REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE IBN KHALDOUN - TIARET

MEMOIRE

Présenté à :

FACULTÉ DES MATHEMATIQUES ET D'INFORMATIQUE
DÉPARTEMENT D'INFORMATIQUE

Pour l'obtention du diplôme de :

MASTER

Spécialité : Réseau et Télécommunication

Par :

KAMLA Bochra Nour Elhoda
GRAICHI Ghalia

Sur le thème

Inspection Approfondie des paquets à l'aide du l'apprentissage en profondeur

---

Soutenu publiquement le 13 / 06 / 2024 à Tiaret devant le jury composé de :

Mr. AID Lahcene	MCA	Université Tiaret	Président
Mr. MOUSTAFAOUI Kadda	MAA	Université Tiaret	Encadrant
Mr MOKHTARI Ahmed	MAA	Université Tiaret	Examinateur

2023-2024

Remerciement

En premier lieu, nous remercions Dieu « ALLAH » le très haut qui nous a donné le courage et la volonté d'accomplir ce modeste travail.

En seconde lieu Mr : MOSTEFAOUI Kadda notre encadreur, il s'est toujours montré à l'écoute et très disponible tout au long de sa réalisation et auquel nous tenons à témoigner nos reconnaissances et notre gratitude les plus sincères.

Outre l'encadreur, nous tenons à remercier Mr. Bouazza Abdel Hamid, notre professeur, qui nous a aidé et nous a fait honneur dans la présentation de cet humble travail.

Nos vifs remerciements vont aux membres du jury d'avoir accepté d'examiner notre travail, à savoir : Monsieur AID Lahcene *et monsieur MOKHTARI Ahmed de notre université. Nous tenons à saisir cette occasion et adresser nos profonds remerciements et reconnaissances à toutes personnes qui nous ont aidés de près ou de loin dans la réalisation de ce mémoire. Enfin nous exprimons notre profonde reconnaissance à tous responsables et enseignants de l'université de Tiaret qui ont contribué à notre formation.*

Dédicace

Je dédie ce modeste travail à mes chers parents, source de mes joies et secret de ma force, vous serez toujours le modèle : mon père dans ta détermination, ta force et ton honnêteté, ma mère dans ta bonté, ta patience et ton dévouement pour nous. Merci pour vos sacrifices. C'est à vous que je dois cette réussite.

A mon frère

Chamsse El dine;

A mes chères sœurs

Imane et wafaa;

A ma chère amie BELKHEIRAT Zineb ;

Sans oubliée mon binôme BOCHRA ;

Tous mes proches.

Ghalia

DÉDICACE

Tout d'abord, je tiens à remercier ***DIEU*** *De m'avoir donné la force et le courage de mener à bien ce modeste travail.*

Je tiens à dédier cet humble travail à :

À ***mes chers parents,*** *merci pour votre soutien constant et vos sacrifices qui ont rendu ma réussite possible.*

À ***ma chère grand-mère*** *qui a toujours prié pour ma réussite,*

À ma sœur ***Hanane*** *, mes frères* ***Hicham*** *et* ***Mohamed***

en reconnaissance de leur soutien indéfectible tout au long de mes études.

À toute ma famille,

À tous mes ***amis*** *,surtout ma copine* ***Zineb*** *et mon binôme* ***Ghalia*** *, pour leur aide précieuse dans la réalisation de ce projet.*

À toutes les personnes qui m'aiment .

NourElhoda

Sommaire

Chapitre 3 : réalisation et implémentation

Liste d'abréviation

IoT : L'internet of Things

QoS : Quality of Service

DPI : Deep Packet Inspection

IDS : Intrusion Detection System

IP : Internet Protocol

IA : Intelligence Artificiel

ML : Machine Learning

DL : Deep Learning

DNN : Deep neural network

CNN : Convolutional neural networks

LSTM : Long short term memory

TE : Taux D'exactitude

TFA : Taux de Fausse alerte

DR : Détection Rate

Liste des figures

Liste des tableaux

Résumé

La classification du trafic réseau est cruciale. Elle permet de catégoriser le trafic, de réaliser des statistiques, d'appliquer une politique de qualité de service adéquate et de détecter les intrusions. Cependant, la généralisation des techniques de chiffrement rend la classification du trafic chiffré difficile avec les approches traditionnelles, compliquant ainsi la gestion du trafic réseau. Pour surmonter ces obstacles, une inspection approfondie des paquets est essentielle. DPI systèmes utilisent des signatures pour repérer le trafic, mais doivent être régulièrement mis à jour en raison des applications fréquentes. Techniques de l'apprentissage autonome, telles que les algorithmes d'apprentissage profond, permettent la création autonome de signatures et l'analyse en temps réel. Data est utilisée pour créer des modèles, avec le modèle le plus efficace utilisé pour la catégorisation du trafic. Les différents algorithmes d'apprentissage utilisés dans notre solution sont ANN, CNN et LSTM. Nous avons constaté que notre modèle basé sur LSTM est plus efficace, avec une résolution de rappel de 0,95, un score F1, une précision et une précision. Cette démonstration met en évidence la capacité du modèle à classer de manière précise des applications pour différents types de trafic, ce qui améliore la qualité générale du service.

Mots clés : inspection approfondie des paquets (DPI), qualité de service (QoS),
Classification de trafic réseau, apprentissage profond.

Abstract

The classification of network traffic is crucial. It enables traffic to be categorized, statistics to be compiled, an appropriate QoS policy to be applied and intrusions to be detected. However, the widespread use of encryption techniques makes it difficult to classify encrypted traffic using traditional approaches, thus complicating network traffic management.

To overcome these obstacles, in-depth packet inspection is essential. DPI systems use signatures to identify traffic, but must be regularly updated due to frequent applications. Autonomous learning techniques, such as deep learning algorithms, enable autonomous signature creation and real-time analysis. Data is used to create models, with the most efficient model used for traffic categorization. The different learning algorithms used in our solution are ANN, CNN and LSTM. We found that our LSTM-based model is more efficient, with a recall resolution of 0.95, F1 score, precision and accuracy. This demonstration highlights the model's ability to accurately classify applications for different types of traffic, improving overall quality of service.

Keywords: deep packet inspection (DPI), quality of service (QoS),
network traffic classification, deep learning.

ملخص

وتجميع الإحصائيات وتطبيق سياسة جودة الخدمة المناسبة واكتشاف الاختراقات. ومع ذلك، فإن الاستخدام الواسع النطاق لتقنيات التشفير يجعل من الصعب تصنيف حركة المرور المشفرة باستخدام الأساليب التقليدية، مما يعقد إدارة حركة مرور الشبكة.

للتغلب على هذه العقبات، من الضروري إجراء فحص متعمق للحزم. تستخدم أنظمة DPI التواقيع لتحديد حركة المرور، ولكن يجب تحديثها بانتظام بسبب التطبيقات المتكررة. تتيح تقنيات التعلم الذاتي، مثل خوارزميات التعلم العميق، إمكانية إنشاء توقيعات مستقلة وتحليلها في الوقت الحقيقي. يتم استخدام البيانات لإنشاء نماذج، مع استخدام النموذج الأكثر كفاءة لتصنيف حركة المرور. خوارزميات التعلم المختلفة المستخدمة في حلنا هي ANN وCNN وLSTM. وقد وجدنا أن نموذجنا القائم على LSTM أكثر كفاءة، مع دقة استرجاع تبلغ 0.95، ودرجة F1 ودقة ودقة. يسلط هذا العرض التوضيحي الضوء على قدرة النموذج على تصنيف التطبيقات بدقة لأنواع مختلفة من حركة المرور، مما يحسن من جودة الخدمة بشكل عام.

الكلمات المفتاحية: الفحص العميق للحزم(DPI) ، جودة الخدمة(QoS) تصنيف حركة مرور الشبكة، التعلم العميق .

Introduction générale

Les systèmes et réseaux de communication modernes, tels que l'Internet des objets et les réseaux cellulaires, génèrent une quantité massive et hétérogène de données de trafic. Dans de tels réseaux, les techniques traditionnelles de gestion de réseau pour la surveillance et l'analyse des données sont confrontées à des défis et à des problèmes importants. L'un des principaux défis est la congestion, où le réseau connaît des volumes de trafic élevés, entraînant des retards, des pertes de paquets et une dégradation des performances. De plus, garantir la qualité de service (QoS) et une classification précise du trafic constitue un autre obstacle.

La QoS garantit un certain niveau de performances en termes de latence, de fiabilité et de bande passante, ce qui devient crucial pour des applications comme le streaming vidéo en temps réel. Une classification correcte du trafic est essentielle pour optimiser les ressources réseau et hiérarchiser les paquets de données critiques. Les techniques de classification basées sur la charge utile, une approche courante pour l'analyse du trafic, souffrent de trois problèmes principaux dans les paradigmes de réseau conventionnels. Premièrement, ces techniques rencontrent des difficultés pour classer le trafic chiffré, limitant la capacité d'analyse des communications sécurisées. Deuxièmement, les politiques de confidentialité peuvent restreindre l'accès au contenu des paquets, empêchant ainsi une analyse détaillée. Enfin, les méthodes de charge utile imposent une lourde charge de calcul aux systèmes de communication, affectant leur efficacité et leur réactivité. Relever ces défis est essentiel pour une gestion efficace du réseau et une optimisation du trafic.

En outre, il est nécessaire d'effectuer une inspection approfondie des paquets pour approfondir le contenu des paquets de données, permettant ainsi aux administrateurs réseau d'identifier des applications, des protocoles ou des menaces de sécurité spécifiques. Les systèmes DPI utilisent des modèles d'octets uniques comme signatures pour détecter le trafic des applications. Les applications mettent fréquemment à jour leur version pour ajouter de nouvelles fonctionnalités et/ou pour contourner les systèmes pare-feu/DPI. Ainsi, un système DPI précis doit vérifier périodiquement les signatures existantes et les mettre à jour si nécessaire. Les techniques d'apprentissage automatique, en particulier les algorithmes DL, font partie des techniques les plus populaires pour le traitement des données de trafic réseau. En tirant parti de ces algorithmes, les systèmes DPI peuvent apprendre et mettre à jour de manière autonome les signatures, garantissant ainsi une analyse précise et efficac

du trafic en temps réel. Cette approche améliore non seulement la précision du DPI, mais contribue également à une infrastructure de sécurité réseau plus résiliente et intelligente.

Dans notre travail, nous inclurons une exploration des différentes architectures d'apprentissage profond utilisées dans le DPI, en trois chapitres.

Le premier chapitre est un aperçu général du trafic réseau. Nous parlerons d'abord de la définition d'un réseau informatique, de la sécurité du réseau et de ses mécanismes. Après cela, nous discuterons de la définition de la qualité de service, la définition du trafic réseau et de ses classifications.

Dans le deuxième chapitre, nous parlons d'intelligence artificielle. Nous avons découvert l'apprentissage automatique et l'apprentissage profond et leurs algorithmes les plus importants.

Le dernier chapitre contient nos travaux appliqués, qui appliquent certaines des architectures d'apprentissage profond utilisées dans le DPI, telles que les réseaux de neurones convolutifs (CNN), les réseaux de neurones profonds (DNN) et la mémoire à long terme (LSTM), pour améliorer ce processus et créer des modèles. Le meilleur modèle sera utilisé pour classer le trafic.

Grâce à notre travail, nous aspirons à contribuer au développement de solutions de classification du trafic qui non seulement améliorent la qualité de service, mais s'adaptent et évoluent également aux côtés du paysage en constante évolution des réseaux de télécommunications modernes.

1. Introduction

Les réseaux informatiques sont devenus beaucoup plus important qu'ils en aient il y a quelques années. De nos jours les entreprises dès leur création n'hésitent pas à mettre en place un réseau informatique pour faciliter la gestion de leur infrastructure, les réseaux informatiques permettent la communication, le partage d'informations et la connectivité dans notre monde interconnecté. Pour assurer un fonctionnement efficace et sécurisé de ces réseaux, deux aspects clés doivent être pris en compte la qualité de service (QoS) et la sécurité. La combinaison de la qualité de service et de la sécurité est essentielle pour garantir des réseaux informatiques performants et protégés. L'inspection approfondie des paquets joue un rôle clé dans l'amélioration de la QoS en permettant une classification et une gestion précises du trafic, ainsi que dans le renforcement de la sécurité en détectant et en prévenant les menaces potentielles. Ce chapitre vous fournira une compréhension approfondie de ces concepts et vous aidera à tirer parti de la DPI pour optimiser la QoS et renforcer la sécurité de votre réseau informatique.

2. Définition d'un réseau informatique

Un réseau informatique est un ensemble de machines et d'équipements électroniques reliés les uns aux autres par le biais d'un câble, d'une liaison radio ou même sans fil. Cependant, l'échange de données entre ces appareils est régi par des protocoles de communication qui établissent des règles. [1]

3. Intérêt d'un réseau

Il y a deux types principaux des objectifs des réseaux

Les objectifs techniques

- Le partage des ressources logicielles (compilateur, système de gestion de base de données) et matérielles (imprimantes, traceurs, scanners,...) permet de diminuer les coûts.
- La fiabilité (un réseau offre la possibilité de duplicate les données et réduit ainsi les pertes de ces données).

Les objectifs des utilisateurs

- La communication entre personnes (via des courriels, des conférences électroniques, des téléphones mobiles, etc..).
- L'accès à l'information depuis l'étranger (banques, bourses, bibliothèques en ligne,...). [2]

4. Les protocoles réseaux

Selon [2] on trouve

- **Le protocole DNS** (Domaine Name Service)

Est une base de données qui sert à convertir les noms d'ordinateurs en adresses IP sur les réseaux IP.

- **Protocole TCP** (Protocole de contrôle de transmission)

C'est un protocole sécurisé, axé sur la connexion qui permet le transfert sécurisé de paquets d'une station à une autre.

- **protocole ICMP** (Internet Control Message Protocol)

Il assure la gestion des erreurs de transmission. Effectivement, étant donné que le protocole IP se limite au transport des paquets et ne permet pas l'envoi de message d'erreur, c'est grâce à ce protocole qu'une machine émettrice peut détecter un incident de réseau.

- **Protocole DHCP** (Dynamic Host Configuration Protocol)

Est un protocole réseau qui permet de configurer automatiquement les paramètres IP d'une station, en lui attribuant automatiquement une adresse IP et un masque de sous-réseau. DHCP a également la possibilité de définir l'adresse par défaut de la passerelle.

- **Protocole FTP** (File Transfert Protocol)

Le protocole de transfert de fichiers (FTP) permet de faire le transfert de fichiers d'une machine à une autre. Il est nécessaire que l'utilisateur se connecte à un serveur FTP pour accéder ou déposer un fichier depuis un poste client, en utilisant un nom et un mot de passe. Si l'utilisateur n'est pas identifié, la connexion ne sera pas mise en place.

- **Le protocole de communication web http** (Hyper Text Transfer Protocol)

Permet d'échanger des documents hypertextes contenant des informations sous la forme de texte, d'images fixes ou animées, ainsi que des sons. Chaque client web utilise le port 80 d'un serveur HTTP pour communiquer.

- **Protocole TFTP** (Trivial File Transfer Protocol)

TFTP (Trivial File Transfer Protocol ou Protocole de transfert de fichiers simplifié) Il s'agit d'un protocole de transfert de fichiers simplifié. Il utilise l'UDP sur le port 69, contrairement au FTP qui utilise TCP. Les logiciels embarqués sur les équipements réseaux (routeurs, pare-feu, etc.) ou le démarrage d'un ordinateur à partir d'une carte réseau sont toujours une pratique fréquente.

5. Sécurité du réseau informatique

5.1.Définition

La sécurité informatique est l'ensemble des moyens techniques, organisationnels, juridiques et humains nécessaire et mis en place pour réduire la vulnérabilité d'un système contre les menaces accidentelles et intentionnelles. [3]

5.2.Objectifs de la sécurité informatique [11]

- **La disponibilité**

Elle consiste à garantir l'accès à un service ou à une ressource. C'est-à-dire, pour que la disponibilité soit garantie, il faut que le système soit capable de maintenir le bon fonctionnement de ses services et de ses ressources.

- **l'intégrité**

L'intégrité spécifie que seules les personnes autorisées peuvent modifier l'information dans le système. Donc il faut pouvoir garantir que les données n'ont pas été altérées durant la communication par une personne non autorisée (de manière fortuite ou).

- **la confidentialité**

La confidentialité est le maintien du secret des informations. C'est-à-dire, les données ne doivent être visibles que pour les personnes autorisées. Elles sont seules qui ont la possibilité de les atteindre. Pour ce faire, il faut les rendre inintelligibles en les chiffrant de telle sorte que les personnes qui ne sont pas autorisés à les déchiffrer ne puissent les utiliser.

- **L'authentification**

L'authentification consiste à assurer l'identité d'un utilisateur, c'est-à dire de garantir à chacun des correspondant que son partenaire est bien celui qu'il croit être. Où des procédures d'identification et d'authentification peuvent être mises en œuvre pour contribuer à réaliser des procédures de contrôle d'accès et des mesures de sécurité permettant d'assures la confidentialité et l'intégrité des données (seule les ayants droit identifiés et authentifiés peuvent accéder aux ressource et les modifier s'ils sont habilités à le faire).

- **La non-répudiation**

La non-répudiation est le processus qui permet de garantir qu'aucun des correspondants ne pourra nier la transaction. A ce critère de sécurité peuvent être associées les notions d'imputabilité (se définit par l'attribution d'un événement à une personne) ou de traçabilité (message électronique, transaction commerciale, transfert de données).

5.3.Les mécanismes de sécurité [5]

Les mécanismes de sécurité sont des mécanismes conçus pour détecter, empêcher ou récupérer suite à une attaque de sécurité.

a. Chiffrement

Le chiffrement transforme tout ou partie d'un texte dit clair en cryptogramme, message chiffré ou protégé. Si une communication utilise des dispositifs de chiffrement, les données sont transmises sous une forme « brouillée », de manière qu'elles ne puissent être comprises par un tiers**.**

b. Signature numérique

Une signature numérique utilise un cryptage irréversible pour transformer un message en un petit bloc de données, rendant impossible sa reconstruction. L'algorithme utilisé est appelé fonction de hachage ou fonction de condensation. La signature est ensuite envoyée et vérifiée par l'expéditeur. Les algorithmes de chiffrement irréversible les plus connus sont MD5 et SHA1. L'intégrité d'une unité de données est assurée par des codes de contrôle cryptographiques et un horodatage.

c. Mots de passe

L'identification des entités homologues peut être effectuée en utilisant un identifiant d'utilisateur et un mot de passe, lorsque les moyens de communication sont sécurisés. La sécurité ne repose exclusivement sur l'identifiant, mais il est difficile de modifier. Les utilisateurs ne connaissent pas le mot de passe inscrit dans une carte magnétique contenant un NIP. Le responsable de la sécurité doit prêter attention au protocole et au système de fichiers pour garantir la sécurité.

d. Liste de contrôle d'accès

Les listes de contrôle d'accès (ACL, Access Control List) déterminent les droits d'accès des entités en utilisant leur identité authentifiée et des informations fiables pour accéder au réseau ou aux ressources sur le réseau. En outre, il est possible de constituer une trace d'audit et de recenser les tentatives d'accès non autorisées. Chaque utilisateur qui fait une erreur de mot de passe laisse une trace. Ainsi, les programmes automatiques qui tentent de pénétrer le système peuvent être repérés en essayant tous les mots de passe. Les données utilisées comprennent les listes de droits d'accès, gérés par des centres, les mots de passe, les jetons de droits d'accès, les divers certificats, ainsi que les labels de sensibilité des données.

e. Bourrage et contrôle de routage par gestion dynamique de la bande passante

Bourrage imite les communications pour dissiper le silence et banaliser les moments réels de communication, évitant ainsi l'attention des pirates lors des démarrages de transmission. Il s'agit d'envoyer des messages utiles entre des séquences non liées, et d'utiliser la fréquence de l'alphabet utilisé.

f.Notarisation

Notarisation garantit l'intégrité et confirme l'origine, date et destination des informations, et doit être obtenue par un tiers de confiance, qui obtient des informations et obtient un certificat numérique. Pour garantir la sécurité, il est nécessaire de superviser les autorités de certification.

g. Le pare-feu

Le pare-feu est un dispositif informatique qui permet le sélectif passage de données entre un réseau interne et un réseau public. Il a également la capacité de filtrer les données échangées avec le réseau et neutraliser les tentatives de pénétration en provenance du réseau public. Les firewalls, également appelés les préservatifs pour les réseaux, sont des technologies de contrôle d'accès qui empêchent l'accès non-autorisés à ressources d'information et les transferts de l'information propriétaire du réseau**.**

h. Les systèmes de détection d'intrusion (IDS)

Un système de détection des intrusions (IDS) est un ensemble de composants logiciels et/ou matériels dont la fonction principale est de détecter et analyser toute tentative d'effraction volontaire.

6. Le trafic réseau

6.1.Définition

Le trafic réseau est la quantité de données qui se déplacent sur un réseau informatique à tout moment. Le trafic réseau, également appelé trafic de données, est divisé en paquets de données et envoyé sur un réseau avant d'être réassemblé par le dispositif ou l'ordinateur destinataire. Le trafic réseau a deux flux directionnels, nord-sud et est-ouest. Le trafic affecte la qualité du réseau, car une quantité anormalement élevée de trafic peut signifier des vitesses de téléchargement lentes ou des connexions Voix sur IP (Voice over Internet Protocol). Le trafic est également lié à la sécurité, car une quantité anormalement élevée de trafic pourrait être le signe d'une attaque. [6]

7. Définition de Qos

La qualité de service est la différence entre le service attendu et la perception de l'offre réelle. Cette différence est négative quand l'offre est inférieure aux attentes, et positive lorsque l'offre est supérieure ou égal aux attentes. [7]

7.1.Les métriques de QoS [8]

- **Bande passante**

La vitesse d'un lien réseau détermine la capacité de transmission de données entre deux points. La Qualité de Service (QoS) permet de gérer et d'optimiser cette bande passante en fonction des besoins spécifiques. Grâce à la QoS, un routeur peut prioriser divers types de trafic en allouant une quantité spécifique de bande passante à différentes files d'attente.

- **Retard**

Le temps nécessaire pour un paquet de partir de sa source jusqu'à sa destination finale. On peut souvent être impacté par le retard de queue, qui survient pendant les périodes de congestion et où un paquet attend dans une queue avant d'être transmis. Le QoS permet aux organisations d'éviter cela en créant une file de priorité pour certains types de trafic.

- **Perte**

Le volume de données perdues à cause de la perte de paquets, qui se produit généralement en raison de la congestion du réseau. QoS permet aux organisations de choisir les paquets à évacuer lors de cet événement.

- **Gigue**

La vitesse instable des paquets sur un réseau en raison de la congestion, ce qui peut entraîner l'arrivée tardive et non chronologique des paquets. Cela peut entraîner des distorsions ou des écarts dans la transmission audio et vidéo.

8. La congestion du réseau

8.1.Définition

La congestion du réseau se produit lorsque trop de paquets sont envoyés vers un seul lien ou nœud du réseau. Dans des cas extrêmes, cela peut entraîner une perte de paquets, des délais d'attente ou même le blocage ou le ralentissement de nouvelles connexions. La cause la plus courante de congestion du réseau est la surcharge d'un ou plusieurs nœuds du réseau. Lorsque cela se produit, les performances globales du réseau en souffrent car la bande passante allouée à chaque

connexion réseau devient limitée. Si cela continue suffisamment longtemps, l'ensemble du réseau finira par atteindre la saturation et cessera de fonctionner correctement. [9]

8.2.Congestion de réseau dans Qos [10]

La congestion du réseau fait référence à une réduction de la qualité de service (QOS) qui entraîne une perte de paquets, un retard dans la file d'attente ou le blocage de nouvelles connexions. Généralement, la congestion du réseau se produit en cas de surcharge de trafic lorsqu'un lien ou un nœud de réseau traite des données au-delà de sa capacité. Pour éviter l'effondrement et réduire les effets de la congestion sur le réseau, les organisations utilisent diverses méthodes d'évitement et de contrôle de la congestion. Ceux-ci inclus

- Réduction de la fenêtre TCP/IP
- File d'attente équitable dans les périphériques réseau tels que les routeurs, les commutateurs et autres périphériques
- Schémas de priorité qui transmettent des paquets de priorité plus élevée avant le reste du trafic
- Allocation explicite des ressources réseau via des contrôles d'admission vers des flux spécifiques.

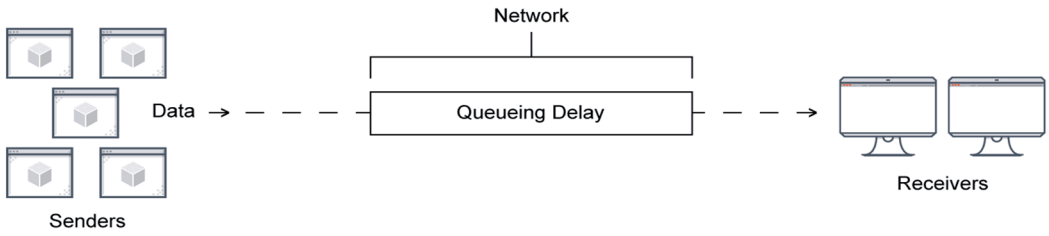


Figure 1.1 la congestion de réseau dans QOS

9. Inspection approfondie des paquets (DPI)

9.1.Définition

L'inspection approfondie des paquets (DPI) est une méthode d'examen du contenu des paquets de données lorsqu'ils passent par un point de contrôle sur le réseau. Avec des types d'inspection dynamique des paquets, le dispositif ne vérifie que les informations dans l'en-tête du paquet, telles que l'adresse IP (Internet Protocol) de destination, l'adresse IP source et le numéro de port. Le DPI examine un plus large éventail de métadonnées et de données connectées à chaque paquet avec lequel le dispositif est connecté. Dans ce sens, le processus d'inspection comprend l'examen de l'en-tête et des données que le paquet transporte. Par conséquent,

le DPI fournit un mécanisme plus efficace pour exécuter le filtrage des paquets réseau. En plus des capacités d'inspection des technologies régulières de renfilage des paquets, DPI peut trouver des menaces autrement cachées dans le flux de données, telles que des tentatives d'exfiltration de données, des violations des politiques de contenu, des logiciels malveillants, etc. [11]

9.2.Les données collectées par DPI

De manière générale les systèmes DPI peuvent collecter toutes les données non chiffrées. Voici les données qui peuvent être collectées et analysés par un système DPI

- **Résolutions DNS** L'en-tête DNS n'est pas chiffré sauf dans le cas de DNSSEC, DNS Over TLS ou HTTPS (Dot et DoH) et DNSCrypt
- **Le protocole réseau utilisé** ils savent que vous utilisez un VPN ou faites du P2P
- **Connectivité d'adresse IP** Ainsi, même si vous utilisez HTTPS pour accéder à un site de vidéos de chats, ils peuvent détecter que vous vous êtes connecté à ce site et que vous avez téléchargé 500 Go de données. Ils ne connaissent pas le contenu exact des données, mais ils peuvent voir le nom de domaine (DNS), l'adresse IP, et la quantité de données transférées sur ce site et sur tous les autres sites que vous visitez.
- **D'autres trafic non-HTTPS** comme UDP, Mail, SNMP, FTP, Telnet, les mises à jour de certains logiciels peuvent ne pas utiliser HTTPS, etc. [12]

9.3.Cas d'utilisation de DPI

L'inspection approfondie des paquets n'est pas seulement une technologie de pointe; c'est un outil polyvalent et essentiel dans le domaine de la gestion et de la sécurité modernes des réseaux. Ses applications couvrent un large éventail de cas d'utilisation critiques qui soulignent son importance.

Explorons certains de ces cas d'utilisation clés, mettant en lumière la façon dont l'inspection approfondie des paquets permet aux organisations d'améliorer la sécurité de leur réseau, d'optimiser leurs opérations et de répondre efficacement aux menaces émergentes.[13]

- **Détection et blocage des logiciels malveillants et des virus**

Une inspection approfondie des paquets peut identifier et bloquer rapidement les logiciels malveillants et les signatures de virus connus, empêchant ainsi les menaces potentielles d'infiltrer le réseau. [13]

- **Surveillance de l'utilisation du réseau et application des politiques**

L'inspection approfondie des paquets fournit des informations granulaires sur le trafic réseau, permettant aux organisations de surveiller les modèles d'utilisation, d'appliquer les politiques réseaux et d'optimiser l'allocation des ressources. [13]

- **Identification des menaces internes et de l'exfiltration de données**

En analysant le comportement du réseau, une inspection approfondie des paquets peut signaler des activités inhabituelles pouvant indiquer des menaces internes ou des tentatives d'exfiltration de données. [13]

- **Analyse du trafic réseau à des fins d'investigation**

En cas d'incident de sécurité, les données d'inspection approfondie des paquets peuvent constituer un outil d'investigation précieux, aidant les organisations à retracer les origines d'une attaque et à comprendre les méthodes employées par les acteurs malveillants. Dans un cas d'utilisation médico-légale, il est particulièrement important de « rejouer » les paquets réseau pour voir ce qu'un utilisateur ou un système a vu. [13]

9.4.L'amélioration de qualité de service par DPI

Voici quelques façons dont la DPI peut contribuer à améliorer la QoS [14]

- **Classification et priorisation du trafic**

La DPI permet d'identifier les protocoles, les applications et les types de trafic spécifiques en analysant le contenu des paquets. Cela permet de classer le trafic en fonction de son importance et de ses exigences en termes de QoS. Par exemple, les applications temps réel telles que la voix sur IP (VoIP) ou la vidéoconférence peuvent être identifiées et priorisées pour garantir une bande passante suffisante et une faible latence, assurant ainsi une expérience utilisateur fluide et de haute qualité**.**

- **Gestion de la bande passante**

En identifiant et en classifiant le trafic à l'aide de la DPI, il devient possible de mettre en place des politiques de gestion de la bande passante plus précises. Les ressources réseau peuvent être allouées de manière intelligente en fonction des besoins spécifiques des applications et des utilisateurs. Par exemple, une application de transfert de fichiers volumineux peut être limitée en termes de bande passante afin de garantir que d'autres applications plus sensibles à la latence, telles que la navigation web ou la diffusion en continu, ne soient pas affectées négativement.

- **Réduction de la latence, de la gigue et de la perte de paquets**

La DPI peut aider à identifier les problèmes de latence, de gigue (variation de délai) et de perte de paquets dans le réseau. En identifiant les goulots d'étranglement, les congestions ou les problèmes de connectivité, la DPI permet aux administrateurs réseau de prendre des mesures correctives pour améliorer ces aspects de la QoS. Par exemple, en identifiant les paquets perdus ou corrompus, des actions peuvent être entreprises pour résoudre les problèmes de connectivité et minimiser les perturbations dans la transmission des données.

- **Adaptation dynamique de la QoS**

La DPI permet une adaptation dynamique de la QoS en fonction des conditions du réseau et des besoins changeants des applications. Par exemple, si la bande passante disponible diminue en raison d'une congestion du réseau, la DPI peut détecter cette situation et ajuster automatiquement les politiques de gestion du trafic pour maintenir des performances acceptables pour les applications essentielles. Cela permet d'optimiser l'utilisation des ressources réseau et de garantir une QoS adaptée aux besoins en constante évolution des utilisateurs.

9.5.Les Outils de dpi [15]

Les pare-feu qui contiennent des fonctionnalités IDS, y compris une inspection du contenu, suivent généralement la méthode DPI. Outre les pare-feu, IDS (système de détection d'intrusion) utilise également la technique DPI. IDS met l'accent sur la protection de réseaux entiers au lieu de détecter des attaques particulières. Cependant, pour arranger les choses, certains outils DPI sont nécessaires

- **Correspondance de motifs ou de signatures**

La correspondance de modèle ou de signature est bonne pour les pare-feu avec les fonctionnalités IDS activées. Il inspecte chaque paquet par rapport à une base de données d'attaques réseau identifiées.

La technique fonctionne très bien pour les attaques familières. Cela signifie que cette approche n'est pas adaptée pour protéger votre système contre les attaques nouvelles ou inconnues.

- **Anomalie de protocole**

Anomalie de protocole également appelée approche de refus par défaut. Les pare-feu avec des fonctionnalités IDS activées peuvent s'appuyer sur ce protocole. L'anomalie de protocole est assez restrictive mais protège le système contre les attaques inconnues. Il rejette le trafic en une seule fois s'il ne correspond pas aux règles du protocole.

- **Solutions IPS**

Les solutions IPS sont également compatibles avec la technique DPI. IPS fait référence au système de prévention des intrusions. DPI et IPS peuvent ensemble détecter et combattre les menaces. L'un des inconvénients de cette approche est le risque de faux positifs, qui peut être légèrement contrôlé par certaines politiques conservatrices.

9.6.Les avantages du DPI [16]

a) Sécurité Internet

DPI peut être utilisé comme système de détection d'intrusion (IDS) ou une combinaison de prévention d'intrusion (IPS) et de détection d'intrusion. Il peut identifier des attaques spécifiques telles que le déni de service et tout autre trafic malveillant provoqué par des virus, des vers ou des ransomwares, que d'autres outils de sécurité pourraient ne pas être en mesure de détecter. DPI fonctionne un peu comme un antivirus, mais il détecte les menaces au niveau de la couche réseau avant même qu'elles n'atteignent l'utilisateur final. Par exemple, dans les grandes entreprises, DPI peut aider à empêcher les virus et les vers de se propager sur l'ensemble du réseau de l'entreprise. Il peut également permettre de détecter les utilisations interdites des applications de votre entreprise.

b) Prévention de la perte de données

DPI peut empêcher la sortie de données dans les entreprises. Par exemple, lors de l'envoi par courrier électronique d'informations confidentielles, DPI inviterait un employé à obtenir l'autorisation nécessaire pour les envoyer.

c) Mise en forme du trafic Internet ou gestion du réseau

Vous pouvez utiliser DPI pour filtrer le trafic et faciliter le flux réseau. Par exemple, vous pouvez le configurer de manière à recevoir en premier les messages hautement prioritaires ou à ralentir ou à donner la priorité à vos téléchargements P2P.Malheureusement, les FAI le font également souvent pour limiter le trafic des utilisateurs. Les titulaires de droits d'auteur peuvent également demander aux FAI, avec l'aide du DPI, d'empêcher le téléchargement illégal de leur contenu.

d) Écoutes clandestines et censure en ligne

Le gouvernement chinois utilise le DPI pour surveiller et contrôler le trafic réseau du pays. Cela les aide à bloquer les sites Web indésirables tels que la pornographie, les plateformes de médias sociaux et l'opposition religieuse ou politique.

e) Publicité ciblée

DPI soulève certains problèmes de confidentialité car il peut creuser suffisamment profondément pour voir l'expéditeur, le destinataire et le contenu du paquet de données. Ces informations peuvent être collectées par des FAI qui surveillent votre trafic et peuvent ensuite être vendues à des sociétés spécialisées dans la publicité ciblée.

9.7.Les limites de DPI [17]

Malgré les avantages de cette technologie quelles peuvent apporter, nous avons identifié trois limites significatives

–Bien qu'il demeure efficace face aux attaques par dépassement de mémoire tampon (Buffer Over flow), aux attaques de déni de service (Dos) et à certains types de programmes malveillants, le DPI peut être utilisé pour générer de nouvelles vulnérabilités en plus de ces protections contre les vulnérabilités déjà existantes et faciliter leur exploitation.

–En outre, le DPI apporte une complexité et une charge supplémentaire aux pare-feu déjà en place ainsi qu'à d'autres systèmes de sécurité. Pour assurer une inspection approfondie des paquets, il est essentiel de les mettre à jour et de les maintenir afin de maintenir une efficacité optimale.

–En raison de l'augmentation de la charge des processeurs de pare-feu, cela peut également entraîner une diminution significative de la vitesse du réseau.

10. La classification du trafic réseau [18]

La classification du trafic réseau (NTC) est l'élément fondamental pour repérer différentes applications et protocoles disponibles sur le réseau. En classant le trafic réseau comme une application ou un protocole, ils sont marqués ou annotés, puis diverses opérations peuvent être réalisées telles que le suivi, la découverte, la détection d'anomalies, le contrôle et l'optimisation, dans le but d'améliorer les performances du réseau. Il joue un rôle majeur dans la sécurité et la gestion des réseaux, notamment en matière de contrôle de la qualité de service (QOS), de détection d'intrusion et d'interception légale. Les réseaux sans fil ont une bande passante restreinte, donc afin de bien organiser différentes applications au sein du réseau, les systèmes de contrôle du QoS utilisent le module de catégorisation du trafic, NTC joue un rôle essentiel dans l'analyse du trafic réseau, en particulier pour filtrer le trafic afin de détecter toute activité malveillante dans le réseau. Différentes stratégies NTC ont été élaborées et créées au cours des vingt dernières années.

NTC est généralement divisé en quatre méthodes :

- Basé sur le port
- Basé sur la charge
- Basé sur la Statistique
- Basé sur le Comportementale

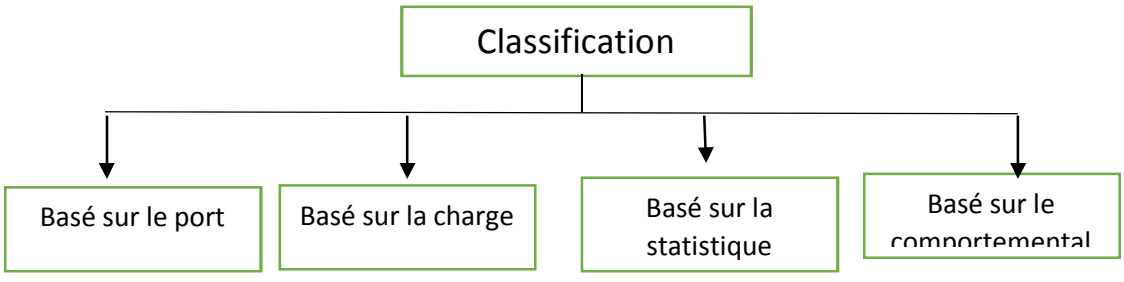


Figure 1.2 Classification du trafic réseau

11. Les types de classification [18]

11.1. Classification basée sur les ports

Les méthodes classiques d'internet utilisent des numéros de port pour identifier le trafic réseau, mais cette méthode est inexacte et utilise souvent des problèmes liés à l'utilisation de port non standard dans les applications actuelles.

11.2. Classification basée sur le poids/Inspection approfondie des paquets

Afin de pallier les lacunes de la classification des ports, une approche alternative a été proposée, à savoir l'inspection profonde des paquets (DPI) ou la détection basée sur le contenu du paquet. « Cette méthode effectue le match entre les contenus du paquet et les compare à un ensemble de signatures accumulées déterministe ».

- **Limite**

• Les résultats de cette méthode sont extrêmement précis. Toutefois, cette méthode échoue lorsqu'il s'agit de gérer le trafic chiffré, ce qui augmente le taux de résultat positif faux lorsque beaucoup de trafic chiffré reste non classé.

• Examiner les contenus du paquet viole la politique de confidentialité des utilisateurs et cette méthode nécessite des coûts de calcul élevés.

11.3. Classification Basé sur la Statistique

L'approche utilise les caractéristiques statistiques du flux de trafic réseau pour repérer les applications réseau, telles que la longueur des paquets, la durée du flux et la déviation standard. On présente un cadre de classification du trafic basé sur les signatures, qui utilise les statistiques de trafic pour classer le trafic en une catégorie de services (Cos). On utilise des techniques de machine Learning pour améliorer les performances de classification.

11.4. Classification Basé sur le Comportementale

L'approche examine les patterns de trafic réseau afin de repérer des applications en vérifiant les numéros d'hôte et de port. Utilisant des heuristiques, il crée des profils comportementaux. Une nouvelle méthode pour regrouper les applications P2P-TV utilise des données comportementales et des systèmes de surveillance virtuelle pour réduire les taux de false alarme.

12. L'état de l'art

Nous présentons maintenant certains travaux antérieurs qui portent sur l'utilisation de l'apprentissage automatique pour la classification du trafic.

Pour faire face au niveau élevé de cryptage [45], de nombreuses études ont été réalisées pour obtenir de véritables résultats en matière de classification du trafic des paquets de réseau [46] [47] [48]. Au cours des dernières années, de nouvelles méthodologies (Yuan et al. 2014) et de nouveaux cadres ont été mis en œuvre pour répondre aux exigences et comprendre la complexité croissante des paquets profonds [49]. Afin de mieux comprendre les caractéristiques des paquets réseau, de nombreuses études ont également démontré la fusion de l'apprentissage profond avec des classificateurs ML (Cai et al, 2010). [50]

Zhanyi Wang a utilisé des réseaux neuronaux pour démontrer leur capacité à identifier les protocoles de réseau via l'apprentissage de modèles. Cela a permis de développer l'utilisation des réseaux neuronaux dans l'extraction et l'apprentissage des caractéristiques [51]. Dans Cuadra-Sanchez et Aracil (2017), Wei Wang et d'autres auteurs ont montré comment un cadre simplifié de réseau neuronal convolutionnel unidimensionnel peut effectuer la classification du trafic de données cryptées de bout en bout. Ils ont démontré que la relation entre les données brutes et la sortie peut être facilement établie et apprise avec leur modèle. [52]

Dans Datta et al. (2015) ont introduit une fusion de l'apprentissage automatique supervisé avec des réseaux neuronaux à entraînement bayésien, ce qui présente l'avantage d'un plus large éventail d'applications [53]. L'approche présentée dans Dorfinger (2010), appelée Seq2Img, capture les comportements statiques et dynamiques de la séquence. Cette approche évite les limitations associées à l'entraînement du modèle avec une poignée de caractéristiques élaborées à la main. [54]

Par ailleurs, Ehlert et al. (2006) et Fu et al. (2016) ont élaboré une approche d'apprentissage profond capable de segmenter le réseau en classes, à savoir F2P et P2P. En outre, le modèle proposé par Ehlert et al. (2006) et Fu et al. (2016) prend également en charge l'identification de l'application de l'utilisateur [55] [56]. Dans une autre recherche présentée par Goo et al. (2016) et Janani et Ramamoorthy (2022), Manuel, Jun et d'autres ont effectué la classification du trafic réseau à l'aide de nouvelles méthodes. [57][58]

Goo et al. (2016) ont amélioré les performances d'un algorithme existant basé sur des seuils normalisés en prenant trois propriétés simples des paquets IP [57]. La recherche présentée par Janani et Ramamoorthy (2022) utilise une information corrélée supplémentaire pour améliorer les performances et surmonter les limites de l'ajustement excessif et de la disponibilité d'un ensemble de données limité lors de la formation [58]. Les données de trafic mobile sont étudiées par Liu et al. (2019). Rahman et al. (2022) utilisent les statistiques des messages pour la classification du trafic. [59] [60]

Avec la demande croissante de classification du trafic des réseaux, de nombreuses études ont incorporé leurs modèles avec des réseaux neuronaux convolutifs à pointes. Une étude récente indiquée par Kumar et Sharma (2016), Lee et al. (2015), Liu et al. (2019) et Park et al. (2008) implique l'utilisation du même modèle. Les réseaux de neurones à pointes ont montré des résultats prometteurs dans de nombreux domaines d'application, notamment les tâches de détection, de calcul et de reconnaissance. Ces modèles ont été introduits dans les problèmes de traitement du signal et ont une portée plus large dans la compréhension du comportement dynamique des paquets de données. [59] [61] [62]

Conclusion

En conclusion, l'intégration de l'inspection approfondie des paquets (DPI) dans les réseaux permet d'améliorer la qualité de service (QoS) en analysant le contenu des paquets de données en temps réel. Cette analyse permet une gestion plus efficace des ressources du réseau, une optimisation des performances et une meilleure répartition de la bande passante. En priorisant certains types de trafic, comme la voix sur IP (VoIP) ou la vidéo en streaming, la DPI assure une expérience utilisateur plus cohérente. Cependant, des préoccupations liées à la vie privée et à la neutralité du réseau doivent être prises en compte. En résumé, la DPI représente un outil crucial pour répondre aux besoins croissants de connectivité et de qualité dans le monde numérique actuel.

1. Introduction

L'apprentissage automatique et l'apprentissage profond sont des domaines en perpétuelle évolution qui offrent de nouvelles opportunités pour résoudre des problèmes complexes et exploiter pleinement le potentiel des données. Il est crucial de bien comprendre les concepts et les méthodes du L'apprentissage automatique et du l'apprentissage profondes afin de profiter de ces avancées technologiques et de contribuer à l'évolution de l'intelligence artificielle dans de multiples domaines d'activité.

Il existe de nombreuses applications réussies du machine Learning dans différents domaines, c'est ce que nous allons aborder dans ce chapitre. Nous explorerons les principes et les techniques fondamentaux du L'apprentissage automatique et du l'apprentissage profond. Nous aborderons également les différents types d'algorithmes utilisés dans la machine Learning, en particulier l'apprentissage avec supervision, qui est l'un des points les plus marquants de notre projet.

2. L'intelligence artificielle

L'IA désigne la possibilité pour une machine de reproduire des comportements liés aux humains, tels que le raisonnement, la planification et la créativité. L'IA permet à des systèmes techniques de percevoir leur environnement, gérer ces perceptions, résoudre des problèmes et entreprendre des actions pour atteindre un but précis. Les systèmes dotés d'IA sont capables d'adapter leurs comportements plus ou moins en analysant les effets produits par leurs actions précédentes, travaillant de manière autonome. [19]

3. Apprentissage automatique

3.1. Définition

L'apprentissage machine (ou apprentissage automatique, Machine Learning en anglais) est un sous-domaine de l'intelligence artificielle, qui donne à un système une capacité de compréhension grâce à ses algorithmes. Il est basé sur l'idée de faire apprendre des algorithmes à partir de données et de faire des prédictions avec ces données et par cela les ordinateurs apprennent à résoudre des tâches spécifiques, sans avoir besoin de les programmer. L'objectif du Machine Learning est de reconnaître parmi des données des structures souvent trop difficiles à détecter ou à mesurer manuellement. À partir de ces structures, on peut chercher à classifier des individus, des objets, à prédire la valeur d'une variable à un certain horizon, à expliquer l'apparition ou non d'une caractéristique. [20]

3.2. Les types d'apprentissage automatique

Il existe différents types d'apprentissage automatique, chacun ayant ses propres caractéristiques et applications. Les principaux types d'algorithmes d'apprentissage automatique sont les suivants

• Apprentissage supervisé.

• Apprentissage non supervisé.

• Apprentissage par renforcement.

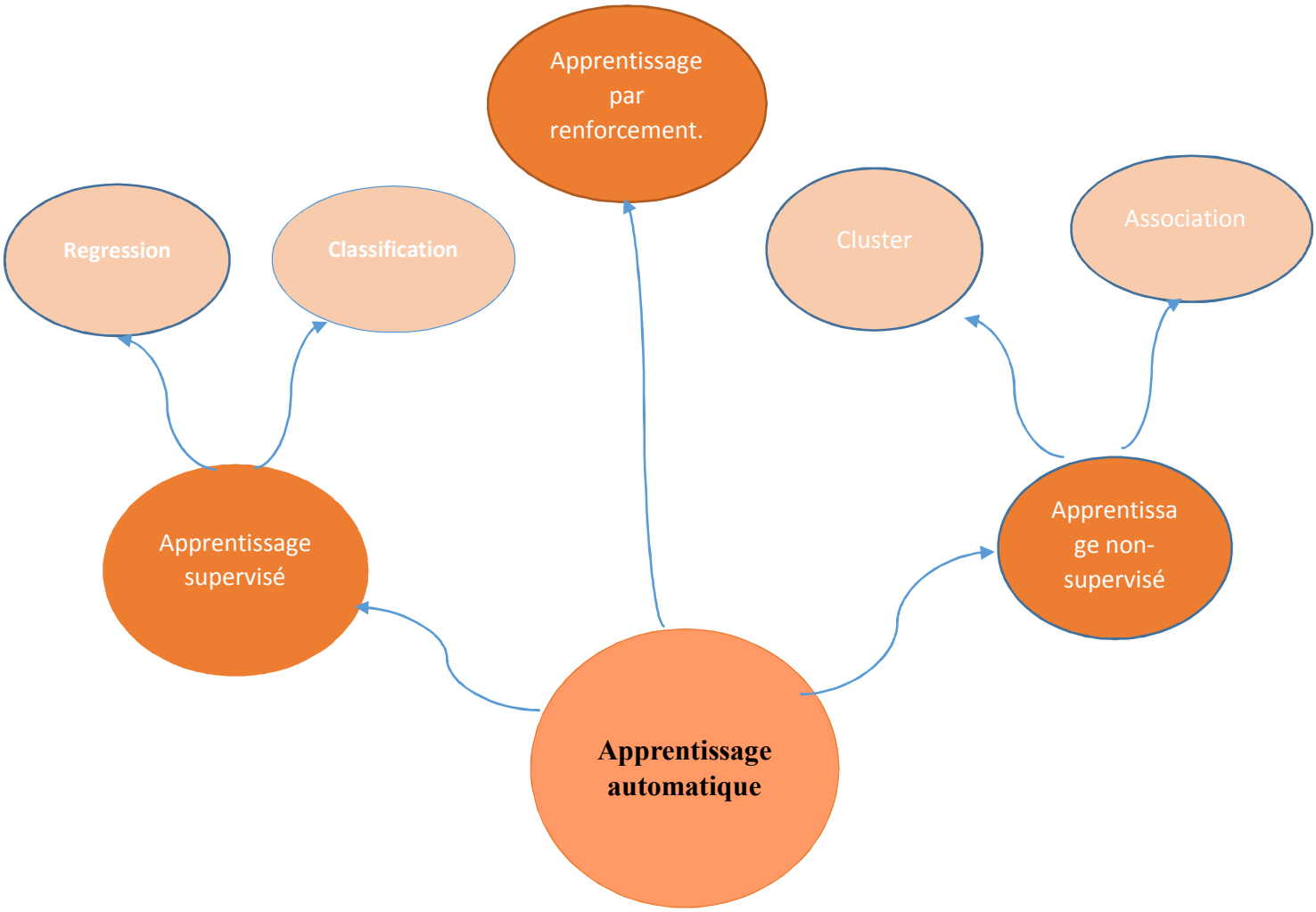


Figure 2.1 les types d'apprentissages automatiques

3.2.1. Apprentissage supervisé

L'apprentissage supervisé est une forme d'apprentissage automatique qui exploite un ensemble de données d'apprentissage étiquetées pour concevoir des modèles d'intelligence artificielle. L'objectif de cette méthode est d'acquérir des connaissances en comparant sa réelle sortie avec les sorties. [21]

L'apprentissage supervisé consiste en des variables d'entrée (x) et une variable de sortie (Y). Vous utilisez un algorithme pour apprendre la fonction de mapping de l'entrée à la sortie Y = f (x).

Le but est d'appréhender si bien la fonction de mapping que, lorsque vous avez de nouvelles données d'entrée (x), vous pouvez prédire les variables de sortie (Y) pour ces données. [22]

Le schéma de la figure 2.2 montre le fonctionnement de l'apprentissage supervisé. [23]

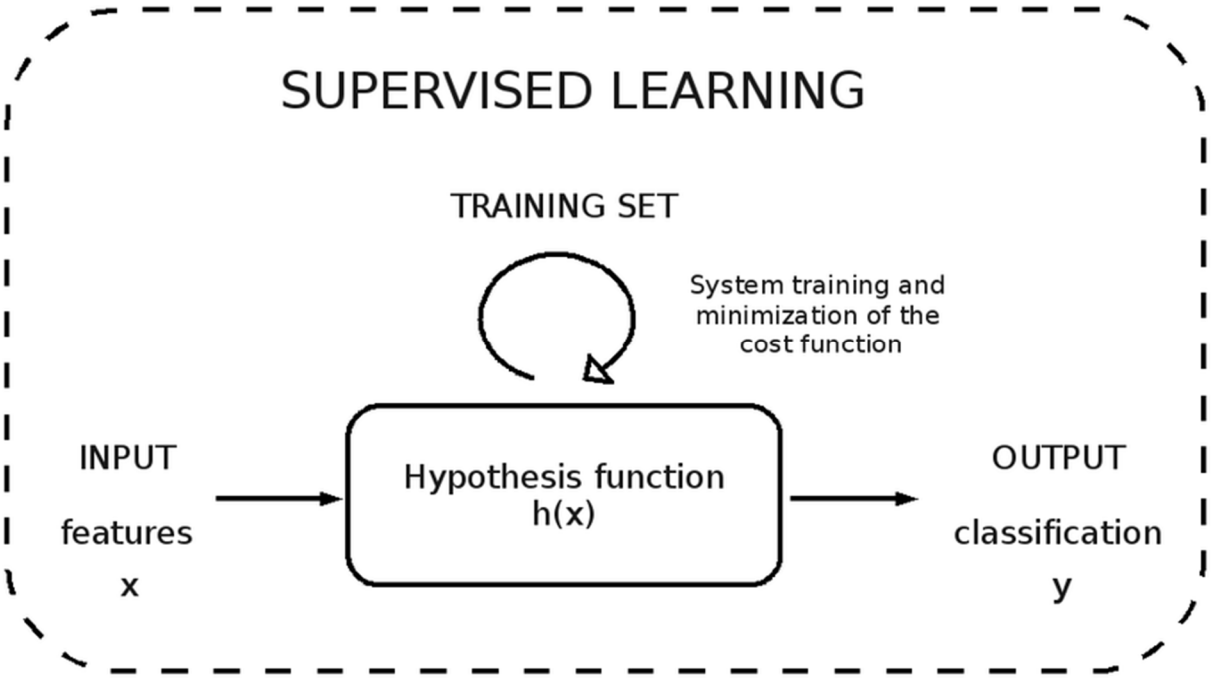


Figure 2.2 le fonctionnement de l'apprentissage supervisé

On peut également classer supervisé en deux catégories

- Classification.
- Régression.

A. Classification

On utilise la classification lorsque la variable sortie est à deux ou plusieurs classes. Par exemple, oui ou non, homme ou femme, vrai ou faux, etc. [22]

B. Régression

On utilise la régression lorsque la variable d'output est une valeur réelle ou continue. Dans cette situation, il existe une corrélation entre deux ou plusieurs variables, c'est-à-dire qu'une variation d'une variable est liée à une variation de l'autre variable. Par exemple, salaire basé sur l'expérience professionnelle ou poids basé sur l'altitude, etc. [22]

La figure 2.3 représente la classification et de la régression. [24]

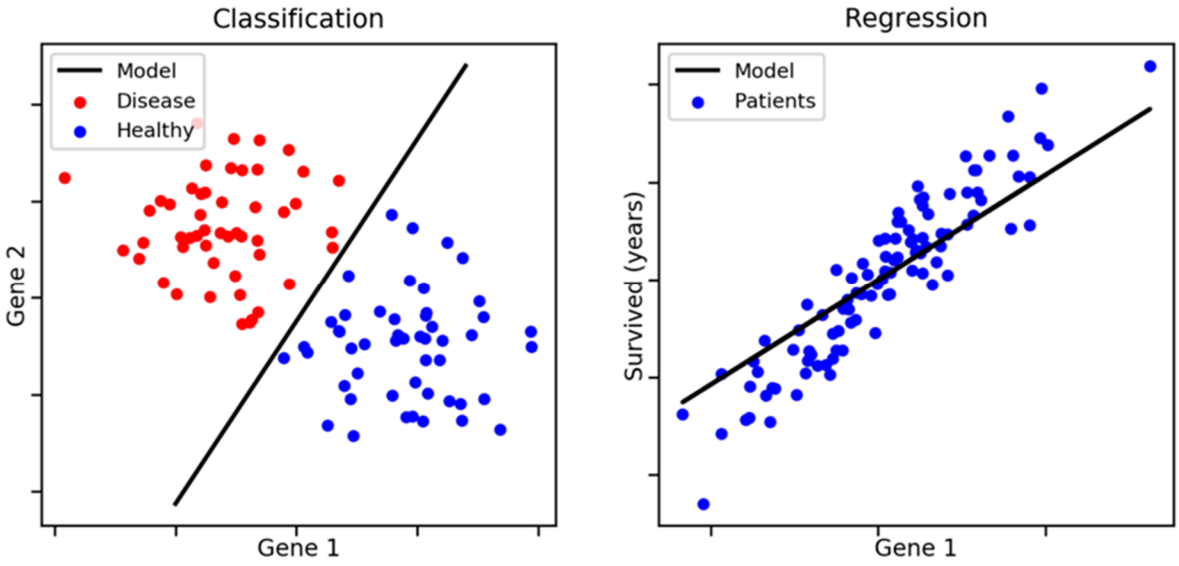


Figure 2.3 la classification et la régression

3.2.2. Apprentissage non supervisé

L'apprentissage non supervisé consiste à ne disposer que de données d'entrée (X) et pas de variables de sortie correspondantes. En ce qui concerne l'apprentissage non supervisé, un ensemble de données d'entrées non étiquetées est utilisé afin de permettre à l'algorithme d'apprentissage de trouver ses propres points communs parmi cet ensemble de données. Les approches d'apprentissage automatique qui simplifient l'acquisition de connaissances les données non étiquetées sont particulièrement bénéfiques, car elles sont plus importantes que celles étiquetées. On peut considérer que l'objectif initial de l'apprentissage non supervisé est aussi simple que de détecter les modèles cachés dans un ensemble de données, mais il peut aussi avoir un objectif d'apprentissage des caractéristiques, ce qui va rendre la machine intelligente capable de découvrir automatiquement les représentations nécessaires pour classer des données brutes. [21]

La figure 2.3 représente le fonctionnement d'apprentissage non supervisé. [24]

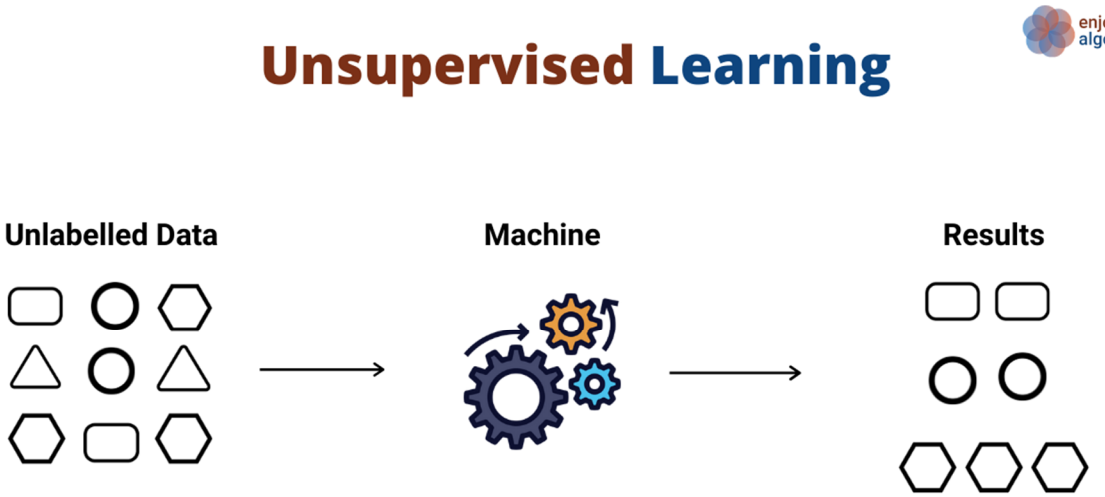


Figure 2.4 Le fonctionnement d'apprentissage non supervisé

On peut également les classer en types

- Cluster.
- Association.

A. Cluster

Le cluster consiste à séparer les objets en clusters qui sont similaires entre eux et différents des objets appartenant à un autre cluster. Par exemple, déterminer les clients qui ont effectué des achats similaires de produits. [25]

B. Association

Association est une méthode d'apprentissage automatique basée sur des règles afin de déterminer la probabilité de cooccurrence d'éléments dans une collection. Par exemple, déterminer quels produits ont été achetés ensemble. [25]

3.2.3. L'apprentissage par renforcement

L'apprentissage par renforcement fait référence à une classe de problèmes d'apprentissage automatique. Il consiste à apprendre à partir d'expériences successives, ce qu'il convient de faire de façon à trouver les meilleures solutions. Autrement dit, les machines intelligentes essaient plusieurs situations afin de pouvoir déterminer les actions les plus avantageuses, et ne se contentent pas de recevoir des instructions sur les actions à appliquer, ce qui distingue cette méthode des autres techniques d'apprentissage. L'apprentissage par renforcement est un modèle d'apprentissage

comportemental. L'algorithme dans ce cas reçoit les informations en analysant des données, pour pouvoir orienter l'utilisateur vers les meilleurs résultats. Dans ce type d'apprentissage, le système n'est pas entraîné à partir d'un ensemble de données mais il apprend par essais et erreurs, ce qui le diffère des autres types d'apprentissage supervisé. [21]

3.3. Apprentissage automatique dans les réseaux de données

L'apprentissage automatique a été largement appliqué aux problèmes de réseautage de données. Cette application a été réalisée dans de nombreux cas sous différents noms au sein de disciplines scientifiques bien connues comme le traitement du signal, la théorie de l'information, la théorie du codage, etc... Des exemples peuvent être observés dans l'application de la régression linéaire et non linéaire, des modèles statistiques, méthodes de compression, etc... Ces applications ont généralement été limitées aux techniques classiques d'apprentissage automatique. Cependant, de nombreuses avancées récentes en matière d'apprentissage automatique ne sont pas appliquées aussi pleinement aux réseaux que dans d'autres domaines (par exemple, traitement d'images, de parole et de vidéo, traitement du langage naturel). ,...).

Les algorithmes d'apprentissage automatique peuvent apprendre directement à partir des données sans tâche de programmation préalable explicite. L'apprentissage automatique, avec sa capacité à apprendre à partir des données, est particulièrement adapté aux problèmes trop complexes pour être entièrement définis ou dont la définition ne peut être effectuée avec précision. C'est précisément le genre de problèmes qui surviennent dans les réseaux.

De nombreux algorithmes d'apprentissage automatique peuvent être appliqués aux divers problèmes générés par les réseaux de données, tels que Random Forest, Gradient Boosting Machine (GBM), Support Victor Machine (SVM), Logistic Regression, Multinomial Régression logistique, Perceptron multicouche (MLP), K-Nearest Neighbours (KNN), Analyse en composantes principales, K-Means, Naïve Bayes et bien d'autres. [26]

3.4. Les algorithmes d'apprentissage automatique

La machine Learning offre un certain nombre d'algorithmes pour traiter des tâches de régression et de classification avec plusieurs variables dépendantes et indépendantes. Parmi ces algorithmes

- La régression linéaire (Linear Régression)

Les algorithmes de régression linéaire modélisent la relation entre des variables prédictives et une variable cible. La relation est modélisée par une fonction mathématique de prédiction.[27]

- La régression logistique (Logistic Régression)

La régression logistique est une méthode statistique pour effectuer des classifications binaires. Elle prend en entrée des variables prédictives qualitatives et/ou ordinales et mesure la probabilité de la valeur de sortie en utilisant la fonction sigmoïde. [27]

- Les machines à vecteurs de support (SVM)

SVM est l'un des algorithmes d'apprentissage supervisé les plus populaires, qui sont utilisé pour les problèmes de classification et de régression.

L'objectif de l'algorithme SVM est de créer la meilleure ligne ou limite de décision capable de séparer l'espace à n dimensions en classes afin que nous puissions facilement placer le nouveau point de données dans la bonne catégorie à l'avenir. Cette frontière de meilleure décision est appelée un hyperplan.

SVM choisit les points/vecteurs extrêmes qui aident à créer l'hyperplan. Ces cas extrêmes sont appelés vecteurs de support. [27]

- L'arbre de décision (Decision Trees)

L'arbre de décision est un algorithme qui se base sur un modèle de graphe (les arbres) pour définir la décision finale. Chaque nœud comporte une condition, et les branchements sont en fonction de cette condition (Vrai ou Faux). Plus on descend dans l'arbre, plus on cumule les conditions. La figure ci-dessous illustre ce fonctionnement.[27]

- Algorithme de forêt aléatoire [28]

Random Forest est un algorithme d'apprentissage automatique populaire qui appartient à la technique d'apprentissage supervisé. Il peut être utilisé pour les problèmes de classification et de régression en ML. Il est basé sur le concept d'apprentissage d'ensemble, qui est un processus de combinaison de plusieurs classificateurs pour résoudre un problème complexe et améliorer les performances du modèle.

- Les algorithmes de similarité

Les algorithmes de similarité évaluent la similarité des nœuds à un niveau individuel en fonction des propriétés des nœuds, des nœuds voisins ou des propriétés des relations.

Il existe deux algorithmes de similarité

- Similitude des nœuds.

- Voisins les plus proches approximatifs.

4. Réseau neuronal (Neural Network)

Un réseau de neurones artificiels est composé de nombreux neurones artificiels reliés entre eux selon une architecture de réseau spécifique. L'objectif du réseau de neurones est de transformer les entrées en sorties significatives. [29]

Les réseaux de neurones sont inspirés des neurones du système nerveux humains. Ils permettent de trouver des patterns complexes dans les données. Ces réseaux de neurones apprennent une tâche spécifique en fonction des données d'entrainement.

Les réseaux de neurones se composent de nœuds (les cercles dans l'image). Dans ces réseaux, on retrouve le tiers d'entrée (Input Layer) qui va recevoir les données d'entrées. L'Input Layer va propager les données par la suite aux tiers cachés (Hidden Layers). Finalement le Tiers de sortie (le plus à droite) permet de produire le résultat de classification. Chaque tiers du réseau de neurones est un ensemble d'interconnexions des noeuds d'un tiers avec ceux des autres tiers. [27]

5. Apprentissage profond

5.1. Définition

L'apprentissage en profondeur en anglais Deep Learning est un sous-domaine de l'intelligence artificielle (IA). Ce terme désigne l'ensemble des techniques d'apprentissage automatique (machine Learning), autrement dit une forme d'apprentissage fondée sur des approches mathématiques, utilisées pour modéliser des données. [30]

5.2. Quelques méthodes d'apprentissage profond

Nous présentons les principales méthodes d'apprentissage profond. La liste suivante n'est pas exhaustive, mais elle représente la grande majorité des algorithmes utilisés.

5.2.1. Machines de Boltzmann Restreintes

Une machine de Boltzmann est composée d'une couche de neurones qui reçoit l'entrée, ainsi que d'une couche de neurones cachée. Si nous supposons que les neurones d'une même couche sont indépendants entre eux, nous appelons cette configuration une machine de Boltzmann restreinte (RBM) [31]. Les machines de Boltzmann restreintes sont un type particulier de réseau de neurones génératifs, où les neurones sont organisés en deux couches, à savoir visible et masquée. Contrairement aux réseaux à retransmission directe, les données d'une RBM peuvent circuler dans les deux sens des unités visibles aux unités cachées, et inversement. La RBM est l'un des outils d'apprentissage en profondeur les plus populaires en raison de sa capacité à connaître la distribution de la probabilité des entrées de manière supervisée et non supervisée. elle a eu une large application

dans diverses tâches telles que l'apprentissage de la représentation, la réduction de la dimensionnalité, la classification, la régression, le filtrage collaboratif (collaborative filtering), l'apprentissage des fonctionnalités (feature Learning) et modélisation des sujets. [32]

5.2.2. Perceptrons multicouches

Des réseaux neuronaux organisés en plusieurs couches (au moins une couche cachée) au sein des quelles une information circule de la couche d'entrée vers la couche de sortie uniquement, il s'agit donc d'un réseau à propagation directe (feedforward) avec propagation anticipée et couches entièrement connectées. Chaque couche est constituée d'un nombre variable de neurones, les neurones de la dernière couche (dite« de sortie ») étant les sorties du système global. Le concept de base du perceptron singulier a été introduit par Rosenblatt en 1958. Le perceptron calcule une sortie unique à partir de multiples entrées à valeurs réelles en formant une combinaison linéaire en fonction de ses poids d'entrée, puis en plaçant éventuellement la sortie via une fonction d'activation non linéaire. [32]

5.2.3. Réseaux de croyance profonde (DBN)

Un réseau DBN est un type de réseau de neurones profonds qui est essentiellement un modèle génératif probabiliste comprenant plusieurs couches de variables cachées. Ces réseaux ont à la fois des bords dirigés et non dirigés. Il est formé à l'aide d'une série de RBM, souvent d'auto-encodeurs, avec une couche supplémentaire formant un réseau bayésien. L'utilisation de RBM signifie la présence d'aucune connexion intra-couche. De plus, les performances d'un DBN dépendent en grande partie de l'initialisation des nœuds. Par conséquent, les couches utilisent un apprentissage préalable non supervisé à l'aide de la procédure d'empilement de RBM, qui intègre une divergence contraste (CD). Un réseau de croyance (BN) est un graphe acyclique dirigé constitué de couches d'unités binaires stochastiques, chaque couche connectée ayant une pondération. Ces unités binaires stochastiques ont l'état 0 ou 1 et la probabilité d'être activé (devenant 1) est déterminée par un biais et une entrée pondérée provenant d'autres unités. [32]

5.2.4. Réseaux neuronaux convolutifs (CNN)

Un CNN est un réseau neuronal spécialisé à action directe utilisé à l'origine dans le traitement d'images mais de plus en plus utilisé dans de nombreux autres domaines. Ce type de réseau applique une collection de filtres pour extraire automatiquement les caractéristiques de l'image, créant finalement une structure hiérarchique de caractéristiques (apprentissage des représentations). Les poids des filtres sont appris directement à partir des données d'entraînement. Normalement, un CNN intègre de nombreuses couches convolutives créant une structure profonde. Un CNN effectue

automatiquement l'ingénierie des fonctionnalités, évitant ainsi l'étape longue et fastidieuse consistant à effectuer manuellement l'ingénierie des fonctionnalités. [26]

5.2.5. Réseaux neuronaux récurrents (RNN)

RNN a été initialement appliqué au traitement du langage naturel, mais comme CNN, il est actuellement incorporé à d'autres domaines. La principale application de RNN concerne les données séquentielles avec dépendances temporelles. Un RNN est capable de traiter de nouvelles données basées sur des données précédentes. Un problème important de RNN a été sa difficulté à être formé avec des données dépendant du temps à long terme (séries chronologiques longues). Afin de résoudre ce problème, une série de variantes RNN ont été créées. [26]

5.2.6. Réseaux de mémoire à long et court terme (LSTM)

Les LSTM proviennent des RNN. Le réseau LSTM est l'une de ces variantes, étant le plus largement utilisé. Ils ont la capacité d'acquérir et de conserver des dépendances sur une période prolongée. Les LSTM maintiennent donc les données mémorisées à long terme. Leur utilité réside dans leur capacité à prédire des séries chronologiques, car ils se souviennent des entrées antérieures. En plus de cette application, les LSTM sont également employés pour créer des notes musicales et détecter des voix. [33]

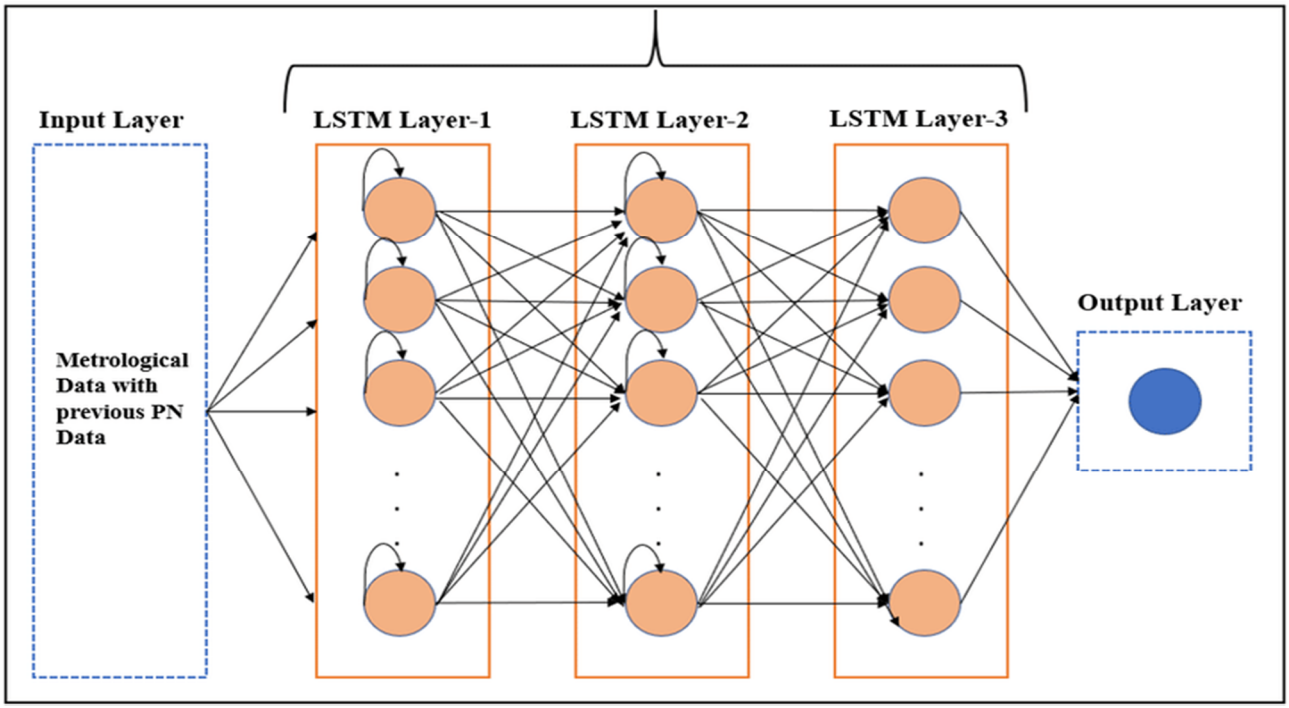


Figure 2.5 Architecture de modèle LSTM

5.2.7. Un réseau neuronal profond (ANN)

Est une branche de l'apprentissage automatique et un mécanisme standard pour résoudre les problèmes de vision par ordinateur. De plus, il s'agit d'un réseau neuronal artificiel avancé avec plusieurs couches entre les couches d'entrée et de sortie. Il fonctionne comme les neurones du cerveau humain et permet de remplacer le travail humain par un travail autonome. Lorsqu'il reçoit une nouvelle image dans le système, il découvre comment agir sur les situations signalées et résoudre les problèmes commerciaux en fonction des caractéristiques extraites de l'image d'entrée. [34]

6. Mesures d'évaluations (performances) des systèmes de détection d'intrusions [35]

La matrice de confusion est utilisée pour visualiser, pour chaque classe de modèle, les vraies classifications et les classifications prédites.

		Prédiction de la classe	
		Classe négative (normale)	Classe positive (attaque)
Classe actuelle	Classe négative (normale)	Vrai négative (VN)	Faux positive (FP)
	Classe positive (attaque)	Faux négative (FN)	Vrai positive (VP)

Tableau 2.1 La matrice de confusion

Les vrais négatifs ainsi que les vrais positifs correspondent à un fonctionnement correct de la technique de data mining, ce qui signifie que la technique de data mining a prédit avec succès respectivement le comportement normal et les attaques. Les faux négatifs sont des attaques incorrectement prédites comme des comportements normaux. Les métriques traditionnelles de classification comprennent le taux d'exactitude, le taux de fausse alerte et le taux d'erreur de la classification, elles sont définies comme suit

1 - Le taux d'exactitude (TE)

Montre à quel point le système est exact, c'est le nombre de type bien classé sur le nombre de type de tout le corpus.

$$Exactitude = \frac{\mathrm{VP} + \mathrm{VN}}{\mathrm{VP} + \mathrm{VN} + \mathrm{FP} + \mathrm{FN}}$$

2 - Le taux de fausse alerte (TFA)

Ce critère mesure le taux de fausses alertes générées par un IDS dans un environnement donné et pendant une durée donnée. C'est le nombre des alertes généré comme attaque sur le nombre des types classés comme normal existants dans le corpus.

$$FAR = \frac{FP}{\mathrm{VP} + \mathrm{FP}}$$

3 - Le taux de détection (DR Détection Rate)

Mesure le taux des attaques détectées par un IDS dans un environnement donné et pendant une durée donnée. C'est le nombre des attaques détecté sur le nombre des attaques existants dans le corpus.

$$DR = \frac{\mathrm{VN}}{\mathrm{VP} + \mathrm{FN}}$$

Dans (Porras & Valdes, 1998), il est défini trois critères pour évaluer l'efficacité des systèmes de détection d'intrusion

• L'exactitude (accuracy)

On parle de l'exactitude quand le système de détection d'intrusion déclare comme malicieux une activité légale. Ce critère correspond au faux Positif.

• La performance (performance)

La performance du système de détection d'intrusion est le taux de traitement des évènements. Si ce taux est faible, la détection en temps réel est donc impossible.

• La complétude (completeness)

On parle de la complétude quand le système de détection d'intrusion ne rate pas la détection d'une attaque. Ce critère est le plus difficile, parce qu'il est impossible d'avoir une connaissance globale sur les attaques.

Ce critère correspond au faux négatif.

• Debar dans (Llorens et al, 2011) a rajouté également ces critères suivant si

• La tolérance aux fautes (Fault tolerance)

Le système de détection d'intrusions doit Lui-même résisté aux attaques, particulièrement au déni de service. Ceci est important, parce que plusieurs systèmes de détection d'intrusion s'exécutent sur des matériels ou logiciels connus comme vulnérables aux attaques.

• La réaction à temps (Timeliness)

Le système de détection d'intrusion doit s'exécuter et propager les résultats de l'analyse le plus tôt possible, pour permettre à l'officier de sécurité de réagir avant que de graves dommages n'aient lieu.

• Rapidité

Un système de détection d'intrusions doit exécuter et propager son analyse d'une manière prompte pour permettre une réaction rapide dans le cas d'existence d'une attaque pour permettre à l'agent de sécurité de réagir.

Conclusion

Le machine Learning est un domaine de recherche très actif, qui ne cesse de progresser afin d'améliorer les performances des résultats.
L'apprentissage profond (Deep Learning) est un outil très puissant qui permet d'effectuer de multiples actions et révolutionner plusieurs domaines technologiques. Traduction automatique moderne, moteurs de recherche, assistants informatiques et plusieurs applications de notre vie quotidienne sont tous alimentés par un apprentissage profond.

1. Introduction

Après avoir exposé toute la théorie nécessaire au développement de notre système, nous abordons maintenant la deuxième partie afin de présenter notre travail.

Dans ce chapitre, nous nous concentrerons sur les différentes étapes clés du projet. Nous présenterons d'abord les différents outils, bibliothèques et langages de programmation utilisés. Ensuite, nous définirons notre ensemble de données en décrivant ses caractéristiques ainsi que les différentes étapes de prétraitement nécessaires pour corriger les valeurs aberrantes et choisir le modèle le plus approprié. Nous passerons ensuite à la sélection des modèles d'apprentissage (DNN, CNN et LSTM) et à la description de leurs architectures respectives. Enfin, nous évaluerons tous les modèles d'apprentissage utilisés et comparerons leurs performances afin de choisir celui qui offre la plus grande précision pour atteindre la qualité de service souhaitée.

2. Environnement d'exécution :

2.1. Google colab :

Google Colab est un environnement de développement en ligne basé sur Jupyter Notebook, qui offre la possibilité d'écrire, d'exécuter et de partager du code Python. Il fournit un accès gratuit à des ressources de calcul puissantes, y compris des unités de traitement graphique (GPU) et des unités de traitement tensoriel (TPU) pour accélérer l'exécution des tâches d'apprentissage automatique et de calcul intensif. [36]

2.2.Définition du langage Python en informatique :

Python est le langage de programmation open source le plus employé par les informaticiens. Ce langage s'est propulsé en tête de la gestion d'infrastructure, d'analyse de données ou dans le domaine du développement de logiciels. En effet, parmi ses qualités, Python permet notamment aux développeurs de se concentrer sur ce qu'ils font plutôt que sur la manière dont ils le font. Il a libéré les développeurs des contraintes de formes qui occupaient leur temps avec les langages plus anciens. Ainsi, développer du code avec Python est plus rapide qu'avec d'autres langages. [37]

2.3.DEFINITION JUPYTER :

Jupyter se présente comme un outil extrêmement simple à mettre en œuvre qui vous permettra de transformer vos Jupyter Notebooks en applications web ou en Dashboard quasiment automatiquement. [38]

3. Bibliothèque supplémentaires

– PANDA :

Pandas est un package Python open source qui est le plus largement utilisé pour la science et l'analyse des données. [39]

– NUMPY :

Le terme Numpy est en fait **l'abréviation de « Numerical Python »**. Il s'agit d'une bibliothèque Open Source en langage Python. On utilise cet outil pour la programmation scientifique en Python, et notamment pour la programmation en Data Science, pour l'ingénierie, les mathématiques ou la science. [40]

– SCIKIT LEARN :

Scikit-learn est une bibliothèque en Python qui offre de nombreux algorithmes d'apprentissage supervisé et non supervisé. Elle repose sur des technologies que vous connaissez peut-être déjà, telles que NumPy, pandas et Matplotlib.

Les fonctionnalités fournies par scikit-learn comprennent :

- Régression, compris la régression linéaire et logistique.
- Classification, compris les voisins les plus proches (K-Nearest Neighbors).
- Sélection de modèles.
- Prétraitement, compris la normalisation Min-Max.

Scikit-learn est une puissante bibliothèque qui facilite l'implémentation de diverses techniques d'apprentissage automatique dans vos projets Python. [41]

– TENSSORFLOW *:*

TensorFlow est une bibliothèque open-source de logiciels pour le flux de données et la programmation différentielle, utilisée pour diverses tâches. De la même manière, TensorFlow est utilisé dans l'apprentissage automatique par les réseaux neuronaux. Développé

par Google en 2011 sous le nom de DistBelief, TensorFlow a été officiellement publié en 2017 gratuitement. La bibliothèque est capable de s'exécuter sur plusieurs CPU et GPU, et est disponible sur différentes plateformes, compris les appareils mobiles. Le nom vient des tableaux multidimensionnels appelés tenseurs, qui sont couramment utilisés dans les réseaux neuronaux.

TensorFlow est une bibliothèque puissante qui permet de créer et d'entraîner des modèles d'apprentissage automatique avancés. Grâce à sa compatibilité avec différentes plates-formes et à sa capacité de tirer parti des ressources matérielles, TensorFlow offre une grande flexibilité pour les projets de machine Learning. [42]

– KERAS :

Keras est une bibliothèque open-source de composants de réseaux neuronaux écrits en Python. Keras est capable de s'exécuter sur TensorFlow, Theano, PlaidML et d'autres plates-formes. Cette bibliothèque a été développée pour être modulaire et conviviale, mais elle a initialement débuté en tant que projet de recherche pour le système d'exploitation intelligent neuro-électronique à réponse ouverte (ONEIROS). L'auteur principal de Keras est François Chollet, un ingénieur de Google qui a également créé le modèle de réseau neuronal profond Exception. Bien que Keras ait été officiellement lancé, il n'a été intégré à la bibliothèque principale TensorFlow de Google qu'en 2017. Un support supplémentaire a également été ajouté pour l'intégration de Keras avec le Microsoft Cognitive Toolkit.

Keras simplifie le processus de création et d'entraînement de réseaux neuronaux en fournissant une interface conviviale et une abstraction des détails complexes. Avec son intégration dans différentes bibliothèques de calcul numérique, Keras offre une flexibilité et une compatibilité étendues pour les projets d'apprentissage profond. [43]

4. Notre contribution

Notre contribution porte sur le développement de solutions avancées pour la gestion et l'analyse des données de trafic dans les systèmes et réseaux de communication modernes, tels que l'Internet des objets (IoT) et les réseaux cellulaires. Ces réseaux génèrent une quantité massive et hétérogène de données de trafic, posant des défis significatifs pour les techniques traditionnelles de gestion, notamment en termes de congestion, de retards, de pertes de paquets et de dégradation des

performances. En outre, garantir la qualité de service (QoS) et une classification précise du trafic représente un autre obstacle majeur.

Pour répondre à ces défis, nous explorons diverses architectures d'apprentissage profond appliquées à l'inspection profonde des paquets (DPI), telles que l les réseaux de mémoire à long terme (LSTM) et les réseaux de neurones profonds (DNN). En exploitant la puissance de l'apprentissage profond, nous visons à développer des algorithmes DPI intelligents capables de classer avec précision les applications en temps réel et de gérer de grandes quantités de données.

Notre travail se décompose en plusieurs étapes clés. Tout d'abord, nous nous concentrons sur le prétraitement de l'ensemble des données, en corrigeant les valeurs aberrantes. Ensuite, nous sélectionnons et décrivons les modèles d'apprentissage (ANN, CNN et LSTM). Enfin, nous évaluons tous les modèles d'apprentissage utilisés et comparons leurs performances afin de choisir celui qui offre la meilleure précision pour atteindre la qualité de service souhaitée.

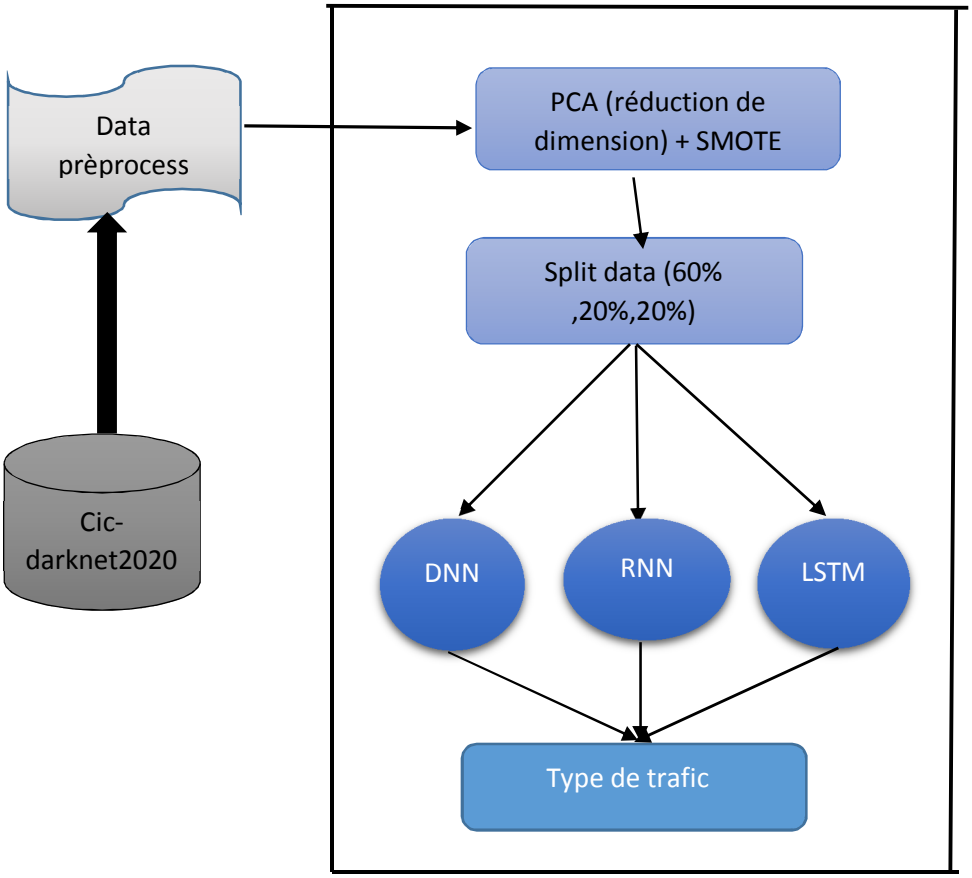


Figure 3.1 Processus global de notre approche basée sur le deep Learning

5. Ensemble de donnée [44]

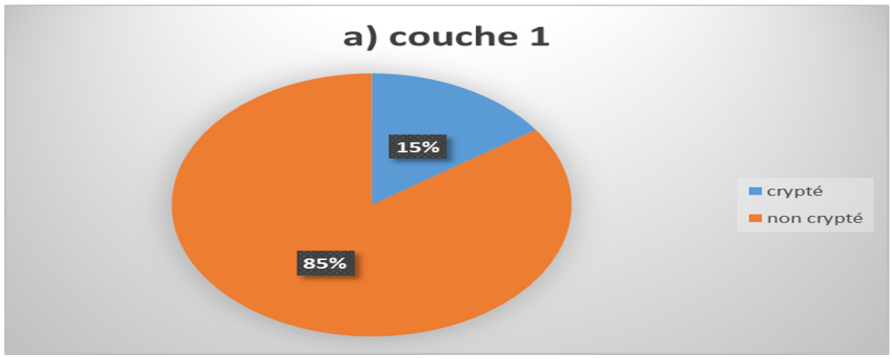
5.1.Description de l'ensemble de donnée CIC-DARKNET2020

Le dataset CIC-Darknet2020 a été développé par l'Institut canadien pour le cyber sécurité de l'Université du Nouveau-Brunswick. Il est conçu pour modéliser des environnements réseau réalistes, incluant des trafics légitimes, et tester de nouvelles méthodes de classification du trafic Darknet. Le Darknet désigne une partie de l'Internet qui n'est pas indexée par les moteurs de recherche traditionnels et qui nécessite des logiciels spécifiques pour y accéder, tels que Tor (The Onion Router) ou I2P (Invisible Internet Project). Il met l'accent sur l'anonymat et la sécurité, permettant aux utilisateurs de naviguer et de communiquer sans révéler leur adresse IP. Le Darknet est souvent associé à des activités anonymes et parfois illégales, bien qu'il puisse également être utilisé pour des communications privées et sécurisées. Dans l'ensemble de données CIC-Darknet2020, une approche à deux niveaux est utilisée pour générer du trafic bénin et du trafic darknet. Au premier niveau, le trafic bénin est créé. Le trafic Darknet, quant à lui, comprend des flux audio, de la navigation, du chat, du courrier électronique, du P2P, du transfert de fichiers, du streaming vidéo et de la VOIP, qui sont générés au second niveau. Pour produire un ensemble de données représentatif, nous avons fusionné nos ensembles de données générés précédemment, à savoir ISCXTor2016 et ISCXVPN2016, en combinant le trafic VPN et Tor respectif dans les catégories de trafic Darknet correspondantes. Les caractéristiques des flux sont disponibles en fichiers CSV et les captures de paquets en fichiers PCAP, facilitant ainsi leur utilisation dans divers outils d'analyse. Le tableau 1 fournit les détails des catégories de trafic DarkNet et des applications utilisées pour générer ce trafic réseau.

Catégorie de trafic	Application utilisée
Flux audio	Vimeo et YouTube
navigation	Firefox et Chrome
chat	ICQ, AIM, SKYPE, Facebook et hangouts
e-mail	SMTPS, POP3S et IMAPS
P2P	uTorrent et Transmission
transfert	Skype, FTP sur SSH (SFTP) et FTP sur SSL (FTPS) utilisent Filezilla et un service externe
Flux video	Vimeo et Youtube
VoIP	Appels vocaux Facebook, Skype et Hangouts

Tableau 3.1 Types de trafic réseau de l'ensemble de données CIC-Darknet2020

La figure 1 (a) présente les détails du nombre d'échantillons de trafic bénin et de trafic darknet au niveau de la première couche, et (b) met en évidence le nombre de flux cryptés dans notre trafic darknet.



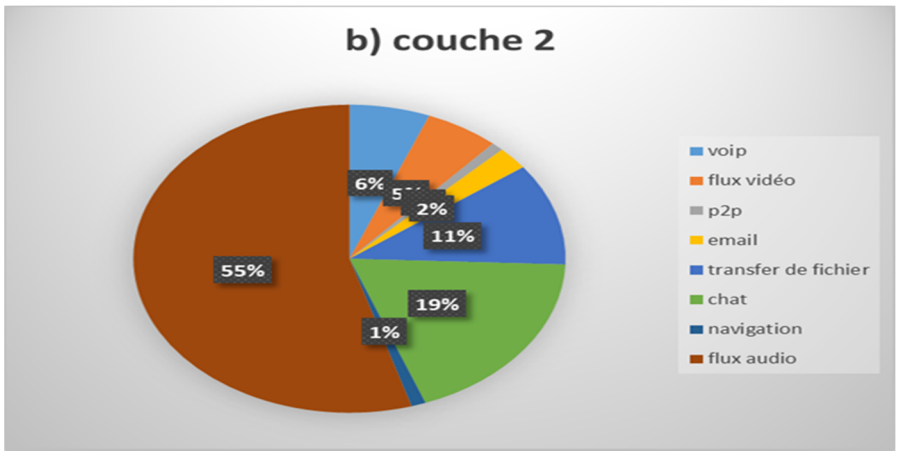


Figure 3.2 Distribution des types de trafic dans l'ensemble de données CIC-Darknet2020

5.2.La préparation des données

Les performances des méthodes de deep Learning dépendent fortement de la quantité et de la qualité des données d'apprentissage : plus il y a de données de qualité, plus les résultats sont précis et performants. Dans notre cas, nous disposons d'une quantité de données très suffisante. Cependant, en raison du déséquilibre des classes de données et du grand nombre de caractéristiques, il est nécessaire de réduire les caractéristiques des données et d'équilibrer les classes.

5.3. La réduction des données

Le volume de notre ensemble de données nous oblige à réduire un nombre important de caractéristiques. Pour ce faire, nous utilisons l'analyse en composantes principales (PCA). La PCA est une technique de réduction dimensionnelle qui transforme les caractéristiques initiales en un ensemble de nouvelles variables non corrélées, appelées composantes principales. Cette méthode permet de conserver l'essentiel de l'information tout en réduisant la complexité des données. En appliquant la PCA, nous pouvons améliorer l'efficacité de nos modèles de Deep Learning, réduire leur complexité et gagner du temps lors de l'apprentissage.

5.4. L'équilibrage des données (Balance de dataset)

Le modèle d'apprentissage risque de prédire principalement les classes majoritaires et de ne pas détecter correctement les classes minoritaires, ce qui introduit un biais dans le modèle. Pour résoudre ce problème, différentes méthodes d'échantillonnage ont été proposées, telles que le sur échantillonnage aléatoire, qui réplique de manière aléatoire les échantillons exacts des classes minoritaires, et le sur échantillonnage en créant des échantillons synthétiques des classes minoritaires à l'aide de techniques comme la technique de sur échantillonnage synthétique des classes minoritaires (SMOTE). Dans cette étude, nous avons utilisé la technique d'échantillonnage SMOTE, car elle est capable de traiter des ensembles de données mixtes comprenant des caractéristiques catégorielles et continues.

5.5. Les prétraitements des données

Afin de construire un modèle très précis, il est important d'effectuer des analyses exploratoires sur l'ensemble des données et ses caractéristiques. Le prétraitement de l'ensemble des données est effectué avant d'être appliqué au réseau neuronal profond. Les étapes de prétraitement sont les suivantes :

– ***Filtrage des données***: Tout d'abord, l'ensemble des données a été filtré afin de supprimer toutes les lignes redondantes représentant les instances de classe. Ensuite, une analyse a été effectuée pour détecter toute valeur 'NAN' (Not A Number) ou 'INF' (Infinite Value). Ces valeurs peuvent être considérées comme des valeurs manquantes. Les algorithmes de deep learning ou de machine learning en général traitent très mal ces valeurs, ce qui affecte directement et négativement les performances des modèles finaux. Il apparaît que les données sélectionnées pour cette étude comportent plusieurs valeurs 'NAN' dans la colonne Flow Bytes. Pour garder cette caractéristique et comme nous avons des données suffisantes, les lignes avec des valeurs NAN ou INF ont été supprimées.

- ***Encodage des caractéristiques catégorielles :*** Il y a un certain nombre de caractéristiques de type catégoriel dans l'ensemble des données qui doivent être encodées. Par exemple, la colonne Flow Paquets a été convertie en une colonne numérique.

– ***Encodage de la colonne Label :*** La colonne Label, qui représente la classe de chaque instance, a été encodée avec une technique populaire appelée "One-Hot-Encoding". Ce codage convertit les lignes contenant des catégories en leur propre colonne, avec une valeur numérique.

– ***Normalisation des données*** : Lorsque nous obtenons des données de qualité, où chaque instance des classes comporte des informations bien décrites, la prochaine étape avant l'apprentissage est la normalisation. Les données d'entrée doivent être normalisées. Cette étape a un effet positif sur la construction du modèle en réduisant le taux d'apprentissage et en accélérant la convergence du modèle. Elle peut aussi avoir un effet de régularisation en réduisant l'erreur de généralisation.

– ***Division des données :*** Les données d'entraînement sont divisées en deux parties : des données pour l'apprentissage et des données pour la validation du modèle (80 % pour l'apprentissage et 20 % pour la validation).

6. Les métriques et évaluation

Dans cette section, nous avons évalué les performances de notre modèle en nous concentrant sur trois mesures : l'exactitude, la précision, le rappel et le f1-score.

7. Résultat obtenu

Nous avons effectué une comparaison des performances des algorithmes que nous avons utilisés, à savoir , DNN (Réseau de neurones profond), CNN (Réseau de neurones convolutif) et LSTM (Long Short-Term Memory). Cette comparaison a été réalisée grâce à des expériences menées sur deux ensembles de données : l'un clair (non chiffré) comme illustré dans le tableau 3.2, et l'autre chiffré par des réseaux VPN comme présenté dans le tableau 3.3.

Classifier	Accuracy	Précision	Recall	F1-Score
DNN	0.92	0.93	0.93	0.93
CNN	0.79	0.79	0.79	0.79
LSTM	**0.95**	**0.95**	**0.95**	**0.95**

Tableau 3.2 Étude comparative des classificateurs sur un ensemble de données non chiffré.

Parmi les résultats présentés dans le tableau 3.2, le modèle LSTM se distingue en surpassant les autres modèles sur toutes les métriques (exactitude, précision, rappel et F1-score). Cela indique que le modèle LSTM est le plus efficace pour capturer les motifs présents dans le jeu de données, grâce à sa capacité à gérer les données séquentielles et à apprendre les dépendances à long terme. Dans le tableau 3.2, nous évaluons le meilleur modèle, LSTM, avec un ensemble de données chiffrées. Avant cela, nous présentons les hyper paramètres de notre modèle dans la section suivante.

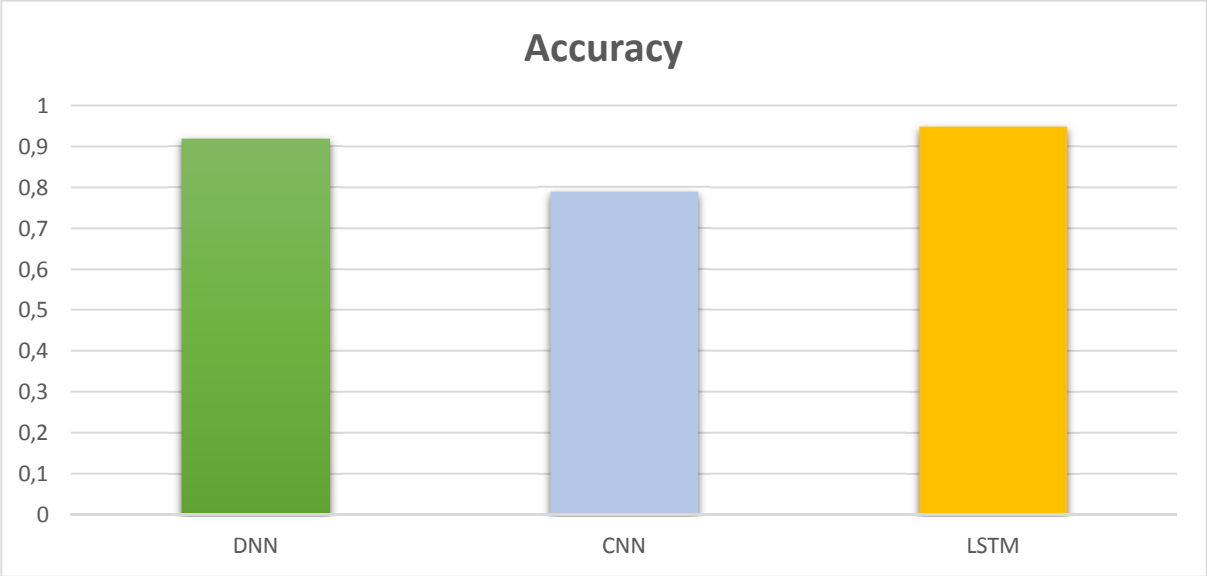


Figure 3.3 Comparaison des résultats d'accuracy obtenus par divers modèles.

7.1.Les hyper paramètres de notre modèle LSTM :

Type de couche	Paramètres
LSTM (LSTM)	Unités: 512
Dropout (Dropout)	0.2
BatchNormalization	
LSTM (LSTM)	Unités: 256
Dropout (Dropout)	: 0.2
BatchNormalization	
LSTM (LSTM)	Unités: 64
Dropout (Dropout)	0.2
BatchNormalization	
Couche de sortie	Unités: Nombre de classes, Activation: softmax
Fonction d'optimisation	Adam (learning_rate=0.001)
Fonction de perte	Sparse Categorical Crossentropy

Tableau 3.4 hyper paramètres de notre modèle LSTM

Nous avons utilisé dans l'architecture BatchNormalization et Dropout, qui sont des techniques complémentaires. BatchNormalization stabilise et accélère l'apprentissage tout en réduisant les risques d'instabilité, tandis que Dropout prévient le sur apprentissage en introduisant de la régularisation et en rendant le modèle plus robuste. Ces deux techniques contribuent à améliorer la performance générale et la capacité de généralisation des réseaux neuronaux.

7.2.L'évaluation de notre modèle avec un ensemble de données chiffrées :

Le tableau 3.5 présente la performance du meilleur modèle, LSTM, avec un ensemble de données chiffrées. La figure 3 montre la matrice de confusion.

Classifier	Accuracy	Précision	Recall	F1-Score
LSTM	0.89	0.90	0.90	0.90

Tableau 3.5 Étude comparative entre les classificateurs sur un ensemble de données chiffré par VPN.

Type de Traffic	précision	Recall	f1-score
VPN-BROWSING	0.81	0.85	0.83
VPN-CHAT	87	0.87	0.87
VPN-MAIL	0.94	0.89	0.91
VPN-P2P	0.81	0.91	0.86
VPN-STREAMING	0.99	0.97	0.98
VPN-VOIP	0.97	0.88	0.92

Tableau 3.5 Performance de notre modèle LSTM en classification multi-classes pour chaque classe sur l'ensemble de données chiffrées par VPN

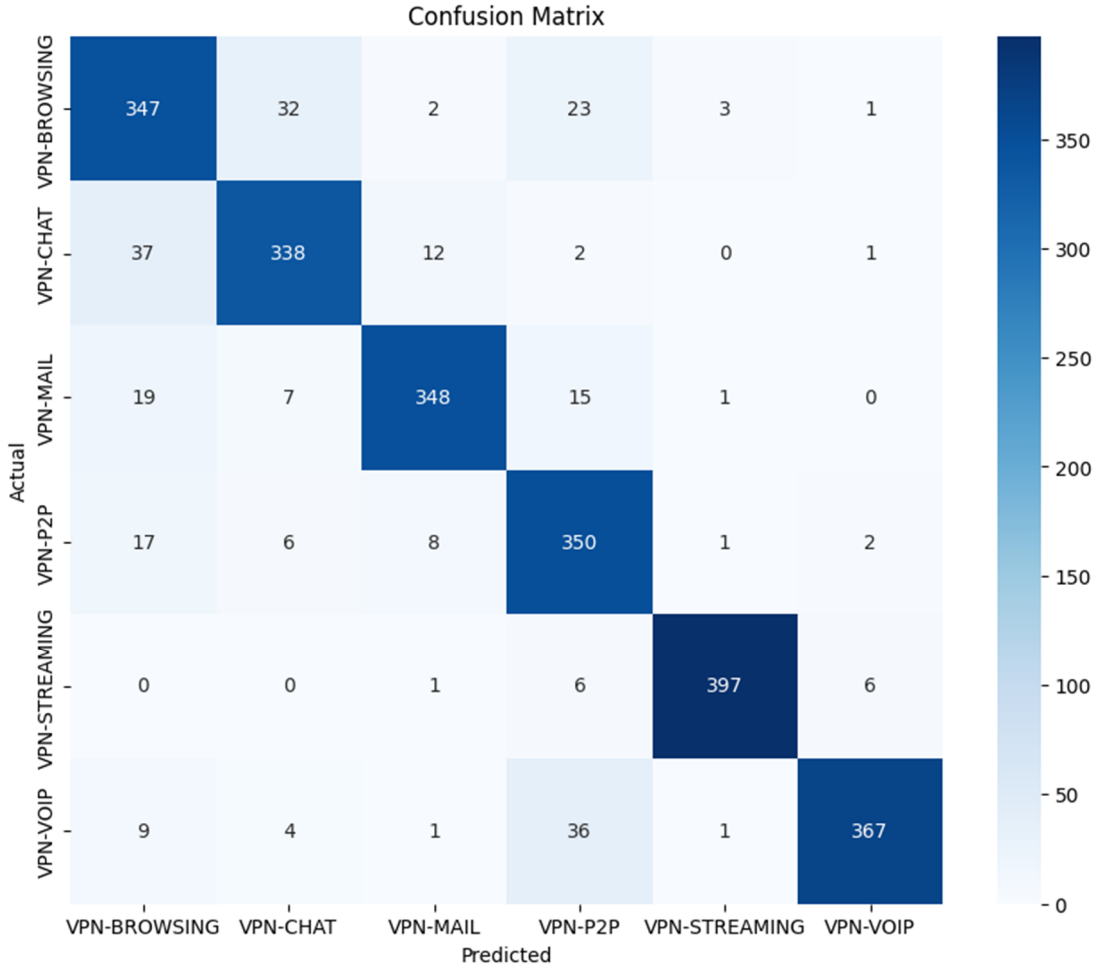


Figure 3.4 Matrice de confusion du modèle LSTM sur l'ensemble de données chiffrées avec VPN

La figure 3.5 présente l'accuracy du modèle durant l'entraînement et la validation au fil des époques. La figure 3.6 présente la perte du modèle durant l'entraînement et la validation au fil des époques.

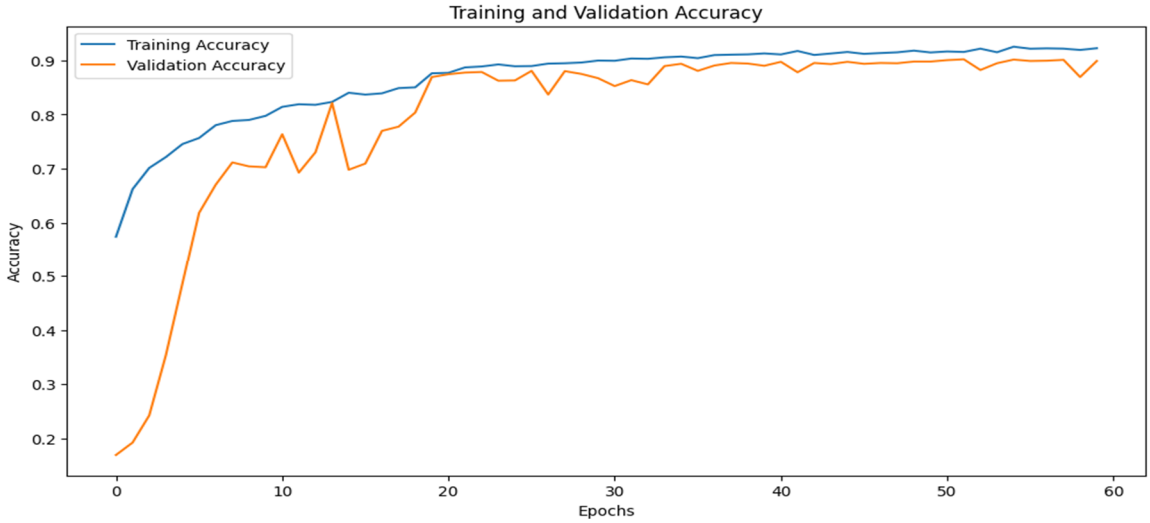


Figure3.5 Accuracy de l'entraînement et de la validation

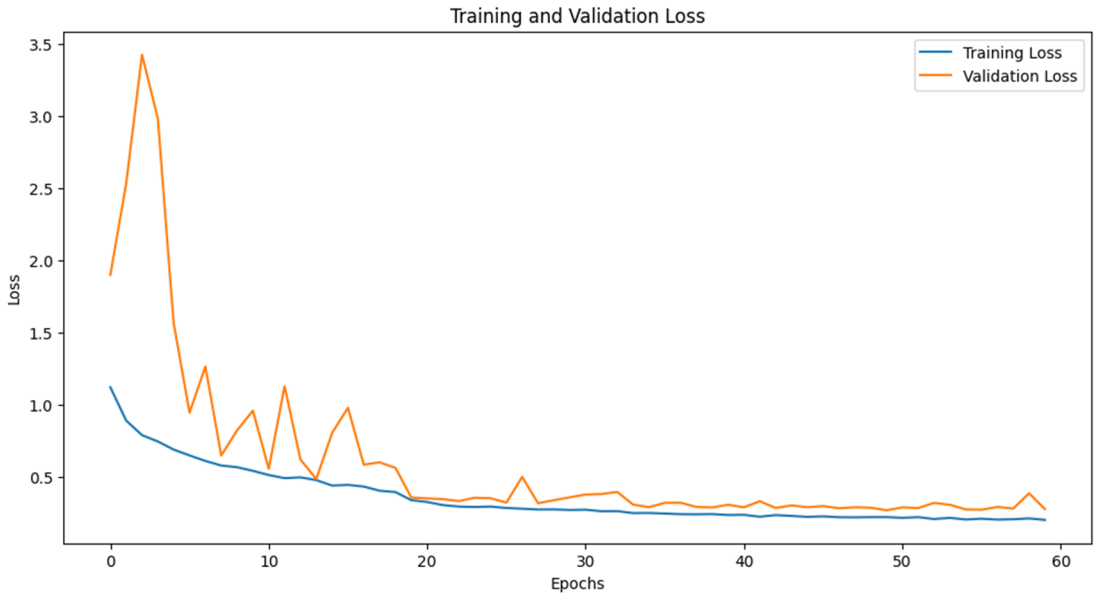


Figure 3.6 perte de l'entraînement et de la validation

L'analyse des résultats des figures 3.5 et 3.6 de notre modèle LSTM sur l'ensemble des données chiffrées montre une performance très prometteuse. L'accuracy de l'entraînement atteint rapidement 90 %, tandis que celle de la validation se stabilise entre 80 % et 85 %, indiquant une bonne généralisation. La diminution constante de la perte d'entraînement témoigne de l'efficacité du modèle à minimiser l'erreur.

Ces résultats indiquent que notre modèle a obtenu des performances supérieures dans la classification du trafic, tant en cas clair qu'en cas chiffré. En conclusion, notre modèle DPI

démontre une performance supérieure et est capable de classer avec précision les applications en temps réel pour différents types de trafic, améliorant ainsi la QoS.

Conclusion

Dans ce chapitre, nous avons présenté l'aspect pratique de notre recherche sur la classification du trafic Internet à l'aide de l'apprentissage automatique. Nous avons détaillé l'architecture proposée de notre système, les algorithmes utilisés, notre ensemble de données et la manière de les exploiter. Nous avons conclu le chapitre par une série d'expériences pour évaluer les performances de notre modèle. Les résultats obtenus montrent que notre modèle basé sur LSTM surpasse les autres modèles. Ces performances supérieures ont été constatées dans la classification du trafic, qu'il soit clair ou chiffré, démontrant ainsi la capacité de notre modèle DPI à classer avec précision les applications en temps réel pour différents types de trafic, améliorant ainsi la QoS.

Conclusion Générale

En conclusion, ce travail a exploré l'application de l'inspection approfondie des paquets (DPI) en utilisant le deep Learning pour garantir la qualité de service (QoS) dans les réseaux de communication.

Nous avons débuté par une exploration des concepts généraux du trafic réseau, de la sécurité réseau, de la QoS et de l'inspection approfondie des paquets (DPI).

Ensuite, nous avons examiné les principes fondamentaux de la machine Learning et du deep Learning, ainsi que les différents algorithmes utilisés dans notre étude.

Nous avons ensuite détaillé l'implémentation de notre travail, mettant en œuvre les algorithmes de deep Learning CNN, DNN et LSTM pour la classification du trafic et l'amélioration de la QoS. Nos résultats ont démontré que le modèle basé sur LSTM surpassait les autres modèles, atteignant des performances exceptionnelles avec un taux de rappel, un F1-Score, une précision et une exactitude proches de la perfection.

En conclusion, notre travail contribue à l'avancement de la gestion du trafic réseau et de la sécurité dans les réseaux de communication en utilisant des techniques de deep Learning pour l'inspection approfondie des paquets. Notre modèle DPI offre une classification précise et efficace du trafic en temps réel, même pour les données chiffrées, ce qui améliore significativement la QoS et renforce la sécurité des réseaux. Ces résultats ouvrent la voie à de futures recherches et applications dans le domaine de la gestion des réseaux et de la sécurité informatique.

Bibliographie

[1] *https://blog.lesjeudis.com/fonctionnement-reseau-informatique.* Accède le 25/06/2024

[2] Soulimane Kamel Eddine et Sebbagh Saad Allah, «Problème et solution de sécurité d'un réseau Wi-Fi,» Mémoire présenté pour l'obtention Du diplôme de Master Académique, Spécialité Réseaux et Télécommunications, Telemcen, 2022

[3] S.Natkin "Les protocoles de sécurité d'Internet", Dunod science sup,

[4] GHERNAOUTI, Solange. Sécurité informatique et réseaux. Dunod, 2013, p. 18

[5] ALEM Abdelkader, « principes de sécurité », Support de cours.

[6] https://www.fortinet.com/fr/resources/cyberglossary/network-traffic.html. [Consulté le 20 mai 2024].

[7] H. Guebailia, "L'impact de la qualité des services sur la satisfaction des clients dans les entreprises algériennes : Cas d'Algérie Télécom Mobilis (A.T.M) Guelma," Mémoire de master, Université du 08 mai 1945, Guelma, Département des sciences de gestion, Faculté des sciences économiques et commerciales et sciences de gestion, 2024

[8] "What Is Quality of Service (QoS) in Networking?," Fortinet, [Online]. Available: https://www.fortinet.com/resources/cyberglossary/qos-quality-of-service. Accessed: May 22, 2024.

[9] phanivedala, "What Is Network Congestion? Causes and How to Fix," Learning Center, Oct. 20, 2022. [Online]. Available: https://www.extnoc.com/learn/general/network-congestion.

[10] "What Is Network Congestion?," Avi Networks, [Online]. Available: https://avinetworks.com/glossary/network-congestion/. Accessed: Apr. 23, 2024.

[11] "Qu'est-ce que l'inspection approfondie des paquets (DPI) ?," Fortinet, [Online]. Available: https://www.fortinet.com/fr/resources/cyberglossary/dpi-deep-packet-inspection.html. Accessed: May 21, 2024.

[12] Malekal, "Qu'est-ce que le DPI (Deep Packet Inspection) ?," Malekal.com, 2024. [Online]. Available: https://www.malekal.com/quest-ce-que-le-dpi-deep-packet-inspection/. Accessed: May 22, 2024.

[13] "Deep Packet Inspection and NetWitness Elevating Network Security," NetWitness.Com, [Online]. Available: https://www.netwitness.com/blog/deep-packet-inspection-dpi-enhancing-network-security-with-netwitness/. Accessed: May 21, 2024.

[14] United States House of Representatives, Committee on Energy and Commerce, United States Congress, "What Your Broadband Provider Knows about Your Web Use Deep Packet Inspection and Communications Laws and Policies," 2019.

[15] Fortinet, [Online]. Available: https://www.fortinet.com/fr/resources/cyberglossary/dpi-deep-packet-inspection.html. Accessed: Apr. 15, 2024.

[16] "What is Deep Packet Inspection? | NordVPN," Jun. 29, 2023. [Online]. Available: https://nordvpn.com/blog/deep-packet-inspection/.

[17] "Deep Packet Inspection (DPI)," ESDAcademy.blogspot.com, [Online]. Available: https://esdacademy.blogspot.com/2019/05/deep-packet-inspection-dpi.html. Accessed: May 21, 2024.

[18] N. Sharma and B. Arora, "Review of Machine Learning Techniques for Network Traffic Classification," SSRN 3747605, Dec. 12, 2020. doi: 10.2139/ssrn.3747605.

[19] Europarl.europa.eu, "Intelligence artificielle : définition et utilisation," Parlement européen, Aug. 27, 2020. [Online]. Available: https://www.europarl.europa.eu/topics/fr/article/20200827STO85804/intelligence-artificielle-definition-et-utilisation.

[20] L. Dekkiche, "Classification des arythmies ECG avec des méthodes de Machine Learning et de Deep Learning," Mémoire de fin d'étude, Université [MOULOUD MAAMERI TIZI-OUZOU, Département d'informatique], 2020.

[21] A. Hadj Mohand and R. Abderrahmani, "Implémentation d'un modèle de Deep Learning basé sur un réseau de neurones sur la carte STM32," Mémoire de fin d'étude, Université Mouloud Mammeri de Tizi Ouzou, Département d'Informatique, Spécialité Réseaux, Mobilité et Systèmes Embarqués, année de soutenance non précisée.

[22] Z. ISMAILI, "Apprentissage Supervisé Vs. NonSupervisé," BrightCape, Jan. 28, 2019. [Online]. Available: https://brightcape.co/apprentissage-supervise-vs-non-supervise/.

[23] "Principe de l'apprentissage supervisé schéma d'une unité logistique," disponible sur ResearchGate, [En ligne]. Disponible sur : https://www.researchgate.net/figure/2-Principe-de-lapprentissage-supervise-schema-dune-unite-logistique_fig7_280735747.

[24] "Introduction to Unsupervised Learning," Bombay Softwares, [En ligne]. Disponible sur : https://www.bombaysoftwares.com/blog/introduction-to-unsupervised-learning.

[25] "Supervised and Unsupervised Learning," Simplilearn, [En ligne]. Disponible sur : https://www.simplilearn.com/tutorials/machine-learning-tutorial/supervised-and-unsupervised-learning. Consulté le 29 avril 2024.

[26] M. Lopez-Martin, "PhD Thesis Novel applications of Machine Learning to Network Traffic Analysis and Prediction," ResearchGate, DOI 10.13140/RG.2.2.18277.76008.

[27] F. Z. GACEM and S. FERNANE, "Vers une approche pour les systèmes de détection d'attaque de réseau basée sur données structurées de graphe," Mémoire de Master, Université Ibn Khaldoun, Tiaret, Algérie, 2022.

[28] "Machine Learning Random Forest Algorithme - Javatpoint," Www.Javatpoint.Com, [En ligne]. Disponible sur : https://www.javatpoint.com/machine-learning-random-forest-algorithm. Consulté le 28 mai 2024.

[29] S. Park and H. Park, "ANN Based Intrusion Detection Model," Suisse, Springer, pp. 433–437, 2019.

[30] C. Deluzarche, "Définition | Deep Learning - Apprentissage profond | Futura Tech," Futura, [En ligne]. Disponible sur : https://www.futura-sciences.com/tech/definitions/intelligence-artificielle-deep-learning-17262/. Consulté le 29 avril 2024.

[31] A. Osseiran et al., "Scenarios for 5G mobile and wireless communications The vision of the metis project," IEEE Commun. Mag., vol. 52, no. 5, pp. 26–35, Mai 2014.

[32] S. Martin, "Anti-IDS Tools and Tactics," SANS Technology Institute, Août 2001.

[33] "Algorithmes de Deep Learning," Jedha, [En ligne]. Disponible sur : https://www.jedha.co/formation-ia/algorithmes-deep-learning. Consulté le 30 avril 2024.

[34] "Deep Neural Network (DNN) : Its Scope and Nature of Complexity," Iotric, [En ligne]. Disponible sur : https://iotric.medium.com/deep-neural-network-dnn-its-scope-and-nature-of-complexity-56af59f87ea4.

[35] S. L. Reguieg, "Étude et Analyse des différentes intrusions réseaux et la détection d'attaque DOS à l'aide des algorithmes d'apprentissage automatique," Mémoire de Master, Université Ibn Khaldoun, Tiaret, 2020.

[36] «https://datascientest.com/google-colab-tout-savoir/ Accède le 09/06/2024»

[37] "Python." https://www.journaldunet.fr/web-tech/dictionnaire-duwebmastering/1445304-python-définition-et-utilisation-de-ce-langage-informatique/.

[38] https://jupyter.org/ Accéde le 25/06/2024

[39] “pandas - Python Data Analysis Library.” https://pandas.pydata.org/

[40] «https://numpy.org/ Accède le 02/06/2024 »

[41]«https://www.inria.fr/fr/lancement-de-linitiative-scikit-learn?fbclid=IwAR1r89W0NsQHyu7BN31qRQJq5YEUS0iORwj37i51Zj0ds35stAwHCL-8N8c Accède le 02/06/2024. »

[42] «https://www.intelligence-artificielle-school.com/ecole/technologies/tensorflow/Accède le 09/06/2024»

[43] «https://datascientest.com/keras/ Accède le 09/06/2024»

[44] « Darknet 2020 | Datasets | Research | Canadian Institute for Cybersecurity | UNB ». Consulté le: 8 juin 2024. [En ligne]. Disponible sur: https://www.unb.ca/cic/datasets/darknet2020.html

[45] Nguyen, T.T.T., & Armitage, G. (2008). A Survey of techniques for internet traffic classification using machine learning. IEEE Communications Surveys & Tutorials, 10(4), 56-76.

[46] Park, K., et al. (2008). Analysis on the TCP SYN Flooding Attacks. Journal of Network and Systems Management, 16(3), 307-323.

[47] Shen, C., & Fan, C. (2008). Improved decision tree algorithm for network traffic classification. International Conference on Computer Science and Software Engineering.

[48] Yoon, C., et al. (2015). Application-level classification using machine learning for large-scale network traffic. International Journal of Computer Science and Network Security, 15(10), 29-35.

[49] Yuan, J., et al. (2014). Deep learning methods for traffic classification. IEEE Transactions on Network and Service Management, 11(3), 327-336.

[50] Cai, H., et al. (2010). An adaptive approach to network traffic classification based on machine learning. IEEE Transactions on Information Forensics and Security, 5(1), 169-179.

[51] Coull, S.E., & Dyer, K.P. (2014). Traffic Analysis of Encrypted Messaging Services: Apple iMessage and Beyond. ACM SIGCOMM Computer Communication Review, 44(5), 5-11.

[52] Cuadra-Sanchez, A., & Aracil, J. (2017). Network traffic classification using one-dimensional convolutional neural networks. International Conference on Network and Service Management.

[53] Datta, S., et al. (2015). Bayesian learning and neural networks for network traffic classification. IEEE Transactions on Neural Networks and Learning Systems, 26(1), 225-238.

[54] Dorfinger, J. (2010). Seq2Img: Sequence learning with convolutional neural networks. International Conference on Machine Learning and Applications.

[55] Ehlert, S., et al. (2006). Deep packet inspection for P2P traffic classification. IEEE Communications Letters, 10(4), 337-339

[56] Fu, X., et al. (2016). Network traffic classification using deep learning. IEEE Transactions on Industrial Informatics, 12(6), 2114-2124.

[57] Goo, J., et al. (2016). Improved network traffic classification method using simple IP packet properties. IEEE International Conference on Communications.

[58] Janani, S., & Ramamoorthy, M. (2022). Enhancing mobile network traffic classification using correlated information. International Journal of Mobile Network Design and Innovation, 12(1), 54-66.

[59] Liu, Y., et al. (2019). Mobile data traffic classification with deep learning. Journal of Network and Computer Applications, 136, 1-10.

[60] Rahman, M., et al. (2022). Traffic classification using message statistics in IoT networks. IEEE Internet of Things Journal, 9(5), 3551-3560.

[61] Kumar, S., & Sharma, A. (2016). Spiking neural networks for network traffic classification. IEEE Transactions on Neural Networks and Learning Systems, 27(12), 2614-2623.

[62] Lee, Y., et al. (2015).