées, on peut en trouver des milliers de fonctionnent, c'est donc une nouvelle opportunité de comprendre qu'on a tout à gagner à vivre en harmonie avec la nature et à protéger la biodiversité, elle nous conduit à des idées contre-intuitives, à des choses auxquelles nous n'aurions pas pensé. C'est intéressant et cela va nous permettre de développer des technologies de rupture. Nous devons identifier nos besoins, l'enjeu principal de ce travail est la création d'une nouvelle technique bio-inspirée qui peut être un bon système de filtrage et qui peut nous protéger contre les spams et contribuer à résoudre l'un des problèmes des êtres humains. Après avoir mené des recherches en bio-inspiration, nous avons choisi d'explorer le monde marin, notre attention s'est portée sur les octopodes, nous avions constaté que les octopodes ont une grande capacité et une grande qualité de défense pour échapper contre les attaques des prédateurs et se protéger, donc son fonctionnement naturel représenté par ses actions lorsqu'il est attaqué peut être un bon protecteur contre les spams, et peut être défini comme un modèle de prédateur-proie. Après un état de l'art sur la bio-inspiration, nous avons constaté qu'il n'y avait pas beaucoup des travaux réalisés dans le domaine du monde marin, l'objectif de ce travail est d'explorer les nouveautés mais d'acquérir les connaissances appropriées pour garantir la protection.

L'un des animaux aquatiques les plus défensifs que nous ayons trouvés est l'octopode, il a une conception de corps incroyable, les octopodes sont capables de se défendre de diverses manières. Nous pouvons adapter ces techniques de défense pour être un bon système de protection. Notre contribution principale dans ce travail est de créer un nouveau modèle bio-inspiré basé sur les techniques de défense de l'octopode pour filtrer les spams des hams. Dans cet esprit, nous avons proposé un modèle heuristique inspiré de la nature pour construire un système de détection des spams en utilisant une méthode de classification basée sur le fonctionnement naturel d'octopode et opérant sur son principe de défense pour un bon filtre de messages, cet algorithme fonctionne à l'aide d'une base d'apprentissage et filtre les messages de la base de test et les étiquette, chaque message prend une des deux classes, soit normale, ou un message spam. Ce modèle vise à résoudre les problèmes causés par les détecteurs classiques et à obtenir un bon détecteur des spams pour assurer la protection.
Nous avons commencé notre travail par quelques travaux connexes effectués, ensuite nous avons défini le fonctionnement naturel de l'octopode et notre proposition pour obtenir le modèle artificiel, ce modèle obtenu est mise en test pour être évaluer, dans la suite nous avons illustré les résultats obtenus sur des tableaux, et à la fin on a sélectionné les meilleurs résultats dans notre expérimentation et les comparer avec d'autres travaux existants.

4.1.2 Détection des spams

Aujourd'hui avec le développement technologique et la multitude des services et des plateformes d'échanges de données, les utilisateurs cherchent à protéger leurs données en garantissant la confidentialité et tous les services de sécurité. Parmi les services d'échanges de données on trouve la messagerie électronique (le courrier électronique), abrégé e-mail. Récemment les e-mails ont pris une grande importance dans la vie personnelle et professionnelle des gens, beaucoup d'informations sensibles et importantes circulent dans la messagerie électronique telle que des secrets professionnels, des idées des projets à réaliser, des secrets intimes dans la vie des gens...etc.
Sécuriser les messagerie électroniques est nécessaire et important pour protéger la confidentialité des utilisateurs car beaucoup de tentatives de fraudes ont été apparus qui cherchent à casser cette plateforme. Certaines entreprises veulent voler des idées des autres entreprises compétitives et voire les futures planes, d'autres pirates veulent menacer la confidentialité des victimes par voler des mots de passe, des secrets intimes ou des clés secrètes de valeur...etc. Il existe des différentes causes pour voler et menacer la plateforme des messageries électroniques, ces menaces se font via des attaques nommées spams.
Les spams sont des courriels électroniques généralement générés de la publicité, ils ciblent les victimes de la messagerie pour prendre des informations ou voler des données, les spams transfèrent des virus à la victime via un e-mail, les spammeurs exploitent les faiblisses de la victime.
Il faut lutter contre les spams et augmenter le taux de sécurité des messageries électroniques pour protéger les données, il existe plusieurs techniques que les utilisateurs peuvent le faire pour éviter les spams, ils peuvent négliger les publicités

en évitant de les consulter sur les sites internet, ils peuvent aussi éviter d'ajouter les adresses mail personnelles dans les sites internet, ils peuvent essayer d'utiliser des e-mails poubelles qui n'ont aucune importance pour les sites qui demandent des informations afin d'éviter de recevoir des spams dans les boîtes mail officielles. Jusqu'à maintenant le problème des spams est un grand problème pour les utilisateurs des messageries électroniques, beaucoup de recherches ont été réalisées pour lutter contre ces attaques et d'autres sont en cours de réalisation. La détection des spams reste un problème à résoudre pour protéger les utilisateurs. Nous pouvons mentionner certaines techniques telles que Zhang a proposé une nouvelle méthode de détection de spam qui se concentrait sur la réduction de l'erreur des faux positifs du mauvais étiquetage du non-spam comme spam, les auteurs ont utilisé PSO pour la sélection des caractéristiques et appliquait l'arbre de décision c4.5 comme algorithme d'apprentissage [146], l'inconvénient de leur travail est que l'auteur a utilisé une méthode wrapper pour la sélection des attributs, cela a entraîné des risques de sur-ajustement et un temps de calcul important lorsque le nombre de variables est important. Santosh Kumar a proposé une approche de classification DMMHSVM (Dual-Margin Multi-Class Hypersphere Support Vector Machine) pour classer automatiquement le web spam par type, il a optimisé les résultats avec un nouveau camoufleur des caractéristiques de spam qui a aidé son modèle de classification à atteindre une haute précision et un taux de rappel élevé, en réduisant le taux des faux positifs [77], le côté négatif de ce modèle est que l'algorithme svm a fait une classification avec plusieurs classes, généralement svm n'est pas bon dans ce cas. José María Gómez Hidalgo a analysé dans quelle mesure les techniques de filtrage bayésien utilisées pour bloquer le courrier électronique spam peuvent être appliquées au problème de la détection et de l'arrêt du spam mobile [64], l'auteur a fait une étude comparative entre les algorithmes de fouille de données. Gordon V. Cormack avait effectuait des expériences sur le filtrage SMS en utilisant des filtres performants des spams courriels sur des messages spams de mobiles en utilisant une représentation de fonctionnalité appropriée avec des résultats soutenants son hypothèse [32], il a également effectué une étude comparative entre les algorithmes de filtrage du spam. Qian Xu a proposé une solution d'utiliser des graphes de fouilles de données pour distinguer les spammeurs des non-spammeurs et détecter le spam sans vérifier le contenu du message [143], le point négatif de cette proposition est que le graphique n'a pas traité tous les cas de données, il a traité juste une petite partie et la classification des instances était basée sur la conclusion, donc le degré d'erreur est plus grand. Dea Delvia Arifin a utilisé naïve bayes comme détecteur des spam sms, les auteurs ont extrait les caractéristiques en utilisant l'algorithme FP-Growth avec trois supports minimum 3%, 6% et 9%, l'inconvénient de ce travail est que si les données sont volumineuses, le FP– algorithme prendra beaucoup de ressources du système et prendre beaucoup de temps de traitement, il pourrait également ne pas tenir en mémoire [9]. Naresh Kumar Nagwani a classé les messages sms en spams ou en messages normaux en utilisant les naïves bayes, SVM, LDA (modèles de documents texte comme mélanges de sujets latents) et NMF (algorithme de factorisation matricielle), les auteurs ont fait une étude comparative entre ces algorithmes, après, ils ont fait une classification de chaque message en utilisant des

techniques de clustering (k-means et NMF text clustering techniques) pour créer des clusters sms des messages non-spam afin de collecter des messages similaires pour l'identification des threads [93]. HassanNajadat a comparé différents classificateurs des sms spam, les auteurs ont mélangé les classificateurs pour améliorer les résultats obtenus par les classificateurs normaux, ils ont utilisé plusieurs algorithmes bien connus dans le domaine de la classification comme : SVM, Naïve Bayes, arbre de décision, KNN, Table de décision ... etc [95]. Inwhee Joe a utilisé SVM comme filtre des sms spam basé sur un thésaurus, le système isole les mots des données d'échantillonnage à l'aide d'un dispositif de prétraitement et intègre la signification des mots isolés à l'aide d'un thésaurus, et génère les caractéristiques des mots intégrés par le biais de la méthode statistiques chi-square, et étudie ces caractéristiques, la limitation de ce travail avec l'utilisation de la méthode chi-square, car cette technique nécessite que les données soient de fréquence, et que le chi-square est sensible à la taille de l'échantillon, la plupart recommandent que le chi-square soit déconseillé d'être utilisé si la taille de l'échantillon est moins de 50 [72]. José M. Bande Serrano a proposé une méthode pour capturer le style d'écriture des spams et des messages non-spam en préservant la séquentialité du texte dans l'espace des caractéristiques, les auteurs ont extrait les caractéristiques des messages en appliquant trois techniques : informations extrinsèques, extraction d'étiquettes séquentielles et regroupement des termes [12]. Camilo Caraveo avait présenté une nouvelle méta-heuristique pour optimisation basée sur le mécanisme d'autodéfense des plantes contre les prédateurs avec trois opérateurs de reproduction, les auteurs ont proposé un modèle prédateur-proie pour l'optimisation appliquée à un corpus des fonctions mathématiques [21]. Leticia Cervantes a proposée d'utiliser une approche basée sur la logique floue pour l'adaptation de la génération de lacunes et la probabilité de mutation dans un algorithme génétique, la performance de cette méthode est présentée avec le problème du contrôle de vol et les résultats montrent comment elle pourrait diminuer l'erreur pendant le vol d'un avion en utilisant la logique floue pour certains paramètres de l'algorithme génétique [24]. Oscar Castillo a présenté une approche généralisée du système de logique uzzy de type 2 (GT2FLS) pour l'adaptation dynamique des paramètres en méta-heuristique et pour la conception optimale du contrôleur floue, les auteurs ont utilisé l'algorithme d'optimisation des colonies d'abeilles pour concevoir de manière optimale un GT2FLS [22]. Leticia Amador-Angulo a présenté une nouvelle méthode d'optimisation des colonies d'abeilles floues pour trouver la distribution optimale des fonctions d'appartenance dans la conception de contrôleurs floue pour les plantes non linéaires complexes. Les auteurs envisagent plusieurs expériences dans la simulation de deux corpus de problème du contrôle avec un type -1 contrôleur de logique floue (T1FLC) et BCO pour minimiser l'erreur dans la simulation des plantes non linéaires pour des problèmes complexes [5]. Frumen Olivas a présenté une comparaison entre l'optimisation des essaims de particules, l'optimisation des colonies d'abeilles et l'algorithme des chauves-souris, les auteurs ont modifié les principaux paramètres de chaque algorithme à travers un système de logique floue d'intervalle du type 2 [101]. Et d'autres beaucoup de recherches pour lutter contre les spams qu'on ne peut pas les mentionner tous.
Les chercheurs été ont orienté vers les méthodes et les analyses heuristiques parce

qu'ils voient que les détecteurs classiques sont limités et ne donnent pas des résultats satisfaisants dans un temps acceptable, les chercheurs voient que les heuristiques cherchent à résoudre les problèmes et exploitent l'espace de recherche pour trouver une solution locale ou globale qui peut être optimale ou pas. Les heuristiques trouvent des solutions dans un temps de calcul réduit par rapport aux systèmes classiques et ils sont praticables et faciles à combiner avec d'autres méthodes, ils cherchent à maximiser ou minimiser la fonction objective selon le problème. Dans les heuristiques il existe une nouvelle discipline connue par le biomimétisme (avec un autre terme la bio-inspiration), ces techniques sont extraites en se basant sur le fonctionnement naturel par l'analyse des interactions et les échanges des actes des êtres vivants dans les différentes espèces. L'inventeur du technique bio-inspiré passe par un processus de biomimétisme pour extraire le modèle artificiel final qui peut être utilisé pour résoudre les besoins humains. Dans notre travail on a utilisé une technique inspirée de la nature, exactement du fonctionnement biologique de la pieuvre, en se basant sur ses techniques de défenses contre une attaque de prédateur.

4.1.3 Notre Approche

L'application Biomimétique

L'approche naturelle de l'octopode

Les octopodes sont considérés comme des mollusques à huit tentacules [56], ils ont un design corporel étonnant. Ils sont capables de se défendre de diverses manières. Le plus courant est le vol, car ils peuvent utiliser la propulsion par jet pour se déplacer rapidement dans l'eau. Leurs corps flexible n'a pas d'os, ils peuvent donc s'échapper dans des petites fissures, rochers, crevasses, et même dans des bouteilles et des canettes qui ont trouvé leur chemin dans le fond de l'eau [142].

Les octopodes sont bien connus par leur capacité de libérer une substance d'encre noire des glandes du corps, comme indiqué sur la figure 4.1. Lorsqu'ils vivent des situations stressantes, ils libèrent ce type d'encre afin de désorienter leurs prédateurs. L'encre réduit la vision et la capacité de sentir, cela laisse le prédateur confus et désorienté tandis que l'octopode fait sa fuite rapide [142].

La morsure d'un octopode contient un venin très puissant, c'est ainsi qu'ils parviennent à paralyser leurs proies pendant qu'ils les consomment, ce venin n'est généralement pas nocif pour l'homme, il n'y a qu'une seule espèce qui a un venin suffisamment puissant pour tuer une personne, c'est le Blue Ring Octopod [142].

Leurs capacités de changer les couleurs sont très importantes, ils peuvent ainsi se fondre dans leurs environnements. Les humains et les prédateurs dans l'eau peuvent passer à côté d'eux sans jamais les voir. En termes simples, ils sont capables de se cacher à la vue [142].

Certains octopodes d'eau profonde sont connus pour émettre de la lumière dans tous leur corps, ce qui est censé être utilisé pour distraire et désorienter les prédateurs potentiels ou pour hypnotiser les proies afin qu'elles puissent s'y déplacer et les capturer [142].

Pour les octopodes mimiques, leurs mécanismes de défense vont encore plus loin, il leur permet d'assumer la coloration et la conception d'une quinzaine de types

d'animaux différents, Ils se déplaceront dans l'eau en agissant comme des anguilles, des étoiles de mer et plus encore afin de rester à l'écart des prédateurs. Ils utiliseront cette défense pour leur permettre également de se rapprocher très près de la nourriture qu'ils souhaitent consommer. Sous une telle forme, ils ne sont considérés pas comme une menace que trop tard [142].

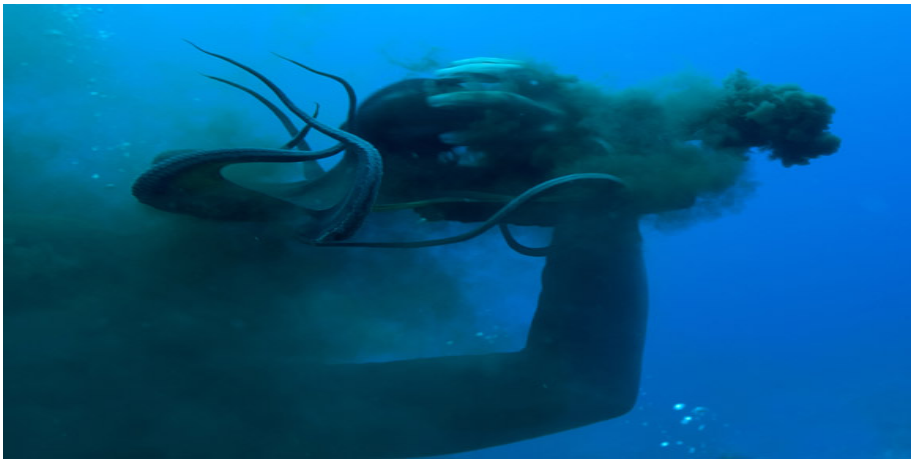


FIGURE 4.1: Octopod dégage l'encre foncé

La conception du corps de l'octopode leur permet d'avoir des différentes façons de se défendre. Parfois, ils sont capturés par des proies par un bras ou deux et cela semble être la fin de la route pour eux. Pourtant, ils peuvent instinctivement permettre à ces bras d'être retirés et ils nagent à grande vitesse. Dans un court laps de temps, ces armes repousseront. Ce type de mécanisme de défense est très fascinant et il fonctionne pour toutes les espèces d'octopodes [142, 87].
Ce qui est également étonnant, c'est qu'ils semblent être capables de s'adapter à leurs différents changements environnementaux. Avec cela, ils finissent par trouver des moyens créatifs de se protéger. Ils feront tout ce qu'ils peuvent pour se défendre. Tout prédateur poursuivant un octopode ferait mieux de se battre. C'est pourquoi ils laissent généralement les plus grandes espèces tranquilles [142, 87].

L'Approche artificielle de l'octopode

Contexte physique

La force est la pression exercée sur un objet pour le mettre en mouvement ou pour accélérer son mouvement. La deuxième loi du mouvement de Newton décrit la relation entre la force, la masse et l'accélération, cette relation est utilisée pour calculer la force. En général, si la masse de l'objet est plus élevée, alors la force nécessaire pour déplacer cet objet est plus élevée [48]. Multiplier la masse par accélération. La force (F) nécessaire pour déplacer un objet de masse (m) avec accélération (a) est donné par la formule : F = m * a

$$Force(N) = Masse(kg) * accélération(m/s^2) \tag{4.1}$$

- La force est mesurée en Newtons, N.
- La masse est mesurée en kilogrammes, kg.

- L'accélération est mesurée en mètres par seconde carrés, m/s^2 [48].
Comment calculer l'accélération, L'accélération est le taux de variation de la vitesse d'un objet lorsqu'il se déplace. Si un objet maintient une vitesse constante, il n'accélère pas.
L'accélération ne se produit que lorsque la vitesse de l'objet change. Si l'objet change de vitesse à une vitesse constante, l'objet se déplace avec une accélération constante. Vous pouvez calculer le taux d'accélération mesuré en mètres par seconde en fonction du temps qu'il vous faut pour passer d'une vitesse à l'autre, ou en fonction de la masse d'un objet [39].

$$a = \Delta v/\Delta t \tag{4.2}$$

Vous pouvez calculer l'accélération moyenne d'un objet sur une période de temps en fonction de sa vitesse (sa vitesse en se déplaçant dans une direction spécifique), avant et après ce temps [139, 76].
Vous devez connaître l'équation de l'accélération : a = Δv / Δt où a est l'accélération, Δv est le changement de vitesse et Δt est le temps qu'il a fallu pour que ce changement se produise. [39].
Vous pouvez également définir Δv et Δt, Δv = vf - vi et Δt = tf - ti où vf est la vitesse finale, vi la vitesse initiale, tf l'heure de fin et ti l'heure de début. Si la vitesse finale est inférieure à la vitesse initiale, l'accélération sera un nombre négatif ou la vitesse à laquelle l'objet ralentit [39, 28].

Dataset

Dans notre travail, nous utilisons la collection SMS Spam, c'est un ensemble de messages SMS qui ont été collectés pour la recherche des SMS Spam. Il contient un ensemble de messages SMS en anglais de 5574 messages, étiquetés en fonction du hams (sms légitime) ou du spam [81, 3]. Les fichiers contiennent un message par ligne. Chaque ligne est composée de deux colonnes : v1 contient l'étiquette (ham ou spam) et v2 contient le texte brut [81, 3]. Cet ensemble de données contient des messages avec un pourcentage différent, 87% pour les messages de "ham" et 13% pour les messages de "spam" [81, 3].

Notre modèle artificiel

L'idée de notre approche artificielle est d'inspirée du fonctionnement naturel de l'octopode et plus précisément de ses techniques de défense contre les attaques des prédateurs. Notre algorithme utilise et applique des fonctions de physique et de probabilité.
L'idée d'utiliser les fonctions de physique dans notre modèle, plus précisément la fonction de la force et du mouvement a été inspirée du fonctionnement biologique de la pieuvre, nous avions observé que l'un des comportements de l'octopode pour échapper contre une attaque de prédateur est de faire une propulsion à réaction pour se déplacer rapidement et s'enfuir. Dans le physique la fonction pour mettre un objet en mouvement est la fonction de la force selon la loi de mouvement de Newton pour le déplacement d'un objet dans une autre position.
Un autre comportement de la pieuvre pour éviter une attaque est la production d'une quantité d'encre noire pour tromper le prédateur afin qu'il ne puisse pas voir et sentir la victime. Dans notre modèle artificiel, pour chaque instance de la

base de test, on calcule deux probabilités une par rapport à la classe des sms spam dans la base d'apprentissage, et la deuxième par rapport à la classe sms ham dans la base d'apprentissage, l'idée d'utiliser la probabilité a été inspiré à partir de la probabilité d'encre noire réalisée par l'octopode dans le modèle naturel.
L'idée dans notre modèle est avec ces deux comportements, avec un autre terme si l'octopode pourrait faire une propulsion à réaction rapide et produire une bonne quantité d'encre contre une attaque le degré de survie sera excellent, donc il pourrait se protéger, son fonctionnement peut être défini comme un modèle de prédateur-proie, ces modèles se définissent par leur capacité d'éviter les attaques et bien défendre de la part de proie, généralement ces systèmes naturels sont inspirants pour des modèles artificiels sécuritaires, en d'autre terme pour trouver un système de sécurité intelligent, dans notre cas pour détecter et lutter contre les attaques des sms spams.
Comme nous l'avons vu dans la fonction naturelle des octopodes, nous sommes intéressés aux deux techniques de défense, la première est la propulsion à réaction pour se déplacer rapidement dans l'eau et s'échapper, et la deuxième technique est la capacité de libérer une substance d'encre sombre des glandes dans le corps afin de désorienter leurs prédateurs. Nous avons vu que la pieuvre utilise la propulsion à réaction pour se déplacer rapidement et s'échapper des prédateurs, il se déplace de la position **P1** à une autre position **P2**, en physique, la force sur un objet pour le mettre en mouvement est définie par la formule (4.3) de Newton.
L'ensemble de données à traiter contient des SMS, chaque sms contient une colonne pour le texte brut du message, et une autre colonne pour la catégorie du sms, soit spam ou ham.
Les algorithmes généralement ne peuvent pas traiter les sms directement dans leur format, il faut faire un prétraitement pour qu'ils puissent traiter les données. Notre algorithme non plus ne peut pas traiter les sms dans leur format original, il traite des données numériques, si pour sa on a utilisé la technique TF-IDF dans notre travail, cette technique est l'une des techniques les plus populaires pour traiter les données textuelles et les convertir en valeurs numériques.
Pour nous, chaque SMS de l'ensemble de données test est représenté par un octopode, l'idée est que si l'octopode peut s'échapper aux prédateurs et les désorienter donc il est en mode protégé et sécurisé et le message est ham, pour le cas contraire, si le prédateur attaque la pieuvre et l'attrape donc la proie est menacée et le message prend la classe des sms spam.
Dans notre proposition, Nous avons deux variables à calculer, la première est la force représentée par une propulsion à réaction pour échapper aux prédateurs, et la deuxième est la probabilité de chaque classe dans la base d'apprentissage représentée par la probabilité de libérer l'encre noire du poulpe.

$$Force(N) = Masse(kg) * accélération(m/s^2) \tag{4.3}$$

$$accélération = (vf - vi)/(tf - ti) \tag{4.4}$$

Paramétrage :
- **Masse** = la valeur moyenne de tous les tf * idf des mots dans chaque message
- On néglige le temps tf - ti = 1

- **vf** : la grande valeur de tf * idf des mots dans chaque message
- **vi** : la petite valeur de tf * idf des mots dans chaque message
- **PH** : probabilité de Ham sur l'ensemble de message dans la base d'apprentissage
- **PS** : probabilité de spam sur l'ensemble de message dans la base d'apprentissage
- **PH**= nombre de messages de ham dans la base d'apprentissage / nombre total de messages
- **PS**= nombre de messages de spam dans la base d'apprentissage / nombre total de messages

Pour chaque message de la base de test et de la base d'apprentissage, nous avons calculé le tf-idf de tous les messages pour avoir la forme numérique de ces derniers, ensuite on a calculé la force de chaque message comme nous l'avons mentionnée. Pour chaque force calculée à partir de la base de test, nous avons calculé deux distances euclidiennes, la première de chaque message de test avec toutes les forces des messages étiquetés ham de la base d'apprentissage, le résultat trouvé est collecté sur **FH (i)**, et la deuxième de chaque message teste avec toutes les forces des messages étiquetés spam de la base d'apprentissage, et le deuxième résultat est collecté sur **FS (i)**.

- **i** : le message sélectionné de la base de test

Après avoir calculé les forces, nous divisons la première force **FH (i)** par la probabilité des hams messages **PH** pour obtenir la première valeur, et nous également divisons la deuxième force **FS (i)** calculée par la probabilité des spams messages.

$$CH(i) = FH(i)/PH(i) \tag{4.5}$$

$$CS(i) = FS(i)/PS(i) \tag{4.6}$$

Si la première valeur est inférieur ou égale à la deuxième valeur (**CH (i)** <= **CS (i)**), alors l'algorithme étiquette le message sélectionné avec Ham, sinon le cas contraire l'algorithme étiquette le message avec spam. Dans notre modèle l'octopode peut s'échapper et être en sécurité si la valeur de **CH (i)** est inférieure ou égale à la valeur de **CS (i)**, sinon l'octopode est attaqué, voir figure 4.2.

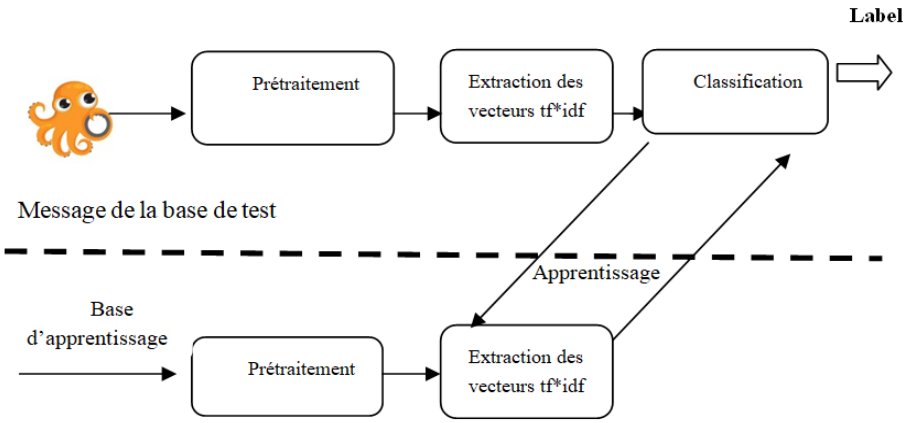


FIGURE 4.2: Illustration de notre modèle

Algorithme

Algorithm **FilterSpam** ;
Begin.
For I = 1 to n do.
Calculate (FH(i)) ;
Calculate (FS(i)) ;
Calculate (PH) ;
Calculate (PS) ;
Calculate (CH(i)) ;
Calculate (CS(i)) ;
If (CH(i)<=CS(i)) then Classe (i) = "ham".
Else Classe (i) = "spam" ;
End.

Modèle de transition du naturel à Notre Approche artificiel

Naturel	***Artificiel***
1.Octopode	**1.**Message de la base de test
2.L'attaque du prédateur sur l'octopode	**2.**Début du processus de détection
3.La force de la propulsion a réaction pour se déplacer rapidement et s'échapper	**3.**Force calculée pour chaque message de la base de test **(F = m*a)**
4.La probabilité de produire l'encre noire pour désorienter les prédateurs	**4.**Probabilité de chaque classe (Ham ou spam) **PH** et **PS**
5.Si l'octopode peut s'échapper	**5.**Le SMS est Ham
6.Si l'octopode est attaqué	**6.**Le SMS est Spam

TABLE 4.1: transition du comportement naturel à notre approche artificiel

4.1.4 Résultats et Expérimentation

Dans notre travail, nous avions pris le jeu de données SMS Spam Collection qui contenait environ 5574 messages, pour obtenir les données de tests et d'apprentissage nous avions utilisé la validation croisée 10-folds, ensuite nous avions comparé nos résultats avec d'autres études obtenues par d'autres auteurs dans leurs récentes recherches appliquées sur le même ensemble de données pour donner plus de crédibilité à nos résultats et évaluer notre modèle comme nous l'avions mentionné dans le tableau 4.3, par exemple : Dea Delvia Arifin a utilisé Naive Bayes comme un détecteur des SMS spam, il a extrait des fonctionnalités en utilisant FP-Growth algorithme avec trois prises en charge minimales 3%, 6% et 9%

[9]. Naresh Kumar Nagwani a également catégorisé les messages SMS en spams ou en messages normaux à l'aide de Naïve Bayes SVM, LDA (modèles de documents texte en tant que mélanges de sujets latents) et NMF (Algorithmes de factorisation en matrices) [93]. Hassan Najadat a comparé différents classificateurs des spams sms comme : AdaBoostM1 *, Table de décision, J48 (D-tree), Random Forest, K-NN, K-Star, Naïve Bayes, NB Multinomial, DMNBtext, SVM , SGD et VotedPerceptron [95].
Nous avions utilisé une méthode bien connue et largement utilisée pour obtenir l'ensemble de test et d'apprentissage qui est la validation croisée k-fold et sélectionnée k avec 10. La validation croisée est une technique très utile pour évaluer les performances des modèles d'apprentissage automatique, il aide à savoir comment le modèle d'apprentissage automatique se généraliserait à un ensemble de données indépendantes. Vous souhaitez utiliser cette technique pour estimer la précision des prévisions de votre modèle dans la pratique. Un type de validation croisée est la validation croisée K-Fold [60]. Dans un cycle de validation croisée, vous devrez diviser votre ensemble de données de formations d'origine en deux parties :
- Ensemble d'apprentissage de validation croisée
- Ensemble de test de validation croisée ou Ensemble de validation [60] Nous avions utilisé l'ensemble de données avec deux opérations, sans nettoyer le texte (en laissant les mots vides) et avec le nettoyage (en supprimant les mots vides), nous avions divisé le corpus en deux parties en utilisant la validation croisée de 10-fold [60, 62] pour obtenir l'ensemble de test et l'ensemble d'apprentissage, nous avions également utilisé deux méthodes de représentation de texte sac de mots et n-gram de mots [6, 23] avec n = 2, n = 3 et n = 4, après cela, nous avions extrait les caractéristiques des bases de test et d'apprentissage, enfin notre algorithme a géré les données numériques et fait la classification, les résultats obtenus de notre modèle sont présentés dans le tableau 4.2 ci-dessous.
Nous avions utilisé la méthode du n-gram car il est possible d'obtenir la fonction de vraisemblance de l'apparition du mot suivant à partir du corpus d'apprentissage, il est donc facile de construire une probabilité de distribution pour le mot suivant avec une taille de n [23].

	Notre approche d'Octopode							
	sans nettoyage des mots vides				*Avec nettoyage des mots vides*			
texte représentation / Mesures	**Sac de mot**	**2-gram**	**3-gram**	**4-gram**	**Sac de mot**	**2-gram**	**3-gram**	**4-gram**
Rappel	1.000	1.000	1.000	0.993	1.000	1.000	0.998	0.987
Précision	0.934	0.938	0.969	0.824	0.925	0.977	**0.994**	0.795
Accuracy	0.943	0.946	0.973	0.842	0.935	0.980	**0.993**	0.813
F-mesure	0.966	0.968	0.984	0.900	0.961	0.988	**0.996**	0.881
Entropie	0.030	0.028	0.014	0.084	0.034	0.010	**0.003**	0.100

TABLE 4.2: Les résultats obtenus par notre modèle en utilisant la validation croisée avec 10-fold

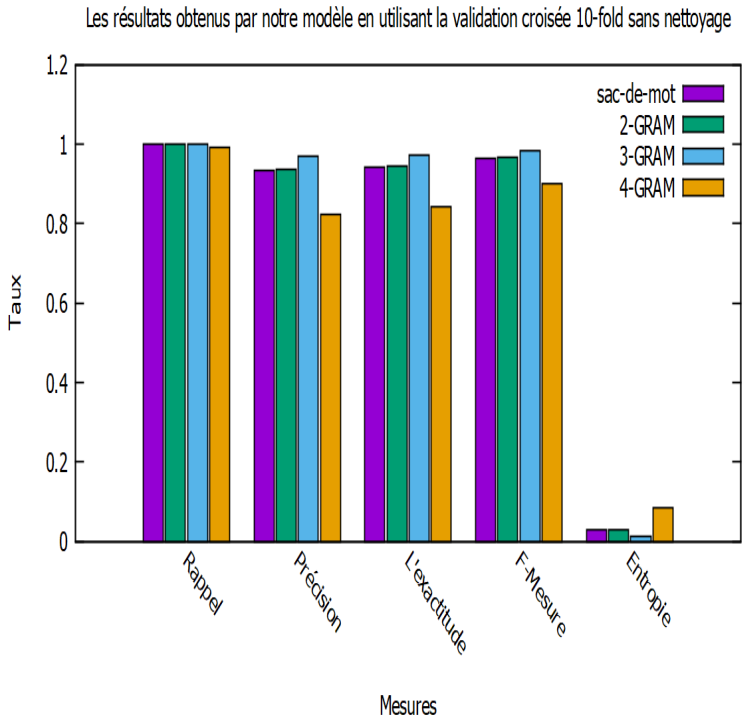


FIGURE 4.3: Illustration des résultats obtenus par notre modèle en utilisant une validation croisée 10-fold sans nettoyer les mots vides

Sur le tableau 4.2, nous avions extrait des vecteurs tf-idf sans supprimer les mots vides du texte (sans nettoyage), et aussi avec la suppression des mots vides (avec nettoyage), et nous avions utilisé deux méthodes de représentation textuelle, la première étant un sac de mots, et la deuxième est n-gram de mots avec n = 2,

n = 3 et n = 4. Pour les lignes nous avions calculé différentes mesures de validation pour évaluer notre modèle, et pour les colonnes nous avions les méthodes de représentation textuelle. La détection des spams SMS est un problème important à résoudre, nous avions proposé un algorithme pour y contribuer. Dans notre travail on a utilisé notre algorithme proposé avec deux opérations sans nettoyer les mots vides et avec le nettoyage des mots vides car parfois les mots vides (comme : les chiffres, les dates, montants...etc.) peuvent faire une différence et être très influentes pour déterminer le type de message, soit spam ou message normal.
Après l'illustration des résultats obtenus dans le tableau 4.2 et la figure 4.3 par notre modèle en utilisant une validation croisée de 10-fold sans nettoyer les mots vides, nous avons remarqué que les résultats de notre algorithme étaient meilleurs avec 3-gram. Notre algorithme a construit une bonne probabilité de distribution pour les trois mots suivants et nous a donné une bonne classification des messages normaux. Il a donné un taux de précision de 0,969 qui est une précision élevée pour un détecteur de spam, cette précision garantie à l'utilisateur de messagerie de ne pas perdre des e-mails importants et pourrait profiter de tous les e-mails envoyés à lui. Notre algorithme a donné un taux de succès (L'exactitude) avec 0.946 et une f-mesure avec 0.984 et un rappel avec 1.000 ce qui donne une excellente détection des messages étiquetés comme spam à partir de la base de test, et cela nous a donné une petite entropie avec 0.014 donc une petite perte de la quantité d'informations et un grand avantage des messages.

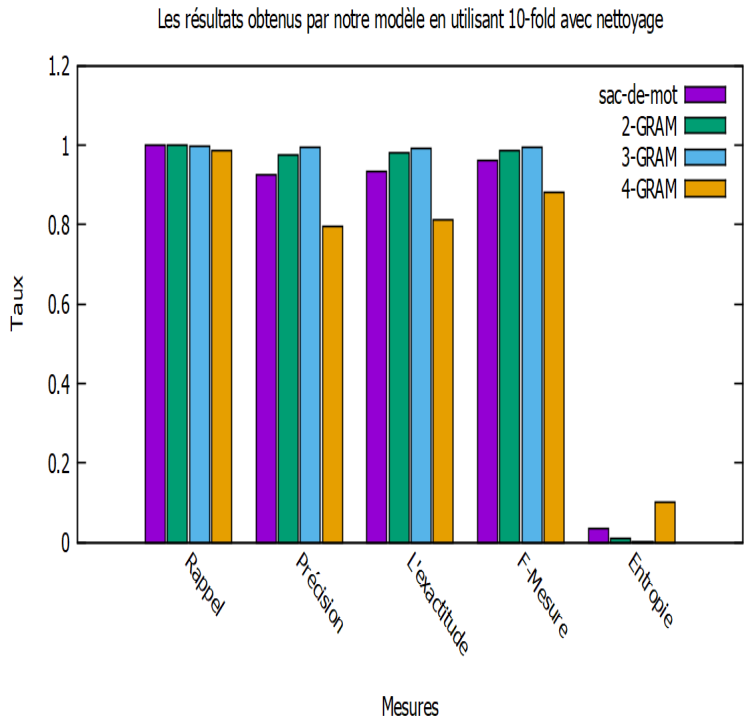


FIGURE 4.4: Illustration des résultats obtenus par notre modèle en utilisant une validation croisée 10-fold avec le nettoyage des mots vides

Après l'illustration des résultats obtenus dans le tableau 4.2 et la figure 4.4 par

notre modèle utilisant une validation croisée de 10-fold avec le nettoyage des mots vides, nous avons remarqué que les résultats de notre algorithme étaient meilleurs avec 3-gram. Notre algorithme a construit une bonne probabilité de distribution pour les trois mots suivants du jeu de données, il a donné le meilleur taux de précision avec 0,994, cette précision est élevée pour le modèle de détection de spam afin que l'utilisateur de messagerie ne perd pas de courriels importants et puisse profiter des courriels, notre modèle a donné aussi un taux de succès avec 0,993 et un rappel avec 0,998 donc une bonne détection des messages étiquetés comme spam à partir de la base de test et cela nous a donné une petite entropie avec 0,003 donc une petite quantité de perte d'informations et un grand avantage des messages.
En analysant les résultats, nous avons remarqué que les résultats de notre approche étaient meilleurs dans le cas du nettoyage des mots vides. Après avoir supprimé les mots vides et convertit les données textuelles en chiffres à l'aide de tf-idf, notre modèle a géré les données numériques, dans ce cas les mots vides n'avaient pas fait de différence, mais avec le nettoyage de ces mots vides, notre algorithme nous a donné de meilleurs résultats. Nous avions observé que les valeurs des mesures augmentent, notre modèle a donnés une haute performance de filtrage des spams et des messages légitimes, il a détecté ces messages d'anomalies pour bénéficier des messages légitimes, notre détecteur des spams aide les utilisateurs des boîtes mails électroniques à sélectionner les spams et les supprimer pour protéger les données des utilisateurs qui circulent dans cette plateforme. Comme nous l'avons vu sur le tableau 4.2, notre modèle a donné les meilleurs résultats avec 3-gram dans le cas de la suppression des mots vides (avec nettoyage), il a donné une performance plus élevée dans le cas de l'obtention de la fonction de vraisemblance de l'apparition des trois mots suivants du jeu de données d'apprentissage et de test. Notre modèle a géré les données numériques et nous a donné des bons résultats, par exemple nous avions obtenu une précision avec 0,994 (99,40%) et un taux de succès avec 0,993 et un f-mesure avec 0,996, et une entropie avec 0,003 qui a minimisé la perte d'information.
Nous avions sélectionné le meilleur résultat obtenu par notre modèle et fait une étude comparative avec des différents algorithmes utilisés par d'autres auteurs dans leur récente enquête. Les résultats sont illustré dans le tableau 4.3.

Mesures / Algorithmes	Rappel	Précision	Accuracy	F-mesure	Entropie
Our Approach Octopod	**0.998**	**0.994**	**0.993**	**0.996**	**0.003**
FP-Growth with NB minsup = 3% [9]	0.882	0.969	0.980	0.925	0.014
FP-Growth with NB minsup = 6% [9]	0.924	0.962	0.984	0.944	0.017
FP-Growth with NB minsup = 9% [9]	0.934	0.954	0.985	0.945	0.020
NMF [93]	0.960	0.960	0.916	0.920	0.018
LDA [93]	0.960	0.960	0.904	0.920	0.018
AdaBoostM1* [95]	0.905	0.912	0.905	0.883	0.040
Decision Table [95]	0.958	0.958	0.958	0.956	0.019
J48 (D-tree) [95]	0.966	0.965	0.966	0.965	0.015
Random Forest [95]	0.972	0.973	0.971	0.970	0.012
K-NN [95]	0.885	0.899	0.885	0.846	0.046
K-Star [95]	0.959	0.961	0.959	0.956	0.017
Naïve Bayes [95]	0.975	0.975	0.975	0.975	0.011
NB Multinomial [95]	0.986	0.986	0.985	0.986	0.006
DMNBtext [95]	0.986	0.986	0.985	0.985	0.006
SVM [95]	0.986	0.986	0.986	0.986	0.006
SGD [95]	0.984	0.984	0.983	0.984	0.007
VotedPerceptron [95]	0.983	0.983	0.983	0.983	0.007

TABLE 4.3: Étude comparative de notre modèle avec différents algorithmes utilisés par d'autres auteurs

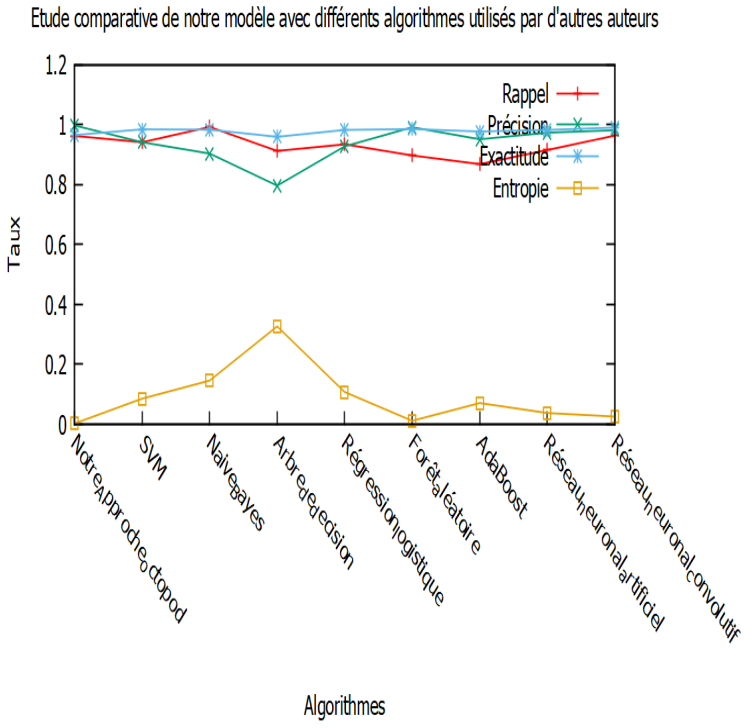


FIGURE 4.5: Étude comparative de notre modèle avec différents algorithmes utilisés par d'autres auteurs

Sur le tableau 4.3 et la figure 4.5, nous avions effectué une étude comparative entre notre modèle et des différents algorithmes utilisés dans d'autres études. Sur les lignes du tableau nous avions cité les algorithmes et sur les colonnes nous avions cité les différentes mesures pour évaluer les résultats de chaque algorithme. Après l'illustration des résultats, notre modèle a donné des meilleurs résultats par rapport aux autres études, commençant par le rappel, notre algorithme a donné un meilleur rappel avec 0,998 et une précision avec 0,994, avec une bonne entropie de 0,003, et un taux de succès (Accuracy) avec 0,993 et un F-mesure avec 0,996. Par rapport aux autres approches, théoriquement notre approche présentait une haute performance en matière de filtrage et de détection des sms spams, nous avions montré que la probabilité de perdre un message normal et l'avait lu comme spam n'existait presque pas, ainsi que le taux de détection des messages légitimes en tant que normal et la protection de ces messages est plus élevée que les autres algorithmes, et la capacité de détection des sms spams à partir de la base de test est plus grande. Hassan Najadat a démontré dans son travail que SVM lui a donné la plus grande précision obtenue jusqu'à présent dans les travaux déjà effectués pour cet ensemble de données (SMS spam collection) avec 98,6% [95], mais notre algorithme nous a donné la meilleure précision jusqu'à présent avec 99,30%.
En général, les expériences ont révélé la supériorité de notre modèle proposé par rapport aux autres études pour la détection du spam. Notre algorithme présentait une bonne qualité de détection et pourrait être très utile en cas d'utilisation de systèmes distribués parallèles, il offre une bonne solution pour les grands problèmes afin de garantir une réponse dans un temps acceptable. Ce modèle proposé pouvait apprendre et prendre des décisions satisfaisantes, et il pouvait être adapté à plusieurs d'autres problèmes pour les résoudre et obtenir des décisions et acquérir des connaissances appropriées.

4.1.5 conclusion

Le courrier électronique est un service important d'échange d'informations pour les internautes, l'utilisation importante de ces courriers en a fait la cible de diverses perturbations sur sa tête les attaques des spams, la plupart du temps sont générés par la publicité, il est donc nécessaire de protéger les e-mails et les messages contre ces spams. Cela a motivé les chercheurs à fournir des nouvelles solutions qui peuvent filtrer et détecter les spams basés sur des techniques heuristiques représentées par l'inspiration de la nature (ou bio-inspiration). La bio-inspiration peut être considérée comme une nouvelle discipline qui étudie les meilleures idées de la nature, puis les imite et applique leurs concepts et processus aux problèmes humains pour les résoudre. Après avoir exploré le monde marin, nous avons opté pour les octopodes a cause de sa haute qualité de défense, et même qu'ils sont peu utilisés dans le monde de la bio-inspiration. Les octopodes sont l'un des animaux les plus défensifs dans leur vie, ils ont une grande capacité pour éviter les attaques des prédateurs, ils créent un système de sécurité intelligent qui est difficile à briser. Dans notre travail, nous avions proposé une technique basée sur la fonction naturelle de l'octopode dans le but de détecter les spams, la technique est basée sur deux fonctions objectives, la première consiste à calculer la force de déplacement de chaque message et la deuxième cherche à calculer la probabilité de messages de

chaque classe de la base d'apprentissage.
Dans notre travail de détection des spams en se basant sur le fonctionnement des pieuvres, on a présenté la détection des spams en général et mentionner quelques études existantes, ensuite on a défini l'approche naturelle des pieuvres et les fonctions du physique utilisé dans notre modèle artificiel et même le paramétrage de ces fonctions, dans la suite on a présenté notre modèle artificiel et le passage du modèle naturel vers le modèle artificiel dans un tableau, et on a détaillé l'adaptation du problème des spams SMS aux comportements des octopodes, et on a aussi défini l'approche artificielle par une illustration et un algorithme.
Dans ce travail, nous avions proposé une technique basée sur la fonction naturelle de l'octopode pour détecter les spams, nous avions évoqué les problèmes des spams et montré quelques travaux réalisés dans ce domaine, ensuite nous avions utilisé la validation croisée avec 10-fold pour générer les bases d'apprentissage et de test à partir du jeu de données, après l'obtention des deux bases on a utilisé le texte avec le nettoyage des mots vides et sans les nettoyer avec deux techniques de représentation de texte sac de mots et n-gram avec n = 2, n = 3 et n = 4 de mots. Dans la suite on a également calculé les différentes mesures de validation de notre algorithme et faire une étude comparative de nos résultats avec des différents algorithmes utilisés par d'autres auteurs dans leurs récentes recherches pour évaluer notre technique. Le mécanisme de sécurité et de défense qui est un modèle prédateur-proie de l'octopode pour éviter les attaques des prédateurs à prouver qu'il peut également être un bon mécanisme de sécurité pour les problèmes humains, et en particulier pour détecter les spams SMS, le système pourrait être utilisé pour résoudre plusieurs problèmes. Pour nos travaux futurs, nous somme intéressés aux techniques inspirées de la nature comme nous l'avions fait dans ce travail, et on cherche à les appliquer dans le domaine de la sécurité informatique. Actuellement nous expérimentons des nouvelles techniques qui peuvent avoir un poids pour résoudre des problèmes humains.

4.2 une étude comparative pour la détection des applications Android malveillantes à l'aide des autorisations

4.2.1 Introduction et problématique

Les appareils mobiles offrent des meilleurs services que les autres appareils, ils ont connu un grand développement ces dernières années, ce qui a conduit à l'émergence des smartphones. Android est le système d'exploitation mobile le plus populaire installé sur des millions d'appareils [85], non seulement pour les mobiles, mais aussi pour les téléviseurs, et pour les montres intelligentes...etc. Il a acquis plus de 50% des ventes de smartphones au troisième trimestre 2011 [78], avec plus d'un milliard d'appareils activés par Android, et plus d'un milliard d'utilisateurs Android actifs chaque mois [133], tels que google Play pilote toute cette économie d'applications mobiles, avec plus de 50 milliards d'applications téléchargées, google Play a généré des revenus dépassant 5 milliards USD en 2013 [133]. Cette évolution s'est accompagnée d'une exploitation négative représentée dans les différents types d'attaques des pirates pour espionner les données des utilisateurs. Les pirates ont pris les limites des techniques de protection et des mécanismes de sécurité des appareils mobiles pour accéder aux données sensibles en utilisant des applications malveillantes sur la plateforme d'Android, pour des différents buts tels que voler le crédit téléphonique de l'utilisateur, accéder à certaines fonctionnalités de l'appareil et obtenir des informations personnelles comme des photos, des contacts ... Etc. Un récent rapport signale également qu'il y a «400% d'augmentation des logiciels malveillants du système Android depuis l'été 2010» [148]. La plate-forme Android a le taux de croissance le plus élevé des logiciels malveillants à la fin de 2011 [40], leurs appareils deviennent la cible de différents types d'attaques. En répondant à ces intrusions, les développeurs ont un grand défi contre ces pirates pour garantir la sécurité des appareils Android, de nombreuses solutions ont été proposées pour détecter toute tentative de violation des données, les chercheurs ont proposé un certain nombre de systèmes de détection d'applications Android malveillantes basées sur l'intelligence artificielle, ou basé sur les algorithmes de la fouille de données, d'autres chercheurs proposent des détecteurs basés sur des méthodes d'apprentissage profond en raison de leur grande capacité de précision, d'autres proposent des méthodes bio-inspirées. Depuis le lancement de la version 6.0 nommée **Marshmallow**, Android a utilisé le système des autorisations des applications, ces derniers demandent des autorisations d'accès aux données lorsqu'elles sont installées pour la première fois, chaque application demande des autorisations spécifiques. Certains pirates profitent de créer des applications illégitimes et demandent des accès via ces applications pour menacer la confidentialité des utilisateurs par voler de crédit, ou voler des mots de passe ou pour d'autre buts. Les chercheurs ont proposé beaucoup de recherches pour lutter contre ces actes malveillants.

Les réseaux de neurones récurrents de mémoire à cour terme sont des approches ont une capacité puissante et facile à adapter aux problèmes, ces techniques d'ap-

prentissage profond donnent des résultats optimaux pour les cas des données séquentielles, ils sont largement utilisés en traduction automatique, reconnaissance vocale, reconnaissance des formes, traitement automatique de langage (TAL)...etc. Les réseaux de neurones récurrents ont un potentiel plus grand que les réseaux de neurones classiques, par exemple pour une séquence de texte de taille fixe les RNNs ont une capacité de prédire le caractère suivant, ils sont faciles à adapter aux problèmes humains pour les résoudre, ces RNNs ont la possibilité de prendre en entrées les données de taille quelconque. Parmi ces RNNs on trouve les LSTMs (Long Short Terme Memory).
Dans notre travail, nous avions pris une base de données des applications Android basée sur les autorisations [37, 136], cette base contient environ 398 applications mélangées du type normal et malveillant, chaque application contient les autorisations demandées par elle [37, 136]. On a essayé d'entraîner des algorithmes de fouille de données tels que SVM, Naïf bayes, Gaussien naïf bayes, KNN avec k=5 et k=3, arbre de décision, classificateur Random forest, classificateur Extratree, Classificateur de renforcement de gradient, AdaBoost [80], et on a entraîné aussi le LSTM RNNs. Les résultats des détecteurs mentionnés ont été comparé et évalué pour trouver un meilleur détecteur des applications Android malveillant.

4.2.2 Le Système d'exploitation mobile Android

Android est un système d'exploitation mobile développé par Google, ce système est facile à utiliser et open source, il équipe plusieurs appareils tels que les téléphones portables, les smart watchs, les tablettes, les téléviseurs...etc. Android vous permet de télécharger des applications chaque jour pour naviguer sur votre appareil, chaque application a un fonctionnement spécifique par exemple des applications qui permet de naviguer sur l'internet, ou connecter avec vos amis sur les réseaux sociaux, ou les applications GPS (MAPS) pour se déplacer et trouver des emplacements...etc. Ces applications peuvent être développées par des ingénieurs de domaine et elles s'augmentent chaque jour. Cette croissance est suivie par des apparitions des nouvelles versions des systèmes Android chaque année pour s'actualiser et faire des mises à jours, et pour améliorer les services [91].
Un système d'exploitation mobile est une plateforme qui permet à une appareille de se fonctionner pour garantir à l'utilisateur de faire des différentes taches telles qu'un appel téléphonique, un message à envoyer, naviguer sur l'internet via un navigateur, télécharger des applications...etc [91].
Android a pris la place numéro une dans le monde des systèmes d'exploitation mobiles, de nombreuses marques de fabrication des appareilles mobiles l'utilisent comme Samsung, LG, Sony, Huawei, Xaomi...etc [91]. Les ingénieurs de développement des applications mobiles joignent ces applications dans des magasins telles que playstore de Google, les utilisateurs d'Android peuvent télécharger et installer les applications sur leurs mobiles à partir de ces magasins. Avec le grand développement dans le monde mobile les utilisateurs peuvent relier leur appareilles mobiles avec plusieurs plateformes tels que de stockage en nuage(Cloud), ou la messagerie électronique ... etc [25]. En 2020 Android est le système d'exploitation le plus utilisé dans les appareils mobiles avec 74.3% dans le monde, suivis par L'IOS d'Apple avec 24,8% [25]. Le système Android est représenté comme un

système d'exploitation en temps réel, il contient des différentes couches telles que la couche application, la couche application framework, la couche libraries, Linux kernel, Android RunTime [121].
La couche **"Applications"** contient toutes les applications de téléphone tel que le navigateur, les contacts, l'appareille photo, calendrier...etc. Elle représente la couche la plus haute [121].
La couche **"Application Framework"** en français cadre d'application, elle contient un ensemble de classes et de services qui permet aux applications d'accéder aux données. On peut trouver des services comme gestionnaire d'activité, gestionnaire de ressource, gestionnaire de notifications...etc [121].
La couche **"Libraries"** en français la couche bibliothèques, elle représente la couche logicielle basse, cette couche contient des différentes bibliothèques écrites en langage C/C++, bibliothèque web, bibliothèque des formats multimédia (image,audio, vidéo)...etc [121].
La couche **"Linux Kernel"** en français la couche Noyau Linux, elle est la couche la plus importante dans l'architecture Android, elle représente le cœur de cette architecture car elle fournit des services importants qui lisent le logiciel avec le matérielle, parmi ces services on trouve la sécurité, la gestion d'alimentation, gestion de la mémoire...etc [121].
La couche **"Android RunTime"** en français la couche d'exécution Android, elle contient la Machine Virtuelle Dalvik (DVM) et les bibliothèques principales [121].

4.2.3 Kit de développement ou SDK

SDK est une abréviation anglaise de Software Development Toolkit, en français kit de développement. Android kit de développement est une plateforme qui permet aux développeurs de programmer des applications Android mobile, elle contient un ensemble de projets, des codes sources, des API et des documentations. Les applications sont écrites en langage de programmation java et exécutées sur la machine virtuelle Dalvik. **Une application Android** est regroupée dans un package Android sous l'extension **.apk** [121, 114].

4.2.4 Bref Historique des versions Android

Android a été officiellement lancé en 2008 avec sa première version Android 1.0, cette version est ancienne, elle contenait des applications comme Gmail, calendrier et Youtube. la version 1.1 a été vu le jour après. En 2009 Android a lancé la version 1.5 nommée **Cupcake**, dans cette version beaucoup d'amélioration en matière d'interface a été introduite, et le clavier a été utilisé pour la première fois à l'écran. Dans la même année en 2009 Android a lancé la version 1.6 appelée **Donut**, cette version a ajouté quelques fonctionnalités telles que la capacité de système d'exploitation de fonctionner avec une variété des résolutions d'écran. Ensuite des versions Android 2.0 à 2.1 nommée **Eclair** ont été lancer, dans ces versions la synthèse vocale a été mise en fonction la première fois. la version 2.2 nommée **Froyo** a vu le jour quatre mois après la version 2.1, dans cette version Google a créé pour la première fois des actions via le vocal en prononçant des commandes à exécuter. En 2010 la version 2.3 nommée **Gingerbread** est née, la

couleur noire et verte du robot Android est utilisée dans l'interface des utilisateurs dans le SE. En 2011 les versions 3.0 à 3.2 nommées **Honeycomb** ont été lancer, ces versions étaient caractérisées par le concept d'une interface spécifique aux tablettes. Dans la même année la version 4.0 nommée **Ice Cream Sandwich** a été apparue, cette version a remis les buttons dans l'écran et l'utilisation des notifications. En 2012 et 2013 successivement deux version 4.1 à 4.3 nommées **Jelly Bean** a été apparu, ces versions se caractérisent par un système de recherche locale et ajoutent des widgets dans l'écran de verrouillage. En 2013 la version 4.4 appelée **KitKat** a été apparue, une barre d'état transparente et des icônes blanches et des arrières plans clairs ont pris leur place dans ce SE. Ensuite la version 5.0 nommée **Lollipop** a vu le jour en 2014, cette version a donné un nouveau style pour les applications et le SE, elle se caractérise aussi par l'utilisation des notifications dans l'écran de verrouillage. Après une année, la version 5.1 **Lollipop** a été apparue. En 2015 la version 6.0 appelée **Marshmallow** est née, cette version a introduit des nouvelles fonctionnalités telles que les autorisations des applications, l'empreinte digitale et l'utilisation d'USB-C. En 2016 les versions 7.0 et 7.1 nommée **Nougat** ont été lancer, ces versions ont ajouté des fonctionnalités telles que le mode d'écran partagée. Les versions 8.0 et 8.1 nommées **Oreo** a été apparu en 2017, la version Oreo a introduit la fonctionnalité d'application qui peut vous alerter et rappeler avec notification. En 2018 la version 9 nommée **Pie** est née, cette version contient beaucoup d'améliorations, elle introduit des différents systèmes intelligents comme la gestion de luminosité, économiseur de la batterie...etc. En septembre 2019 la version 10 a été apparue, cette version n'est pas nommée comme les versions précédentes, elle contient beaucoup d'améliorations, le système d'autorisations est mise à jour, ceci vous offre un bon contrôle sur les autorisations des applications sur vos données. La dernière version 11 est née en février 2020, elle n'a pas de nom comme la version 10, beaucoup d'améliorations étaient faites, dans cette version Google se localise sur la confidentialité, surtout sur la mise à jour des autorisations sur un accès limité sur l'appareille photo, le microphone et l'emplacement de l'appareil, seules les applications légitimes selon Google peuvent accéder aux emplacements des appareilles d'utilisateurs [114].

4.2.5 Les autorisations des applications Android

Android a utilisé le système des autorisations depuis le lancement de la version 6.0 nommée **Marshmallow**, les utilisateurs contrôlent les autorisations des applications. Une autorisation d'application signifie ce que votre application est autorisée à faire et à accéder, l'accès est aux données trouvées dans votre mobile par exemple l'accès aux contacts, aux fichiers multimédias, à l'emplacement, à l'appareil photo ou au microphone. Donner l'accès à une autorisation permet à une application d'utiliser les données de cette fonctionnalité, refuser l'accès permet d'empêcher d'utiliser les données. Les applications ne peuvent pas prendre des autorisations d'accès aux données de façon automatique, l'utilisateur doit accorder l'accès et approuver l'autorisation pour que les données soient utilisées par l'application. Une certaine ancienne application peut se bloquer en cas de manque des mises à jour, ou elle ne peut pas fonctionner de façon correcte à cause d'un refus des autorisations. Une application demande d'approuver ou refuser leurs autorisa-

tions l'or de la première fois de son lancement via une fenêtre qui contient les types d'autorisations. Le but d'utiliser les autorisations est de garantir la confidentialité des données des utilisateurs, donc refuser les permissions pour les applications douteuses est essentielle pour protéger vos données contre les attaques des applications malveillantes [135].
Il faut toujours lire les définitions des applications avant de les installer en consultant la description dans Play Store. Il faut toujours connaître les permissions nécessaires pour chaque application, par exemple une application de messagerie nécessite des autorisations des contacts et SMS, mais elle n'est pas besoin des autorisations d'accès aux données de santé [135].

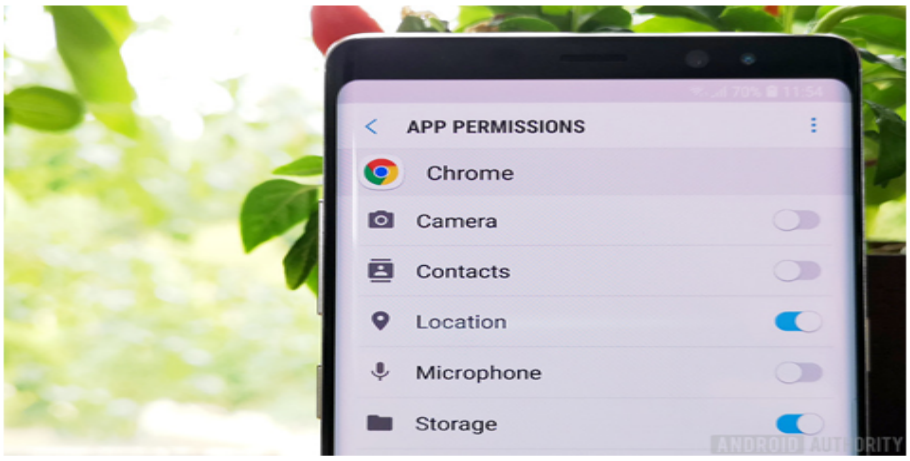


FIGURE 4.6: Exemple des Permissions pour l'application Google chrome

quelques autorisations

— **Capteurs corporels** : cette permission permet d'accéder aux données de santé, par exemple : le nombre de pas, les fréquences cardiaques, capteur de suivi de la condition physique et d'autres capteurs [135].

— **Calendrier** : cette permission permet de modifier, créer, lire ou supprimer un événement dans le calendrier d'utilisateur [135].

— **Appareil photo** : cette permission permet de prendre des photos et enregistrer des vidéos [135].

— **Contacts** : cette permission permet d'accéder à la liste des contacts et modifier, ou ajouter, ou supprimer un contact [135].

— **Localisation** : cette permission permet d'accéder à votre position à l'aide du GPS pour une grande précision de détection, ou à l'aide du WI-FI ou données cellulaires pour une détection approximative [135].

— **Microphone** : cette permission permet d'enregistrer un audio (mémo vocal) [135].

— **Téléphone** : cette permission permet l'accès aux données de numéro de téléphone, elle permet de rediriger les appels, modifier les journaux d'appels, l'accès à la messagerie vocale...etc [135].

— **SMS** : cette permission permet de lire, écrire ou envoyer des messages SMS ou MMS [135].

— **Stockage** : cette permission permet de lire, écrire des fichiers dans l'espace du stockage interne ou externe dans l'appareil mobile [135].

4.2.6 Détection des applications Android malveillantes

La plateforme android a évolué rapidement, elle a été le système d'exploitation le plus populaire et le plus utilisé dans de nombreux appareils, cette croissance rapide en a fait la cible de nombreuses applications malveillantes qui visent à voler des informations et des données sensibles. De nombreuses recherches ont été faites pour détecter ces intrusions, Justin Sahs a présenté un système qui utilise l'apprentissage automatique pour la détection des applications malveillantes sur les appareils Android, son système extrait un certain nombre de caractéristiques et entraîne un SVM à une classe de manière hors ligne (hors périphérique) afin de tirer parti de la puissance de calcul plus élevée d'un serveur ou d'un cluster de serveurs [124]. Asaf Shabtai a présenté un cadre de détection des applications malveillantes sur les appareils mobile Android, son cadre proposé réalise un système de détection de logiciels malveillants basés sur l'hôte qui surveille en permanence diverses fonctionnalités et événements obtenus à partir de l'appareil mobile, puis il applique des détecteurs d'anomalies d'apprentissage automatique pour classer les données collectées comme normales (bénignes) ou anormales (malveillantes) [118]. Naser Peiravian a proposé de combiner les autorisations et les appels API, il a utilisé des méthodes d'apprentissage automatique pour détecter les applications android malveillantes [109]. Zhenlong Yuan a proposé une méthode basée sur L'AA qui utilise plus de 200 caractéristiques extraites à la fois de l'analyse statique et de l'analyse dynamique de l'application Android pour les logiciels malveillants. La comparaison des résultats de la modélisation démontre que la technique d'apprentissage profond est particulièrement adaptée à la détection des logiciels android malveillants et peut atteindre un niveau élevé de taux de succès 96% avec des ensembles d'applications Android réelles [145]. Gianluca Dini a décrit un détecteur d'anomalies à plusieurs niveaux pour android. Ce détecteur surveille simultanément android au niveau du noyau et au niveau de l'utilisateur pour détecter de véritables infections de logiciels malveillants en utilisant des techniques d'apprentissage automatiques pour distinguer les comportements standard des comportements malveillants [40]. Jaemin Jung a proposé une méthode d'apprentissage automatique de détection de logiciels malveillants qui identifie le sous-ensemble d'API Android qui est efficace en tant que caractéristiques, et classent les applications android comme applications bénignes ou malveillantes. La méthodologie proposée construit d'abord deux listes d'API android populaires, l'une contient des applications bénignes et l'autre contient des applications malveillantes, ensuite il applique l'algorithme Random Forest sur un ensemble de données en utilisant chaque liste comme caractéristiques du classificateur pour évaluer la méthodologie proposée [71]. Shaikh Bushra Almin a proposé un système pour détecter et supprimer les malwares présents dans l'appareil Android de l'utilisateur [19]. Chenglin Li a proposé un classificateur nouveau

et très fiable pour la détection des malwares android basés sur l'architecture de la machine de factorisation et l'extraction des caractéristiques des applications Android à partir des fichiers manifestes et du code source [27]. Hossein Fereidooni a proposé un système pour détecter les applications Android malveillantes en analysant statiquement les comportements des applications, il offre une couverture plus complète des comportements de sécurité, ensuite il a construit un cadre de détection basé sur l'apprentissage automatique avec une détection de haute performance et un taux de faux positifs acceptable [63]. Tieming Chen a proposé un nouveau modèle de détection statique léger(TinyDroid) en utilisant la simplification des instructions et la technique d'apprentissage automatique, et un classificateur est entraîné pour les tâches de détection et de classification des logiciels malveillants [130].

4.2.7 Notre Contribution

Dans notre travail, on a analysé et détecté les logiciels malveillants en utilisant des algorithmes de fouille de données d'une part et un algorithme d'apprentissage profond d'autre part, ensuite on a comparé les résultats obtenus pour montrer le meilleur détecteur des applications malveillantes en fonction des autorisations. Si une appareille utilise la version Android 6.0 ou une version supérieure, les applications installées dans ces systèmes demandent des accès d'autorisation, l'utilisateur doit autoriser ou refuser les autorisations d'accès aux données sensibles sur le mobile, le but d'une autorisation est de protéger la confidentialité d'un utilisateur Android.
Les étapes de processus de détection des applications malveillantes sont illustrées dans la figure 4.7.

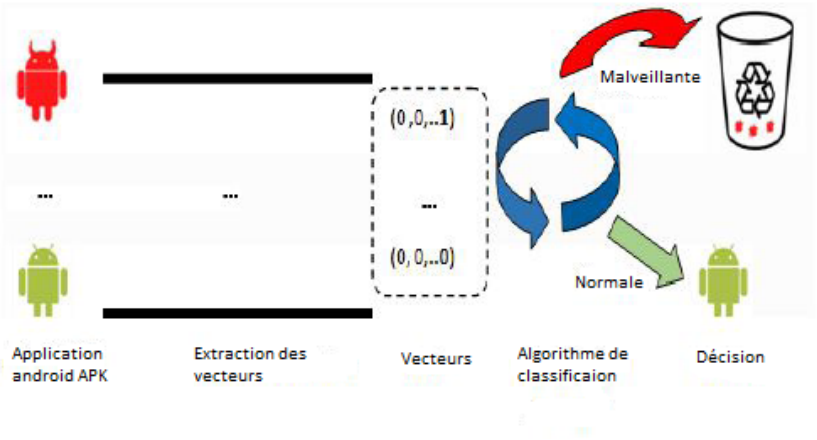


FIGURE 4.7: Processus de détection des applications malveillantes

Dans notre travail, on a entraîné un ensemble d'algorithmes de la fouille de données tel que : SVM, Naïf bayes, Gaussien naïf bayes, KNN avec k=5 et k=3, arbre de décision, classificateur Random forest, classificateur Extratree, classificateur de renforcement de gradient, AdaBoost [80], et on a aussi entraîné un RNN

LSTM, tous ces algorithmes sont utilisés pour la classification des applications Android en se basant sur les autorisations. Pour l'apprentissage on a utilisé une base de données qui contient environ 398 applications des deux types, soit une application normale, soit une application malveillante. Chaque application veut accéder à un ensemble des autorisations. Les algorithmes déjà cités font une classification binaire pour obtenir le type d'application qui arrive à partir de la base de test (normale ou malveillante). À la fin les résultats obtenus sont soumis à une étude comparative pour voir le meilleur classificateur qui peut être le meilleur détecteur des applications Android dans notre cas.

Dataset

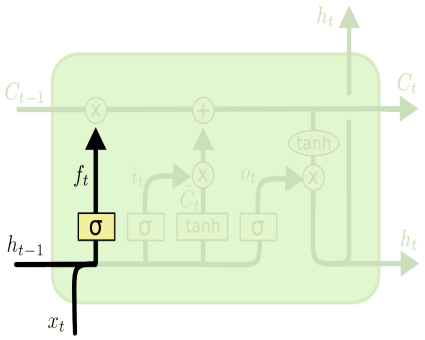
Le corpus utilisé dans notre travail est nommé **Dataset malware/beningn permissions Android**, il contient environ 398 applications mélangées du type normal ou malveillant. Le corpus contient 199 applications normales et 199 applications malveillantes, les applications malware prennent la valeur "1", et les applications normales prennent la valeur "0". Ce corpus est créé par Urcuqui Christian grâce à ses recherches en apprentissage automatique et en sécurité des systèmes Android [37, 136]. Le dataset contient 331 colonnes la dernière contient le type d'application Android, les autres 330 colonnes contiennent des différents types d'autorisations [37, 136].

LSTM RNNs

Récemment Les RNNs ont été utilisés beaucoup pour des différents problèmes tels que la reconnaissance vocale, la traduction automatique, TAL...etc. Ils ont donné des résultats incroyables et performants pour résoudre plusieurs problèmes artificiels, parmi ces RNNs on trouve le LSTM (Long Short Terme Memory). Le LSTM est un type particulier des RNNS, cet algorithme est inventé en 1997 par Hochreiter et Schmidhuber, ensuite il a été popularisé par d'autres chercheurs, il donne des résultats meilleurs pour le cas des données en séquence. Parfois pour résoudre une tache courante on a besoin de voir des informations récentes, et même pour les autres problèmes le LSTM a montré son succès dans ses taches de classification, il est largement utilisé et il fonctionne sur une variété des problèmes [30, 4, 61].
Les LSTMs ont une structure simple qui peut être répétitive, la structure est représentée dans la suite. Les LSTMs contiennent des blocs mémoires, ces derniers contiennent des cellules de mémoire et des portes pour contrôler les flux d'informations, chaque bloc contient une porte d'entrée et une porte de sortie, une autre porte appelée la porte d'oubli est ajoutée au bloc mémoire [30, 4, 61].
Le principe d'algorithme LSTM est lié à un ensemble d'états, chaque état a un rôle spécifique, l'état caché (Hidden state) qui prend l'information importante a court terme, cet état joue un rôle de mémoire à court terme. L'état de la cellule qui prend l'information qui est importante à long terme, cet état joue un rôle de mémoire à long terme. L'idée de LSTM en premier lieu est de décider quelle information à oublier et jeter de l'état de cellule, cette tache est faite par la porte d'oubli (forget gate), un élément courant de la séquence x_t et un état caché de la cellule précédente h_{t-1} est examiné, un nombre est généré entre 0 et 1 pour

chaque état de la cellule C_{t-1}, un 0 signifie se débarrasser complètement de cela, un 1 signifie garder complètement cela [30, 4, 61].

$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f) \tag{4.7}$$



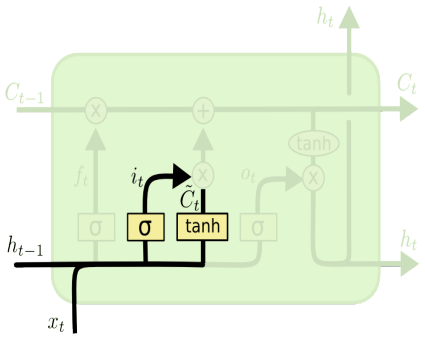
$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

FIGURE 4.8: Calcule de f_t

Dans la suite, l'algorithme choisit quelles nouvelles informations veulent stocker dans l'état de la cellule, la porte d'entrée (input gate) décide de mettre à jour des valeurs, la couche tanh crée un vecteur de la nouvelle valeur candidate $\tilde{C}_t$ qui pourrait être ajouté à l'état, ensuite les deux éléments seront combinés pour mettre à jour l'état, voir 4.9 [30, 4, 61].

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \tag{4.8}$$

$$\tilde{C}_t = tanh(W_C.[h_{t-1}, x_t] + b_C) \tag{4.9}$$

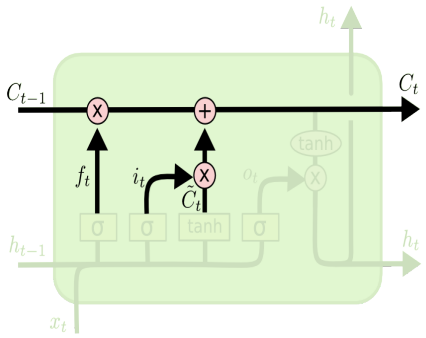


$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

FIGURE 4.9: Calcule de i_t et $\tilde{C}_t$

Dans cette étape, l'ancien état de cellule C_{t-1} est mis à jour dans le nouvel état de cellule C_t, en oubliant les informations que l'algorithme a choisies d'oublier, l'ancien état C_{t-1} est multiplié par f_t, en ajoutant $i_t * \tilde{C}_t$, dans la fin d'opération l'algorithme trouve les nouvelles valeurs candidates, voir 4.10 [30, 4, 61].

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{4.10}$$



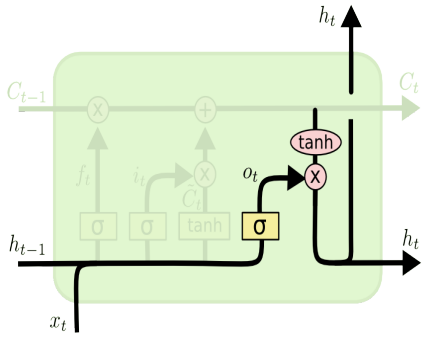
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

FIGURE 4.10: Calcule de C_t

Dans la fin l'algorithme produit une sortie basée sur l'état de la cellule, une couche d'activation sigmoïde génère les parties de l'état de cellule, ensuite l'état de cellule est mis en couche tanh pour localiser les valeurs entre 1 et -1, le résultat est multiplié par la sortie de la couche sigmoïde, voir 4.11 [30, 4, 61].

$$o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o) \tag{4.11}$$

$$h_t = o_t * tanh(C_t) \tag{4.12}$$



$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

FIGURE 4.11: Calcule de o_t et h_t

4.2.8 Expérimentation et résultats

La plateforme Android a évolué rapidement, elle a été le système d'exploitation le plus populaire et le plus utilisé dans de nombreux appareils, cette croissance rapide en a fait la cible de nombreuses applications malveillantes qui visent à voler des données sensibles, de nombreuses recherches ont été faites pour détecter ces intrusions. Dans ce travail, nous voulons analyser et détecter les logiciels malveillants en utilisant des algorithmes de fouilles de données et un algorithme d'apprentissage profond, ensuite les résultats obtenus sont comparés pour montrer le meilleur détecteur d'application Android malveillante en fonction des autorisations.

Dans notre travail, nous avions pris l'ensemble de données nommé Android malware/benign permissions, ce jeu de données est le résultat d'une recherche en apprentissage automatique et en sécurité d'android [37, 136]. Les données ont été

Mesures / Algorithmes	Rappel	Précision	Accuracy	F-mesure	Entropie
SVM	0.882	0.909	0.895	0.908	0.049
Naïve Bayes	0.881	0.788	0.841	0.832	0.104
Gaussian Naïve Bayes	0.883	0.791	0.842	0.835	0.102
5-NN	0.861	**0.939**	0.894	0.899	**0.027**
3-NN	0.884	0.924	0.902	0.904	0.034
Decision Tree	0.897	0.924	0.909	0.910	0.034
Random forest Classifier	0.882	0.896	0.888	0.889	0.048
ExtraTree Classifier	0.912	**0.939**	**0.925**	**0.925**	**0.027**
Gradient Boosting Classifier	**0.938**	0.909	0.924	0.923	0.041
AdaBoost	0.922	0.894	0.909	0.908	0.049
LSTM (RNN)	0.882	0.909	0.895	0.896	0.041

TABLE 4.4: Résultats de classification obtenus par les algorithmes utilisés

obtenues par un processus qui consistait à créer un vecteur binaire des autorisations utilisées pour chaque application analysée malware / bénigne divisée par "type" 1 malware et 0 non-malware. Ce jeu de données contient 398 applications réparties à parts égales par 199 applications pour malware, et 199 pour applications bénignes. Le corpus contient 331 attributs, la dernière colonne contient le type de l'application soit malware(Malveillante) ou bénigne(normale) [37, 136].
Nous avions pris des différents classificateurs pour détecter les applications malveillantes, et nous avions utilisé la validation croisée avec 3-folds pour obtenir les données d'apprentissage et les données de test, ensuite les algorithmes utilisés faisaient la classification et obtenaient le type d'application soit malveillante ou une application normale. Les résultats obtenus sont illustrés dans le tableau 4.4 ci-dessous. Le LSTM RNN utilisé dans notre travail est pris à partir de la bibliothèque "Keras" en utilisant l'environnement python [75], et les autres algorithmes mentionnés ont été utilisés à partir de la bibliothèque Scikit-Learn également de l'environnement python [80]. Scikit-Learn fournit un accès facile à de nombreux algorithmes de classification différents, dans notre cas nous avions utilisé les algorithmes suivants : SVM, Naïve Bayes, Gaussian Naïve Bayes, KNN avec k = 5 et k = 3, Decision Tree, Random Forest Classifier, ExtraTree Classifier, Gradient Boosting classifier, ADA Boost.

Paramètres des algorithmes utilisés

1) Gradient boosting classifier (n_estimators=100, learning_rate=1.0,max_depth=1, random_state=0)
2) ExtraTreesClassifier (n_estimators=10, max_depth=None,min_samples_split=2, random_state=0)
3) LSTM RNN (function d'activation= sigmoid, droupout=0.5,hidden layer=64, loss='binary_crossentropy', optimizer='adam', epochs=500, batch_size=1,verbose=0)

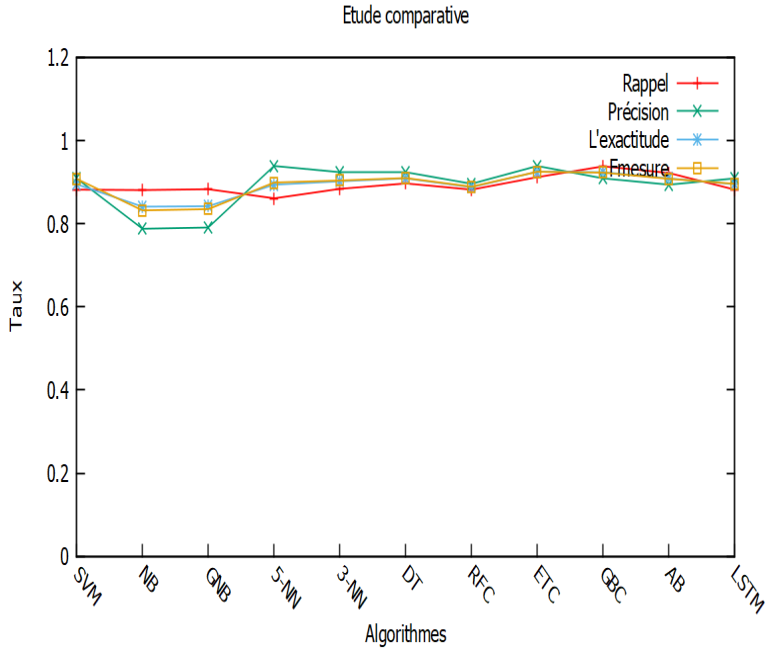


FIGURE 4.12: Résultats de classification obtenus par les algorithmes utilisés

Comme nous l'avons vu sur le tableau 4.4 et la figure 4.12, pour le SVM, l'algorithme n'a pas donné les meilleurs résultats, il incluait le fait que l'algorithme est sujet à un sur-ajustement car le nombre de fonctionnalités est beaucoup plus proche du nombre d'échantillons, pour le naive bayes et le gaussienne naïve bayes, l'indépendance des attributs fait perdre la capacité d'exploiter les interactions entre les caractéristiques de sorte qu'elles affectent la tâche de classification, pour le KNN, il faut identifier la valeur de k, dans notre travail nous avons testé cet algorithme avec k = 3 et k = 5 mais les résultats n'étaient pas assez bons, dans ce cas nous avons dû à sélectionner la métrique et les attributs à utiliser par l'algorithme, pour le Random forest les résultats ne sont pas assez bons en raison du nombre élevé des échantillons, ceux-ci n'amélioreront pas la précision. Le LSTM n'a pas donné les meilleurs résultats, il est un excellent outil pour les données en séquences, mais pas dans ce cas car l'algorithme a une mémoire qui enregistre les actions passées, pour le cas des attributs qui dépendent l'une des autres le LSTM donne des meilleurs résultats, pour l'algorithme adaboost les résultats sont bons. Nous avions observé que le classificateur ExtraTree a donné les meilleurs résultats en matière de précision avec 93,90%, et en terme taux de succès (Accuracy) avec 92,50%, et une F-mesure avec 92,50%, et une entropie avec 0,027 (2.7%), l'algorithme ExtraTree est un type de technique d'ensemble d'apprentissage qui agrège les résultats de plusieurs arbres de décisions décorrélées collecter dans une «forêt» pour produire son résultat de classification, il randomise certaines décisions et certains sous-ensembles de données pour minimiser le sur-apprentissage à partir des données, et pour améliorer la précision prédictive et contrôler le sur-ajustement. Le classificateur gradient Boosting a donné un meilleur rappel avec 93,80%, il construit des arbres un à la fois, où chaque nouvel arbre aide à corriger les erreurs faites par un arbre précédemment formé, avec chaque arbre ajouté, le

modèle devient encore plus expressif et donne un meilleur rappel.

4.2.9 Conclusion

Récemment les systèmes d'exploitation Android sont largement utilisés dans des différents appareilles, chaque application installée dans ces appareils a un but spécifique d'utilisation. Depuis le lancement de la version android 6.0, le système des autorisations des applications est né, une autorisation d'accès est demandé l'or de l'installation d'une application pour la première fois. Selon la large utilisation des systèmes d'exploitation Android et le nombre important d'appareils qui fonctionnent avec ce système, ils ont été ciblée par des différents types d'accès malveillants, de nombreuses tentatives de violation ont semblé voler et pirater les données sensibles des utilisateurs via les autorisations des applications, certains pirates cherchent à atteindre leurs buts via ces accès illégitimes. Les chercheurs ont été motivés pour lutter contre ces attaques des pirates et garantir le bon fonctionnement des systèmes android par présenter des différents détecteurs en répondants à ces tentatives pour les dissuader. De nombreuses recherches ont été effectuées par de nombreux auteurs, certains sont cités dans ce travail, et en tant que continuation de ces recherches on a aussi présenté des détecteurs des applications malveillantes dans ce travail. Dans notre travail de détection des applications Android malveillantes, on a commencé par la définition des SE Android et on a vu les différentes versions d'Android, ensuite on a défini les autorisations et mentionné les différents détecteurs utilisés pour la classification des applications Android (des algorithmes d'exploration de données et un algorithme d'apprentissage profond LSTM RNN), dans la fin les résultats trouvés sont mis à une étude comparative, dans notre cas le classificateur Extra Trees a donné une meilleure accuracy (taux de succès) avec 92,50%. Les recherches sur ce domaine ne sont pas terminées, il faut trouver des solutions plus efficaces pour stopper ces tentatives d'intrusion.

CONCLUSION GÉNÉRALE ET PERSPECTIVES

Conclusion générale et perspectives

Nos travaux réalisés dans cette thèse touchent de façon générale la sécurité des données. Notre contribution en général dans ce travail est de lutter contre les tentatives des intrus qui vise les courriels électroniques et les systèmes d'exploitation android. On a proposé des filtres pour détecter les tentatives des attaques des intrus et augmenter le taux de sécurité. Nos études sont fortement localisé sur le domaine de la sécurité des messageries électroniques d'une part et sur les systèmes android par la détection des applications malveillantes d'autre part. Notre main objective dans cette thèse est qu'on a essayé de proposer des détecteurs qui aident à améliorer les résultats des travaux déjà faites. D'une part on a essayé d'inspirer et d'utiliser des nouvelles techniques biomimétiques, et les adapter comme des filtres des e-mails qui circulent dans les messageries électroniques afin de lutter contre les attaques spams, et ensuite on a évalué les résultats obtenus par notre détecteur par rapport aux détecteurs existants. D'autre part on a essayé d'utiliser les réseaux de neurones récurrents LSTM et les adapter dans le domaine de la sécurité des données comme des détecteurs des applications android malveillantes, afin de lutter contre ce type d'attaque des pirates et identifier ces applications intruses.

Notre première contribution dans cette thèse est de proposer une nouvelle technique bio-inspirée basée sur les octopodes pour le filtrage des spams, dans ce travail on a présenté une nouvelle technique qui est une technique heuristique inspirée du fonctionnement biologique des octopodes. La nature est une source vaste des idées et des inspirations innombrables d'où les causes et les modalités sont très variées. La bio-inspiration permet d'imiter des procédés et matériaux présentés dans la nature, ou des formes et structures qui se trouve dans la nature, ou des fonctionnements et des interactions des êtres vivants qui forment les écosystèmes et les espèces dans la nature, afin d'utiliser toutes ces fonctionnements et les adapter pour trouver des solutions à des problèmes humains artificiels. Depuis long temps l'être humain retourne à la nature pour trouver des réponses à ses questions, donc on peut gagner des millions des idées qui se trouvent dans la nature, il faut seulement identifier nos besoins. Après mener avoir fait plusieurs recherches, nous avons

opté d'explorer le monde marin, précisément nous avons opté pour le fonctionnement des octopodes, ces derniers nous ont attirés par sa grande qualité de défense contre les attaques des prédateurs et ses actions naturelles qui peuvent forment un modèle prédateur-proie pour se protéger. Ce modèle a été imiter pour être adaptable afin de lutter contre les attaques spams. Cette technique est nouvelle, elle cherche à détecter et filtrer les courriers électroniques afin de les classer soit des e-mails normaux, ou soit des spams. Cette technique proposée est basée sur deux actes naturels de l'octopode qui les réagit quand il est attaqué, le premier est de faire un jet propulsion pour s'échapper rapidement, et le deuxième acte est de libérer une quantité d'encre noire pour désorienter les prédateurs, ces deux actes sont imités, et convertis dans notre modèle par une fonction de force pour mettre un objet en mouvement, et la probabilité de chaque classe à partir de la base d'apprentissage, d'où chaque message de la base de test représente un octopode, l'idée principale de notre modèle est que si un octopode peut réagir de façon rapide et libérer une quantité d'encre grande, donc il peut s'échapper, et le mail est considéré comme un mail normal (Ham), si le cas contraire donc l'octopode est attaquer et le mail est considérer comme spam. Cette technique proposée a été expérimenté et testée plusieurs fois en utilisant plusieurs paramètres pour essayer d'obtenir des résultats meilleurs et fiables. Les résultats obtenus par cette technique ont montré qu'elles sont bonnes par rapport aux détecteurs des autres études, notre algorithme a améliorer les résultats trouvés par rapport aux recherches existantes.
Dans le deuxième travail dans cette thèse, on a essayé de tester le réseau de neurones récurrents LSTM dans le domaine de la sécurité des données, plus précisément dans la détection des applications android malveillantes. Cet algorithme n'était pas utilisé beaucoup comme un détecteur, ces réseaux de neurones récurrents donnent des résultats de haute qualité dans le cas des données liées entre eux (données en séquences) telles que TAL, reconnaissance vocale...etc. Dans notre travail, on a paramétré le LSTM et l'adapter pour la détection des applications Android malveillantes, ensuite on a testé un ensemble d'algorithmes basé sur l'apprentissage automatique pour détecter les applications android malveillantes, ces détecteurs sont des algorithmes de la fouille de données connues dans la littérature. Le but d'utiliser ces détecteurs est de comparer les résultats obtenus avec les résultats du réseau de neurones récurrent LSTM et évaluer le meilleur détecteur des applications intruses. Dans cette expérimentation le LSTM n'a pas donné des bons résultats par rapport aux ses résultats quand il est appliqué aux données en séquences.
Comme perspective nous intéressons dans le futur de toucher plusieurs problématiques qui cherchent à compromettre la sécurité des données et en essayant d'offrir et d'utiliser des méthodes pour augmenter la confidentialité des données pour les protéger. Parmi ces problématiques nous intéressons aux attaques deanonymisation des utilisateurs dans les réseaux sociaux, en espérant de toucher ce sujet et améliorer les recherches existants pour offrir des idées abordables et fiables.

E LISTE DES PUBLICATIONS

E.1 Revues Scientifiques

Mokri Miloud Aboubakeur El Sadek , Hamou Reda Mohamed & Amine Abdelmalek. A new bio inspired technique based on octopods for spam filtering. Applied Intelligence 49, 3425–3435 (2019). `https://doi.org/10.1007/s10489-019-01463-y`, (Springer).

E.2 Conférences Internationales

Mokri Miloud Aboubakeur El Sadek , Hamou Reda Mohamed & Amine Abdelmalek. A new meta-heuristic based on social bees for intrusion detection. 6th International Conference on Computer Intelligence and Its Applications CIIA'2018, Doctorial symposium, USTO-MB Oran University, Oran, Algeria, May 6-8, 2018.

Mokri Miloud Aboubakeur El Sadek , Hamou Reda Mohamed & Amine Abdelmalek. Machine learning methods and deep learning for android malware detection using permission. The First International Conference on Inovative Trends in computer Science CITCS'2019, University 8 Mai 1945, Guelma, Algeria, November 20-21, 2019.

Mokri Miloud Aboubakeur El Sadek , Hamou Reda Mohamed & Amine Abdelmalek. A Comparative Study of Android Malware Detection. The 8th International Conference on Inovation and New Trends in Information Technology (INTIS'2019), Tangier, Morocco, December 20-21, 2019.

F ANNEXE

F.1 A new bio inspired technique based on octopods for spam filtering

```
sc = SparkContext(appName="TFIDFExample")  # SparkContext to read the Data set
spark = SparkSession.builder \
    .master("local") \
    .appName("Word Count") \
    .config("spark.some.config.option", "some-value") \
    .getOrCreate()
documents = sc.textFile("F:/spam.csv")
```

FIGURE 6.1: Lecture du dataset avec apache spark

```
dData = documents.map(lambda line: line.split(',')).map(lambda line: Row(label=line[0], sentence=str(line[1:len(line)])))
sentenceData = spark.createDataFrame(dData)
countTokensD = udf(lambda words: words.strip("\""), StringType())
sentenceData = sentenceData.select("label", "sentence") \
    .withColumn("label", countTokensD(col("label")))    # convert the data into dataframe
```

FIGURE 6.2: Conversion des données au dataframe

```
tokenizer = Tokenizer(inputCol="sentence", outputCol="words")
wordsData = tokenizer.transform(sentenceData)
countTokens = udf(lambda words: tokenize(words), ArrayType(StringType()))
newDF = wordsData.select("label", "sentence", "words") \
    .withColumn("tokens", countTokens(col("sentence")))  # Get The Tokens from Sentences
ngram = NGram(n=3,inputCol="tokens", outputCol="ngrams")
DataFrameWithNgram = ngram.transform(newDF)   # Calculate ngram of words with n=3 and add it to the dataframe
```

FIGURE 6.3: Calculer ngram avec n=3

```
hashingTF = HashingTF(inputCol="ngrams", outputCol="rawFeatures", numFeatures=262144)
featurizedData = hashingTF.transform(DataFrameWithNgram) # Calculate TF of 3-grams Calculated
idf = IDF(inputCol="rawFeatures", outputCol="features")
idfModel = idf.fit(featurizedData)
rescaledData = idfModel.transform(featurizedData)  # Calculate IDF of 3-grams Calculated
```

FIGURE 6.4: Calculer les valeurs de TF*IDF

```
countTokens1 = udf(lambda x: float(np.sum(x.values)/np.size(x.values)), FloatType())
DataFrameWithMasse = rescaledData.select("label", "sentence", "words", "tokens", "rawFeatures", "features") \
    .withColumn("masse", countTokens1(col("features"))) #Calculate Masse

countTokens2 = udf(lambda x: float(np.amax(x.values)-np.amin(x.values)), FloatType())
DataFrameWithAcceleration = DataFrameWithMasse.select("label", "sentence", "words", "tokens", "rawFeatures", "features", "masse") \
    .withColumn("Acceleration", countTokens2(col("features"))) #Calculate Acceleration

comp = udf(comp, FloatType())
DataFrameWithForce = DataFrameWithAcceleration.withColumn("Force", comp("masse", "Acceleration"))# Calculate Force

total = DataFrameWithForce.select('label').count()
DataFrameWithprobability = (DataFrameWithForce.groupBy('label').count()
         .withColumn('total', F.lit(total))
         .withColumn('Probability', F.expr('count/total'))) # calculate the probability of each class (Ham and Spam)

DataframeDiv = DataFrameWithForce.join(DataFrameWithprobability,"label")
divide = udf(divide, FloatType())
FinaleDataframe = DataframeDiv.withColumn("Division", divide("Force", "Probability"))# calculate Finale value Of division of the force on probability of each class
```

FIGURE 6.5: Calculer la valeur finale force

```
for Training, Test in k_fold_cross_validation(FinaleDataframe.collect(), K=10): # kfold with k=10
    Trainingdata = sc.parallelize(Training)
    TestData = sc.parallelize(Test)
    ham=Trainingdata.filter(lambda x : x[0]=="ham")
    spam=Trainingdata.filter(lambda x : x[0]=="spam")
    test=TestData.filter(lambda x : x[0])
    testc=test.map(lambda x : x[12])
    testl=test.map(lambda x : x[0])
    hamc=ham.map(lambda x : x[12])
    spamc=spam.map(lambda x : x[12])
```

FIGURE 6.6: Obtenir les deux bases d'apprentissage et de test avec 10 fold validation croisé

```
hamd=map(lambda v : sum(map(lambda w : distance.euclidean(v,w), x)), y)# Calculate the difference value with Ham class
spamd=map(lambda v : sum(map(lambda w : distance.euclidean(v,w), z)), y) # Calculate the difference value with spam class

for x_values, y_values in zip(hamd, spamd):
    if x_values <= y_values:
        classe.append("ham")
    else:
        classe.append("spam")
                                    # Affecte the class for the test instance
```

FIGURE 6.7: Affecter la classe pour l'instance du base de test par l'algorithme

```
for L1, L2 in zip(classe, testl.collect()):
    if L1 == "ham" and L2 == "ham":
        VP+=1.0
    else:
        if L1 == "spam" and L2 == "spam":
            VN+=1.0
        else:
            if L1 == "spam" and L2 == "ham":
                FP+=1.0
            else:
                FN+=1.0

Rappel=VP/(VP+FN)
Precision=VP/(VP+FP)
Accuracy=(VP+VN)/(VP+VN+FP+FN)
Fmesure=(2*Rappel*Precision)/(Rappel+Precision)
Entropie= - math.log(Precision)
```

FIGURE 6.8: Calculer la matrice de confusion et les mesures d'évaluation

F.2 Machine learning methods and deep learning for android malware detection using permission

```
def RNN():
                inputs = Input(name='inputs',shape=[1])
                layer = Embedding(1,50,input_length=1)(inputs)
                layer = LSTM(64)(layer)
                layer = Dense(256,name='FC1')(layer)
                layer = Activation('relu')(layer)
                layer = Dropout(0.5)(layer)
                layer = Dense(1,name='out_layer')(layer)
                layer = Activation('sigmoid')(layer)
                model = Model(inputs=inputs,outputs=layer)
                return model


model = Sequential()
model.add(Dense(4, input_dim=330, activation='relu'))
model.add(Dense(4, activation='relu'))
model.add(Dense(1, activation='sigmoid'))
model.compile(loss='binary_crossentropy', optimizer='adam')
model.fit(data,target, epochs=500, batch_size=1,verbose=0)
predictions=model.predict_classes(x_test)
```

FIGURE 6.9: L'algorithme RNN LSTM

BIBLIOGRAPHIE

[1] Alain, P. (2017). Biomimétisme et biosinpiration. www.alain-pave.fr. (accessed : 27.03.2020 at 13 :52).

[2] Allard, O. (2012). *Biomimétisme : Comment les entreprises peuvent-elles intégrer le biomimétisme dans leur stratégie d'innovation ?* ESIEE PARIS.

[3] Almeida TA, H. J. (2018). Sms spam collection v.1. "http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/". (accessed : 05.01.2018 at 22 :51).

[4] Alouini, S. (2019). Les réseaux de neurones récurrents : des rnn simples aux lstm. "https://blog.octo.com/les-reseaux-de-neurones-recurrents-des-rnn-simples-aux-lstm/". (accessed : 14.05.2020 at 16 :11).

[5] Amador-Angulo L, C. O. (2018). *A new fuzzy bee colony optimization with dynamic adaptation of parameters using interval type-2 fuzzy logic for tuning fuzzy controllers.* Soft Comput 22(2) : 571–594.

[6] Amine, A. (2012). *Text Mining Course.* Department of Computer Science, Dr. Tahar Moulay University of Saida, Saida, Algeria.

[7] Amine, A. (2013). *Data Mining.* Laboratoire GeCoDe, Universté de Saida.

[8] Arciszewski T, C. J. (2006). *Bio-inspiration : Learning Creative Design Principia.* Conference paper EG-ICE : Intelligent Computing in Engineering and Architecture pp 32–53, Part of the Lecture Notes in Computer Science book series (LNCS, volume 4200).

[9] Arifin DD, M. S. (2016). *Enhancing Spam Detection on Mobile Phone Short Message Service (SMS) Performance using FP-Growth and Naive Bayes Classifier.* School of Computing, Telkom University, Bandung, Indonesia, The IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob).

[10] Arockia Panimalar.S, Varnekha Shree.S, V. K. (2017). *The 17 V's Of Big Data.* International Research Journal of Engineering and Technology (IRJET).

[11] Atif, J. (2015-2016). *Data Mining/ML Validation*. Université Paris-Dauphine.

[12] Bande Serrano JM, P. J. (2014). *The Evaluation of Ordered Features for SMS Spam Filtering*. Conference paper CIARP 2014 : Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications pp 383–390, Part of the Lecture Notes in Computer Science book series (LNCS, volume 8827).

[13] Bernard, I. (2015). Cryptanalyse de chiffrement par flot. "https://www.supinfo.com/articles/single/1290-cryptanalyse-chiffrement-flot". (accessed : 18.02.2020 at 10 :31).

[14] Bernard ESPINASSE, P. B. (2017). *Introduction au Big Data - Opportunités, stockage et analyse des mégadonnées*. Technologies de l'information, Technologies logicielles, Architectures des systèmes.

[15] Bonnefoi, P.-F. (2018). Cours de sécurité informatique. "https://www.coursehero.com/file/36331255/Cours-securite-informatiquepdf/". (accessed : 28.11.2018 at 16 :55).

[16] Boussaid, I. (2013). *Perfectionnement de méta-heuristiques pour l'optimisation continue*. Université paris-est Créteil école doctorale (ED 532) mathématiques et technologies de l'information et de la communication (MSTIC).

[17] Bremme, L. (2016). Qu'est-ce que le big data ? "https://www.lebigdata.fr/definition-big-data". (accessed : 23.10.2019 at 15 :41).

[18] Bruwer, H. J. (2014). *An Investigation of developments in web 3.0 :opportunities, risks, safeguards and governance*. (Computer Auditing), at Stellenbosch University.

[19] B.Shaikh, M. (2015). *A novel approach to detect android malware*. Procedia Computer Science 45 , Peerreview under responsibility of scientific committee of International Conference on Advanced Computing Technologies and Applications (ICACTA-2015), Elsevier B.V.

[20] C-Marketnig (2018). Du web 1.0 au web 4.0. "https://c-marketing.eu/du-web-1-0-au-web-4-0/". accessed : 25.07.2020 at 23 :57.

[21] Caraveo C, V. F. (2018). *A new optimization metaheuristic algorithm based on self-defense mechanism of the plants with three reproduction operators*. Soft Comput 22(15) :4907–4920.

[22] Castillo O, A.-A. L. (2018). *A generalized type-2 fuzzy logic approach for dynamic parameter adaptation in bee colony optimization applied to fuzzy controller design*. Inf Sci 460–461 : 476–496.

[23] Cavnar WB, T. J. (1991). *N-Gram-Based Text Categorization*. Environmental Research Institute of Michigan.

[24] Cervantes L, C. O. (2018). *Fuzzy dynamic adaptation of gap generation and mutation in genetic optimization of type 2 fuzzy controllers.* Adv Oper Res 2018 : 9570410.

[25] Chen, J. (2020). Android operating system. "https://www.investopedia.com/terms/a/android-operating-system.asp". (accessed : 03.05.2020 at 00 :10).

[26] Claude, J. (2017). Biomimétisme. "https://www.dev.scienceenlivre.org/index.php/2017/05/09/biomimetisme/". (accessed : 28.03.2020 at 19 :17).

[27] C.Li, K. (2016). *Android malware detection based on factorization machine.* Cryptography and Security (cs.CR).

[28] Clintberg's, M. (2019). *Acceleration.* studyphysics.ca.

[29] Clodera (2014/2015). *Guide du Big Data l'annuaire de référence à destination des utilisateurs.*

[30] Colah (2015). Understanding lstm networks. "http://colah.github.io/posts/2015-08-Understanding-LSTMs/". (accessed : 14.05.2020 at 18 :31).

[31] Consulting, S. (2020). Contrôle d'accès. "https://www.sartagas.fr/outils-de-la-ssi/controle-dacces/". (accessed : 04.02.2020 at 21 :05).

[32] Cormack GV, H. J. (2007). *Feature engineering for mobile (SMS) spam filtering.* SIGIR '07 proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval pages 871–872, Amsterdam.

[33] Cornuéjols, A. (2020). *Apprentissage par renforcement.* AgroParisTech & L.R.I., Université d'Orsay.

[34] Cottrell, M. (2020). Les réseaux de neurones historique, méthodes et applications. "https://samos.univ-paris1.fr/archives/ftp/preprints/samos174.pdf". (accessed : 04.04.2020 at 13 :07).

[35] Cousin, B. (2011). Sécurité des réseaux informatiques. "https://fr.scribd.com/doc/63869087/Securite-des-reseaux". (accessed : 31.12.2019 at 18 :17).

[36] Cryptage (2020). La cryptographie à clé publique. "http://www.cryptage.org/cle-publique.html". (accessed : 18.02.2020 at 12 :20).

[37] C.Urcuqui, A. (2016). *Machine learning classifiers for android malware analysis.* In Communications and Computing (COLCOM), 2016 IEEE Colombian Conference on (pp. 1-6). IEEE.

[38] Daniel Barsky, G. D. (2010). *Cour Cryptographie.* Cryptographie Paris 13.

[39] de Verdière, A. C. (2018). *Cinématique : Déplacement, vitesse, accélération.* Universite de Bretagne Occidentale.

[40] D.Gianluca, M. (2012). *A multi-level anomaly detector for android malware.* International Conference on Mathematical Methods, Models, and Architectures for Computer Network Security, MMM-ACNS 2012 : Computer Network Security Springer.

[41] Digabel, S. L. (2018). *Introduction aux métaheuristiques.* MTH6311, Ecole Polytechnique de Montréal.

[42] Douib, A. (2019). *Algorithmes bio-inspirés pour la traduction automatique statistique.* Université de Lorraine.

[43] Dumont, R. (2009-2010). *de cours provisoires Cryptographie et Sécurité informatique INFO0045-2.* Université de Liège, Faculté des Sciences Appliquées.

[44] Durand, N. (2004). *Algorithmes génétiques et autres outils d'optimisation appliqués à la gestion du trafic aérien.* L'institut national polytechnique de Toulouse, Laboratoire d'Optimisation Globale CENA - ENAC.

[45] e learning, . A. (2019). Le big data : un peu d'histoire. "https://www.26academy.com/le-big-data-un-peu-dhistoire/". (accessed : 05.01.2020 at 20 :15).

[46] Environment and Ecology (2020). What is biomimicry ? "http://environment-ecology.com/biomimicry-bioneers/367-what-is-biomimicry.html". (accessed : 07.04.2020 at 10 :42).

[47] Evidian (2001). *Etude de la sécurité informatique auprès de 250 entreprises européennes.* Europe des menaces informatiques croissantes.

[48] Fawzy, M. B. (2019). *4 Dynamics : force and newton's laws of motion.* researchgate.

[49] Feng, W. (2020). *Learning Apache Spark with Python.* Apache Spark.

[50] Ferradi, H. (2016). *Initiation à la cryptographie : théorie et pratique.* Université Paris 13 Villetaneuse.

[51] Ferreira JD, R. L. (2014). *Challenges and Properties for Bio-inspiration in Manufacturing.* Conference paper DoCEIS : Technological Innovation for Collective Awareness Systems pp 139–148, Part of the IFIPAdvances in Information and Communication Technology book series (IFIPAICT, volume 423).

[52] Fessant, F. (2006). *Apprentissage non supervisé.* TECH/SUSI,recherche & développement.

[53] Fillatre, L. (2014-2015). *Cryptologie et signature électronique.* Université Nice Sophia Antipolis, Polytech Nice Sophia.

[54] Fouque, P.-A. (2020). *Introduction à la cryptographie*. Université Rennes 1 et Institut Universitaire de France (IUF).

[55] FuturaTech (2018a). Introduction à la bioinspiration. "https://www.futura-sciences.com/tech/dossiers/robotique-robotique-inspiree-nature-816/page/2/". (accessed : 28.03.2020 at 20 :31).

[56] FuturaTech (2018b). Poulpe. "https://www.futura-sciences.com/planete/definitions/zoologiepoulpe-9551/". (accessed : 05.09.2018 at 22 :41).

[57] Garfinkel, S. L. (2015). *De-Identification of Personal Information*. U.S. Department of Commerce , National Institute of Standards and Technology.

[58] Gayatri Kapil, Alka Agrawal, R. A. K. (2016). *A Study of Big Data Characteristics*. SIST-Department of Information Technology, Babasaheb Bhimrao Ambedkar University (A Central University), Lucknow, India.

[59] Gilles Gasso, K. Z. (2020). *Apprentissage semi supervisé via un SVM parcimonieux : calcul du chemin de régularisation L1-Norm regularization path for the sparse semi-supervised Laplacian SVM*. LITIS EA 4108 - INSA de Rouen, Avenue de l'Université, 76801, Saint Etienne du Rouvray, France.

[60] H, S. (2018). What is k-fold cross validation ? "https://magoosh.com/data-science/k-fold-cross-validation/". (accessed : 08.09.2018 at 23 :11).

[61] Hasim Sak, A. S. (2014). *Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling*. Google, USA.

[62] Hastie, T. (2009). *K-fold cross validation, SLDM III Cross-v Alidation and bootstrap*.

[63] H.Fereidooni, M. (2016). *Android malware detection using static analysis of applications*. IFIP International Conference on New Technologies, Mobility and Security (NTMS) IEEE.

[64] Hidalgo JMG, B. G. (2006). *Content based SMS spam filtering*. DocEng '06 proceedings of the ACM symposium on document engineering pages 107–114, Amsterdam.

[65] Ikram Chraibi Kaadoud, T. V. (2018). Reprenons les bases : Neurone artificiel, neurone biologique. "http://www.scilogs.fr/intelligence-mecanique/reprenons-bases-neurone-artificiel-neurone-biologique/". (accessed : 04.04.2020 at 14 :50).

[66] Jacques, J. (2020). *Fouille de données - Data Mining*. Université Lumière Lyon 2.

[67] Jain, S. (2017). Introduction to genetic algorithm & their application in data science. "https://www.analyticsvidhya.com/blog/2017/07/introduction-to-genetic-algorithm/". (accessed : 27.03.2020 at 12 :39).

[68] JDN (2019a). Les fichiers log, des indicateurs utiles. "https://www.journaldunet.com/web-tech/developpeur/1008712-les-fichiers-log-des-indicateurs-utiles/". (accessed : 01.02.2020 at 13 :13).

[69] JDN (2019b). Script informatique : définition simple et pratique. "http://www.journaldunet.com/web-tech/dictionnaire-du-webmastering/1203599-script-definition/". accessed : 30.10.2017 at 20 :10.

[70] Jin-Kao Hao, C. S. (2013). *Méta-heuristiques et intelligence artificielle*. LERIA - Laboratoire d'Etudes et de Recherche en Informatique d'Angers et LIRIS - Laboratoire d'Informatique en Image et Systèmes d'information.

[71] J.Jaemin, K. (2018). *Android malware detection based on useful api calls and machine learning*. First International Conference on Artificial Intelligence and Knowledge Engineering (AIKE) IEEE.

[72] Joe I, S. H. (2010). *An SMS Spam Filtering System Using Support Vector Machine*. Conference paper FGIT : Future Generation Information Technology pp 577–584, Part of the Lecture Notes in Computer Science book series (LNCS, volume 6485).

[73] Jourdan, L. (2003). *Méta-heuristiques pour l'extraction de connaissances : application à la génomique*. Université des sciences et technologies de lille, U.F.R. D'I.E.E.A.

[74] Kanona, R. M. (2017). *Viruses and Anti-Virus*. Researchgate.

[75] Keras (2019). The python deep learning library. "https://keras.io". (accessed : 06.07.2019 at 11 :16).

[76] Kout A, L. S. (2018). *a new bio-inspired routing protocol based on cuckoo search algorithm for mobile ad hoc networks*. Wirel Netw 24(7) :2509–2519.

[77] Kumar S, G. X. (2016). *A Machine Learning Based Web Spam Filtering Approach*. IEEE 30th International Conference on Advanced Information Networking and Applications (AINA), 23–25 , Print ISSN : 1550-445X, Conference Location : Crans-Montana, Switzerland.

[78] K.Y.Lok, Y. (2012). *Droid scope : seamlessly reconstructing the os and dalvik semantic views for dynamic android malware analysis*. Syracuse University Air Force Research Laboratory Syracuse, , New York, USA.

[79] Laurent Bloch, C. W. (2009). *Sécurité informatique Pour les DSI, RSSI et administrateurs*. Eyrolles.

[80] learn, S. (2019). Machine learning library in python. "https://scikitlearn.org/stable/". (accessed : 07.07.2019 at 14 :23).

[81] Learning, U. M. (2016). Sms spam collection. "https://www.kaggle.com/uciml/sms-spam-collection-dataset/home". (accessed : 05.01.2018 at 20 :50).

[82] LeMagIT (2019). Apprentissage par renforcement. "https://www.lemagit.fr/definition/Apprentissage-par-renforcement". (accessed : 31.03.2020 at 19 :07).

[83] Limousin, S. (2018). Les différentes attaques web. "https://www.supinfo.com/articles/single/7085-differentes-attaques-web". (accessed : 15.01.2020 at 02 :08).

[84] Éléonore Marmion, M. (2011). *Recherche locale et optimisation combinatoire : De l'analyse structurelle d'un problème a la conception d'algorithmes efficaces*. Centre de Recherche INRIA Lille - Nord Europe, Laboratoire d'Informatique Fondamentale de Lille (UMR CNRS 8022), École Doctorale Sciences Pour l'Ingénieur Université Lille Nord-de-France, Université Lille 1.

[85] L.Yongfeng, S. (2015). *Detection classification and characterization of android malware using api data dependency*. Laboratory for Novel Software Technology, Department of Computer Science and Technology, Nanjing University, China.

[86] Malek, M. (2020). *Introduction Fouille des données*. EISTI.

[87] Mawdoo3 (2018). information on octopod. "https://mawdoo3.com/Information_on_octopod". (accessed : 07.08.2018 at 11 :37).

[88] Médini, L. (2018). *World Wide Web*. Université Lion 1.

[89] Michel, L. C. (2017). *Applications Client/Serveur et Web*, chapter 4. Licence Pro SIL.

[90] Michel Gendreau, Cirrelt, M. (2015). *Introduction à la recherche avec tabous*. École Polytechnique de Montréal.

[91] Muhammad Haris, Basit Jadoon, F. H. K. (2017). *Evolution of Android Operating System : A Review*. International Conference on Advanced ResearchAt : Melbourne, Australia.

[92] Multimedia, I. (2019). Qu'est-ce que le développement web ? "https://www.iesamultimedia.fr/actualite/news/definition-developpement-web". (accessed : 15.10.2019 at 11 :34).

[93] Nagwani NK, S. A. (2017). *SMS spam filtering and thread identification using bi-level text classification and clustering techniques*. Department of Computer Science and Engineering, National Institute of Technology Raipur, India. J Inf Sci 43(1) :75–87.

[94] Naima Hadj-Said, A. A.-P. (2018). *Nouveau Mode Opératoire pour la Cryptographie*. Laboratoire SIMPA (Signal-Image-Parole), Université des Sciences et de la Technologie d'Oran USTO.

[95] Najadat H, A. N. (2014). *Mobile SMS spam filtering based on mixing classifiers.* Department of Computer Information Systems, Faculty of Computer and Information Technology, Jordan University of Science and Technology, Irbid, Jordan. Int J of Adv Comput Res 1.

[96] Natasha Chayaamor-Heil, F. G. e. N. H.-B. (2018). *Biomimétisme en architecture. État, méthodes et outils Biomimicry in Architecture : State, methods and tools.* Les Cahiers de la recherche architecturale urbaine et paysagère.

[97] Nerzic, P. (2019). *Outils pour le BigData.* IUT de Lannion, Dept Informatique, Univ rennes.

[98] Network Associates, I. s. f. (1990-1998). *De Introduction à la cryptographie.* Network Associates.

[99] Niblo, G. A. (2008). *Quelques méthodes de chiffrement.* University of Southampton.

[100] Nizar, C. (2019). *Cours sécurité informatique.* Institut Supérieur des études technologique de Siliana, TUNIS.

[101] Olivas F, A.-A. L. (2017). *Comparative study of Type-2 fuzzy particle swarm, bee Colony and bat algorithms in optimization of fuzzy controllers.* special issue Extensions to Type-1 Fuzzy Logic : Theory, Algorithms and Applications.

[102] Oracle (2019). Qu'est-ce que le big data ? `"https://www.oracle.com/fr/big-data/guide/what-is-big-data.html"`. (accessed : 21.10.2019 at 18 :09).

[103] Oracle (2020a). Qu'est-ce que l'apprentissage automatique ? `"https://www.oracle.com/ca-fr/artificial-intelligence/what-is-machine-learning.html"`. (accessed : 26.02.2020 at 18 :24).

[104] Oracle (2020b). Qu'est-ce que le spoofing ? `"https://www.oracle.com/fr/security/spoofing-usurpation-identite-ip.html"`. (accessed : 17.01.2020 at 17 :20).

[105] Paper, O. W. (2014). *Information Management and Big Data.* A Reference Architecture.

[106] Papini, O. (2019). *Cours 5 : Détection d'intrusions, Sécurité des Systèmes d'information.* ESIL, Universit'é de la méditerranée.

[107] Piette, M.-A. (2013). Biomimétisme. `"https://www.affairesdegars.com/page/article/4156049774/biomimetisme-les-10-inventions-les-plus-impressionnantes-inspirees-de-la-nature.html"`. (accessed : 08.04.2020 at 19 :33).

[108] Pine, J. A. (2019-2020). *Apprentissage automatique.* Laboratoire Université de Lyon 2.

[109] P.Naser, Z. (2013). *Machine learning for android malware detection using permission and api calls*. IEEE 25th International Conference on Tools with Artificial Intelligence.

[110] Poinsot, L. (2019). Chap. i : Introduction à la sécurité informatique. https ://lipn.univ-paris13.fr/p̃oinsot/save/INFO%203/Cours/Cours%201.pdf. (accessed : 31.12.2019 at 17 :13).

[111] Preux, P. (2011). *Notes de cours Fouille de données*. Université Université de Lille 3.

[112] Rajeev Sobti, G. (2012). *Cryptographic Hash Functions : A Review*. IJCSI International Journal of Computer Science.

[113] Rakesh Kumar, Bhanu Bhushan Parashar, S. G. (2014). *Apache Hadoop, NoSQL and NewSQL Solutions of Big Data*. International Journal of Advance Foundation and Research in Science & Engineering (IJAFRSE).

[114] Raphael, J. (2020). Android versions : A living history from 1.0 to 11. "https://www.computerworld.com/article/3235946/android-versions-a-living-history-from-1-0-to-today.html". (accessed : 09.05.2020 at 05 :14).

[115] Revel, A. (2020). *Apprentissage Semi-Supervisé et Apprentissage Transductif*. Laboratoire L3i, Faculté des Sciences et Technologies, Bâtiment Pascal, Avenue Michel Crépeau, 17042 La Rochelle Cedex 1 – France.

[116] Richer, J.-M. (2008). *Développement Web Introduction générale*. Université angers faculté des sciences unité de formation et de recherche.

[117] Sareh Aghaei, Mohammad Ali Nematbakhsh, H. K. F. (2012). *Evolution of the world wide web : from web 1.0 to web 4.0*. International Journal of Web & Semantic Technology (IJWesT).

[118] S.Asaf, K. (2012). *A behavioral malware detection framework for android devices*. Journal of Intelligent Information Systems Springer.

[119] Search, S. (2020). *Machine Learning Intelligent Connections : How Machine Learning is Revolutionizing Marketing*. Synapse Search.

[120] SecuritéInfo (2020). Les fonctions de hachage en cryptographie. "https://www.securiteinfo.com/cryptographie/hash.shtml". (accessed : 18.02.2020 at 21 :24).

[121] Shibly, F. (2016). *Android Operating System : Architecture, Security Challenges and Solutions*. Lecturer in IT, South Eastern University of Sri Lanka, Oluvil, Sri Lanka.

[122] Sidi Mohamed Douiri, S. E. (2020). *Cours des Méthodes de Résolution Exactes Heuristiques et Métaheuristiques*. Laboratoire de Recherche Mathématiques, Informatique et Applications, Université Mohammed V, Faculté des Sciences de Rabat.

[123] Sidorenko, S. (2017). *Improvement of Creativity via the Six-Step Bio-Inspiration Strategy*. Ss Cyril and Methodius University of Skopje, Faculty of Mechanical Engineering, Karpos II bb, Skopje, Republic of Macedonia.

[124] S.Justin, K. (2012). *A machine learning approach to android malware detection*. European Intelligence and Security Informatics Conference IEEE.

[125] Smola, A. and Vishwanathan, S. (2008). *Introduction to Machine Learning*. Cambridge University Press.

[126] Spark, A. (2017). Apache spark. "https://spark.apache.org/". (accessed : 30.12.2019 at 19 :53).

[127] Tahir AM, N. G. (2017). *Robotizing the Bio-inspiration*. Conference paper RiTA : Robot Intelligence Technology and Applications 5 pp 313–334, Part of the Advances in Intelligent Systems and Computing book series (AISC, volume 751).

[128] Talbi, E.-G. (2020). *Fouille de données - Un tour d'horizon*. Laboratoire d'Informatique Fondamentale de Lille OPACOPAC.

[129] Tarpy, D. R. (2004). *The Honey Bee Dance Language*. North Carolina Cooperative Extension Service.

[130] T.Chen, Q. (2018). *Tiny droid : a lightweight and efficient model for android malware detection and classification*. Mobile Information Systems.

[131] Tech, F. (2020a). Backdoor. "https://www.futura-sciences.com/tech/definitions/informatique-backdoor-2047/". (accessed : 13.01.2020 at 23 :53).

[132] Tech, F. (2020b). Déni de service. "https://www.futura-sciences.com/tech/definitions/internet-deni-service-2433/". (accessed : 17.01.2020 at 19 :03).

[133] T.Kimberly, J. (2015). Copper droid : automatic reconstruction of android malware behaviors. *Internet Society*, 322(4) :8–11.

[134] Toolbox, B. (2020). Methods. "https://toolbox.biomimicry.org/methods/". (accessed : 12.04.2020 at 00 :14).

[135] Triggs, R. (2018). What android app permissions mean and how to use them. "https://www.androidauthority.com/app-permissions-886758/". (accessed : 09.05.2020 at 18 :21).

[136] U.Christian (2016). Dataset malware/beningn permissions android. "https://www.kaggle.com/xwolf12/datasetandroidpermissions". (accessed : 05.06.2019 at 22 :41).

[137] Unitec (2019). Bio-inspiration : s'inspirer du vivant pour penser demain. "https://medium.com/@veilleunitec/bio-inspiration-sinspirer-du-vivant-pour-penser-demain-4e7960d791d". (accessed : 06.04.2020 at 16 :42).

[138] VertigoLab (2020). Le biomimétisme est dépassé. "http://vertigolab.eu/le-biomimetisme-est-depasse-vive-la-bioinspiration-six-principes-cles-inspires-du-vivant-pour-une-innovation-durable-au-service-de-la-transition-ecologique-et-energetique-des-territoires-et-des-or/". (accessed : 06.04.2020 at 16 :36).

[139] Wang G-G, D. S. (2018). *A new monarch butterfly optimization with an improved crossover operator*. Oper Res 18(3) :731–755.

[140] WayToLearnX (2018). Différence entre le chiffrement par bloc et le chiffrement par flot. "https://waytolearnx.com/2018/07/difference-entre-le-chiffrement-par-bloc-et-le-chiffrement-par-flot.html". (accessed : 18.02.2020 at 10 :25).

[141] Widmer, M. (2001). *Les Métaheuristiques : des outils performants pour les problémes industriéls*. 3e Conférence Francophone de Modélisation et Simulation "Conception, Analyse et Gestion des Systèmes Industriels" MOSIM'01 – du 25 au 27 avril- Troyes (France), Université de Fribourg, Département d'informatique Rue Faucigny 2 CH – 1700 Fribourg (Suisse).

[142] Worlds, O. (2018). Octopus defenses. "https://www.octopusworlds.com/octopus-defenses/". (accessed : 18.08.2018 at 01 :26).

[143] Xu Q, X. E. (2012). *SMS spam detection using noncontent features*. IEEE Intell Syst 27(6).

[144] Xumei Fan, W. S. (2020). *Review and Classification of Bio-inspired Algorithms and Their Applications*. Journal of Bionic Engineering,Springer.

[145] Y.Zhenlong, Y. (2014). *Droid-sec : deep learning in android malware detection*. SIGCOMM '14 Proceedings of the 2014 ACM conference on SIGCOMM Chicago, Illinois, USA.

[146] ZhangY, Wang S, P. P. (2014). *Binary PSO with mutation operator for feature selection using decision tree applied to spam detection*. Knowl-Based Syst 64 : 22-31.

[147] Zhenfang, Z. (2015). *Study on Computer Trojan Horse Virus and Its Prevention*. International Journal of Engineering and Applied Sciences (IJEAS).

[148] Z.Yajin, J. (2012). *Dissecting android malware : characterization and evolution.* Department of Computer Science North Carolina State University, Yajin Zhou. Under license to IEEE Computer Society.