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 DE l'INFORMATIQUE
DÉPARTEMENT D'INFORMATIQUE

Pour l'obtention du diplôme de :

MASTER

Spécialité : Réseaux et Télécommunication

Par :

Ouadah Bochra
Boughaddou Bochra

Sur le thème

Classification du Trafic Réseau :Vers une Amélioration de la Qualité de Service à l'Aide de Deep Learning

Soutenu publiquement le .. / 06/ 2025 à Tiaret devant le jury composé de :

Mr MEGHAZI Hadj Madani	MCB Université de Tiaret	Président
Mr MOSTEFAOUI Kadda	MCB Université de Tiaret	Encadrant
Mr LAHCEN Aid	MCA Université de Tiaret	Examinateur

2024-2025

Remerciements

بسم الله الرحمن الرحيم

قال الله تعالى: فَاذْكُرُونِي أَذْكُرْكُمْ وَاشْكُرُوا لِي وَلَا تَكْفُرُونِ

Sourate Al-Baqarah, Ayah 152

Au nom de Dieu, le Tout Miséricordieux, le Très Miséricordieux.

Nous remercions avant tout Dieu pour Sa guidance et la force qu'Il nous a donné pour mener à bien ce travail.

Nous exprimons notre profonde gratitude à ***Monsieur Mostefaoui Kadda****, notre encadreur, pour son écoute attentive, sa disponibilité constante, et son accompagnement précieux tout au long de ce projet. Que ses efforts soient récompensés.*

Nous adressons nos plus vifs remerciements à ***Monsieur Bouazza Abdel Hamid****, notre professeur, pour son soutien indéfectible, sa patience, ses précieux conseils, ainsi que son aide continue tout au long de la préparation et la présentation de ce mémoire. Sa disponibilité et son professionnalisme ont été essentiels à la réussite de ce travail.*

Nos sincères remerciements vont aussi aux membres du jury, ***Monsieur Aid Lahcene*** *et* ***Monsieur Meghazi Hadj Madani****, pour avoir accepté d'évaluer ce travail et pour le temps qu'ils lui ont consacré.*

Enfin, nous tenons à exprimer notre reconnaissance à toutes les personnes qui ont contribué, de près ou de loin, à la réalisation de ce mémoire, ainsi qu'à l'ensemble des enseignants et responsables de l'Université de Tiaret pour leur rôle important dans notre formation.

Dédicace

*À mes **parents adorés**, piliers de ma vie,*
merci pour votre amour inépuisable et votre soutien indéfectible.

*À ma précieuse **grand-mère**,*
pour ses prières et sa sagesse.

À mes merveilleuses tantes,
***Tarari Naima** et **Tarari Rachida**,*
pour leur présence et leur affection.

*À ma chère sœur, **Chahinez**, complice de cœur,*
merci pour ta présence rassurante et ton amour constant.

*À mes formidables frères, **Saad** et **Abed**,*
merci pour votre énergie et votre encouragement sincère.

*À mon amie et binôme exceptionnelle, **Bochra**, pour*
sa collaboration et son soutien.

Je vous remercie tous du fond du cœur.

Ouadah

Dédicace

*Pour mes **parents**,*
Merci pour votre amour et vos sacrifices. Votre soutien m'a porté et cette réussite est autant la vôtre que la mienne.

*Pour mes frères **Ilyes** et **Walid**, et ma sœur **Hadjer**,*
Merci pour votre présence, votre affection, et tous ces moments partagés qui rendent la vie plus belle. Je vous dédie cette réussite avec amour.

*Pour mes **amies**,*
Merci pour votre amitié, votre soutien, et tous les souvenirs précieux. Vous êtes un vrai trésor.

*Pour mon amie et ma collègue **Bochra**,*
Merci Bochra pour ton aide, ton sérieux, et ta bonne humeur tout au long de notre travail ensemble. Avec toute mon affection.

Boughaddou

Table des matières

Chapitre 02 : Généralité sur L'intelligence Artificielle

Chapitre 03 :Contribution et Implémentation

Liste des Abréviations

IP Protocol	Internet Protocol
TCP Protocol	Transmission Control Protocol
UDP Protocol	User Datagram Protocol
HTTP Protocol	HyperText Transfer Protocol
HTTPS Protocol	HyperText Transfer Protocol Secure
FTP Protocol	File Transfer Protocol
QoS	Quality of Service
VoIP	Voice over Internet Protocol
FIFO	First In, First Out
WFQ	Weighted Fair Queuing
DPI	Deep Packet Inspection
FAI	Fournisseur d'Accès Internet
IDS	Intrusion Detection System
IA	Intelligence Artificielle
ANN	Artificial Neural Network
SVM	Support Vector Machine
ARL	Active Reinforcement Learning
MLP	Multi-Layer Perceptron
TD Learning	Temporal Difference Learning
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GAN	Generative Adversarial Network
P2P	Peer-to-Peer
IANA	Internet Assigned Numbers Authority
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
DPI	Deep Packet Inspection

Liste des figures

Liste des tableaux

Résumé

L'augmentation significative du trafic réseau résultant de la multiplication des services numériques rend sa gestion de plus en plus complexe. Identifier et classer précisément les flux Internet selon leurs caractéristiques devient essentiel pour assurer une bonne Qualité de Service (QoS), renforcer la sécurité, ainsi que faciliter la gestion et la planification des réseaux.

Les approches classiques, telles que la classification par numéros de ports ou l'analyse approfondie des contenus des paquets, présentent désormais d'importantes limites face aux applications modernes et aux flux cryptés. En réponse à ces défis, les techniques d'apprentissage automatique apparaissent comme des solutions prometteuses, exploitant efficacement les caractéristiques statistiques et temporelles des flux réseau.

Dans ce contexte, notre étude propose une méthodologie hybride hiérarchique novatrice nommée Deep HyCLASS-Net, combinant des approches de machine learning et des réseaux de neurones profonds. Cette approche intègre un filtrage initial du trafic à l'aide du classificateur CatBoost, réputé pour sa robustesse face aux données complexes, suivi d'une classification fine par des architectures optimisées CNN et CNN-LSTM. Cette stratégie hybride garantit une gestion efficace et fiable de la QoS dans les infrastructures réseau contemporaines.

Mots-Clés : Internet, Classification Du Trafic Réseau, Apprentissage Automatique, Apprentissage Profond, Qualité De Service (Qos), Darknet.

Abstract

The significant increase in network traffic resulting from the proliferation of digital services makes its management increasingly complex. Accurately identifying and classifying Internet traffic according to their characteristics is becoming essential to ensure good Quality of Service (QoS), enhance security, and facilitate network management and planning.

Traditional approaches, such as classification by port number or in-depth analysis of packet content, now have significant limitations when faced with modern applications and encrypted traffic. In response to these challenges, machine learning techniques are emerging as promising solutions, effectively exploiting the statistical and temporal characteristics of network traffic.

In this context, our study proposes an innovative hybrid hierarchical methodology called Deep HyCLASS-Net, combining machine learning approaches and deep neural networks. This approach integrates initial traffic filtering using the CatBoost classifier, renowned for its robustness to complex data, followed by fine-grained classification using optimized CNN and CNN-LSTM architectures. This hybrid strategy ensures efficient and reliable QoS management in contemporary network infrastructures.

Key Words : Internet, Network Traffic Classification, Machine Learning, Deep Learning, Quality Of Service (Qos), Darknet.

الملخص

إن الزيادة الكبيرة في حركة مرور الشبكة الناتجة عن انتشار الخدمات الرقمية تجعل إدارتها أكثر تعقيدًا. وأصبح تحديد حركة مرور الإنترنت وتصنيفها بدقة وفقًا لخصائصها أمرًا ضروريًا لضمان جودة الخدمة (QoS الجيدة، وتعزيز الأمان، وتسهيل إدارة الشبكة وتخطيطها.

تواجه الأساليب التقليدية، مثل التصنيف حسب رقم المنفذ أو التحليل المتعمق لمحتوى الحزمة، قيودًا كبيرة الآن عند مواجهة التطبيقات الحديثة وحركة المرور المشفرة. واستجابةً لهذه التحديات، تبرز تقنيات التعلم الآلي كحلول واعدة، تستغل بفعالية الخصائص الإحصائية والزمنية لحركة مرور الشبكة.

في هذا السياق، تقترح دراستنا منهجية هرمية هجينة مبتكرة تُسمى Deep HyCLASS-Net، تجمع بين أساليب التعلم الآلي والشبكات العصبية العميقة. يدمج هذا النهج تصفية حركة المرور الأولية باستخدام مصنف CatBoost، المعروف بمتانته للبيانات المعقدة، متبوعًا بتصنيف دقيق باستخدام هياكل CNN وCNN-LSTM المُحسّنة. تضمن هذه الاستراتيجية الهجينة إدارة جودة الخدمة بكفاءة وموثوقية في البنى التحتية للشبكات المعاصرة.

كلمات مفتاحية: الإنترنت ، تصنيف حركة مرور الشبكة ، التعلم الآلي ، التعلم العميق ، جودة الخدمة (QoS) ، الشبكة المظلمة.

Introduction Générale

Les réseaux de communication occupent une place centrale dans la société numérique contemporaine, assurant l'échange permanent de données entre des milliards d'appareils connectés et soutenant des services essentiels tels que la visioconférence, le cloud computing, les applications mobiles ainsi que l'accès à Internet. Cependant, la croissance constante de ces services numériques entraîne une augmentation exponentielle du trafic réseau, à la fois en termes de volume et de diversité, rendant leur gestion opérationnelle de plus en plus complexe.

Dans ce contexte, les approches classiques de classification du trafic réseau, fondées principalement sur l'analyse des numéros de ports ou sur l'inspection approfondie du contenu des paquets, montrent aujourd'hui leurs limites. Ces méthodes traditionnelles s'avèrent inefficaces face aux flux chiffrés, sont confrontées à des enjeux importants de confidentialité des données, et nécessitent des ressources computationnelles élevées, impactant négativement la réactivité et l'efficacité des systèmes de gestion de réseaux. Cette situation met en évidence la nécessité d'intégrer des solutions innovantes capables de répondre dynamiquement à la complexité croissante du trafic réseau moderne.

Face à ces défis, les techniques d'apprentissage automatique, particulièrement celles basées sur l'apprentissage profond (Deep Learning), se révèlent comme des solutions prometteuses. Ces approches avancées permettent d'exploiter efficacement les caractéristiques statistiques et temporelles des flux réseau, apportant ainsi une réponse pertinente aux limites rencontrées par les méthodes conventionnelles. En particulier, elles renforcent les capacités des systèmes d'inspection approfondie des paquets (Deep Packet Inspection, DPI), leur permettant d'apprendre et de mettre à jour dynamiquement les signatures d'applications. Ainsi, l'intégration de l'apprentissage profond contribue à développer une infrastructure réseau intelligente, résiliente, et capable de s'adapter rapidement à l'évolution des usages et aux nouvelles menaces.

Notre étude propose une méthodologie hybride hiérarchique innovante, intitulée Deep HyCLASS-Net, combinant judicieusement les avantages des algorithmes traditionnels d'apprentissage automatique et des architectures modernes de réseaux neuronaux profonds. Ce modèle utilise notamment l'algorithme CatBoost pour effectuer un filtrage initial robuste du trafic, suivi d'une classification fine réalisée par des réseaux neuronaux convolutifs (CNN) ainsi qu'une architecture hybride CNN-LSTM spécifiquement optimisée pour cette tâche. Cette approche garantit ainsi une gestion efficace et fiable de la Qualité de Service (QoS) au sein des réseaux actuels.

Ce mémoire est structuré en trois chapitres principaux :

- **Chapitre 1** : Présentation des techniques et des enjeux liés à la classification du trafic réseau, accompagnée d'une revue des technologies existantes.

- **Chapitre 2** : Exploration complète des concepts fondamentaux d'apprentissage automatique et d'apprentissage profond, définissant précisément ces notions et leurs applications spécifiques dans le contexte de la classification du trafic réseau.

- **Chapitre 3** : Réalisation pratique de l’approche Deep HyCLASS-Net proposée, détaillant l’environnement de développement, la conception architecturale ainsi qu’une analyse des résultats obtenus, accompagnée d’une discussion comparative vis-à-vis de l’état de l’art.

L'objectif ultime de ce travail est de développer une approche robuste et adaptative capable de classifier efficacement le trafic Internet, améliorant significativement la gestion de la QoS tout en répondant efficacement aux exigences dynamiques et complexes des réseaux modernes.

Chapitre 01 : Notion de base sur le trafic réseau

1 Introduction

La montée en puissance des réseaux internet au fil des années a révolutionné le domaine technologique, devenu un pilier fondamental de la communication, des échanges économiques et des activités quotidiennes. Ces réseaux, composés de millions de dispositifs interconnectés, transportent des volumes de données importants. La diversité des applications qui s'appuient sur Internet, telles que le streaming vidéo, les jeux en ligne, les services en cloud, ou encore les communications en temps réel, a généré des besoins variés et exigeants en matière de performances réseau.

Pour garantir une expérience utilisatrice satisfaisante, il est crucial d'assurer une gestion efficace de la Qualité de Service (QoS). La QoS englobe des paramètres comme la latence, la bande passante, la perte de paquets et la gigue, qui influencent directement les performances des applications. Toutefois, le succès de cette gestion repose sur la capacité à comprendre et classifier le trafic réseau.

2 Réseaux

2.1 Définition d'un réseau

Un réseau se compose de deux ordinateurs ou plus qui sont reliés pour partager des ressources (comme des imprimantes et des CD), échanger des fichiers ou permettre les communications électroniques. Les ordinateurs d'un réseau peuvent être reliés par des câbles, des lignes téléphoniques, des ondes radio, des satellites ou des faisceaux de lumière infrarouge. [1]

Figure I 1: Un exemple de réseau informatique

2.2 Intérêt des réseaux

Un réseau peut répondre à plusieurs objectifs distincts :

- Le partage de ressources (fichiers, applications ou matériels, connexion à internet, etc.)

- La communication entre personnes (courrier électronique, discussion en direct, etc.)
- La communication entre processus (entre des ordinateurs industriels par exemple)
- La garantie de l'unicité et de l'universalité de l'accès à l'information (bases de données en réseau).
- Le jeu vidéo multijoueur

Les réseaux permettent aussi de standardiser les applications, on parle généralement de logiciel collaboratif pour qualifier les outils permettant à plusieurs personnes de travailler en réseau. Par exemple la messagerie électronique et les agendas de groupe permettent de communiquer plus efficacement et plus rapidement.

Voici un aperçu des avantages qu'offrent de tels systèmes :

- Diminution des coûts grâce aux partages des données et des périphériques.
- Standardisation des applications.
- Accès aux données en temps utile.
- Communication et organisation plus efficace.[2]

2.3 L'architecture générale du réseau

On simplifie souvent la description d'un réseau en disant qu'il comporte un « serveur» (ou poste « maître ») et des « clients » (ou postes « esclaves»). Cette distinction fonctionnelle recouvre des configurations très différentes. On identifie deux grandes catégories de réseaux locaux : les réseaux d'égal à égal ou « poste à poste » et les réseaux client/serveur. Chacune présente des avantages et des inconvénients [3]

2.3.1 Le réseau pair-à-pair (P2P)

Dans une architecture « pair à pair », aucun des postes n'est spécialisé. Tous peuvent être à la fois serveurs et clients. C'est l'utilisateur du poste qui décide de partager tout ou partie de ses ressources et qui paramètre lui-même ce partage, la connexion entre les postes ne passant pas par un serveur central. Une limite existe, toutefois, pour la majorité des applications : des données en cours de modification par un utilisateur ne seront pas disponibles pour les autres tant que le premier sera en train de travailler dessus, ou seront alors disponibles en consultation (mode «lecture seule »)[3]

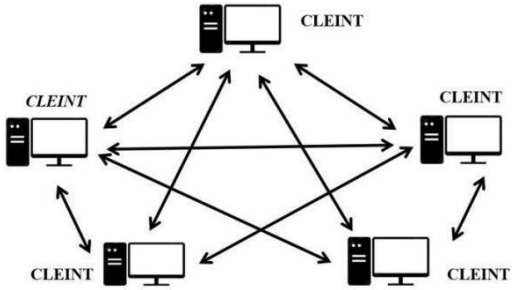

Figure I 2: Architecture de réseau paire a paire

2.3.2 L'architecture client/serveur

Dans le cas d'un réseau « client/serveur», les fonctions de services sont assurées par deux ou plusieurs postes précis, n'ayant pas d'autre fonction (serveurs dédiés). À l'inverse du réseau pair-à-pair, il est possible pour plusieurs utilisateurs de travailler simultanément sur les mêmes données, avec des limitations d'accès moindres. Pour que ce mode de travail en réseau soit possible, il est toutefois nécessaire que le logiciel d'application soit conçu et écrit pour une architecture client/serveur. Le développement et la mise au point de telles applications sont plus complexes et plus coûteux que ceux d'un logiciel mono-utilisateur. Ces logiciels clients/serveurs sont très souvent déployés en interne par le service informatique d'une organisation et utilisés exclusivement pour répondre à ses propres besoins [3]

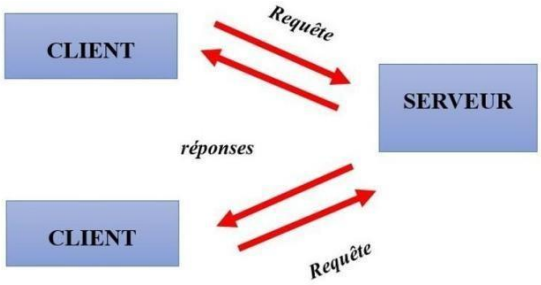

Figure I 3: Architectures réseau Client/Serveur

3 Données dans les réseaux d'internet

3.1 Définitions des données

Le mot "données" n'est pas d'usage courant pour désigner un ensemble ou bloc d'un ou plusieurs caractères alphabétiques et numériques sous forme numérique communiquée entre

deux dispositifs .Ces données constituent une série numérique ou, le contenu d'un fichier informatique contenant un document conservé.[4]

3.2 Type de données

3.2.1 Données d'utilisateur

La donnée d'utilisateur, qui sont généralement des données locales dont les individus ont besoin pour effectuer leurs tâches spécifiques. Ces données doivent être conservées dans le système de fichiers /home ou dans des systèmes de fichiers créés spécifiquement pour les données d' utilisateur. [5]

3.2.2 Données multimédias

Une donnée qui combine différents formats d'information tels que l'image fixe (y compris l'écriture manuscrite), l'image animée, le son, la vidéo et le texte alphanumérique est appelée le multimédia. [6]

3.2.3 Données textuelles

Les données textuelles sont des informations variées, allant de textes formels à des messages courts et informels, destinées à la lecture et à la compréhension humaine, parfois au sein de groupes spécifiques. Elles représentent une source riche pour l'exploration de données. [7]

3.2.4 Données applicatives

Il y a des applications données qui mélangeaient de la visualisation donnée des applications web, ce qui permet aux usagers finaux (les décideurs, les experts et même les consommateurs) de voir et de gérer facilement une masse de données. [8]

3.2.5 Données de contrôle

Le contrôle des données est la supervision managériale des politiques d'information d'une organisation. Contrairement à la qualité des données axée sur la résolution, il observe, rapporte le fonctionnement des processus et gère les problèmes via inspection, validation, notification, documentation, signalement et suivi. Il assure la conformité et l'efficacité des processus de gestion de l'information. [9]

3.2.6 Données en temps réel

Les données en temps réel sont des données regroupées, communiquées à l'heure à laquelle ils sont collectées accélérant l'obtention des informations et la prise de décision. À la différence des données traditionnelles, accompagnées d'un intervalle de temps, avant de pouvoir les récupérer, les données en temps réel offrent un instantané des événements ou des activités. [10]

3.2.6.1 VoIP

VoIP est l'abréviation de Voice Over Internet Protocol, ou en d'autres mots, l'envoi de la voix via Internet. C'est un outil qui permet de transmettre des messages (vocal ou audio-vidéo) sur le réseau Internet (IP). [11]

3.2.6.2 Jeux en ligne

Un jeu vidéo en ligne disponible sur Internet permet soit de jouer seul contre l'ordinateur, soit d'affronter des adversaires du monde entier. Un jeu en ligne (ou online) permet de jouer sur un réseau, en étant connecté à d'autres ordinateurs via Internet. Les joueurs peuvent ainsi jouer avec ou contre d'autres internautes. [12]

3.2.6.3 Streaming multimédia

Le streaming est un protocole permettant la lecture instantanée de contenus audio et vidéo directement dans le navigateur web, sans téléchargement préalable. Il envoie un flux continu de données en temps réel, idéal pour les vidéos en direct et la diffusion continue. Cette méthode est privilégiée pour sa diffusion en direct et pour limiter le téléchargement et la distribution illégale de fichiers compressés pour minimiser la bande passante. [13]

3.3 Transmission de données

3.3.1 Définition

La transmission de données est le transfert de données d'un appareil numérique à un autre. Ce transfert s'effectue via des flux ou canaux de données point à point. Auparavant, ces canaux étaient constitués de fils de cuivre, mais ils sont désormais beaucoup plus susceptibles de faire partie d'un réseau sans fil.[14]

3.3.2 Protocole de transmissions de données

Les protocoles de transmission de données sont ceux qui permettent à deux entités de communiquer à travers un réseau de télécommunications. Un protocole est un ensemble de règles

à respecter pour que ces deux entités puissent s'échanger de l'information. Ces règles peuvent être simples comme, par exemple, la technique de codage à utiliser pour reconnaître un caractère ou très complexes comme les protocoles acheminant des blocs d'information d'une extrémité à l'autre du réseau. [15]

3.3.2.1 Protocole IP

Le protocole IP (Protocole Internet) achemine les données sur Internet en adressant et en fragmentant les paquets d'informations. La programmation Internet crée des applications web accessibles via des navigateurs ou des applications, s'appuyant sur des protocoles comme IP pour la communication entre appareils et serveurs. Elle permet de développer des solutions interactives et dynamiques pour le web [16]

3.3.2.2 Protocole TCP

TCP (Protocole de contrôle de transmission) est un protocole de transport fiable et orienté connexion, essentiel pour des applications comme le web, l'email et le transfert de fichiers. Il assure la fiabilité par une connexion en trois étapes, la livraison ordonnée des données avec accusé de réception et retransmission, et la gestion du flux. [17]

3.3.2.3 Protocole UDP

L'UDP (Protocole de datagramme utilisateur) est un protocole léger qui permet l'envoi rapide de datagrammes sur IP sans connexion préalable, privilégiant la vitesse à la fiabilité et l'ordre. Il est adapté aux applications tolérantes aux pertes ou gérant elles-mêmes les erreurs, comme le streaming, les jeux en ligne, la VoIP et le DNS, où la faible latence est cruciale. Sa faible surcharge le rend idéal pour les transmissions unidirectionnelles et les réseaux peu chargés. [18]

3.3.2.4 Protocole HTTP

HTTP (Protocole de transfert hypertexte) est un protocole applicatif sans état, fondamental pour le web depuis 1990, permettant l'échange de données hypermédia. Sa flexibilité, grâce à ses codes, en-têtes et méthodes extensibles, le rend utile pour diverses applications au-delà de l'hypertexte, comme les serveurs de noms. Il facilite la communication entre systèmes en permettant la saisie et la négociation des données. [19]

3.3.2.5 Protocole HTTPS

HTTPS (Protocole sécurisé de transfert hypertexte) est une version sécurisée de HTTP qui chiffre les données échangées entre le navigateur et le site web, garantissant confidentialité et intégrité. La présence du "S", du cadenas vert et de la mention "Sécurisé" dans la barre d'adresse

indique l'utilisation d'un certificat SSL/TLS obtenu auprès d'une Autorité de Certification. Ce protocole empêche l'espionnage et la modification des informations transmises. [20]

3.3.2.6 Protocole FTP

File Transfer Protocol (protocole de transfert de fichier), ou FTP, est un protocole de communication destiné au partage de fichiers sur un réseau TCP/IP. Il permet, depuis un ordinateur, de copier des fichiers vers un autre ordinateur du réseau, ou encore de supprimer ou de modifier des fichiers sur cet ordinateur. Ce mécanisme de copie est souvent utilisé pour alimenter un site web hébergé chez un tiers. [21]

3.3.2.7 Protocole Bluetooth

Bluetooth est une norme de communication sans fil à courte portée utilisant les ondes radio UHF, développée par Ericsson en 1994. Elle permet la connexion et l'échange bidirectionnel de données entre plusieurs périphériques sans câbles. Son principal avantage est d'établir des connexions sans fil entre appareils. [22]

3.3.2.8 Protocole WIFI

Wi-Fi (fidélité sans fil) est un ensemble de protocoles de communication sans fil régis par les normes du groupe IEEE 802.11 (ISO/CEI 8802-11). Un réseau Wi-Fi permet de relier sans fil plusieurs appareils informatiques (ordinateur, routeur, décodeur Internet, etc.) au sein d'un réseau informatique. [23]

4 Les flux du réseau

Il s'agit d'un terme clé dans le domaine de l'informatique et des réseaux. Il s'agit d'une quantification du volume d'informations véhiculées ou transmises sur un réseau au cours d'un intervalle de temps. Le trafic peut être induit par l'une ou l'autre d'une multitude de sources différentes, par exemple des appareils, des applications ou des utilisateurs. Son analyse et sa gestion appropriées font partie de l'optimisation du réseau et de la fourniture d'une expérience utilisateur de qualité. Au départ, les flux de réseaux sont modélisés comme des flux de paquets de données qui circulent de l'adresse source à l'adresse de destination. Les paquets de données contiennent des informations relatives au protocole utilisé, au port source, au port de destination, etc. qui est pertinent pour l'analyse du réseau. [24]

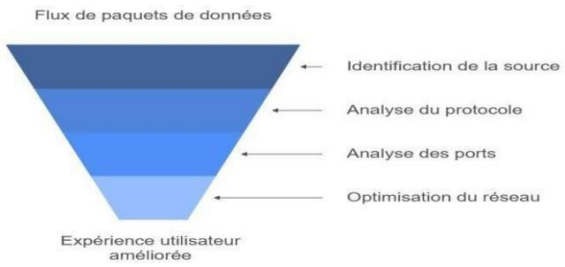

Figure I 4: Processus de gestion du trafic réseau

5 Trafic réseau

5.1 Definition

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 .Le trafic est également lié à la sécurité, car une quantité anormalement élevée de trafic pourrait être le signe d'une attaque. [25]

5.2 Les caractéristiques de trafic

Les caractéristiques du trafic réseau désignent les propriétés essentielles qui influencent la circulation des données, telles que la vitesse, la latence, et la fiabilité des connexions.

5.2.1 La bande passante

La bande passante désigne la capacité maximale de transmission d'un réseau de communication, mesurée en bits par seconde (bps). Elle représente la quantité de données qui peut être transmise d'un point à un autre dans un laps de temps donné. En d'autres termes, c'est la largeur de l'autoroute numérique par laquelle les données circulent. Plus la bande passante est large, plus le volume de données pouvant circuler simultanément est important. [26]

5.2.2 La latence

La latence fait référence au délai qui se produit entre le moment où un utilisateur effectue une action sur un réseau ou une application Web et le moment où il reçoit une réponse. Une

autre définition de la latence est le temps total ou « aller-retour» nécessaire pour qu'un paquet de données se déplace. [27]

5.2.3 La gigue

La gigue est la mesure de la variation de la latence au fil du temps. Plutôt que d'être une mesure du délai total, la gigue exprime l'incohérence de ce délai. C'est essentiellement la mesure de la variabilité des délais d'arrivée des paquets. [28]

5.2.4 La perte de paquets

5.2.4.1 Définition d'un Paquet

Dans le domaine des réseaux, un paquet est un petit segment d'un message plus important. Les données envoyées sur les réseaux informatiques, tels qu'Internet, sont divisées en paquets. Ces paquets sont ensuite recombinés par l'ordinateur ou le dispositif qui les reçoit. [29]

5.2.4.2 Définition d'un paquet de données

Un paquet de données est essentiellement une petite unité de données structurée utilisée pour transmettre des informations sur des réseaux numériques. Lorsque l'envoi d un e-mail ou diffusion d'une vidéo, les informations sont divisées en parties plus petites, appelées paquets. Chaque paquet contient non seulement une section des données principales mais également des métadonnées importantes, telles que l'adresse de destination et le numéro de séquence. Ces métadonnées garantissent que les paquets sont correctement réassemblés à leur destination. En segmentant les données en paquets, les réseaux peuvent gérer le flux de trafic plus efficacement et réduire le risque d'erreurs lors de la transmission. [30]

5.2.4.3 Comment se manifeste la perte de paquets?

Lorsque vous envoyez des informations sur un réseau, le protocole TCP/IP sépare ces informations en petits paquets afin de les fractionner pour une transmission plus rapide. Les paquets sont étiquetés avec des informations d'en-tête afin que les informations puissent être reconstituées lorsqu'elles atteignent le destinataire. Lorsqu'une partie importante de ces paquets est corrompue ou perdue en transit, la perte de paquets entraîne un échec de la communication. La perte de paquets peut se produire sur n'importe quel réseau, mais elle est beaucoup plus fréquente lorsque la communication est envoyée sur de longues distances, comme sur Internet. [31]

5.2.5 Débit

Le débit, ou taux de transfert de données, mesure la quantité de données pouvant être transférées d'un point à un autre en un temps donné, généralement exprimé en mégabits par seconde (Mbps) ou gigabits par seconde (Gbps). Il représente la capacité réelle du réseau à transmettre des données. [32]

5.3 Analyse du trafic

L'analyse du trafic est le processus d'interception et d'examen des messages afin de déduire des informations à partir des modèles de communication. Elle joue un rôle crucial dans la compréhension du flux de données à travers les réseaux, en aidant à identifier les risques de sécurité , à optimiser la performance du réseau et à contribuer aux enquêtes judiciaires. Cette technique relie divers aspects de l'architecture du réseau, des protocoles, des zones de sécurité et des vulnérabilités, fournissant des informations sur la fois la fonctionnalité et la posture de sécurité d'un réseau. [33]

5.4 Surveillance du trafic réseau

La surveillance du trafic réseau est le processus qui consiste à suivre le trafic et la bande passante du réseau en analysant les flux sur la base de différentes technologies, notamment NetFlow (Flux réseau de Cisco), sFlow (Flux échantillonné) et J-Flow (Flux réseau de Juni- per). Grâce aux statistiques sur les pics de trafic, les principales applications et les principales conversations, vous pouvez obtenir une visibilité complète de la tendance du trafic réseau et de la bande passante. [34]

5.5 Les Types de trafic réseau

Il existe différents types de trafic réseau dans un réseau. Ces types de trafic ont des caractéristiques et des exigences distinctes. Ainsi, dans un réseau, le comportement varie en fonction des types de trafic. Cela est rendu possible grâce aux configurations de la Qualité de Service (QoS). Alors, quels sont ces types de trafic réseau ?
Il y a trois types de trafic : le trafic vocal, le trafic vidéo et le trafic de données.

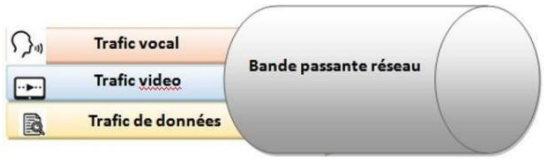

Figure I 5: Les types de trafic reseau

5.5.1 Trafic vocal

Le trafic vocal est sensible et nécessite une faible latence. Les paquets vocaux doivent arriver simultanément pour être compréhensibles. Le retard et la gigue doivent être minimisés, car des délais variables affectent la qualité. La perte de paquets peut être tolérée, mais pas de manière excessive. Pour éviter ces problèmes, il est nécessaire d'utiliser la Qualité de Service (QoS) sur les routeurs, priorisant le trafic vocal. Le protocole UDP est généralement utilisé pour sa rapidité dans ce type de trafic. [35]

5.5.2 Trafic de données

Le trafic de données constitue une autre catégorie importante de trafic réseau, moins vulnérable à la perte de paquets. Ce type utilise un mécanisme de retransmission en cas de perte de paquet. Il est utilisé pour les e-mails, les transferts de fichiers, les pages web, etc. Ainsi, la garantie est un terme clé pour ce type de trafic. Pour assurer un transfert sécurisé et garanti, le protocole TCP est utilisé avec le trafic de données. Le mécanisme de retransmission de TCP garantit cette fiabilité et le trafic de données est envoyé avec une perte minimale.
Encore une fois, le délai n'est pas aussi critique pour ce type de trafic. Par exemple, si un retard se produit, vous recevrez un e-mail avec un léger délai, mesurable en secondes. [35]

5.5.3 Trafic vidéo

Est l'un des types de trafic les plus utilisés dans le monde d'aujourd'hui. Avec l'augmentation des temps de visionnage sur YouTube, la formation en ligne et les trafics similaires, la vidéo est devenue très importante comme type de trafic. Le trafic vidéo est un trafic de volume élevé qui n'est pas aussi sensible que le trafic vocal, car le trafic vidéo n'est généralement pas en temps réel. Il peut tolérer la perte de paquets et les retards. Et le retard sur les paquets ne cause aucun malentendu. Cela ne provoque qu'un peu plus de temps. A côté de toute perte peut être tolérée car il s'agit d'un trafic très élevé et si nous perdons une petite quantité de ce

trafic, la vidéo peut encore être compréhensible et claire. [35]

6 La qualité de service (Qos)

6.1 Definition

QoS est le sigle de « Quality of Service » que l'on traduit par « qualité de service ». D'un point de vue technique la QoS représente la capacité d'un réseau à offrir un service de transmission de données avec un niveau de performance élevé. Cela inclut la gestion de la bande passante, le contrôle du délai, la réduction de la gigue (variation de la latence) et la minimisation de la perte de paquets.
Par la configuration QoS et l'utilisation de protocoles spécifiques, les routeurs et les commutateurs peuvent prioriser certains types de trafic comme la voix sur IP ou les applications critiques pour l'entreprise. Cela permet de s'assurer que ces services fonctionnent de manière optimale même en période de congestion du réseau. [36]

6.2 Importance de la qualité de service (Qos)

La qualité de service (QoS) réseau est essentielle pour garantir la performance optimale des applications sensibles comme la VoIP, la visioconférence et le streaming, en évitant latence et décalage. Elle permet de hiérarchiser le trafic, les ressources, les applications, les flux de données et les utilisateurs pour assurer le niveau de performance souhaité. Cruciale pour les applications interactives et en temps réel nécessitant une bande passante élevée et une faible latence, la QoS aide à prioriser les applications "inélastiques" sensibles à la gigue. [37]

6.3 Paramètres de qualité de service

Les exigences de QoS pour les applications sont fréquemment exprimées en termes de :

- **Bande passante**
- **Latence**
- **Gigue**
- **Perte de paquets**
- **Fiabilité**

6.3.1 Bande passante

La bande passante est le débit de données requis pour la performance d'une application réseau, mesuré en paquets ou bits par seconde (minimal, maximal, moyen, rafales). Bien que rarement problématique en LAN, elle est cruciale en WAN et MAN aux besoins plus élevés. Les gestionnaires doivent l'adapter aux applications et à la topologie via des choix d'équipements, services et technologies.[38]

6.3.2 Latence et Délai

La latence désigne le délai que rencontrent les paquets en traversant le réseau. Bien que les termes « latence » et « délai » soient souvent utilisés de manière interchangeable, la latence est généralement associée aux équipements et services réseau (commutateurs, routeurs), tandis que le terme « délai » est utilisé pour les communications de données (transmission, propagation).[38]

6.3.2.1 Latence Réseau

La latence réseau (NL) représente la somme des délais de bout en bout, produits par les équipements réseau (hubs, commutateurs, routeurs) et les services réseau, incluant les délais de propagation dans les supports de communication (fibre optique, sans fil, etc.) et le délai de traitement dans la pile réseau des hôtes.[38]

6.3.2.2 Délais de traitement de l'application

Le délai de traitement de l'application (AD) correspond au délai introduit par le programme de l'application, incluant les composants de protocole et l'interface de programmation (API) utilisée (par exemple, API TCP/UDP).[38]

6.3.2.3 Délai de bout en bout

Le délai de bout en bout (EE) est le temps total qu'un paquet met pour aller de la source à la destination, en excluant les délais dans les couches supérieures de l'hôte. Ce paramètre est utilisé pour spécifier les exigences de QoS pour les applications de type utilisateur-à-utilisateur, telles que la VoIP.[38]

6.3.2.4 Temps de réponse

Le temps de réponse (RT) est une mesure des exigences de QoS pour les applications client/serveur. Il dépend de la latence aller-retour (NL) et des délais de traitement aux deux extrémités

de la communication :

$$RT = \Sigma(NL_{aller-retour}, AD_{source}, AD_{destination})$$

La latence réseau est le principal composant du temps de réponse. Les composants de la latence incluent :

- Délais de propagation
- Vitesse de transmission
- Latence des équipements
- Latence des services de communication [38]

6.3.2.5 Composants de la latence réseau

- **Délai de propagation**

  Les délais de propagation représentent le temps nécessaire à un signal pour se propager dans le réseau, ce qui peut être critique pour les réseaux longue distance et satellites. Dans les LANs, ces délais sont généralement négligeables.[38]

- **Vitesse de transmission**

  Le délai de transmission est le temps nécessaire pour transmettre un paquet à travers un lien réseau. La vitesse de transmission est un paramètre gérable lors de la conception du réseau, mais les liens longs distances peuvent introduire des délais significatifs.[38]

- **Latence des équipements**

  La latence d'équipements est le délai introduit par les équipements réseau comme les routeurs et les commutateurs. Elle dépend du type d'équipements et de sa conception interne. Les commutateurs LAN ont un impact minimal, tandis que les routeurs peuvent introduire des délais plus significatifs.[38]

- **Latence des services de communication**

  La latence des services de communication est le délai ajouté par les services publics, comme les liens Frame Relay (relais de trames), qui peuvent introduire des délais de plusieurs centaines de millisecondes. Les services de communication publics doivent être considérés comme un composant de délai dans la conception de réseaux avec garanties QoS.[38]

6.3.3 Gigue

La gigue est la variation du délai d'arrivée des paquets, critique pour les applications temps réel comme la VoIP. Causée par des temps de traitement variables, la latence réseau, la congestion et d'autres irrégularités, une forte gigue peut entraîner des retards ou une réception désordonnée. Alors que le protocole TCP corrige cela, l'UDP nécessite des mécanismes supplémentaires comme RTP. Un tampon de gigue est utilisé pour stocker temporairement les paquets et les délivrer de manière régulière.[38]

6.3.4 Perte de paquets

La perte de paquets survient principalement à cause de :

- la congestion et les limitations de ressources dans les routeurs (causes principales);
- les erreurs de transmission (moins fréquentes grâce aux technologies comme la fibre optique) ;
- un mauvais acheminement dû à des erreurs de traitement du routage (rare) ;

Certaines applications, comme la VoIP, peuvent tolérer une perte allant jusqu'à 5

6.3.5 Disponibilité et Fiabilité

La disponibilité et la fiabilité sont deux paramètres clés de la QoS, souvent pris en compte dès la conception du réseau. Ils garantissent un fonctionnement continu et sans échec des équipements et des services.

Ces paramètres dépendent notamment :

- de la fiabilité des équipements (réseaux privés) ;
- de la disponibilité des services publics, souvent définie dans un accord de niveau de service (SLA) ;

Dans les secteurs critiques (banque, e-commerce, industrie...), une disponibilité supérieure à 99,9 % est généralement exigée pour assurer la continuité des services comme le Web, la VoIP ou les systèmes client/serveur.[38]

6.4 Les mécanismes de QoS

6.4.1 Mise en file d'attente du trafic

La QoS peut être mise en œuvre en utilisant des schémas de priorité basés sur la mise en file d'attente. Nous examinons ensuite différents schémas de file d'attente et comment ils peuvent être utilisés pour la QoS.[39]

6.4.1.1 Premier entré, premier sorti (First-In-First-Out)

Dans ce schéma, les paquets arrivant dans le tampon sont délivrés sur la base du premier arrivé, premier servi. Il est utile lorsque les paquets doivent être stockés avant d'être transmis. C'est un schéma simple nécessitant très peu de traitements. Cependant, comme aucune priorité n'est définie, différents types de trafic sont traités de la même manière. Ce schéma n'est pas adapté lorsqu'une véritable QoS est nécessaire.[39]

6.4.1.2 File d'attente à priorité (Priority Queuing)

La file d'attente à priorité est un mécanisme de QoS en deux étapes consistant à identifier les classes de trafic puis à leur attribuer une priorité (élevée, moyenne, normale ou basse), selon divers critères comme le protocole ou la taille des paquets. Elle permet de traiter en priorité le trafic critique, comme les commandes clients, et est généralement configurée statistiquement. Bien qu'elle améliore la gestion du trafic important, elle présente des limites, telles qu'une certaine lenteur, une forte consommation de ressources CPU et un risque de famine du tampon lorsque le trafic prioritaire est trop important, au détriment du trafic normal. [39]

6.4.1.3 Mise en file d'attente par classe (Class-Based Queuing)

La file d'attente personnalisée est une variante plus équitable de la mise en file à priorité. Elle impose des limites de traitement pour chaque niveau de priorité afin d'éviter que le trafic prioritaire ne monopolise pas les ressources. Chaque classe de trafic reçoit un espace dédié dans la file et est desservie de manière cyclique (round robin). Cette méthode permet une différenciation du trafic, mais elle peut devenir difficile à gérer lorsque la répartition entre les classes change fréquemment, et elle pose un problème d'évolutivité en raison de la charge de traitement élevée qu'elle implique. [39]

6.4.1.4 Mise en file d'attente équitable pondérée (Weighted Fair Queuing)

La mise en file d'attente équitable pondérée priorise le trafic à faible volume et garantit que le trafic à haut volume ne consomme pas toutes les ressources. Elle programme le trafic interactif

en priorité pour réduire les délais de réponse et ajuste automatiquement la bande passante selon les besoins du réseau. Cependant, elle peut manquer de paramètres configurables pour certaines applications, et son pondérale dépend des spécifications du fournisseur.[39]

6.4.2 Algorithme du seau percé (Leaky Bucket)

L'algorithme du seau percé permet d'assurer un flux de sortie constant, même lorsque l'entrée varie fortement. Bien que la taille du seau et le taux de transmission soient configurables, il peut ne pas être efficace pour gérer les trafics en rafale ou pour utiliser les ressources de manière optimale en cas de faible volume de trafic .[39]

6.4.3 Seau à jeton (Token Bucket)

Le seau à jetons contrôle le trafic en stockant des jetons, qui sont nécessaires pour le traitement des données. Ce système permet d'obtenir une transmission de données régulière tout en gérant efficacement les pics de trafic. Si les jetons ne sont pas suffisants, les données peuvent être traitées en mode best-effort ou être abandonnées.[39]

6.5 Qualité de service : Principaux avantages

6.5.1 Amélioration des performances des applications

la qualité de service améliore les performances des applications critiques en leur accordant une teneur supérieure prioritaires, métiers de trafic.[40]

6.5.2 Optimisation de l'acheminement du trafic

Lorsqu'un trafic provient d'un site et que les applications ciblées sont diverses, une seule approche d'acheminement peut s'avérer inefficace et entraîner des retards. En identifiant l'application associée à une adresse IP et en appliquant des règles spécifiques à chaque flux, une organisation peut diriger le trafic réseau de manière plus efficace vers sa destination.[40]

6.5.3 Diminution de l'encombrement du réseau

Le réseau de l'organisation peut facilement se retrouver encombré par du trafic à faible priorité. La qualité de service (QoS) permet de soulager le réseau en bloquant ou en limitant ce type de trafic pendant les périodes de forte utilisation. Cela permet d'économiser la bande passante dans les zones critiques du réseau et de privilégier un routage spécifique au type d'application afin de réduire efficacement la congestion. [40]

7 La classification du trafic réseau

La classification du trafic permet de grouper les trafics réseaux en différentes catégories, et intervient en premier lieu dans de nombreuses applications comme le contrôle qualité de service (QoS), la facturation, la gestion des ressources, la détection de logiciels malveillants et de parasites. Devant l'importance de l'empilement et à raison de cette évolution constante des besoins de nombreuses utilités, différentes techniques ont vu le jour au fil du temps pour répondre à toutes les situations nécessitant toujours une intensité particulière. Mais ces progrès récents dans la communication, en particulier le secret et l'offuscation des ports rendent encore plus difficile, la classification. Par conséquent, les techniques de classification du trafic ont connu une évolution significative au fil du temps.

La classification du trafic réseau repose généralement sur trois approches principales :

- Basée sur le port ;
- Basée sur l'inspection approfondie (DPI);
- Basée sur les caractéristiques statistiques.

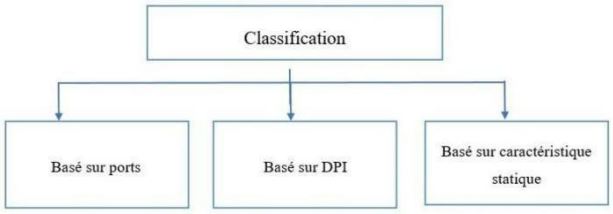

Figure I 6: La classification du trafic réseau

7.1 Les avantages de la classification du trafic réseau

- Identifier les types de trafic présents;
- Organiser les flux pour appliquer des politiques QoS différenciées;
- Allouer intelligemment les ressources pour garantir les performances;
- Améliorer l'efficacité des applications critiques. [41]

7.2 Les types de classification

7.2.1 Classification basée sur le port

Les premières méthodes de classification du trafic réseau reposaient sur l'utilisation des numéros de port des paquets pour identifier les protocoles associés. L'IANA (Internet Assigned Numbers Authority) a défini des numéros de port standards pour différencier les services et protocoles. Ces ports sont répartis en trois catégories : les ports système (0–1023), les ports utilisateur (1024–49 151) et les ports dynamiques (49 152–65 535). Les ports système (0–1023) sont spécifiquement réservés aux protocoles standards. Cette méthode de classification examine généralement les numéros de port des paquets TCP et UDP en les associant aux ports préassignés par l'IANA. Par exemple, un port 80 correspond au protocole HTTP. Les avantages de la solution de classification basée sur les ports se reflètent dans sa simplicité de mise en œuvre, les faibles besoins en ressources informatiques et le processus de classification à haute vitesse. D'autre part, elle présente deux inconvénients majeurs. Premièrement, de nombreuses applications déploient aujourd'hui le masquage, c'est-à-dire l'utilisation de ports standard pour acheminer le trafic d'autres protocoles, tel que le trafic de logiciels malveillants sur HTTP. Deuxièmement, de nombreuses applications sont déployées de manière aléatoire, ce qui fait référence à l'utilisation de ports non standard/dynamiques pour acheminer leur trafic réseau, telles que les applications VoIP.[42]

7.2.2 Classification basée sur DPI (Deep Packet Inspection)

L'inspection approfondie des paquets (DeepPacket Inspection), 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 normaux d'inspection dynamique des paquets, le dispositif ne vérifie que les informations dans l'en-tête du paquet, telles que l'adresse IP 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 classiques d'analyse des paquets, le DPI peut détecter des menaces autrement dissimulées dans le flux de données, telles que des tentatives d'exfiltration de données, des violations des politiques de contenu, des logiciels malveillants, etc.[43]

7.2.2.1 Fonctionnement de Fonctionnement de l'inspection approfondie des paquets DPI

DPI examine le contenu des paquets de données à l'aide de règles spécifiques prépro-

grammées par l'utilisateur, un administrateur ou un fournisseur de services Internet (FAI). Ensuite, il décide comment gérer les menaces qu'il découvre. Les DPI peuvent non seulement identifier l'existence de menaces, mais en utilisant le contenu du paquet et son en-tête, ils peuvent également déterminer d'où il provient. De cette manière, DPI peut identifier l'application ou le service qui a lancé la menace. Les DPI peuvent également être configurés pour fonctionner avec des filtres qui lui permettent d'identifier et de réacheminer le trafic réseau provenant d'un service en ligne ou d'une adresse IP spécifique. [43]

7.2.2.2 Technique d'inspection approfondie des paquets

Les pare-feux dotés de fonctionnalités IDS et les systèmes IDS destinés à la protection du réseau utilisent le DPI. Les techniques qu'ils utilisent comprennent l'anomalie de protocole, les solutions IPS et la correspondance de modèles ou de signatures.[43]

- **Anomalie du protocole**

L'anomalie du protocole utilise une approche appelée « refus par défaut ». Avec le refus par défaut, le contenu est autorisé à passer conformément aux protocoles prédéfinis. Seul le contenu qui correspond au profil acceptable peut être consulté. Cela est différent de permettre à tout ce qui n'est pas identifié comme malveillant de passer, ce qui peut toujours permettre aux attaques inconnues de pénétrer le réseau.[43]

- **Solutions IPS**

Les solutions IPS peuvent bloquer les menaces en temps réel, et certaines d'entre elles utilisent le DPI. Cependant, l'un des défis réside dans le fait que les solutions IPS peuvent parfois générer des faux positifs. L'utilisation de règles prudentes peut réduire l'impact d'un IPS qui a tendance à indiquer des alertes faussement positives.[43]

- **Correspondance de modèle ou signature**

Avec la mise en correspondance des modèles ou des signatures, le contenu d'un paquet de données est analysé et comparé à une base de données de menaces précédemment identifiées. Si le système est constamment mis à jour avec des renseignements sur les menaces, il peut s'agir d'une défense très efficace contre les attaques. Cependant, si l'attaque est nouvelle, le système peut la manquer.[43]

7.2.2.3 Avantages du DPI

Le DPI offre un contrôle avancé du trafic réseau en analysant les données en transit pour en déterminer la nature, la provenance et la destination. Il permet de renforcer la sécurité en bloquant les données provenant de sources malveillantes, en identifiant les paquets dangereux et en hiérarchisant certains types de trafic, comme les communications VoIP. En dehors de la sécurité, le DPI peut être utilisé

pour la censure Internet ou la surveillance des communications en analysant et filtrant le trafic selon des critères prédéfinis.[43]

7.2.2.4 Limites du DPI

Le DPI s'accompagne d'au moins trois limitations importantes. Premièrement , elle peut créer de nouvelles vulnérabilités tout en protégeant des failles existantes. Si le DPI se montre efficace face à des attaques de type dépassement de mémoire tampon ou de déni de service et s'il sait faire face à certains types de logiciels malveillants, cette technologie peut également être détournée et faciliter les attaques dans ces mêmes catégories. Deuxièmement , le DPI complique la mise en œuvre et la nature déjà difficile des pare-feu et autres logiciels de sécurité existants. Le DPI exige des mises à jour et des révisions régulières pour fonctionner de manière optimale. Troisièmement, DPI peut ralentir un ordinateur en augmentant la charge qui pèse sur le processeur. Malgré ces limitations, de nombreux administrateurs réseau ont misé sur la technologie DPI dans l'espoir de faire face à une hausse prévisible de la complexité et de la diversification des dangers liés à Internet. [43]

7.2.3 Classification basée sur caractéristiques statistiques

La solution de classification statistique repose sur l'extraction de caractéristiques statistiques enchaînées à des algorithmes d'apprentissage automatique pour passer en revue le trafic du réseau. Le processus commence avec l'encapsulation du trafic réseau sous la forme de flux, c'est-à-dire la combinaison d'éléments adresses IP source et destination, numéros de port source et destination, et le protocole (TCP ou UDP). des caractéristiques statistiques sont extraites à deux niveaux :[44]

7.2.3.1 Au niveau des paquets

lors de l'analyse d'un paquet ou agrégé pour extraire les données comme la taille d'un paquet et l'intervalle entre arrivées.

7.2.3.2 Au niveau des flux

en examinant l'ensemble du flux pour extraire des informations globales comme le nombre total de paquets, le volume total de données (en octets) et la durée du flux.Ces caractéristiques servent ensuite à alimenter les algorithmes d'apprentissage pour la classification du trafic.

7.2.3.3 Avantages de la classification basée sur caractéristique statique

Les techniques mises en œuvre sur la base de la classification statistique sont capables de percevoir les

comportements des flux attendus à travers les observations. Les méthodes statistiques, combinées à des méthodes basées sur des règles, peuvent offrir une évolutivité, adaptabilité, flexibilité et robustesse. De plus, pour différencier le trafic présentant des anomalies du trafic normal, des mesures statistiques peuvent être utilisées.

7.2.3.4 Limites de la classification basée sur caractéristique statique

- **Surajustement**
- Susceptibilité au surajustement et mauvaise généralisation sur de nouvelles données.

- **Qualité des données**
- Dépend de la qualité et de la représentativité des données d'entraînement.

- **Biais algorithmique**
- Risques associés à la prise de décision biaisée basée sur la distribution des données d'entraînement. leurs performances dépendent fortement des caractéristiques conçues par des experts, ce qui limite leur généralisabilité. [44]

8 Conclusion

En conclusion, la gestion efficace du trafic réseau repose sur des stratégies de classification robustes et adaptatives. L'évolution constante des menaces, ainsi que les exigences croissantes en matière de qualité de service (QoS), imposent le recours à des approches hybrides combinant inspection approfondie, analyse statistique et intelligence artificielle. Une politique de QoS bien définie, alliée à une classification précise du trafic, garantit des performances optimales, une meilleure réactivité et une sécurité renforcée du réseau.

Chapitre 02 : Généralité sur L'intelligence Artificielle

1 Introduction

L'intelligence artificielle (IA) et l'apprentissage automatique jouent un rôle clé dans la gestion des réseaux informatiques en optimisant leur gestion, en analysant en continu de grandes quantités de données. Ils permettent de détecter les anomalies, d'anticiper les problèmes et de réagir en temps réel sans intervention humaine. En automatisant les tâches complexes, l'IA améliore la sécurité, réduit les temps d'arrêt et diminue les coûts d'exploitation. Elle facilite aussi la maintenance préventive, aidant notamment les PME à mieux gérer leur réseau. À terme, ces avancées ouvrent la voie à des réseaux auto-réparateurs et à une gestion entièrement autonome.

Dans ce chapitre, nous explorerons les notions essentielles dans ce domaine, afin de poser une base solide pour comprendre les approches modernes. Cela nous permettra d'aborder ensuite les travaux existants et les résultats obtenus.

2 Intelligence Artificielle

L'intelligence artificielle est l'intelligence d'une machine qui peut comprendre ou apprendre n'importe quelle tâche intellectuelle qu'un être humain peut accomplir. C'est un objectif principal de certaines recherches sur l'intelligence artificielle et un sujet commun dans la science- fiction et les études prospectives. AGI peut également être appelée IA forte, IA complète ou action intelligente générale.

Certaines sources académiques réservent le terme "IA forte" aux machines qui peuvent ressentir la conscience. D'autres distinguent IA forte de l'IA appliquée (appelée aussi IA étroite ou IA faible), qui se limite à des tâches spécifiques. Contrairement à l'IA forte, l'IA faible ne tente pas de reproduire toute la gamme des capacités cognitives humaines.[45]

3 Apprentissage automatique

L'apprentissage automatique (machine learning) emprunte son terme "apprentissage" à la psychologie cognitive pour décrire la capacité des systèmes d'IA à s'adapter et à apprendre de l'expérience. Inspiré des théories psychologiques, l'apprentissage est vu comme la mémorisation et l'adaptation à l'information. L'objectif est de construire des systèmes capables de s'ajuster à de nouvelles données. Les techniques principales incluent : l'apprentissage supervisé pour la prise de décision discrète et la prédiction continue, l'apprentissage par renforcement pour la prise de décision séquentielle, et l'apprentissage non supervisé. L'apprentissage automatique consiste à ajuster un modèle à un ensemble de données en minimisant l'erreur de

prédiction à l'aide de paramètres non linéaires et hautement paramétrés. [46]

4 Apprentissage profond

L' apprentissage profond est une branche d'apprentissage automatique utilisée pour former des systèmes informatiques appelés réseaux de neurones artificiels (ANN).Ces techniques reposent sur des algorithmes capables d'imiter les comportements du cerveau humain. Les RNA peuvent résoudre des problèmes complexes que les langages de programmation traditionnels ne peuvent pas résoudre, comme la reconnaissance d'images, la reconnaissance vocale et le traitement du langage naturel. Les algorithmes d'apprentissage profond mettent en œuvre plusieurs couches de neurones artificiels ou de nœuds avec diverses connexions entre eux. Les couches de nœuds sont reliées par plusieurs types de connexions. Ces connexions sont entraînées pour reconnaître et comprendre les caractéristiques d'un ensemble de données donné.Cette structure permet aux algorithmes d'apprendre de leurs expériences et d'améliorer la manière dont ils effectuent leurs tâches.[47]

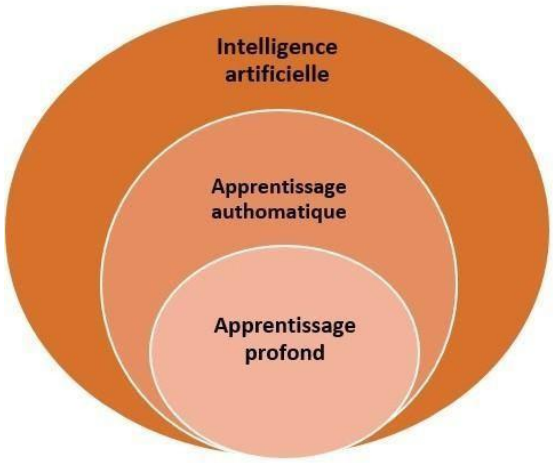

Figure II 1: Les niveaux de l'intelligence artificielle

5 Les types d'apprentissage automatique

5.1 Apprentissage automatique supervisé

L'apprentissage supervisé est une tâche de l'apprentissage automatique consistant à apprendre une fonction qui associe une entrée à une sortie à partir de paires exemple entrée sortie.

Il déduit une fonction à partir de données d'entraînement étiquetées, consistant en un ensemble d'exemples d'entraînement.

Les algorithmes d'apprentissage supervisé nécessitent une assistance externe. Le jeu de données est divisé en deux parties :

- un jeu de données d'entraînement contenant la variable de sortie à prédire ou à classifier .
- un jeu de données de test servant à évaluer la performance du modèle.[48]

5.1.1 Les Algorithmes d'apprentissage supervisé

5.1.1.1 Arbre de décision

L'arbre de décision est un graphe représentant des choix et leurs résultats sous forme d'un arbre. Les nœuds dans le graphe représentent un événement ou un choix, et les arêtes représentent les règles de décision ou les conditions. Chaque arbre est composé de nœuds et de branches. Chaque nœud représente un attribut dans un groupe à classifier, et chaque branche représente une valeur que le nœud peut prendre.[48]

5.1.1.2 Naive Bayes

C'est une technique de classification basée sur le théorème de Bayes, avec une hypothèse d'indépendance entre les prédicteurs. Un classificateur Naive Bayes suppose que la présence d'une caractéristique dans une classe est sans rapport avec les autres caractéristiques. Naive Bayes est largement utilisé dans la classification de texte, le clustering, et la prédiction basée sur la probabilité conditionnelle.[48]

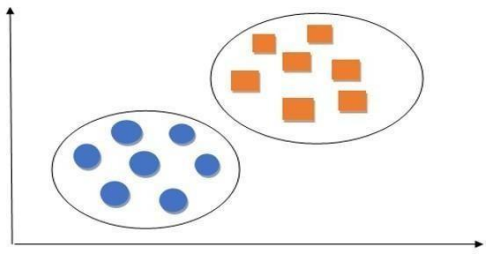

Figure II 2: Naive Bayes Classifieur

5.1.1.3 SVM (Support Vector Machine)

Le SVM est une méthode d'apprentissage supervisé utilisée pour la classification et la régression. Elle trace une marge optimale entre différentes classes afin de maximiser leur séparation. Grâce au kernel trick, elle peut aussi gérer des données non linéaires en les projetant dans un espace de plus haute dimension.

L'objectif principal est de minimiser l'erreur de classification en maximisant la distance entre les classes et la marge.[48]

5.2 Apprentissage automatique non supervisé

Les algorithmes d'apprentissage non supervisés vont apprendre certaines parties des données. Lors d'un nouveau feed, l'algorithme boude les appris précédent pour comprendre le classe de cette donnée.Mais il est surtout utilisé avec le clustering (regroupement) et la suppression des caractéristiques.[49]

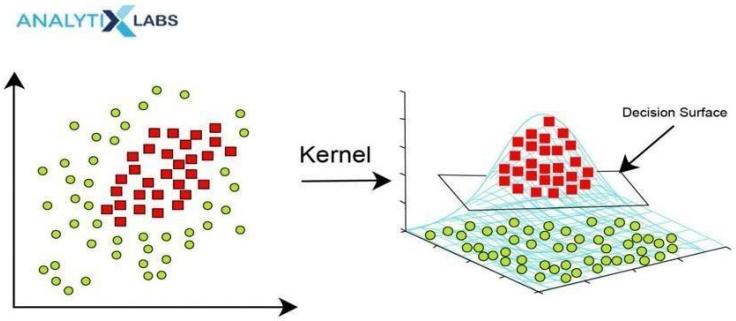

Figure II 3: SVM algorithme

5.2.1 les algorithmes d'apprentissage non supervisé

5.2.1.1 Clustering (Regroupement)

Le clustering (ou regroupement) c'est de classer des objets selon certaines similitudes entre ceux-ci. Il y a plusieurs sortes de clustering :

- **Clustering exclusif** :Toutes les données d'un élément participent à un seul groupe. L'algorithme K-means est un exemple d'approche de cette sorte.
- **Clustering hiérarchique** Construction d'une hiérarchie de groupes. Il existe deux types :
  - **Agglomératif** : Commence par considérer chaque point de données comme un groupe unique, puis renforce progressivement leur regroupement.
  - **Divisif** : Commence tous les points dans un groupe et divise progressivement..
- **Clustering chevauchant** :Un point de données peut appartenir à plusieurs de groupes en même temps avec des degrès d'appartenance différentes.
- **Clustering probabiliste** : utilise des modèles de probabilité pour former les groupes.

[50]

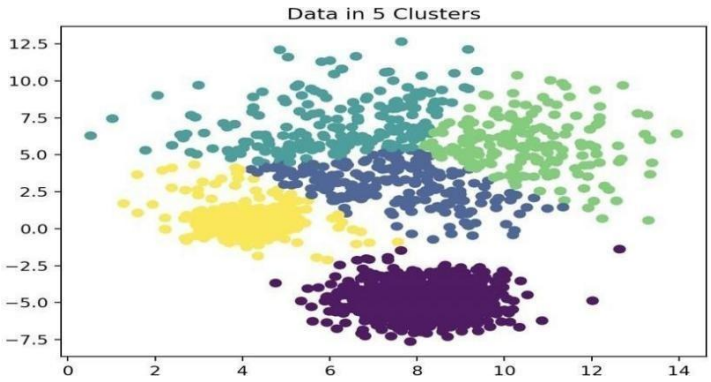

Figure II 4: K-Means Clustering

5.2.1.2 Association Rule Learning (ARL)

est un algorithme d'apprentissage non supervisé qui est utilisé pour détecter les relations négatives entre un ensemble important de données. Il est le contraire des techniques qui peuvent être utilisées avec uniquement des données numériques. Par exemple, il peut grimper à des liens comme celui entre l'achat d'une moto et d'un casque. Ces cascades peut-être exploités commercialement comme insister un produit à un client selon ses commandes anciennes. Les règles d'association sont appréciées en fonction de deux critères : le support (la fréquence avec laquelle une relation apparaît), et la confiance (la légitimité d'une relation séquelle). Cette technique est employée dans des cas d'application tels que l'analyse des shopping-cart et la navigation des mœurs en ligne.[50]

5.2.1.3 Détection d'anomalies

est un processus qui identifie les valeurs aberrantes dans un ensemble de données, ce qui peut indiquer des activités inhabituelles, comme un trafic réseau anormal ou un capteur défectueux. Lorsqu'un modèle de données s'écarte du comportement habituel, il est considéré comme une anomalie. Cette technique est utilisée dans des domaines tels que la détection d'intrusions, la prévention des fraudes, l'assurance et la surveillance militaire.[50]

5.2.1.4 Autoencodeurs

Les autoencodeurs sont un type de réseau de neurones qui apprennent à résumer les données. Ils compriment d'abord les informations pour en garder l'essentiel, puis essaient de les reconstruire. Cela aide à découvrir des liens cachés entre les données, comme des champs qui se ressemblent ou qui changent ensemble. Au lieu de garder toutes les données d'origine, le réseau apprend à reconnaître leur structure pour pouvoir les traiter plus facilement.[50]

5.3 Apprentissage par renforcement

L'apprentissage par renforcement est un type d'apprentissage automatique qui traite les données par essais et erreur , de la même façon que l'être humain traite un problème . Il peut se faire sans interaction d'humain .Il permet a un algorithme d'apprentissage automatique d'apprendre par plusieurs essaies en attribuant une valeur positive ou négatives a chaque action d'après le résultat.[51]

5.3.1 les algorithmes d'apprentissage par renforcement

5.3.1.1 Apprentissage par différence temporelle (TD Learning)

Le TD Learning est une méthode d'apprentissage par renforcement où un agent apprend seul, sans avoir besoin de réponses toutes faites. Il essaye de prédire des récompenses futures en se basant sur ce qu'il observe petit à petit. Par exemple, pour deviner le temps qu'il fera samedi, il ajuste ses prévisions chaque jour en fonction des nouvelles infos. Il compare ce qu'il attendait avec ce qu'il voit réellement et s'améliore ainsi progressivement.[52]

5.3.1.2 Apprentissage Q (Q-Learning)

Apprentissage Q est issue de l'apprentissage par renforcement , est un modèle qui permet à un agent d'apprendre comment agir de manière optimale dans une situation donnée. Il est utile dans le cas où l'agent doit prendre une série de décisions pour maximiser les récompenses au fil du temps. Le principe du Q-Learning est l'utilisation d'une table Q, qui stocke l'utilité attendue d'une action dans un état spécifique . En mettant à jour de manière itérative ce tableau en fonction des expériences de l'agent, Q-Learning permet à l'agent de se diriger vers une politique optimale, qui montre la meilleure action à prendre dans chaque état.[53]

6 Algorithmes d'apprentissage profond

6.1 Perceptron multicouche (MLP)

Le Multilayer Perceptron est un réseau de neurones à propagation avant (feed-forward). Les informations circulent de l'entrée vers la sortie à travers des couches dites "cachées".

Il existe également des versions feed-back, où les connexions sont bidirectionnelles. Ces architectures sont plus complexes et dynamiques, car l'état du réseau évolue jusqu'à atteindre un équilibre pour chaque entrée.[54]

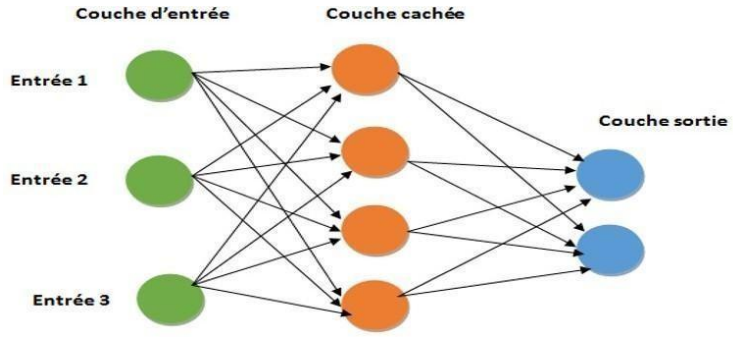

Figure II 5: Architecture MLP

6.2 Réseaux neuronaux convolutifs (CNN)

Les réseaux neuronaux convolutifs (CNN) sont un type de réseau utilisé surtout pour traiter des images. Ils apprennent automatiquement les caractéristiques importantes, des détails simples aux motifs plus complexes. Les premières couches utilisent des filtres pour détecter des éléments dans l'image, comme une loupe qui glisse dessus. Ces filtres créent des cartes de caractéristiques, qui résument ce que le réseau a détecté. Après ces étapes, le CNN utilise des couches finales pour faire la classification.[55]

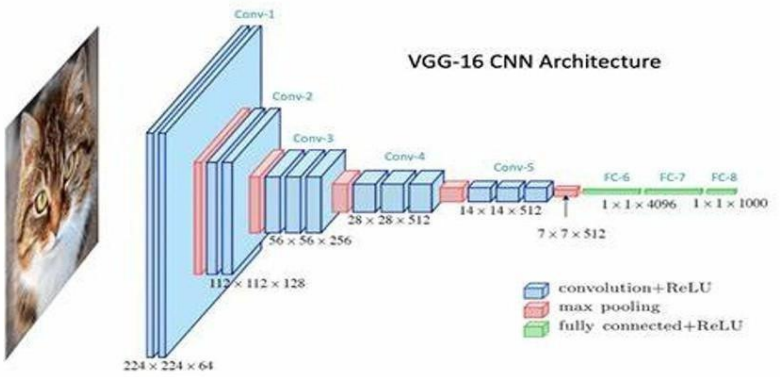

Figure II 6: Architecture CNN

6.3 Réseaux neuronaux récurrents (RNN)

Les réseaux neuronaux récurrents (RNN) sont un type de réseau qui peut mémoriser des informations dans le temps grâce à un état caché récurrent. Contrairement aux réseaux clas-

siques, ils prennent en compte les activations précédentes, ce qui les rend utiles pour traiter des données séquentielles comme la parole. Grâce à leurs connexions de rétroaction, ils peuvent retenir le contexte. Les RNN peuvent être plus ou moins connectés : certains relient tous les neurones entre eux, ce qui permet de transmettre l'état d'un moment à l'autre, même sans couches d'entrée et de sortie distinctes.[56]

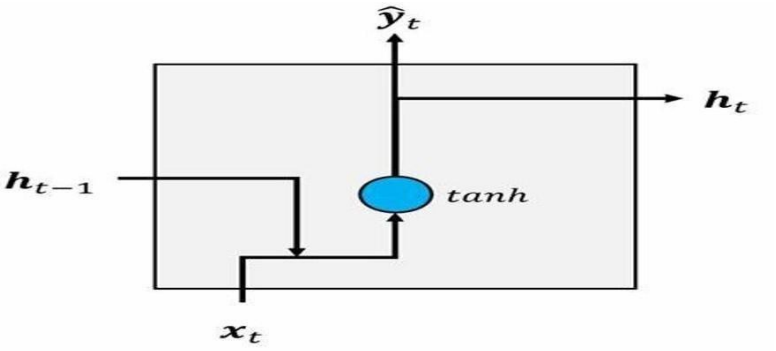

Figure II 7: Architecture RNN

6.4 Réseaux de mémoire à long et court terme (LSTM)

Les réseaux LSTM (Long Short-Term Memory) sont une version spéciale des RNN capables de retenir des informations sur une longue durée. Ils utilisent une mémoire interne pour décider quoi garder, oublier ou mettre à jour. Grâce à un système de portes contrôlées par des poids appris, les LSTM gèrent les informations importantes et ignorent celles qui ne le sont pas. Cela les rend très efficaces pour traiter des données comme des textes ou des séquences audio.[57]

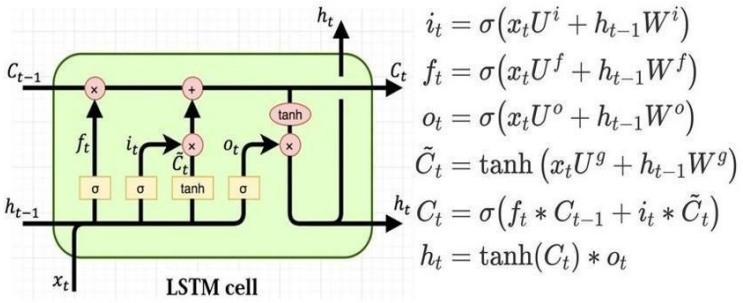

Figure II 8: Architecture LSTM

6.5 Réseaux antagonistes génératifs (GAN)

Les GAN (Generative Adversarial Networks) sont des réseaux neuronaux capables de créer de nouvelles données qui ressemblent à celles d'origine, comme des images de visages inventés.

Ils fonctionnent avec deux parties :

- Le générateur, qui crée des données (images, sons, etc.)
- Le discriminateur, qui essaie de repérer si ces données sont vraies ou fausses

Les deux réseaux s'affrontent : le générateur essaie de tromper le discriminateur, et ce jeu les rend tous deux meilleurs au fil du temps. [58]

7 Différences entre apprentissage automatique et apprentissage profond

L'apprentissage profond est un sous-ensemble de l'apprentissage automatique. Voici leurs principales différences :

- L'apprentissage automatique nécessite souvent la sélection manuelle des variables;
- L'apprentissage profond extrait automatiquement les caractéristiques ;
- Le matériel requis est plus exigeant pour l'apprentissage profond (GPU, TPU);
- L'apprentissage automatique fonctionne avec des jeux de données plus petits, tandis que le deep learning exige de grands volumes ;
- Le deep learning est mieux adapté à la vision par ordinateur et au traitement du langage naturel. [59]

Caractéristique	**Apprentissage Automatique**	**Apprentissage Profond**
Origine	Branche originale de l'IA	Sous-ensemble de l'AA
Sélection des caractéristiques	Manuelle	Automatique
Architecture	Peu de couches	Réseaux profonds
Modélisation	Statistique	Optimisation numérique
Supervision	Données étiquetées requises	Peut être non supervisé
Traitement des données	Données prétraitées	Données brutes acceptées
Matériel requis	Peu exigeant	Besoin de GPU/TPU
Applications	Statistiques, classification	Vision, NLP, audio
Auto-amélioration	Limitée	Optimisation automatique possible
Volume de données	Faible volume suffisant	Besoin de grands volumes

Tableau II 1: Caractéristiques fondamentales entre l'apprentissage automatique traditionnel et l'apprentissage profond

8 Métriques d'évaluation en classification

8.1 Matrice de confusion

La matrice de confusion est un tableau qui compare les prédictions du modèle avec les vraies étiquettes. Elle permet d'évaluer la performance du modèle à travers quatre résultats possibles :

- **Vrais positifs (TP)** : bonnes prédictions de la classe positive ;
- **Faux positifs (FP)** : prédictions incorrectes de la classe positive ;
- **Vrais négatifs (TN)** : bonnes prédictions de la classe négative ;
- **Faux négatifs (FN)** : cas où le modèle a raté la classe positive.

8.2 Précision (Precision)

La précision correspond au pourcentage des prédictions positives qui sont réellement correctes :

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

8.3 Rappel (Recall)

Le rappel est le pourcentage des vrais positifs détectés correctement par le modèle :

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

8.4 F-mesure (F1-score)

La F-mesure (F1-score) est la moyenne harmonique entre la précision et le rappel. Elle est utile lorsque les classes sont déséquilibrées :

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

8.5 Exactitude (Accuracy)

L'exactitude représente la proportion de prédictions correctes (positives et négatives) sur l'ensemble des prédictions :

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

9 État de l'art

Nous présentons maintenant certains travaux antérieurs portant sur l'utilisation de techniques classiques et d'apprentissage automatique pour la classification du trafic réseau.

9.1 Travaux connexes

9.1.1 Études basées sur le Machine Learning

Moore et Papagiannaki [60] ont proposé une approche simple de classification du trafic réseau basée sur les numéros de ports. Leur méthode repose sur l'hypothèse que chaque protocole utilise des ports standards bien définis. Cependant, leur étude a montré que cette technique atteignait difficilement les 70 % de précision, et ce, quel que soit le volume de paquets analysés. Elle ne tient pas compte des applications utilisant des ports dynamiques ou chiffrés, ni des stratégies d'obfuscation employées par certains types de trafic comme le P2P ou le Darknet.

Madhukar et Williamson [60] ont mené une analyse approfondie du trafic P2P et ont démontré que l'approche par ports ne permettait pas d'identifier entre 30 % et 70 % du trafic total. Ce taux d'échec élevé s'explique par la nature dynamique et décentralisée des applications P2P, qui utilisent souvent des ports aléatoires ou déguisés, rendant l'identification par numéro de port inefficace.

Sen et al. [60] ont également critiqué la méthode basée sur les ports, en montrant que seuls 30 % du trafic P2P utilisaient des ports standards. Cela signifie que 70 % de ce trafic échappait à la détection via cette méthode, soulignant l'importance d'approches plus intelligentes basées sur le contenu ou les caractéristiques du trafic.

Al Rawi et al. [61] ont utilisé une combinaison de caractéristiques statistiques (durée, taille des paquets, inter-arrival time) et de techniques de machine learning comme SVM et Random Forest. Leur approche hybride a atteint une précision de 98 %, prouvant que des modèles bien entraînés sur des features soigneusement sélectionnées peuvent offrir d'excellentes performances.

Pookpun et al. [62] ont examiné la classification du trafic réseau Darknet en utilisant des techniques d'apprentissage automatique classiques sur le dataset CIC-Darknet2020. Initialement, les auteurs ont testé ces modèles sans suréchantillonnage, ce qui a conduit à des résultats faibles en raison du déséquilibre entre les classes, les flux malveillants étant sous-représentés. Random Forest a montré une précision inférieure à 80 % dans cette configuration, limitant sa capacité à détecter efficacement les flux malveillants, notamment ceux obfusqués.

Face à ces limitations, les auteurs ont appliqué des techniques de suréchantillonnage, comme SMOTE et ADASYN, et ont constaté une amélioration significative. Avec Random Forest combiné à SMOTE, la précision a atteint 91,3 %, démontrant l'impact positif de ces techniques sur l'équilibre des classes et la qualité de la classification.

Zangeneh Nezhad et Baniasadi [63] ont mené une étude comparative sur la détection du trafic VPN et Tor dans le contexte du Darknet, en s'appuyant sur des algorithmes de machine learning supervisé. Leur travail se base sur le jeu de données CIC-Darknet2020, qui comprend différents types

de trafic chiffré, y compris Tor, VPN, et trafic web classique.

Ils ont testé quatre modèles d'apprentissage classiques : Random Forest, Decision Tree (J48), Naive Bayes, et Support Vector Machine (SVM). L'originalité de leur approche repose sur la robustesse de la validation croisée : 5-fold, 10-fold, split 66/34, split 80/20.

Parmi tous les modèles testés, J48 s'est distingué avec une précision maximale de 99,6 %, obtenue lors de la validation 10-fold, et un temps d'exécution réduit à 15 secondes, ce qui est notable pour une détection en quasi temps réel. Random Forest a également obtenu d'excellents résultats avec une précision pouvant atteindre 98,74 %, mais avec un coût computationnel légèrement plus élevé que J48.

9.1.2 Études basées sur le Deep Learning

K. Ali, A. Tariq et H. Abbas [61] ont proposé un modèle basé sur des Convolutional Neural Networks (CNN) pour classifier du trafic chiffré. Leur approche se concentre sur l'extraction automatique de caractéristiques à partir des flux de paquets. Les résultats ont montré une précision atteignant 97 %, prouvant que les CNN sont capables de capturer des motifs discriminants même dans des données non lisibles.

Lu et al. [61] ont développé un modèle hybride combinant Bidirectional LSTM (réseaux de neurones récurrents bidirectionnels) pour modéliser les dépendances temporelles, et Random Forest pour la classification finale. Ce modèle a atteint 96,7 % de précision, mettant en évidence l'efficacité des architectures mixtes qui exploitent à la fois les séquences temporelles et les arbres de décision.

Zhou et al. [61] dans cette même logique, ont proposé une combinaison de Bi-LSTM pour capturer les relations temporelles dans les séquences de paquets, et K-Nearest Neighbors (KNN) pour la classification. Leur modèle a obtenu 96,4 % de précision, illustrant la complémentarité entre une analyse séquentielle profonde et une classification non supervisée.

Wang [61] a introduit un modèle hybride qui fusionne un LSTM pour l'analyse séquentielle et un CNN2D pour l'extraction de motifs visuels/spatiaux dans les données réseau. Ce modèle a permis d'atteindre une précision de 92,2 %, avec une bonne capacité à gérer des données bruyantes ou partiellement chiffrées.

Xinyi Hu et al. [61] ont introduit CLD-Net, un modèle combinant CNN et LSTM pour classifier le trafic chiffré. Leur approche a obtenu 98 % de précision pour les flux VPN, et 92 % pour le trafic Skype. Le modèle tire parti des forces des CNN pour l'extraction de caractéristiques locales, et des LSTM pour la modélisation des séquences.

Lulu Guo et al. [61] ont testé deux modèles : un AutoEncodeur Convolutif (CAE) et un CNN classique pour la classification de trafic VPN en temps réel. Le CAE a atteint une précision remarquable de 99,8 %, tandis que le CNN a obtenu 92,92 %. Le CAE a montré une meilleure capacité de compression et de reconstruction de motifs complexes.

Pathmaperuma et al. [61] ont exploré la détection de l'usage d'applications mobiles à partir du

trafic réseau en utilisant des techniques d'apprentissage profond, notamment les réseaux de neurones convolutifs (CNN). L'objectif était de construire un modèle capable de reconnaître automatiquement les applications utilisées sur des appareils mobiles à partir de leurs empreintes réseau.

Les auteurs ont développé une architecture CNN entraînée à partir de données collectées sur diverses applications mobiles. Le modèle a d'abord atteint une précision moyenne de 88 %, démontrant une capacité notable à distinguer les flux réseau selon l'application génératrice.

Cependant, pour améliorer davantage la précision et réduire le bruit présent dans les données, les chercheurs ont intégré une phase de filtrage dans le prétraitement. Ce filtrage consistait à nettoyer les données de trafic et à extraire les caractéristiques les plus pertinentes avant de les introduire dans le modèle.

Grâce à cette optimisation, la précision du CNN a augmenté, atteignant 92 %, prouvant que le traitement préalable des données joue un rôle déterminant dans la performance du modèle. Dans une approche orientée Deep Learning, Etyang et al. [62] ont proposé un modèle avancé pour la classification du trafic chiffré sur le Darknet. Leur travail repose également sur le dataset CIC-Darknet2020 et se distingue par la comparaison de plusieurs architectures neuronales : FFNN, CNN, LSTM, et un modèle hybride CNN + LSTM + Attention.

Le modèle hybride a surpassé tous les autres, atteignant des scores de performance élevés : précision de 91,5 %, recall élevé, et F1-score de 91,5 %. Les performances du modèle s'expliquent par sa capacité à capturer les caractéristiques spatiales via les couches CNN, à modéliser la dynamique temporelle grâce au LSTM, et à focaliser l'apprentissage sur les séquences les plus discriminantes via le mécanisme d'attention.

Les modèles simples ont obtenu des scores inférieurs : CNN seul (90,8 %), FFNN (88,7 %), et LSTM seul (87,65 %).

9.2 Synthèse des travaux

La comparaison des études présentée dans le tableau ci-dessous repose sur la précision (accuracy).

Auteur(s)	**Précision obtenue**
Moore & Papagiannaki	≤70%
Madhukar & Williamson	30–70% de trafic non identifié
Sen et al.	30%
K. Ali et al.	97%
Lu et al.	96,7%
Zhou et al.	96,4%
Wang et al.	92,2%
Al Rawi et al.	98%
Xinyi Hu et al.	98% / 92%
Lulu Guo et al.	99,8% / 92,92%
Pathmaperuma et al.	88% / 92%
Pookpun et al.	91,3%
Zangeneh Nezhad & Baniasadi	99,6%
Etyang et al.	91,5%

Tableau II 2: Précisions obtenues dans les différentes études

10 Conclusion

Avec l'explosion du trafic et la complexité croissante des réseaux, les anciennes méthodes de classification montrent leurs limites. Aujourd'hui, grâce à l'intelligence artificielle, notamment au deep learning, il est possible d'analyser le trafic de manière plus efficace, rapide et précise. Ces techniques ouvrent la voie à une meilleure gestion de la Qualité de Service (QoS), tout en répondant aux nouveaux défis techniques tels que le chiffrement ou les flux inconnus. Ce chapitre a permis de montrer l'évolution des approches et de poser les bases pour une solution plus moderne et performante.

Chapitre 03 :Contribution et Implémentation

1 Introduction

Ce chapitre est consacré à la phase pratique de notre travail, à savoir la réalisation et l'implémentation du projet. Après avoir défini précisément les besoins fonctionnels et techniques dans les chapitres précédents, nous abordons maintenant la mise en œuvre concrète des solutions proposées.

Dans un premier temps, nous présenterons l'environnement de développement ainsi que les outils utilisés. Ensuite, nous détaillerons méthodiquement les différentes étapes de réalisation, allant de la conception initiale des modules jusqu'à leur implémentation effective.

2 Environnement d'exécution

2.1 Google Colab

Google Colab est un outil puissant pour le développement en Python, en particulier dans le domaine de l'apprentissage automatique. Il offre un environnement de développement interactif, un accès aux GPU et TPU, un stockage sur Google Drive, la collaboration en temps réel et de nombreuses autres fonctionnalités avancées [64].

Figure III: 1 Logo Google Collab

2.2 Jupyter

Jupyter est un projet open source qui offre un environnement interactif accessible via le Web. Il permet aux utilisateurs de concevoir et de partager des documents intégrant du code exécutable, des équations, des visualisations et des textes explicatifs. Son nom, « Jupyter », provient des langages de programmation Julia, Python et R, qu'il prend en charge. Grâce à cette compatibilité, Jupyter est largement adopté par les data scientists, les statisticiens et les chercheurs en quête d'une plate-forme souple pour l'analyse et la visualisation de données. [65]

Figure III : 2 Logo Jupyter

2.3 Language python

Jupyter est un projet open source qui offre un environnement interactif accessible via le Web. Il permet aux utilisateurs de concevoir et de partager des documents intégrant du code exécutable, des équations, des visualisations et des textes explicatifs. Son nom, « Jupyter », provient des langages de programmation Julia, Python et R, qu'il prend en charge. Grâce à cette compatibilité, Jupyter est largement adopté par les data scientiste, les statisticiens et les chercheurs en quête d'une plate-forme souple pour l'analyse et la visualisation de données. [66].

Figure III : 3 Logo Python

2.3.1 Bibliothèques de Python

2.3.1.1 Pandas

Pandas est une bibliothèque Python open source incontournable en Data Science, utilisée pour la manipulation et l'analyse de données. Basée sur NumPy, elle tire son nom de « panel data » et s'intègre facilement avec d'autres outils de l'environnement Python.[67]

2.3.1.2 NumPy

NumPy, abréviation de « Numerical Python », est une bibliothèque open source en Python. Elle est largement utilisée pour la programmation scientifique, notamment en Data Science, ingénierie, mathématiques et autres domaines scientifiques. [68]

2.3.1.3 Scikit-Learn

Scikit-Learn est une bibliothèque open source Python pour l'apprentissage automatique, basée sur NumPy, SciPy et Matplotlib. Elle propose des algorithmes de classification, régression, clustering et réduction de dimension, ainsi que des outils de prétraitement des données.

[69].

2.3.1.4 Matplotlib

Matplotlib est une bibliothèque de Data Visualization en Python, inspirée de Matlab, elle permet de créer des visualisations statiques, animées et interactives. C'est un outil prisé par les utilisateurs de Python et de NumPy, souvent utilisé dans des serveurs d'applications web, des shells et des scripts Python. [70].

2.3.1.5 TensorFlow

TensorFlow est un framework open source complet pour créer des applications d'apprentissage automatique. Il utilise un calcul symbolique pour entraîner et exécuter des réseaux neuronaux profonds grâce à des flux de données et à la programmation différentiable. TensorFlow fournit

aux développeurs une vaste gamme d'outils, bibliothèques et ressources communautaires pour construire des modèles d'apprentissage automatique. Créé par Google, c'est aujourd'hui l'un des frameworks de deep learning les plus populaires, utilisé dans de nombreux produits Google comme la recherche, la traduction, les sous-titres automatiques et les recommandations.

[71].

Figure III : 4 Logo TensorFlow

2.3.1.6 Keras

Keras, écrite en Python, est une bibliothèque open source conçue pour le prototypage rapide de modèles de deep learning. Accessible aux débutants en intelligence artificielle, elle propose une API de haut niveau compatible avec plusieurs frameworks de réseaux de neurones, tels que TensorFlow, Microsoft Cognitive Toolkit, PlaidML ou Theano.

Créée en 2015 par François Chollet, développeur chez Google, Keras vise à fournir un environnement simple et rapide pour construire des réseaux de neurones artificiels. Elle fait partie du projet Oneiros. (Open-ended Neuro-Electronic Intelligent Robot Operating System). [72].

Figure III : 5 Logo keras

3 Notre contribution

Notre contribution s'articule autour de l'exploration et du développement de méthodes novatrices pour la classification intelligente du trafic réseau, en exploitant la puissance des

techniques d'apprentissage profond. Face à la complexité croissante des flux de données sur les réseaux modernes, caractérisée par une diversité des applications, des protocoles chiffrés et des exigences de qualité de service variées, les approches traditionnelles de classification montrent leurs limites.

Pour relever ces enjeux, nous explorons différentes architectures d'apprentissage profond appliquées à l'inspection approfondie des paquets (DPI), telles que les réseaux de neurones à mémoire longue (LSTM) et les perceptrons multicouches (MLP). En tirant parti de la puissance du deep learning, notre objectif est de concevoir des algorithmes DPI intelligents, capables de classer avec précision les applications en temps réel tout en traitant de grands volumes de données.

Notre approche se divise en plusieurs étapes clés. Nous commençons par le prétraitement des données, notamment la correction des valeurs aberrantes. Ensuite, nous sélectionnons et présentons différents modèles d'apprentissage (MLP, CNN, LSTM , GRU , CNN- LSTM). Enfin, nous évaluons et comparons les performances de ces modèles afin d'identifier celui qui offre la meilleure précision pour garantir une qualité de service optimale.

3.1 Ensemble de données

3.1.1 Description de l'ensemble de données CIC-DARKNET2020

Le jeu de données CIC-Darknet2020 [73] , développé par l'Institut canadien pour la cybersécurité de l'Université du Nouveau-Brunswick, a été conçu pour simuler des environnements réseau réalistes intégrant à la fois du trafic légitime et du trafic issu du Darknet. Son objectif est de permettre l'évaluation de nouvelles approches de classification du trafic réseau.

Le Darknet désigne une zone de l'Internet non indexée par les moteurs de recherche classiques, accessible uniquement via des outils spécifiques comme le réseau Tor, qui est un système de communication conçu pour assurer l'anonymat des utilisateurs lors de leur navigation sur Internet. Son nom, "The Onion Router", fait référence à sa méthode de chiffrement en couches, comparable aux pelures d'un oignon. [74].

Les flux VPN (Virtual Private Network), quant à eux, permettent de chiffrer le trafic réseau et de masquer l'adresse IP réelle de l'utilisateur, offrant ainsi une navigation sécurisée et privée.

Bien que souvent associé à des activités illégales, le Darknet peut aussi être utilisé pour garantir la confidentialité de certaines communications dans des contextes légitimes. [75]

Dans CIC-Darknet2020, une architecture en deux niveaux est utilisée pour générer le trafic. Le premier niveau produit du trafic bénin, tandis que le second génère les flux liés au Darknet, répartis en plusieurs catégories : audio, navigation web, messagerie instantanée, e-mail, P2P, transfert de fichiers, streaming vidéo et VOIP.

Pour constituer un jeu de données complet et représentatif, les ensembles ISCXTor2016 et ISCXVPN2016 ont été fusionnés, regroupant les flux VPN et Tor dans les catégories de trafic darknet appropriées. Les caractéristiques extraites sont fournies au format CSV, tandis que les captures de

paquets (PCAP) sont également disponibles, facilitant leur exploitation dans diversoutils d'analyse.

Le Tableau III 1 récapitule les catégories de trafic Darknet et les applications utilisées pour leur génération.

Type 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 (BitTorrent)
Transfert	Skype, FTP sur SSH (SFTP) et FTP sur SSL utilisant Filezilla et un service externe
Flux-vidéo	Vimeo et Youtube
VoIP	Appels vocaux Facebook, Skype et Hangouts

Table III 1 : Types de trafic réseau de l'ensemble de données CIC-Darknet2020Table

La Figure III 6 présente :

- les détails du nombre d'échantillons de trafic bénin et de trafic darknet à la première couche
- le nombre de flux chiffrés présents dans notre trafic darknet.

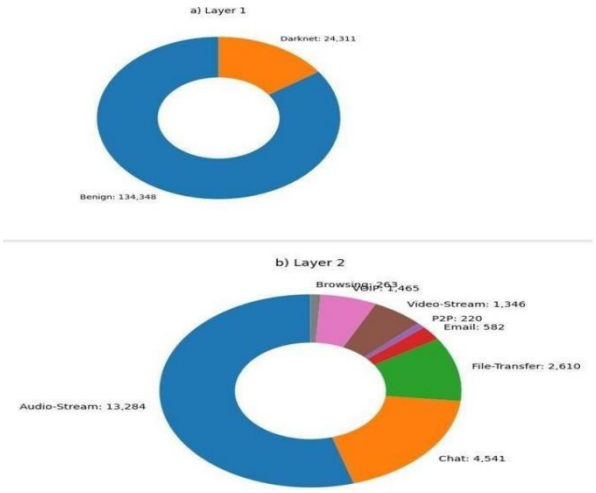

Figure III 6 Distribution des données dans darknet2020

3.1.2 Caractéristiques du Jeu de Données CIC-Darknet2020

Afin d'effectuer une classification efficace du trafic réseau, le jeu de données CIC-Darknet2020 fournit un ensemble riche de caractéristiques statistiques extraites de chaque flux réseau. Ces attributs permettent de capturer des informations essentielles sur le comportement temporel, les tailles de paquets, les taux de transmission et les délais d'inactivité, entre autres.

Le tableau suivant présente une description détaillée des 40 principales caractéristiques extraites du dataset, qui ont été utilisées comme variables d'entrée pour l'entraînement des modèles de classification :

Numéro	Caractéristique et Description
1	Flow IAT Max : Le temps d'intervalle maximal entre deux paquets dans le flux.
2	Fwd IAT Total : Somme des intervalles de temps entre paquets envoyés en avant.
3	Packet Length Mean : La taille moyenne des paquets.
4	Bwd Packets/s : Nombre de paquets reçus par seconde dans le sens arrière.
5	Average Packet Size : La taille moyenne des paquets dans le flux.
6	Fwd IAT Mean : Intervalle moyen entre paquets envoyés en avant.
7	Idle Max : Durée maximale d'inactivité dans le flux.
8	Bwd Segment Size Avg : Taille moyenne des segments reçus en arrière.
9	Idle Mean : Durée moyenne d'inactivité dans le flux.
10	Bwd Packet Length Mean : Taille moyenne des paquets reçus en arrière.
11	Packet Length Max : La taille maximale d'un paquet.
12	Fwd Header Length : Taille de l'en-tête des paquets envoyés en avant.
13	Flow Duration : La durée totale du flux.
14	Total Length of Bwd Packet : Somme des tailles des paquets reçus en arrière.
15	Idle Min : Durée minimale d'inactivité dans le flux.
16	Bwd Packet Length Max : Taille maximale des paquets reçus en arrière.
17	Flow IAT Mean : Temps d'intervalle moyen entre deux paquets.
18	Bwd Header Length : Taille de l'en-tête des paquets reçus.
19	Fwd Packets/s : Paquets envoyés par seconde dans le sens avant.
20	Fwd IAT Min : Intervalle minimal entre paquets envoyés.
21	Flow Packets/s : Paquets par seconde dans le flux.
22	Fwd Packet Length Min : Taille minimale des paquets envoyés.
23	Total Length of Fwd Packet : Somme des tailles des paquets envoyés.
24	Packet Length Min : Taille minimale des paquets.

Numéro	Caractéristique et Description
25	Fwd IAT Max : Intervalle maximal entre paquets envoyés.
26	Subflow Fwd Bytes : Octets envoyés dans le sous-flux avant.
27	Fwd Packet Length Max : Taille maximale des paquets envoyés.
28	FWD Init Win Bytes : Taille de la fenêtre d'initialisation avant.
29	Flow Bytes/s : Octets transférés par seconde dans le flux.
30	Subflow Bwd Bytes : Octets reçus dans le sous-flux arrière.
31	Fwd Segment Size Avg : Taille moyenne des segments envoyés.
32	Bwd Packet Length Min : Taille minimale des paquets reçus.
33	Fwd Packet Length Mean : Taille moyenne des paquets envoyés.
34	Bwd Init Win Bytes : Taille de la fenêtre d'initialisation arrière.
35	Packet Length Variance : Variance des tailles de paquets.
36	Flow IAT Std : Écart-type des intervalles entre paquets.
37	Packet Length Std : Écart-type des tailles des paquets.
38	Bwd IAT Max : Intervalle maximal entre paquets reçus.
39	Flow IAT Min : Intervalle minimal entre paquets dans le flux.
40	Fwd IAT Std : Écart-type des intervalles envoyés.

Table III 2 : Caractéristiques du jeu de données Darknet2020

3.2 Préparation et gestion du jeu de données CIC-DARKNET2020

Avant d'entraîner un modèle d'apprentissage profond, une phase essentielle de prétraitement des données a été réalisée afin d'améliorer la qualité de l'apprentissage et de réduire la complexité du modèle. Dans un premier temps, les colonnes non informatives ou redondantes ont été supprimées du jeu de données. C'est notamment le cas des adresses IP source et destination (Source IP, Destination IP), qui ne contribuent pas à la généralisation du modèle et peuvent introduire un biais ou du surapprentissage, car elles sont souvent spécifiques à une session ou à un environnement particulier.

Cette étape permet également de réduire la dimensionnalité des données, ce qui diminue les temps de traitement et améliore la stabilité de l'entraînement. D'autres colonnes jugées non pertinentes pour la tâche de classification ont également été supprimées, telles que les identifiants uniques de flux ou certains drapeaux de contrôle réseau ne contenant que des valeurs nulles ou constantes.

3.3 Méthodologie

La méthodologie proposée dans cette étude repose sur une approche hiérarchique hybride nommée Deep HyCLASS-Net, structurée en deux phases distinctes. La première phase s'appuie sur des modèles de machine learning pour réaliser un filtrage préliminaire du trafic Internet, tandis que la seconde phase utilise des réseaux de neurones profonds pour effectuer une classification fine du trafic identifié. Cette démarche permet de traiter efficacement les défis liés à la complexité de classification du trafic Internet, en distinguant clairement les flux VPN, Tor et non chiffrés, tout en assurant une haute précision dans l'identification des sous-catégories telles que la V, et autres contenus spécifiques, comme illustré à la figure 7. Ce processus hiérarchique initial de filtrage permet de réduire la complexité globale du modèle et d'améliorer ses performances globales.

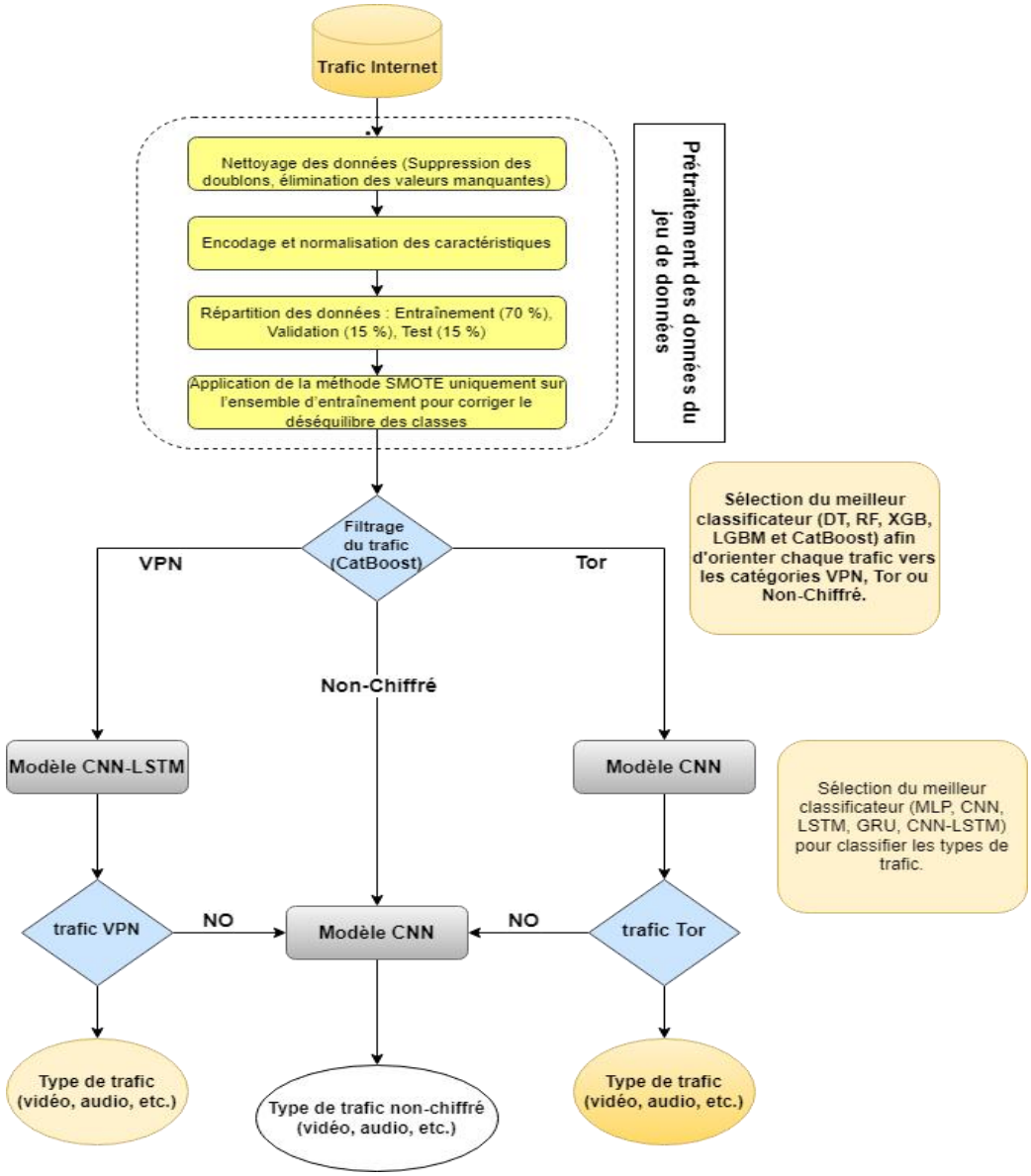

Figure III 7 : schéma representant la methodologie Deep HyCLASS-Net

3.3.1 Le prétraitement des données

Le prétraitement représente une étape essentielle pour garantir la qualité et l'exploitabilité des données d'entrée. Il comprend les étapes suivantes :

3.3.1.1 Nettoyage des données

Les valeurs infinies (+inf, -inf) ont été remplacées par NaN, puis les lignes contenant des valeurs manquantes ont été supprimées dans les ensembles d'entraînement, de validation et de test.

3.3.1.2 Encodage et normalisation

Transformation des variables catégorielles par encodage (ex. : one-hot encoding) et normalisation des variables numériques afin d'optimiser l'apprentissage.

3.3.1.3 Répartition des données

Séparation des données en ensembles d'entraînement (70 %), validation (15 %) et test (15 %), suivant les bonnes pratiques de l'apprentissage supervisé.

3.3.1.4 Équilibrage des classes

Application de l'algorithme SMOTE (Synthetic Minority Oversampling Technique) uniquement sur l'ensemble d'entraînement afin de compenser efficacement le déséquilibre des classes.

3.3.2 Classification hiérarchique initiale (Filtrage du trafic)

La première phase consiste en un filtrage préliminaire du trafic Internet via le modèle CatBoost, un algorithme de gradient boosting reconnu pour sa robustesse face à des données complexes et catégorielles. L'objectif principal de cette phase est de répartir efficacement le trafic en trois grandes classes : VPN, Tor et Non-Chiffré.

- **Sélection du modèle :**

La sélection du modèle CatBoost a été effectuée après une comparaison approfondie des performances de plusieurs classificateurs, notamment les arbres de décision (DT), Random Forest (RF), XGBoost (XGB) et LightGBM (LGBM). CatBoost s'est distingué par ses excellentes performances en termes de précision, rapidité d'exécution et faible complexité, particulièrement adaptées à la classification préliminaire en trois catégories seulement.

3.3.3 Classification fine par réseaux de neurones profonds

Après le filtrage initial, la deuxième phase consiste en une classification fine, spécifiquement adaptée à chaque catégorie préalablement identifiée :

a) Trafic VPN (CNN-LSTM) :

L'architecture hybride CNN-LSTM combine des couches convolutionnelles (CNN) pour extraire des caractéristiques spatiales pertinentes et des couches récurrentes LSTM pour capturer les dépendances temporelles spécifiques au trafic VPN, assurant ainsi une classification optimale.

b) Trafic Tor et Non-Chiffré (CNN) :

Ces types de trafic présentent principalement des motifs spatiaux distincts ; ils sont donc traités efficacement à l'aide d'un modèle CNN classique.

Pour chaque type de trafic, le choix définitif du modèle optimal a été réalisé après une comparaison approfondie des performances obtenues par plusieurs architectures de deep learning, incluant les modèles MLP, CNN, GRU, LSTM et CNN-LSTM. Ce choix s'est fondé sur des critères rigoureux d'évaluation des performances telles que l'exactitude, la précision, le rappel et le score F1.

3.3.4 validation des performances

L'évaluation des modèles repose sur des critères rigoureux : exactitude, précision, rappel et score F1, assurant ainsi une validation robuste, complète et reproductible des résultats.

3.3.5 Description détaillée des architectures retenues

a) **Algorithmes de filtrage initial (Machine Learning)** :

CatBoost, XGBoost, LightGBM, Random Forest, Decision Tree avec leurs paramètres par défaut.

b) **Architectures des modèles Deep Learning** :

Le tableau ci-dessous présente un résumé des différentes architectures de modèles de deep learning explorées dans cette étude, en détaillant leur structure de couches, l'optimiseur utilisé ainsi que les techniques complémentaires mises en œuvre pour améliorer les performances de classification.

Modèle	Structure des couches	Optimiseur	Techniques complémentaires
CNN	Conv1d (512 Et 256 Filtres), Maxpooling, Dropout, Dense Finale (Softmax)	Adam (Lr=0,0001)	Régularisation L2, Dropout
CNN-LSTM	Conv1d (Extraction Spatiale), Couches LSTM (128-256 Unités), Dropout, Dense Finale (Softmax)	Adam (Lr=0,0001)	Dropout, Régularisation L2
MLP	3 Couches Denses (512, 256, 128 Unités), Dropout, Dense Finale (Softmax)	Adam (Lr=0,0001)	Dropout
LSTM	3 Couches Lstm (512, 256, 128 Unités), Dropout, Dense Finale (Softmax)	Adam (Lr=0,0001)	Dropout
GRU	3 Couches Gru (512, 256, 128 Unités), Dropout, Dense Finale (Softmax)	Adam (Lr=0,0001)	Dropout

Table III 3 : Architectures des modèles de deep learning

c) **Paramètres complémentaires :**

- **Dropout**

Entre 0,2 et 0,5 pour limiter le surapprentissage.

- **Régularisation L2**

Utilisée principalement sur les couches convolutionnelles.

- **Batch Size & Epochs**

Ajustement adapté à la taille du jeu de données (typiquement batch size = 128, epochs = 100).

- **Early stopping et réduction adaptative du taux d'apprentissage**

pour optimiser la convergence et éviter le surapprentissage.

En résumé, notre approche hiérarchique hybride Deep HyCLASS-Net utilise efficacement CatBoost pour le filtrage initial et une combinaison optimisée des architectures CNN-LSTM et CNN pour la classification fine du trafic, assurant ainsi une gestion optimale de la Qualité de Service (QoS) dans les réseaux contemporains.

4 Résultats et Discussion

4.1 Résultats obtenus

Cette section présente les résultats des différentes expérimentations réalisées avec divers algorithmes d'apprentissage automatique pour la classification du trafic Internet.

4.1.1 Classification globale sur Darknet2020

Une première expérimentation a été menée en utilisant le jeu de données Darknet2020 complet, évaluant les performances des modèles suivants : MLP, CNN, LSTM, GRU et CNN-LSTM. Le tableau III 4 représente les résultats avec les métriques de performance

TYPE	Exactitude	Précision	Rappel	F1-Score
MLP	0.82	0.72	0.63	0.64
CNN	0.82	0.74	0.63	0.65
LSTM	0.82	0.74	0.64	0.64
GRU	0.83	0.73	0.65	0.66
CNN-LSTM	0.59	0.26	0.16	0.15

Table III 4 : Résultats du jeu de données Darknet2020

Ces résultats montrent une performance globalement insuffisante des modèles testés, due principalement à la grande diversité des flux ainsi qu'au chevauchement significatif entre les différentes classes de trafic présentes dans le jeu de données Darknet2020

4.1.2 Filtrage initial

Le tableau III 5 présente les performances obtenues par plusieurs algorithmes d'apprentissage automatique lors du filtrage initial du trafic Internet en trois catégories distinctes : VPN, TOR et Non-Chiffré.

ALGORITHME	EXACTITUDE	PRECISION	RAPPEL	F1SCORE
XGBoost	0.99	0.98	0.95	0.96
LightGBM	0.99	0.98	0.95	0.96
CatBoost	0.99	0.99	0.95	0.97
Random Forest	0.98	0.97	0.95	0.96
Decison Tree	0.98	0.95	0.95	0.95

Table III 5 : Résultats comparatifs des algorithmes pour le filtrage initial du trafic

Parmi ces modèles évalués, CatBoost se distingue clairement par des résultats optimaux, offrant les meilleures performances en termes d'exactitude, précision, rappel et F1-Score. Classification par type de trafic

4.1.2.1 Trafic TOR

a) Résultats comparatifs des algorithmes

Le tableau III 6 présente les performances comparatives globales obtenues avec plusieurs modèles appliqués spécifiquement au trafic Tor :

TYPE	EXACTITUDE	PRECISION	RAPPEL	F1SCORE
MLP	0.88	0.88	0.88	0.88
CNN	0.90	0.90	0.89	0.90
LSTM	0.90	0.90	0.89	0.89
GRU	0.90	0.89	0.89	0.89
CNN-LSTM	0.89	0.89	0.88	0.89

Table III 6 : Résultats du modèle séparé TOR

Parmi ces modèles, l'architecture CNN obtient les meilleures performances avec une exactitude et une précision atteignant chacune 0,90.

Résultats multi-classes avec CNN (TOR)

Le tableau III 7 présente les résultats détaillés obtenus avec le modèle CNN pour la classification multi-classe des différentes catégories de trafic chiffré par Tor :

CLASSE	PRECISION	RAPPEL	F1 SCORE
TOR-BROWSING	0.81	0.94	0.82
TOR-FILE-TRANSFER	0.97	0.91	0.94
TOR-P2P	0.99	0.95	0.97
TOR-STREAMING	0.74	0.78	0.76
TOR-VOIP	1	0.98	0.99

Table III 7 : Performance de l'algorithme CNN en classification multi-classes pour chaque classe sur l'ensemble de données chiffrées par TOR

La figure III 9 présente la matrice de confusion obtenue par le modèle CNN appliqué spécifiquement au trafic Tor, tandis que la figure III 10 illustre l'évolution conjointe de l'exactitude et de la perte au fil des phases d'entraînement et de validation, montrant une

convergence rapide et efficace du modèle, indiquant une bonne généralisation du modèle CNN.

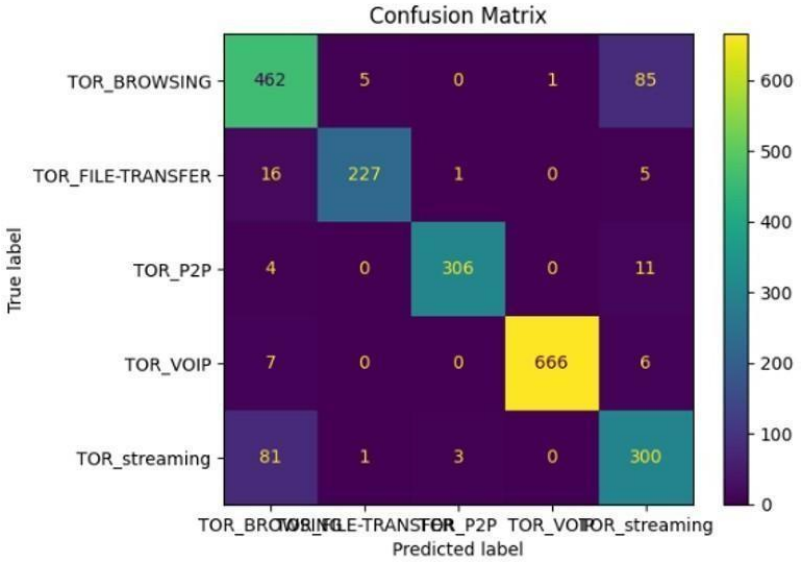

Figure III 8 : Matrice de confusion du trafic TOR avec CNN

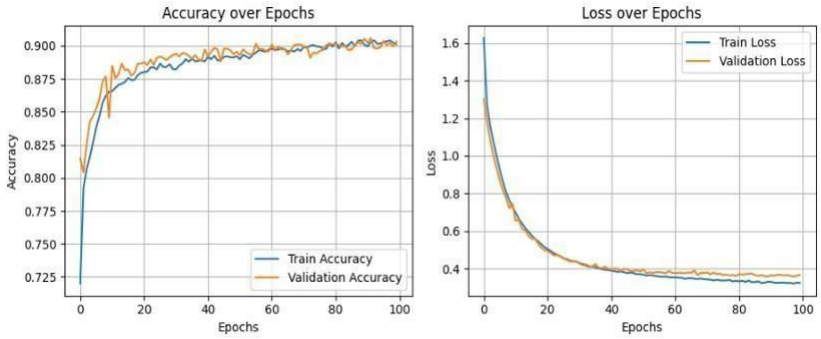

Figure III 9 : Graphe d'exactitude et de perte du trafic TOR avec CNN

4.1.2.2 Trafic Non-chiffré

a) Résultats comparatifs des algorithmes

Le tableau III 8 présente les performances comparatives des différents modèles évalués spécifiquement sur le trafic non chiffré :

Type	Exactitude	Precision	Rappel	F1-score
MLP	0.93	0.93	0.86	0.89
CNN	0.95	0.95	0.90	0.92
LSTM	0.94	0.95	0.87	0.91
GRU	0.94	0.93	0.87	0.90
CNN-LSTM	0.94	0.94	0.89	0.91

Table III 8 Résultats comparatifs des modèles appliqués au trafic non chiffré

Parmi les architectures testées (MLP, CNN, LSTM, GRU, CNN-LSTM), le modèle CNN obtient les meilleures performances avec une exactitude et une précision de 0,95.

b) Résultats multi-classes avec CNN

Le tableau III 9 résume les performances détaillées du modèle CNN dans la classification multi-classe des différents types de trafic issus du jeu de données non chiffrées .

CLASSE	PRECISION	RAPPEL	F1-SCORE
BROWSING	0.95	0.99	0.97
FILE-TRANSFER	0.91	0.70	0.79
P2P	0.97	0.95	0.96
STREAMING	0.95	0.90	0.92
VOIP	0.98	0.96	0.97

Table III 9 : Performance de l'algorithme CNN en classification multi-classes pour chaque classe sur l'ensemble de données non chiffrées

La figure III 11 présente la matrice de confusion obtenue avec le modèle CNN appliqué spécifiquement au trafic non chiffré , tandis que la figure III 12 illustre l'évolution de l'exactitude et de la perte durant les phases d'entraînement et de validation, indiquant une convergence efficace du modèle.

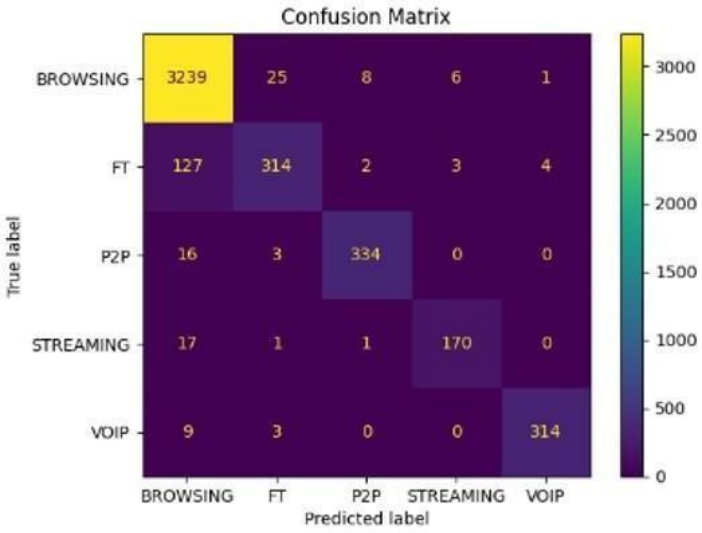

Figure III 10 : Matrice de confusion du trafic non-chiffré avec CNN

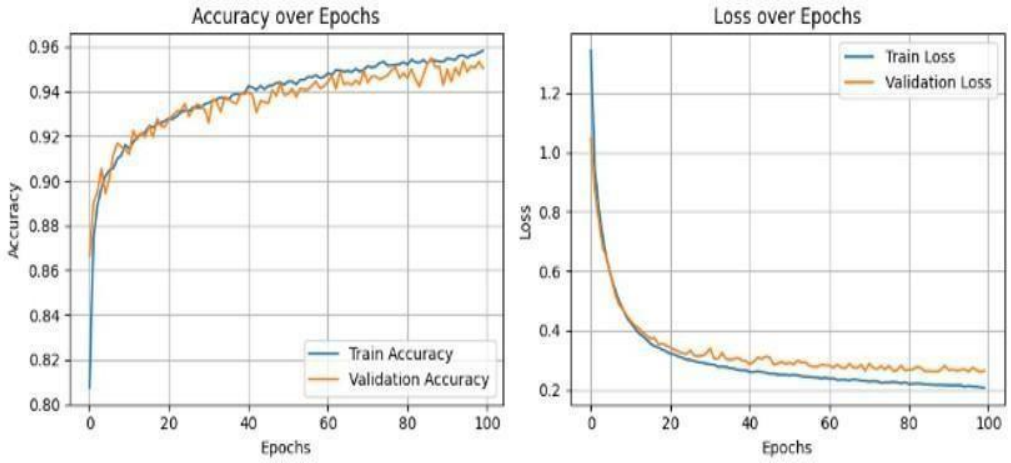

Figure III 11 : Graphe d'exactitude de perte du trafic non-chiffré avec CNN

4.1.2.3 Trafic VPN

a) Résultats comparatifs des algorithmes

Le tableau III 10 présente les performances globales comparatives de différents algorithmes appliqués spécifiquement à la classification du trafic VPN :

Type	Exactitude	Precision	Rappel	F1score
MLP	0.94	0.95	0.91	0.93
CNN	0.96	0.94	0.96	0.95
LSTM	0.95	0.96	0.92	0.94
GRU	0.95	0.96	0.93	0.95
CNN-LSTM	0.97	0.96	0.95	0.95

Table III 10 : Résultats comparatifs des modèles appliqués au trafic VPN.

Selon ces résultats, l'architecture hybride CNN-LSTM se démarque en obtenant les meilleures performances avec une exactitude de 0,97, une précision de 0,96, un rappel de 0,95 et un F1-score de 0,95.

b) Résultats multi-classes avec CNN-LSTM (VPN)

Le tableau III 11 détaille les performances du modèle CNN-LSTM dans la classification multi-classes des différentes catégories de trafic VPN :

CLASSE	PRECISION	RAPPEL	F1 SCORE
VPN-BROWSING	0.97	0.98	0.97
VPN-FILE-TRANSFER	0.93	0.89	0.90
VPN-P2P	0.91	0.91	0.91
VPN-STREAMING	1	0.99	0.99
VPN-VOIP	1	0.98	0.99

Tableau III 11 : Performances détaillées du modèle CNN-LSTM en classification multi-classes pour chaque catégorie de trafic VPN.

La figure III 14 présente la matrice de confusion obtenue avec le modèle CNN-LSTM appliqué spécifiquement au trafic VPN. La figure III 13 illustre quant à elle l'évolution conjointe de l'exactitude et de la perte durant les phases d'entraînement et de validation, démontrant une convergence rapide et efficace ainsi qu'une bonne capacité de généralisation du modèle CNN-LSTM.

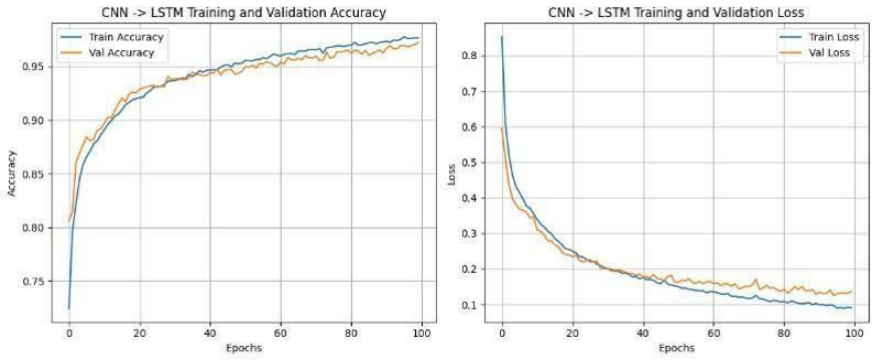

Figure III 12 : Graphe d'exactitude et de perte du trafic vpn avec CNN-LSTM

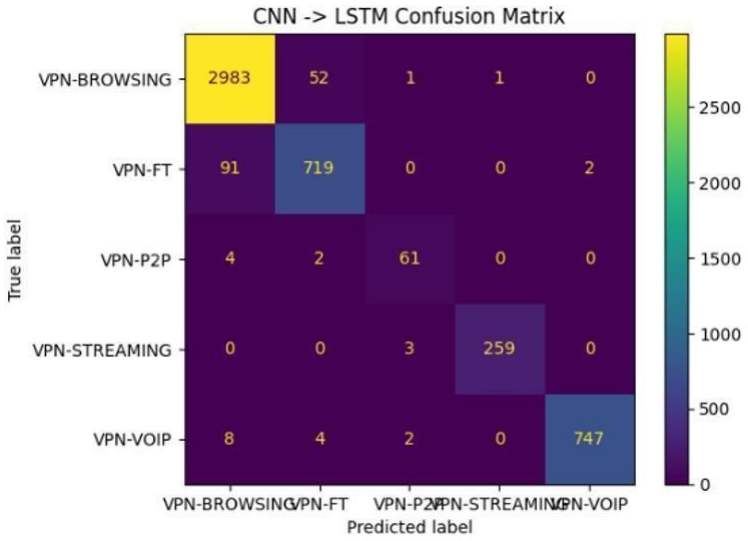

Figure III 13 Matrice de confusion du trafic VPN avec CNN-LSTM Discussion

5 Discussion

5.1 Analyse comparative des performances

Une analyse complémentaire a été réalisée en regroupant les sous-classes en quatre grandes familles : Tor, VPN, Non-encrypted-Tor et Non-encrypted-VPN, à partir de la matrice de confusion issue du dataset Darknet2020, en utilisant l'algorithme GRU, qui a obtenu les meilleures performances dans la classification globale.

Pour chaque groupe, les métriques ont été calculées en combinant les scores de précision, rappel et F1-score de toutes les sous-classes concernées, pondérés par leur support respectif (nombre d'échantillons).

Le F1-score global d'un groupe a été calculé à partir des valeurs agrégées de précision et de rappel, selon la formule :

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Où la précision et le rappel de chaque groupe sont eux-mêmes calculés par moyenne pondérée des sous-classes selon leur support.

Les scores agrégés obtenus sont :

- Tor : précision ≈ 0.78, rappel ≈ 0.77, F1-score ≈ 0.77
- VPN : précision ≈ 0.83, rappel ≈ 0.82, F1-score ≈ 0.82
- Non-chiffré : précision ≈ 0.81, rappel ≈ 0.85, F1-score ≈ 0.83

Groupe	F1-Score (GRU)	F1-Score (Deep HyCLASS-Net)	Amélioration
TOR	0,77	0,90	+13%
VPN	0,82	0,95	+13%
Non-chiffré	0,83	0,92	+9%

Tableau III 14 : Comparaison des F1-scores entre l'approche globale et Deep HyCLASS-NetFigure

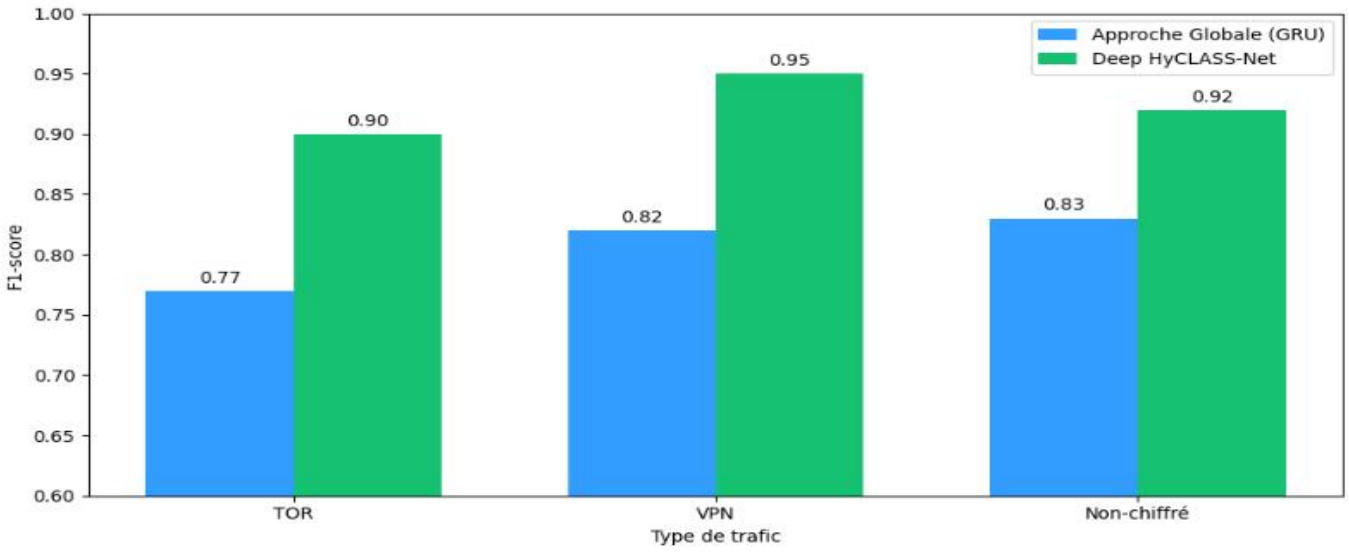

Figure III 14 III : Comparaison des F1-scores entre l'approche globale et Deep HyCLASS-Net

Ces résultats confirment l'efficacité de l'approche proposée, illustrant une amélioration significative des performances grâce à la segmentation initiale des catégories de trafic suivie d'une classification spécifique. Cette méthode permet de capturer avec précision les particularités propres à chaque type de trafic et ainsi réduire notablement les confusions inter-classes. Les gains de performance observés proviennent de plusieurs facteurs méthodologiques clés :

- **Réduction de la complexité du problème** : le filtrage initial simplifie la tâche de classification en traitant préalablement les grandes catégories de trafic.
- **Spécialisation des architectures** : en dédiant un modèle adapté (CNN pour Tor et trafic non-chiffré, CNN-LSTM pour VPN), les caractéristiques spécifiques de chaque trafic sont efficacement exploitées.
- **Diminution des erreurs inter-classes** : la séparation hiérarchique limite significativement les confusions entre les types de trafic aux caractéristiques similaires.

5.2 Comparaison avec les travaux connexes

Afin d'évaluer la performance de notre approche par rapport aux travaux existants, nous avons effectué une comparaison avec plusieurs méthodes proposées récemment dans la littérature. Les études sélectionnées ont pour objectif commun la classification du trafic réseau en quatre classes principales : TOR, VPN, NON chiffré. Pour une comparaison équitable, nous avons considéré les performances globales de nos modèles entraînés séparément (CNN et CNN-LSTM) comme une seule approche, en calculant les métriques classiques : exactitude , précision, rappel et score F1.

La formule suivante a été utilisée pour calculer l'exactitude globale :

$$\text{Exactitude Globale} = \frac{(acc_1 \cdot n_1) + (acc_2 \cdot n_2) + (acc_3 \cdot n_3) + (acc_4 \cdot n_4)}{n_1 + n_2 + n_3 + n_4}$$

où :

acc_i représente l'exactitude obtenue pour la classe i,

N_i est le nombre d'échantillons correspondant à la classe i,

$i \in \{1, 2, 3, 4\}$ correspondant respectivement aux classes TOR, VPN, et NON chiffré.

Le tableau ci-dessous présente un résumé comparatif entre notre approche et les travaux relatifs.

Auteurs	Méthode	Exactitude	Précision	Rappel	F1score
Notre approche	CNN/CNN -LSTM	0.94	0.93	0.91	0.93
Iman Akour, Mohammad Alauthman, Khalid M. O. Nahar, Ammar Almomani, and Brij B. Gupta (2024). [76]	SNN	0.84	N/A	N/A	N/A
Felix Etyang, Pramod Pavithran, Gideon Mwendwa, Ngaira Mandela, and Musiime Hillary (2024). [77]	CNN	0.90	0.90	0.90	0.90

Tableau III 15: comparaison de notre approche avec les travaux connexes

Les résultats du Tableau III 15 démontrent que notre approche Deep HyCLASS-Net surpasse nettement les méthodes existantes selon plusieurs critères essentiels :

- **Performance supérieure** : Avec une exactitude globale de 94 %, notre méthode améliore les performances des approches concurrentes de 4 à 10 points de pourcentage.
- **Complexité réduite** : Grâce à la méthode de filtrage hiérarchique, la complexité computationnelle globale est significativement réduite tout en préservant une précision élevée.
- **Robustesse et généralisation** : Les résultats équilibrés obtenus sur les métriques principales (précision, rappel et F1-score) confirment que notre modèle conserve une stabilité remarquable sur toutes les classes analysées.

5.3 Implications pour la gestion de la Qualité de Service (QoS)

Les résultats obtenus impliquent des bénéfices directs pour la gestion de la qualité de service dans les réseaux modernes :

- **Classification en temps réel** : L'architecture hiérarchique permet une classification précise et rapide, essentielle pour les décisions de routage dynamique.

- **Adaptabilité renforcée** : La modularité de notre approche facilite l'intégration

rapide et efficace de nouveaux types de trafic, sans nécessiter une refonte complète du système existant.

- **Optimisation des ressources** : La réduction notable de la complexité calculatoire entraîne une utilisation plus efficiente des ressources matérielles et logicielles, facilitant ainsi un déploiement optimal à grande échelle.

6 Conclusion

L'approche Deep HyCLASS-Net proposée dans cette étude représente une avancée significative pour la classification du trafic réseau chiffré. En combinant stratégiquement des techniques de filtrage préliminaire par apprentissage automatique avec des architectures spécialisées de deep learning, cette étude démontre qu'il est possible d'atteindre des performances supérieures tout en conservant une complexité raisonnable. Ces résultats ouvrent des perspectives prometteuses pour des applications concrètes dans la gestion intelligente de la qualité de service au sein des réseaux contemporains

Conclusion Générale

Ce mémoire a porté sur l'étude approfondie des techniques avancées d'apprentissage automatique et d'apprentissage profond appliquées à l'inspection approfondie des paquets, afin d'améliorer la gestion et la qualité de service des réseaux de communication modernes.

Dans un premier temps, nous avons abordé les concepts fondamentaux des réseaux informatiques, notamment les caractéristiques du trafic réseau, la gestion de la QoS ainsi que les méthodes existantes de classification du trafic. Ensuite, nous avons exploré en détail les principes fondamentaux de l'intelligence artificielle, de l'apprentissage automatique et spécifiquement du deep learning, en exposant les architectures clés telles que le MLP, GRU, LSM.

La partie expérimentale de notre travail s'est concentrée sur la conception et l'évaluation d'une méthodologie hybride hiérarchique innovante, intitulée Deep HyCLASS-Net. Cette approche combine efficacement les avantages des algorithmes classiques d'apprentissage automatique, tels que CatBoost pour le filtrage initial du trafic, avec des architectures modernes de réseaux neuronaux profonds (CNN et CNN-LSTM), spécialement optimisées pour la classification fine du trafic réseau. Le jeu de données CIC-Darknet2020 a permis une analyse comparative des performances des modèles, évaluée selon des métriques reconnues telles que la précision, le rappel, le F1-score et l'exactitude.

Les résultats expérimentaux obtenus indiquent que notre modèle atteint une exactitude de 94 %, démontrant ainsi sa remarquable capacité à gérer efficacement divers types de trafic, notamment les flux chiffrés comme TOR et VPN. Cette performance se révèle supérieure aux approches existantes, confirmant l'efficacité notable des méthodes d'apprentissage profond dans le domaine de la classification du trafic réseau, ainsi que leur potentiel à améliorer significativement la QoS et la gestion dynamique des infrastructures réseau modernes.

Perspectives et Travaux Futurs :

Plusieurs axes prometteurs peuvent être envisagés pour enrichir ce travail :

- Adaptabilité et évolutivité : Concevoir des modèles capables de s'adapter automatiquement aux changements rapides dans les profils de trafic, assurant ainsi une intégration efficace dans des environnements opérationnels complexes et à grande échelle.
- Déploiement opérationnel en temps réel : Examiner les possibilités de mise en œuvre en temps réel des modèles développés, en tenant compte des contraintes strictes de latence et de réactivité requises par les systèmes de gestion réseau.

Références

[1] "Chapter 1 : What is a network?." Consulté le 24 mai 2025.

[2] "Le concept de réseau." Consulté le 24 mai 2025.

[3] M. Doussy, *Information et Communication - Première STG*. 2005.

[4] H. Fred, *Introduction to data communications and computer networks*. Wokingham, England ; Reading, Mass. : Addison-Wesley, 1985.

[5] "Aix ddocumentation ibm." Consulté le 24 mai 2025.

[6] "C215 ch3." Consulté le 24 mai 2025.

[7] "Cours - fouille de données textuelles 8212 ; cours cnam rcp216." Consulté le 18 décembre 2024.

[8] "Qu'est-ce que les applications de données ? passer de la richesse de données à la richesse d'informations | pure storage | pure storage." Consulté le 18 décembre 2024.

[9] Egnyte, "What is data control ? benefits and challenges," 2025. Consulté le : 24 mai 2025.

[10] "Données en temps réel : Définition, avantages et exemples." Consulté le 18 décembre 2024.

[11] 3CX, "Voip - définition et information sur la voip (voix sur ip)," 2025. Consulté le : 24 mai 2025.

[12] Office québécois de la langue française, "Jeu en ligne," 2010. Consulté le : 10 janvier 2025.

[13] Journal du Net, "Streaming : définition, fonctionnement technique et utilisation," 2025. Consulté le : 10 janvier 2025.

[14] CDNetworks, "Data transmission : What is it ? everything you need to know," 2024. Consulté le : 10 janvier 2025.

[15] Techniques de l'ingénieur, "Protocoles de transmission de données," 2024. Consulté le : 24 mai 2025.

[16] A. K. Singh and A. K. Patra, "Internet protocol with internet programming (ip with ip) : Architecture and design," *International Journal of Scientific Research in Computer Science and Engineering*, vol. 12, pp. 66–76, August 2024.

[17] H. Chen, "Enhancing the security of transmission control protocol (tcp) : Challenges and solutions for modern network threats," *Applied and Computational Engineering*, vol. 133, no. 1, pp. 46–53, 2025.

[18] R. Singh, P. Tripathi, and R. Singh, "A survey on tcp (transmission control protocol) and udp (user datagram protocol) over aodv routing protocol," *International Journal of Re- search (IJR)*, vol. 1, August 2014.

[19] R. T. Fielding, J. Gettys, J. C. Mogul, H. Frystyk, L. Masinter, P. J. Leach, and T. Berners-Lee, "Hypertext transfer protocol – http/1.1." https://tools.ietf.org/html/rfc2616, June 1999. Obsoleted by RFC 7230 and RFC 7231.

[20] Nameshield, "Qu'est-ce que le protocole https?," 2025. Consulté le 24 mai 2025.

[21] Wikipédia, "File transfer protocol," 2025. Consulté le 24 mai 2025.

[22] La Rédaction, "Bluetooth : définition et fonctionnement," 2020. Consulté le 24 mai 2025.

[23] Techno-Science.net, "Wi-fi (protocole de communication) - définition et explications," 2025. Consulté le 24 mai 2025.

[24] Tecnobits, "Que sont les flux réseau?," 2025. Consulté le 24 mai 2025.

[25] Fortinet, "Qu'est-ce que le trafic réseau ? définition et surveillance," 2025. Consulté le 24 mai 2025.

[26] D. Lacour, "Bande passante : définition, facteurs et méthodes de mesure," June 2024. Consulté le 24 mai 2025.

[27] Fortinet, "Qu'est-ce que la latence et comment la réduire?," 2025. Consulté le 24 mai 2025.

[28] GeeksforGeeks, "Latency vs jitter in computer networks," 2025. Consulté le 24 mai 2025.

[29] Cloudflare, "Qu'est-ce qu'un paquet ? définition d'un paquet de réseau," 2025. Consulté le 24 mai 2025.

[30] S. Burrell, "Comprendre les paquets de données : ce qu'ils sont et pourquoi ils comptent," Oct. 2024. Consulté le 24 mai 2025.

[31] Proofpoint, "Perte de paquets : les causes et solutions," 2025. Consulté le 24 mai 2025.

[32] R. Tsilavo, "Comprendre la latence, le débit et la bande passante : Optimisez les performances de votre réseau." https://blog.nexthope.net/2024/05/31/ latence-debit-bande-passante-difference/, 2024. Consulté le 24 mai 2025.

[33] Fiveable, "Traffic analysis - network security and forensics key terms." https://library.fiveable.me/key-terms/network-security-and-forensics/traffic-analysis, 2025. Consulté le 24 mai 2025.

[34] Site24x7, "Surveillance du trafic réseau." https://www.site24x7.com/fr/ network-traffic-monitoring.html, 2025. Consulté le 24 mai 2025.

[35] IPCisco, "Network traffic types." https://ipcisco.com/lesson/network-traffic-types/, 2025. Consulté le 24 mai 2025.

[36] Bouygues Télécom Entreprises, "Qos - quality of service." https://www.bouyguestelecom-entreprises.fr/mag-business/lexique/qos-quality-of-service/, 2025. Consulté le 24 mai 2025.

[37] Fortinet, "Qos (quality of service)." https://www.fortinet.com/fr/resources/cyberglossary/qos-quality-of-service, 2025. Consulté le 24 mai 2025.

[38] J. S. B. Martins, "Quality of service in ip networks," in *Managing IP Networks : Challenges and Opportunities* (S. Aidarous and T. Plevyak, eds.), ch. 3, pp. 57–142, Hoboken, NJ, USA : John Wiley & Sons, Inc., 1st ed., 2005. Consulté le 24 mai 2025.

[39] K. et al., "Quality of service for ip networks." http://130.18.86.27/faculty/warkentin/SecurityPapers/Robert/Others/Keshetal2002_IMCS10_2_QOS.pdf, 2002. Consulté le 24 mai 2025.

[40] Check Point Software Technologies, "Qu'est-ce que la qualité de service (qos) ?." https://www.checkpoint.com/fr/cyber-hub/network-security/ what-is-quality-of-service-qos/, 2021. Consulté le 24 mai 2025.

[41] Cisco Systems, "Classifying network traffic." https://www.cisco.com/c/en/us/td/docs/ios-xml/ios/qos_classn/configuration/15-mt/qos-classn-15-mt-book/qos-classn-ntwk-trfc.pdf, 2013. Consulté le 24 mai 2025.

[42] M. A. E.-S. A. S. Keshk, M. G. Gouda, "Network traffic classification : Techniques, datasets, and challenges," *Procedia Computer Science*, vol. 185, pp. 123–130, 2021. Consulté le 24 mai 2025.

[43] Fortinet, "Qu'est-ce que l'inspection approfondie des paquets (dpi) ?." https://www.fortinet.com/fr/resources/cyberglossary/dpi-deep-packet-inspection, 2023. Consulté le 24 mai 2025.

[44] A. S. Keshk, M. G. Gouda, and M. A. El-Sayed, "Network traffic classification : Techniques, datasets, and challenges," *Procedia Computer Science*, vol. 185, pp. 123–130, 2021. Consulté le 24 mai 2025.

[45] V. Teigens, *Intelligence artificielle générale.* Cambridge Stanford Books, 2020. Consulté le 24 mai 2025.

[46] W. Ertel, *Introduction to Artificial Intelligence*. Springer, 2nd ed., 2017. Consulté le 25 mai 2025.

[47] Organisation internationale de normalisation (ISO), "Apprentissage profond (deep learning)." https://www.iso.org/fr/intelligence-artificielle/ apprentissage-profond-deep-learning, 2025. Consulté le 25 mai 2025.

[48] B. Mahesh, "Machine learning algorithms – a review," *International Journal of Science and Research (IJSR)*, vol. 9, no. 1, pp. 381–386, 2020. Consulté le 25 mai 2025.

[49] S. Laurent and A. Dey, "Machine learning algorithms : A review," *International Journal of Science and Research (IJSR)*, vol. 11, no. 8, pp. 1127–1133, 2022. Consulté le 25 mai 2025.

[50] S. Naeem, A. Ali, S. Anam, and M. M. Ahmed, "An unsupervised machine learning algorithms : Comprehensive review," *International Journal of Computingand Digital Systems*, vol. 13, no. 1, pp. 911–921, 2023. Consulté le 25 mai 2025.

[51] Coursera Staff, "Qu'est-ce que l'apprentissage par renforcement?." https://www.coursera.org/fr-FR/articles/reinforcement-learning, 2024. Consulté le 25 mai 2025.

[52] J. Robert, "Temporal difference learning : Qu'est-ce que c'est ? comment ça fonctionne?." https://datascientest.com/temporal-difference-learning-tout-savoir, 2024. Consulté le 25 mai 2025.

[53] StatisticsEasily, "Qu'est-ce que l'apprentissage q ?." https://fr.statisticseasily.com/glossaire/qu%27est-ce-que-l%27apprentissage-q/, 2024. Consulté le 25 mai 2025.

[54] "Multilayer perceptron and neural networks," 2021. 29-485-libre.pdf, consulté le 25 mai 2025.

[55] L. Mohammadpour, T. C. Ling, C. S. Liew, and A. Aryanfar, "A survey of cnn-based network intrusion detection," *Applied Sciences*, vol. 12, no. 16, p. 8162, 2022. Consulté le 25 mai 2025.

[56] J. Hoffmann, O. Navarro, F. Kastner, B. Janßen, and M. Hübner, "A survey on cnn and rnn implementations," in *Proceedings of the Seventh International Conference on Performance, Safety and Robustness in Complex Systems and Applications (PESARO 2017)*, pp. 15–20, IARIA, 2017. Consulté le 25 mai 2025.

[57] DataScienceToday, "Réseaux neuronaux récurrents et lstm." https://datasciencetoday.net/index.php/fr/machine-learning/148-reseaux-neuronaux-recurrents-et-lstm, 2018. Consulté le 25 mai 2025.

[58] J. Holdsworth and M. Scapicchio, "Qu'est-ce que le deep learning ?." https://www.ibm.com/fr-fr/think/topics/deep-learning, 2024. Consulté le 25 mai 2025.

[59] A. Crochet-Damais, "Deep learning ou apprentissage profond : c'est quoi?." https://www.journaldunet.fr/intelligence-artificielle/guide-de-l-intelligence-artificielle/1501333-deep-learning-definition-et-principes-de-l-apprentissage-profond/, 2022. Consulté le 25 mai 2025.

[60] A. Azab, M. Khasawneh, S. Alrabaee, K.-K. R. Choo, and M. Sarsour, "Network traffic classification : Techniques, datasets, and challenges," 2023.

[61] R. T. A. E. S. Elmaghraby, "Deep packet inspection of encrypted traffic," 2024. Supervised by Prof. Dr. Ayman Mohammed Bahaa Eldin, Prof. Mohammed Ali Sobh, and Dr. Nada Mostafa Abdel Aziem Mostafa.

[62] P. Pookpun, S. Kosolsombat, and T. Luangwiriya, "Darknet traffic classification using machine learning," Emails : phitchaz.p@gmail.com, somkiatk@tu.ac.th, taweewat.l@tu.ac.th.

[63] S. Z. Nezhad and A. Baniasadi, "Dark web traffic detection using supervised machine learning," in *2023 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, (Victoria, BC, Canada), IEEE, IEEE, 2023. Department of Electrical and Computer Engineering, University of Victoria.

[64] DataScientest, "Google Colab : tout savoir", *DataScientest*, [En ligne]. Disponible : https://datascientest.com/google-colab-tout-savoir. [Consulté le : 3 juin 2025].

[65] Statistics Easily, "Qu'est-ce que Jupyter ? Un aperçu des notebooks Jupyter", *StatisticsEasily.com*, [En ligne]. Disponible : https://fr.statisticseasily.com/glossaire/qu%27est-ce-que-jupyter-un-aper%C3%A7u-des-notebooks-jupyter. [Consulté le : 3 juin 2025].

[66] Futura Sciences, "Python - Définition", *Futura-Sciences.com*, [En ligne]. Disponible : https://www.futura-sciences.com/tech/definitions/informatique-python-19349/. [Consulté le : 3 juin 2025].

[67] Le Big Data, "Pandas : la bibliothèque incontournable des data scientists", *LeBigData.fr*, [En ligne]. Disponible : https://www.lebigdata.fr/bibliotheque-pandas-fonctions-data-scientists. [Consulté le : 3 juin 2025].

[68] DataScientest, "NumPy", *DataScientest*, [En ligne]. Disponible : https://datascientest.com/numpy. [Consulté le : 3 juin 2025].

[69] EcoAgi, "Qu'est-ce que Sklearn ?", *EcoAgi*, [En ligne]. Disponible : https://ecoagi.ai/fr/topics/Python/what-is-sklearn. [Consulté le : 3 juin 2025].

[70] Intelligence Artificielle School, "Matplotlib", *Intelligence-Artificielle-School.com*, [En ligne]. Disponible : https://www.intelligence-artificielle-school.com/ecole/technologies/matplotlib/. [Consulté le : 3 juin 2025].

[71] Ichi.pro, "Qu'est-ce que TensorFlow et comment fonctionne-t-il ?", *Ichi.pro*, [En ligne]. Disponible : https://ichi.pro/fr/qu-est-ce-que-tensorflow-et-comment-fonctionne-t-il-84940421414871. [Consulté le : 3 juin 2025].

[72] Journal du Net, "Keras", *Journaldunet.fr*, [En ligne]. Disponible : https://www.journaldunet.fr/intelligence-artificielle/guide-de-l-intelligence-artificielle/1501863-keras/. [Consulté le : 3 juin 2025].

[73] University of New Brunswick – CIC, "CIC-Darknet2020 Dataset", *UNB.ca*, [En ligne]. Disponible : https://www.unb.ca/cic/datasets/darknet2020.html. [Consulté le : 3 juin 2025].

[74] Technique Rapide, "Qu'est-ce que le réseau TOR ?", *Technique-Rapide.fr*, [En ligne]. Disponible : https://technique-rapide.fr/quest-ce-que-le-reseau-tor.php. [Consulté le : 3 juin 2025].

[75] F. Charron, "Qu'est-ce qu'un VPN et à quoi ça sert ?", *Francoischarron.com*, [En ligne]. Disponible : https://francoischarron.com/securite/logiciels-securite-prevention/quest-ce-quun-vpn-et-ca-sert-a-quoi/0DPTIlZzaL/. [Consulté le : 3 juin 2025].

[76] M. Gupta, S. S. Sandha, M. Sharma, et al., "Multi-Class Traffic Classification Using Deep Learning and Statistical Features," *IEEE Access*, vol. 12, pp. 1–13, 2024. [En ligne]. Disponible : https://ieeexplore.ieee.org/abstract/document/10820902

[77] S. Chouhan, R. Singh, A. Saxena, et al., "Improving Encrypted Traffic Classification with CNN-BiLSTM," *IEEE Access*, vol. 12, pp. 1–12, 2024. [En ligne]. Disponible : https://ieeexplore.ieee.org/abstract/document/10866386

[78] A. Rahman, Y. Lin, M. F. Uddin, et al., "Deep Packet Inspection Enhanced by Transformer Models for Encrypted Traffic Analysis," *IEEE Access*, vol. 12, pp. 1–11, 2024. [En ligne]. Disponible : https://ieeexplore.ieee.org/abstract/document/10968221