الجـــــــــــــــــــــمهـوريـــــة الجـــــــــــــزائـريـة الديمـــــــقـراطيـة الشـــــــــعبيـة

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

وزارة التـــــــــــــــــــــــــــعـليـم العـــــــالي و البـــــــــحث العـــــلـمـــي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

جـــــــــــــامعة أبي بكـــر بـلـقـايد– تـلمســـــــــــــان –

Université Aboubakr Belkaïd– Tlemcen –

Faculté de TECHNOLOGIE

MEMOIRE

Présenté pour l'obtention du **diplôme** de **MASTER**

En : Télécommunications

Spécialité : Réseaux et Télécommunications

Par : Belamri Nour Djihan

Sujet

Modèles d'apprentissage automatique supervisé pour l'identification du trafic du darknet

Soutenu publiquement, le 12/06/2024, devant le jury composé de :

Mr. MARZOUGUI Rachid		Université de Tlemcen	Président
Md. TALEB Sara		Université de Tlemcen	Examinateur
Mr. MOUSSAOUI Djilali		Université de Tlemcen	Encadreur
Md. FERHI Wafaa	Doctorante	Université de Tlemcen	Co-Encadreur

Année universitaire : 2023/2024

وَمَا أُوتِيتُم مِّنَ الْعِلْمِ إِلَّا قَلِيلًا

[سورة الإسراء، الآية 85]

Remerciements

Après avoir remercié ALLAH Tout-Puissant et Bienveillant, nous souhaitons exprimer notre sincère gratitude à tous ceux qui ont participé à la réalisation de cette thèse.

Nous tenons également à remercier notre superviseur, M. Moussaoui Djilali, pour son soutien et son orientation inébranlables tout au long de cette recherche. Son expertise et ses connaissances ont été inestimables pour façonner la direction de cette thèse.

Nous sommes également reconnaissants à Mme Ferhi Wafaa, notre coprésidente, pour ses contributions à ce travail. Ses contributions et ses commentaires ont contribué à nous aider à perfectionner et à améliorer notre recherche.

Enfin, nous tenons à remercier les membres du jury pour leur intérêt pour notre recherche.

Dédicace

Je remercie ALLAH le tout-puissant de m'avoir accordé, ainsi que mon amie Zahira, l'opportunité de compléter ce mémoire.

Je tiens c'est avec grande plaisir que je dédie ce modeste travail

A mes chers parents ma mère et mon père pour leur patience, leur amour, leur soutien et leurs encouragements pendants tout mon parcours universitaire.

A l'âme de ma cher, grand-mère, Fatma qui elle signifié et continuent de représenter tant de choses pour moi, bien qu'ils ne soient plus de ce monde, leurs souvenirs continuent de faire partie de ma vie. Qu'Allah leurs accorde son vaste paradis.

A ma chère famille Kawther, Latifa, Souhila, Mohammed et tante Hafida qui m'ont toujours encouragé, et à qui je souhaite plus de succès

NOUR DJIHAN

Résumé

Le darknet fait référence à un réseau caché d'ordinateurs qui opère sur Internet et qui n'est pas accessible via les navigateurs web traditionnels. Ce réseau est souvent utilisé pour des activités illégales, telles que la vente de drogues, d'armes et de matériel pédopornographique. L'identification du trafic du darknet est une tâche importante pour les forces de l'ordre et les agences de sécurité, car elle peut aider à perturber ces activités illégales. L'apprentissage automatique supervisé est une branche de l'apprentissage automatique qui utilise des algorithmes pour apprendre à partir d'exemples étiquetés. Dans le contexte de l'identification du trafic du darknet, les exemples étiquetés peuvent être des paquets réseau ou des flux de trafic étiquetés comme étant légitimes ou illégaux. Les algorithmes d'apprentissage automatique supervisé peuvent ensuite être utilisés pour apprendre à identifier de nouveaux paquets ou flux de trafic comme étant légitimes ou illégaux. Les modèles d'apprentissage automatique supervisé sont un outil prometteur pour l'identification du trafic du darknet. Cependant, il existe un certain nombre de défis qui doivent être résolus avant que ces modèles puissent être déployés à grande échelle.

Mots clé : Apprentissage automatique supervisé, Darknet, Trafic illégal, Sécurité informatique, Analyse de réseau.

Abstract

The darknet refers to a hidden network of computers operating on the Internet that is not accessible via traditional web browsers. This network is often used for illegal activities, such as the sale of drugs, weapons and child pornography. Identifying darknet traffic is an important task for law enforcement and security agencies, as it can help disrupt these illegal activities. Supervised machine learning is a branch of machine learning that uses algorithms to learn from labelled examples. In the context of identifying darknet traffic, the labelled examples can be network packets or traffic streams labelling as legitimate or illegal. Supervised machine learning algorithms can then be used to learn how to identify new packages or traffic streams as legitimate or illegal. Supervised machine learning models are a promising tool for identifying darknet traffic. However, there are a number of challenges that need to be addressed before these models can be deployed on a large scale.

Keywords: Supervised machine learning, Darknet, Illegal Trafficking, Computer Security, Network analysis.

الملخص

يشير الإنترنت المظلم إلى شبكة مخفية من أجهزة الكمبيوتر التي تعمل على الإنترنت والتي لا يمكن الوصول إليها عبر متصفحات الويب التقليدية. غالباً ما تُستخدم هذه الشبكة في الأنشطة غير القانونية، مثل بيع المخدرات، الأسلحة، والمواد الإباحية للأطفال. يُعد التعرف على حركة المرور في الإنترنت المظلم مهمة مهمة لوكالات إنفاذ القانون والأمن، حيث يمكن أن يساعد في تعطيل هذه الأنشطة غير القانونية. التعلم الآلي المُشرف هو فرع من فروع التعلم الآلي الذي يستخدم الخوارزميات للتعلم من الأمثلة المسمّاة. في سياق التعرف على حركة المرور في الإنترنت المظلم، يمكن أن تكون الأمثلة المسمّاة هي حزم الشبكة أو تدفقات المرور المصنفة على أنها شرعية أو غير قانونية. يمكن بعد ذلك استخدام خوارزميات التعلم الآلي المُشرف للتعلم كيفية التعرف على الحزم الجديدة أو تدفقات المرور على أنها شرعية أو غير قانونية. تُعد نماذج التعلم الآلي المُشرف أداة واعدة للتعرف على حركة المرور في الإنترنت المظلم. ومع ذلك، هناك عدد من التحديات التي يجب معالجتها قبل أن يتم نشر هذه النماذج على نطاق واسع

الكلمات المفتاحية: التعلم الآلي المُشرف، الإنترنت المظلم، الاتجار غير القانوني، أمن الحاسوب، تحليل الشبكات.

TABLE DES MATIERES

LISTE DES FIGURES

LISTE DES TABLEAUX

Acronymes	
ANN	**Artificial Neural Network**
ARP	**Address resolution protocol**
CIC	**Canadian Institute for Cybersecurity**
CNN	**Convolution Neural Networks**
DNN	**Deep Neural Network**
DDoS	**Distributed Denial of Service**
DNS	**Domain Name System**
DoS	**Denial of Service**
DRDOS	**Distributed Reflection Denial of Servise**
FBI	**Federal Bureau of Investigation**
IA	**Intelligence Artificielle**
KNN	**K-Nearest Neighbour**
MLP	**Multilayer perception**
NRL	**Naval research laboratory**
RBF	**Réseau à function radiale**
RNA	**Artificielle neural networks**
TDG	**Traffic Dispersion Graphs**

Introduction Générale

Dans un monde de plus en plus connecté, la cyber sécurité est devenue une préoccupation majeure pour les individus, les entreprises et les gouvernements. Parmi les défis les plus complexes et les plus urgents se trouve la surveillance et l'analyse du darknet, une partie de l'internet souvent associée à des activités illicites telles que le trafic de drogue, la fraude et le commerce de données volées. La nature anonyme et cryptée du darknet rend la détection et l'analyse de son trafic particulièrement difficile, nécessitant des approches sophistiquées et innovantes. Les avancées récentes en intelligence artificielle (IA) et en apprentissage automatique ont ouvert de nouvelles perspectives pour la détection et l'identification du trafic du darknet. L'apprentissage automatique, et plus particulièrement l'apprentissage supervisé, offre des outils puissants pour analyser de vastes ensembles de données et identifier des schémas de comportement anormaux ou suspects. Grâce à des algorithmes capables d'apprendre et de s'adapter à partir de données étiquetées, il est possible de développer des systèmes capables de détecter automatiquement le trafic malveillant en temps réel. Ce mémoire explore l'application des modèles d'apprentissage automatique supervisé pour l'identification du trafic du darknet. Nous examinerons les différentes techniques de prétraitement des données, les méthodes de sélection des caractéristiques pertinentes, et les algorithmes de classification les plus efficaces pour cette tâche. En outre, nous évaluerons les performances de ces modèles à l'aide de métriques standard et discuterons des défis et des opportunités liées à leur déploiement dans des environnements réels. L'objectif de ce travail est de contribuer à l'amélioration des méthodes de détection du trafic du darknet, en proposant des approches basées sur l'apprentissage automatique qui soient à la fois robustes, précises et rapides. En fin de compte, notre recherche vise à renforcer la cyber sécurité et à fournir des outils pratiques pour aider les experts à surveiller et à sécuriser les réseaux contre les activités malveillantes sur le darknet. Ce mémoire se présente sous forme de trois chapitres :

- ✓ Le premier chapitre : L’intelligence artificiel et le Deep Learning
- ✓ Ensuite Le chapitre 2 : Darknet
- ✓ Le chapitre 3 : Résultat de dataset

Nous terminerons ce mémoire par une conclusion générale.

Chapitre 01

L’intelligence Artificial et Le Deep Learning

1.1 Introduction

L'avènement d'Internet a créé un lien hyperactif entre le monde entier. Cela implique que chaque objet, y compris les réseaux sociaux, les maisons et les voitures, génère des millions de données supplémentaires quotidiennement. Ces informations sont incluses dans un océan de données sans fin, qui peut être utilisé pour offrir des services personnalisés et immédiats. Toutefois, la question clé est de savoir comment transformer cet océan de données en flux régulier. La technologie de l'intelligence artificielle (IA) est la solution à cette question. Il s'agit d'une technologie qui permet aux machines d'apprendre et d'utiliser des techniques pour s'adapter à de nouvelles circonstances sans intervention humaine directe. L'IA fait usage des algorithmes sophistiqués pour extraire des données utiles et prédire des tendances et faire des choix judicieux. De nombreux domaines peuvent utiliser l'intelligence artificielle à savoir. les secteurs de la médecine, de la finance, de la production industrielle, du transport et de la sécurité, etc. L'IA permet le traitement de grandes quantités de données complexes, cela ne serait pas possible pour un être humain. Cependant, l'IA a des inconvénients et une limite. Par conséquent, il est crucial de trouver un équilibre entre les avantages et les inconvénients [1].

1.2 Intelligence Artificielle

1.2.1 Définition de l'intelligence artificielle

L'intelligence artificielle est une technologie qui permet aux machines (ordinateurs et programmes informatiques) d'exécuter des tâches qui nécessitent généralement une forme d'intelligence humaine. De nos jours, les ordinateurs sont plus capables que le cerveau humain de calculer, de comprendre, d'adapter, de communiquer et d'apprendre profondément. L'intelligence artificielle comprend également des dispositifs informatiques ou robotiques ainsi qu'un ensemble d'algorithmes pour la prise de décision ou la résolution de problèmes. L'IA peut donc être définie comme :

- Un domaine de l'informatique qui vise à créer une technologie comparable à l'intelligence humaine.
- Un ensemble de méthodes destinées à permettre aux machines de reproduire une forme d'intelligence réelle.
- L'ensemble des théories et des algorithmes utilisés pour construire des machines capables de reproduire l'intelligence humaine [2].

1.2.2 Principe de l'intelligence artificielle

Même si nous ne nous en rendons pas toujours compte, l'intelligence artificielle est présente dans notre vie quotidienne. Par exemple, nous l'utilisons pour commander des applications vocales, jouer à des jeux sur ordinateur ou utiliser des outils de domotique. Le but de l'intelligence artificielle est de reproduire le fonctionnement du cerveau humain sur les machines en utilisant un ensemble d'algorithmes d'apprentissage qui permettent à ces machines d'acquérir une forme d'intelligence [3].

1.2.3 Utilisation de l'intelligence artificielle

À travers diverses applications, l'intelligence artificielle an effectivement un impact significatif sur notre quotidien. De plus, nous pouvons citer d'autres exemples d'utilisation de l'IA dans notre vie quotidienne, tels que :

a. Les assistants vocaux : Siri d'Apple, Google Assistant et Amazon Alexa, utilisent l'IA pour faciliter l'interaction homme-machine.

b. Les systèmes de recommandation : Les systèmes de recommandation, comme ceux utilisés par Netflix, Amazon et Spotify analysent les habitudes de consommation des utilisateurs et proposent des biens ou des services pertinents.

c. Les voitures autonomes : Elles sont une illustration de l'utilisation de l'IA pour rendre la conduite plus efficace et plus sûre grâce à l'utilisation de capteurs et d'algorithmes pour prendre des décisions en temps réel et identifier les obstacles.

d. Les robots domestiques : Les robots pour la maison comprennent des aspirateurs robots et les assistants de cuisine utilisent l'IA pour comprendre leur environnement et accomplir des tâches particulières de manière autonome.

e. Les systèmes de surveillance : Les caméras de surveillance les systèmes de sécurité et de détection d'intrusion utilisent IA pour détecter les comportements des utilisateurs suspects et les alerter en temps réel [4].

1.2.4 Domaines d'application de l'intelligence artificielle

Il est fascinant de constater que l'IA an une variété d'applications dans divers domaines de la société contemporaine. Par exemple, de nombreux systèmes basés sur l'IA peuvent diagnostiquer des maladies, y compris certains types de cancer, avec une précision équivalente, voire supérieure, à celle des experts. L'IA peut également contribuer à une détection précoce des maladies et une mise sur le marché rapide de nouveaux produits pharmaceutiques.

a. Le secteur bancaire et financier

L'intelligence artificielle crée des agents conversationnels capables de répondre aux questions des clients en utilisant des milliers de conversations enregistrées et analysées. Elle est capable d'aider les conseillers à prendre des décisions concernant l'octroi de prêts ou gestion des e-mails urgents des clients et facilitation de la détection des opérations frauduleuses qui deviennent de plus en plus complexes.

b. Le secteur des transports

L'IA est largement employée dans le domaine des transports, en particulier dans les voitures autonomes qui ont des fonctionnalités d'assistance à la conduite basées sur l'IA. De plus, elle permet d'optimiser les applications de gestion du trafic et de réduire les temps d'attente, les émissions et la consommation d'énergie. Afin de faire face aux défis de la main-d'œuvre que rencontrent les supermarchés, les robots dans le commerce sont capables d'effectuer plusieurs tâches, telles que la vérification de l'inventaire ou la surveillance des étiquettes de prix.

c. L'industrie

L'IA est essentielle à la robotique contemporaine. En réduisant les coûts de production et les défaillances et en garantissant une cadence de production élevée, elle permet d'optimiser les performances industrielles. Enfin, dans l'agriculture, l'IA peut aider les agriculteurs à surveiller leurs champs grâce à l'utilisation de capteurs et de techniques d'apprentissage automatique, à détecter les ravageurs sur les feuilles et à augmenter les rendements en se basant sur divers paramètres tels que le climat, l'état du sol et les niveaux d'irrigation.

d. Les médias

L'intelligence artificielle est utilisée pour analyser le contenu multimédia audiovisuel, y compris les films, les programmes télévisés, les vidéos publicitaires et le contenu créé par les utilisateurs. Les solutions impliquent souvent l'utilisation de la vision par l'ordinateur pour l'analyse d'images à l'aide de techniques de reconnaissance d'objets ou de reconnaissance faciale, ou l'analyse de vidéos pour la reconnaissance de scènes. L'utilisation de l'analyse des médias basée sur l'IA a de nombreux avantages, y compris la facilité de recherche multimédia ou créer du contenu spécifique aux utilisateurs [5].

1.2.5 Avantages et inconvénients de l'intelligence artificielle

L'intelligence artificielle présente de nombreux avantages, principalement dans le domaine du travail en raison de ses hautes performances fascinantes. Cependant, il existe toujours certains inconvénients et limites.

1.2.5.1 Les avantages de l'intelligence artificielle

Parmi les avantages de l'intelligence artificielle :

- Elle est capable de remplacer l'homme dans des tâches difficiles et dangereuses sans limites physique.
- En permettant des calculs plus rapides et efficaces, les erreurs humaines sont réduites.
- Elle permet aux véhicules autonomes de se déplacer plus facilement.
- Elle a des avantages en médecine, comme le suivi à distance des patients et les traitements individualisés.
- Les machines fonctionnent sans interruption et ont moins besoin de pauses et de rafraîchissements que les humains.
- En permettant aux joueurs de défier des adversaires plus forts et plus expérimentés, elle améliore l'expérience de jeu.

1.2.5.2 Les inconvénients de l'intelligence artificielle

L'intelligence artificielle (IA) présente certains inconvénients, tels que :

- **Risque d'erreurs de programmation** : L'IA utilise des algorithmes complexes qui peuvent inclure des erreurs de programmation. Ces erreurs peuvent avoir des répercussions catastrophiques, en particulier lorsque l'IA est utilisée dans des domaines tels que la santé, la finance, les transports ou la sécurité.
- **Perte d'emplois :** L'automatisation accrue par l'IA peut entraîner la suppression d'emplois pour les travailleurs et la substitution des machines aux travailleurs. Cela

pourrait entraîner une augmentation du taux de chômage et une insécurité économique pour les employés.

- **Coût élevé de développement :** Il faut beaucoup d'argent pour développer l'IA. Les gouvernements et les entreprises doivent dépenser beaucoup d'argent pour la recherche, le développement et la mise en œuvre de l'IA. Cela pourrait rendre l'IA difficile pour les petites entreprises et les pays moins développés.
- **Confidentialité et sécurité des données :** Pour apprendre et s'améliorer, les systèmes de l'IA nécessitent beaucoup de données. Cependant, il existe des inquiétudes quant à la confidentialité et à la sécurité de la collecte et de l'utilisation de ces données. Les violations de données peuvent nuire aux entreprises et à la vie privée des individus [6].

1.2.6 Les limites de l'intelligence artificielle

Malgré son potentiel, l'intelligence artificielle a des limites. Trois limites principales sont connues:

1.2.6.1 Les limites matérielles Bien que l'IA utilise des transistors pour communiquer rapidement, le cerveau humain contient beaucoup plus de neurones que l'IA, bien que la communication soit plus lente.

1.2.6.2 Les limites émotionnelles Contrairement aux machines, les êtres humains ont des sentiments et des émotions difficiles à reproduire car ils sont liés à la nature plutôt qu'à la programmation.

1.2.6.3 Les limites cognitives Bien que les ordinateurs développent de plus en plus de capacités de traitement des données, leur système binaire a des lacunes par rapport au cerveau humain, qui utilise un réseau complexe et étendu de neurones pour la pensée [7].

1.3. Approches principales de l'IA

Nous pouvons citer deux principales approches de l'intelligence artificielle : programmation symbolique et apprentissage machine .

1.3.1. Programmation symbolique :

Elle consiste à coder les problèmes à l'aide d'un ensemble d'instructions qui doivent être vérifiées afin de donner une décision. Ce type de solutions est simple à créer mais difficile à généraliser car il peut être utilisé pour des problèmes simples mais pas pour des problèmes complexes. Par exemple, si une image contient une queue en plus de quatre pattes, on peut conclure qu'il s'agit d'un animal. Par contre, pour déterminer la race de l'animal, il serait impossible d'analyser tous les pixels de l'image, ce qui est irréalisable en termes de nombre de possibilités.

1.3.2 Apprentissage machine « Machine Learning » :

Nous pouvons utiliser les techniques d'apprentissage automatique qui s'appuient sur l'apprentissage de données plutôt que sur des règles pour faire face au problème de la programmation symbolique. Au contraire, les algorithmes de « machine learning » repèrent des modèles dans les données avant de prendre les meilleures décisions en utilisant des techniques statistiques issues d'exemples fournis. Les algorithmes d'apprentissage sont généralement basés

sur trois étapes : apprentissage, validation et test.

Entrainement : Le modèle sera testé et modifié après chaque itération pour obtenir le meilleur résultat à partir d'un ensemble de données suffisamment grand.

Validation : Le modèle est testé avec un jeu de données différent après chaque itération. Cela évitera que le modèle soit entrainé de manière aveugle sur les données d'apprentissage, ce qui entraînerait des problèmes de sous-apprentissage (underfitting) et de sur apprentissage (overfitting).

Test : Après la génération du modèle, un nouveau jeu de test peut être utilisé pour le tester. Des résultats satisfaisants prouvent que la stratégie suggérée est efficace et vice versa. Les étapes de collecte et de préparation des données sont des étapes cruciales qui nécessitent beaucoup de ressources humaines. Il est également crucial d'avoir suffisamment de données pertinentes et variées. De plus, il est fréquemment nécessaire de passer par l'étape d'annotation, qui consiste à étiqueter les données avec le nom de la catégorie à laquelle elles appartiennent ou la valeur numérique qu'elles représentent. Il est également nécessaire de spécifier aux données les caractéristiques à prendre en compte pour l'algorithme. L'apprentissage supervisé, non supervisé, semi-supervisé et par renforcement sont les types d'apprentissage les plus courants.

1.3.2.1. Apprentissage supervisé : Il consiste à créer des algorithmes capables d'apprendre à partir d'un ensemble de données ou à étiqueter chaque entrée (X) par un label. Les labels aident à évaluer le modèle et à réduire l'erreur pendant l'entraînement [8]. Les arbres de décision, la régression linéaire, la régression logistique, KNN (K Nearest Neighbors), les réseaux neuronaux sont quelques-uns des algorithmes les plus populaires pour ce type d'apprentissage.

1.3.2.2 Apprentissage non supervisé : On ne connaît que les données non labellisées (X) dans ce type d'apprentissage. Chaque groupe (cluster) de données doit être analysé et examiné par un modèle d'apprentissage non supervisé afin d'identifier les caractéristiques les plus représentatives [9].

1.3.2.3 Apprentissage semi supervisé : Ce type d'apprentissage entraîne les modèles à l'aide de données étiquetées (X, Y) et non étiquetées. Ce type d'apprentissage permet à la fois d'utiliser une approche supervisée et de bénéficier d'une grande quantité de données non supervisées, qui sont généralement moins chères et plus simples à collecter [10].

1.3.2.4 Apprentissage par renforcement : Il s'agit d'un apprentissage par essai/erreur qui consiste à apprendre des actions à faire à partir d'expériences pour optimiser une récompense quantitative au fur et à mesure du processus. L'algorithme multiplie les tentatives pour tenter de trouver les actions qui maximisent une fonction objective calculée grâce aux récompenses. Dans ce contexte, les données d'apprentissage sont directement obtenues de l'environnement. Cet apprentissage est largement utilisé. Basé sur des réseaux de neurones en réseau.

1.4 Le deep Learning

L'deep learning, également connu sous le nom d'apprentissage profond, est une branche de l'intelligence artificielle qui s'appuie sur les principes de l'apprentissage automatique. John McCarthy a inventé cette méthode en 1955 pour créer des algorithmes capables d'apprendre et de s'améliorer de manière autonome. Contrairement à la programmation traditionnelle, la machine

apprend à partir des données qu'elle traite plutôt que d'exécuter des règles préétablies. L'architecture en couches d'unités de traitement non linéaires utilise l'apprentissage profond pour extraire ou transformer les caractéristiques des données. Les résultats de chaque couche sont utilisés comme entrée pour la couche suivante. Ces algorithmes peuvent être supervisés pour effectuer une analyse de modèle ou non supervisés pour classer les données. La reconnaissance faciale et vocale, ainsi que la reconnaissance d'images ou la vision robotique, ont été considérablement améliorées grâce au deep learning. Cette méthode repose sur les réseaux de neurones artificiels, qui sont inspirés du fonctionnement des neurones du cerveau humain. Ces réseaux sont constitués de nombreux neurones artificiels connectés les uns aux autres, et plus le nombre de neurones artificiels est important, plus le réseau est profond. En conséquence, le deep learning a la capacité d'absorber une quantité considérable de données et a même dépassé les capacités humaines dans certaines tâches cognitives [11].

1.5 Réseaux de neurones artificiels

1.5.1 Définition

Un réseau de neurones est un groupe de neurones qui sont connectés les uns aux autres par des connexions pondérées. Le type d'unités utilisées et sa topologie le caractérisent principalement. Dans un réseau, on distingue fréquemment deux catégories de neurones distinctes : les neurones d'entrée reçoivent les informations provenant du monde extérieur et les neurones de sortie fournissent les résultats du traitement effectué. En général, les autres unités sont appelées cachées. Cependant, il n'est pas nécessaire de faire cette distinction et tous les neurones ont la capacité de communiquer dans les deux sens avec le monde extérieur [12].

1.5.2 Fonctionnement

Une règle de calcul appelée modèle de neurone formel permet d'associer une sortie à x entrées : c'est donc une fonction à x variables et à x valeurs réelles. Un poids synaptique, c'est-à-dire une valeur numérique notée de w1 pour l'entrée 1 jusqu'à wx pour l'entrée x, est associé à chaque entrée x. Le neurone formel effectue d'abord la somme des valeurs obtenues en entrées, qui sont pondérées par les coefficients synaptiques, appelée somme [13].

$$w_1x_1 + w_2x_2 + \ldots + w_nx_n = \sum_{j=1}^{n} w_jx_j \quad (1.1)$$

Un seuil θ est comparé à cette grandeur. Une fonction d'activation non linéaire F est utilisée pour modifier la sortie, qui est liée aux entrées x1 à xm de cette manière :

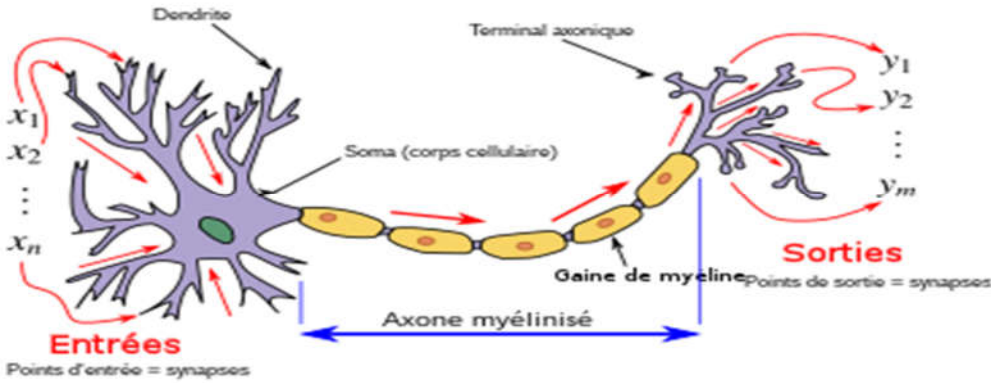


Figure 1.1 : Modèle général d'un neurone.

1.5.3 Fonction d'activation

La fonction d'activation, également appelée fonction de seuillage ou de transfert, a pour but d'introduire une non linéarité dans le fonctionnement duneurone. En général, les fonctions de seuillage présentent trois intervalles : Si le neurone est en dessous du seuil, il n'est pas actif (souvent sa sortie vaut 0 ou -1). Une période de transition se produit autour du seuil. Le neurone est actif au-dessus du seuil (souvent, sa sortie vaut 1) [14].

Le tableau 1.1 englobe l'ensemble des fonctions d'activations qui sont généralement utilisées.

Nom de La fonction	**Type**	**Equation**
Seuil	Binaire (fonction de Heaviside)	f (x)=0si x < 0 f (x)=1si x >= 0
	Signe	*f (x) =1six >0* *f(x)=-1 si x<= 0*
Léniare	Identité	f (x) =x
	Saturé positif	*f (x ,k) =0six<0* *f(x,k)=1six>=1/k*

		$f(x,k)=kx$ sinon
	Saturé symétrique	$f(x,k)=1$ six<-1 /k $f(x,k)=-1$six>=1/k $f(x,k)$sinon
Sigmoide	Positive(type logistique)	$f(x,k)=\frac{1}{1+e^{-ks}}$
	Symétrique (type tanh)	$f(x,k)=\frac{2}{1+e^{-ks}}-1$

Tableau 1.1: Fonctions d'activation

1.5.4 Architectures des réseaux de neurones

La topologie ou architecture d'un réseau de neurones artificielles (RNA) est la façon dont les neurones sont organisés et connectés. Les RNA peuvent généralement être d'une connectivité totale, où chaque neurone est connecté à tous les autres neurones du réseau, ou d'une connectivité locale, où chaque neurone est connecté à ses voisins. Les réseaux de neurones non bouclés, également connus sous le nom de réseaux de neurones « à l'avant » et les réseaux de neurones bouclés, également connus sous le nom de réseaux de neurones « à l'arrière » [15].

1.5.4.1 Les réseaux de neurones non bouclés

Les réseaux "proactifs", "de type perceptron" et "feed-forward" ont également été utilisés. Un réseau non bouclé n'utilise pas de boucle de rétroaction pour propager les signaux de la couche d'entrée à la couche de sortie. Comme les entrées et les sorties sont indépendantes du temps, il est généralement statique. Ce type de RNA est principalement utilisé pour effectuer des tâches telles que l'approximation de fonctions non linéaires, la classification et la modélisation de processus statiques non linéaires [16].

1.5.4.2 Les réseaux de neurones bouclés

La présence, au moins, d'une boucle de rétroaction des neurones de sorties vers les neurones d'entrée distingue les réseaux « récurrents » ou « de retour ». Ces réseaux sont fréquemment utilisés pour assigner des tâches de modélisation de systèmes dynamiques et de commande de processus car ils sont dynamiques et sont régulés par des équations aux différences non linéaires en raison des retards associés aux connexions.

1.5.4.3 Les réseaux de neurones convolutifs (CNN)

Les ConvNets sont des outils courants pour l'apprentissage en profondeur. Ils sont principalement destinés aux images en tant qu'entrées, mais ils peuvent également être utilisés pour d'autres applications comme le texte, les signaux et d'autres réponses continues. Il existe de nombreuses façons dont ces types de réseaux de neurones se distinguent des autres types [10]. La structure biologique du cortex visuel, qui contient des arrangements de cellules simples et

complexes, est la source des réseaux de neurones convolutionnels [17]. Les sous-régions d'un champ visuel déclenchent l'activité de ces cellules. Ces zones sont connues sous le nom de champs réceptifs. Selon les résultats de cette étude, les neurones d'une couche convolutive se connectent aux sous-régions des couches avant cette couche plutôt qu'ils ne se connectent entièrement comme dans d'autres types de réseaux neuronaux.

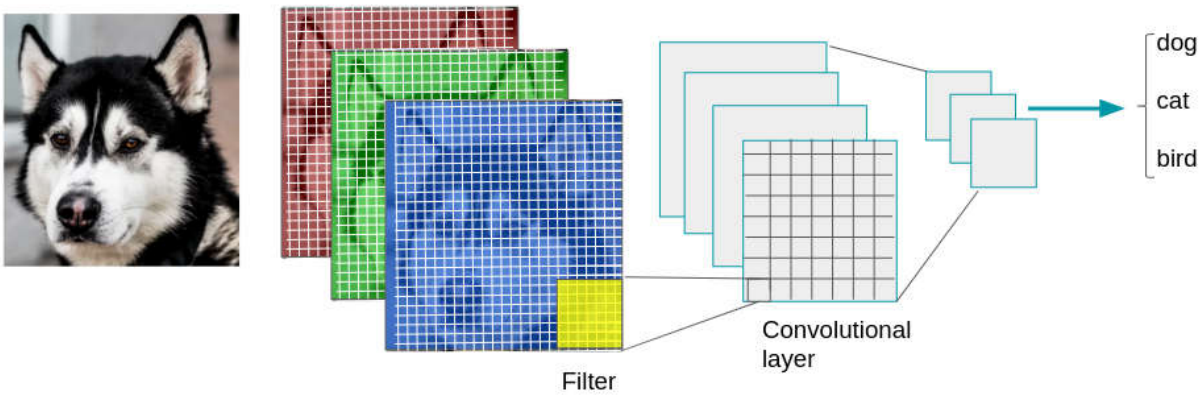


Figure 1.2 : Structure d'un réseau CNN.

1.5.5 Types de réseaux à entraîner

1.5.5.1 Le perceptron multicouche MLP

Ce type de réseau est un réseau de "propagation vers l'avant", c'est-à-dire que l'information se propage dans un sens unique, des entrées aux sorties, sans aucune rétroaction. Son apprentissage est supervisé et correctif. Le signal d'erreur est rétro-propagé aux entrées dans ce cas uniquement pour mettre à jour le poids des neurones. Le principal objectif de ce type de RN est de rassembler les neurones en une couche. Les neurones des deux couches adjacentes sont ensuite connectés complètement après avoir relié plusieurs couches. Ainsi, les sorties des neurones de la première couche sont les entrées des neurones de la deuxième couche. Les neurones de la première couche ont tous le même vecteur d'entrée et sont connectés au monde extérieur. Après cela, ils calculent leurs sorties qui sont ensuite transmises aux neurones de la deuxième couche. Le RN sort des neurones de la dernière couche. Les neurones d'une même couche ne sont pas connectés les uns aux autres [18] (Voir Figure 1. 3).

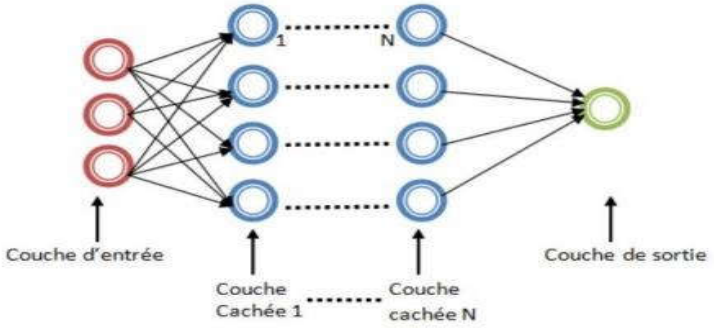


Figure 1.3 : Structure d'un réseau MLP.

1.5.5.2 Réseaux à fonction radiale (RBF)

L'architecture RBF (Fonctions de base radiales) est similaire à celle du PMC ; les fonctions de base utilisées ici sont des fonctions gaussiennes et il se compose d'une seule couche intermédiaire. Par conséquent, les RBF seront utilisés pour résoudre les mêmes problèmes que les MLP. Le mode hybride est le mode d'apprentissage RBF le plus couramment utilisé, et les règles utilisées sont soit la règle d'apprentissage par compétition, soit la règle de correction de l'erreur [19]

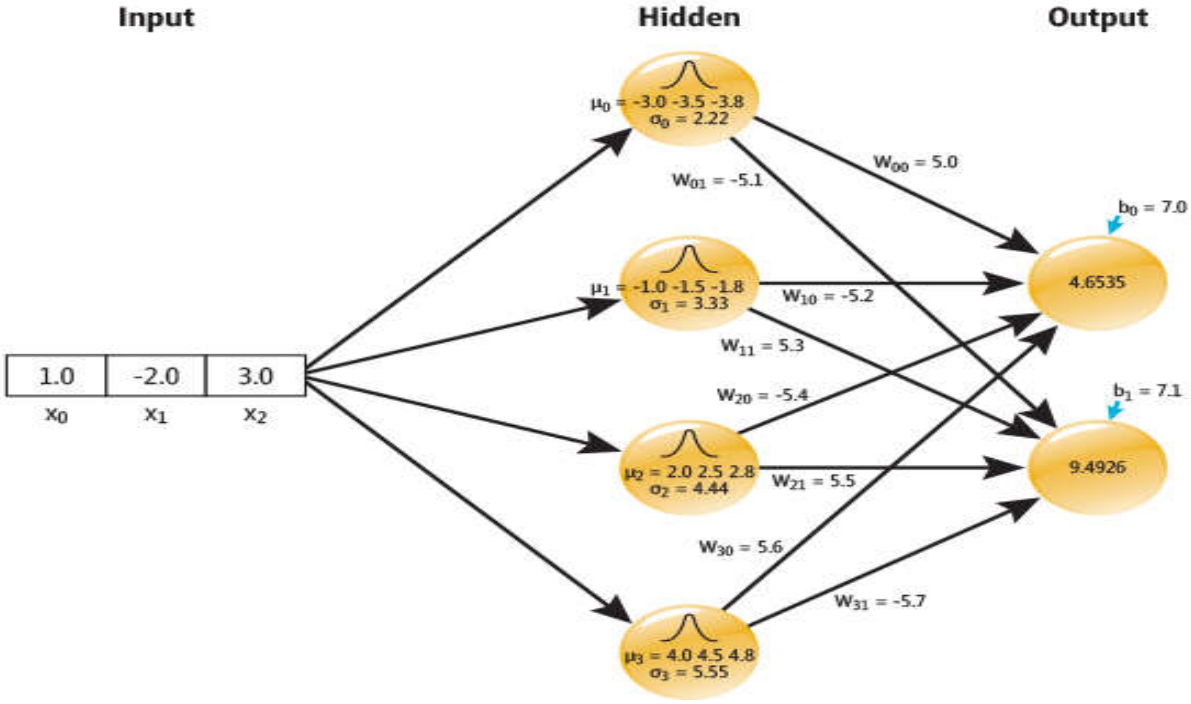


Figure 1.4 : Structure d'un réseau RBF.

1.6 Conclusion

Ce chapitre traite les principes fondamentaux de l'intelligence artificielle, en fournissant des définitions clés pour une meilleure compréhension de cette discipline émergente et de ses applications. Par la suite, il s'est penché sur le concept global du machine Learning, en examinant ses différentes catégories. Ensuite, une attention particulière a été portée à l'apprentissage profond (deep Learning), l'un des types d'intelligence artificielle. Les réseaux de neurones, qui forment la base du deep Learning, ont été examinés en détail, ainsi que leurs variations. Le chapitre suivant abordera le darknet.

Chapitre 02

Darknet

2.1. INTRODUCTION :

L'Internet peut être largement divisé en Web de surface ou Web clair et Web profond. Le web de surface ou web clair se compose de pages Web ou de contenu Web qui seront indexés par le moteur de recherche populaire comme Google ou Yahoo et accessibles via un navigateur standard sans avoir besoin de logiciels et de configurations spéciales [20]. Le contenu de Deep Web ne peut pas être indexé par les moteurs de recherche tels que Google, Yahoo et Bing et, Darknet est situé dans le Deep Web et DarkNet est une petite section de celui-ci. Darknet a été préférentiellement caché à l'intérieur du deep web et ne peut pas être accédé par l'intermédiaire de navigateurs Web standards parce que le contenu de deep web n'est pas indexé par n'importe quel navigateur Web populaire. La plupart du contenu de Darknet se trouve sur le réseau TOR qui est un réseau anonyme et ce contenu peut être accédé via le navigateur TOR.TOR est un réseau de tunnels virtuels chiffrés qui permet aux gens d'utiliser Internet de manière anonyme, cachant leur identité et leur trafic réseau. En utilisant le protocole de service caché de TOR, les gens peuvent également héberger des sites Web anonymement qui ne sont accessibles que par ceux sur le réseau TOR [21].Le Deep Web est également appelé réseau à accès limité [22], qui comprend des sites privés (qui nécessitent obligatoirement des informations d'identification pour y accéder), des sites non liés, des sites bloqués (qui requièrent de répondre à un CAPTCHA pour l 'accès), des pages web dynamiques (qui exigent une URL complète pour y avoir accès), des contenus non HTML ou scripts, et un réseau qui n'est pas ouvert à tous les utilisateurs.

Les sites Darknet sont hébergés avec Domain Name System (DNS) root tels que les domaines. N'importe qui peut partager, communiquer et diffuser des idées via Internet mais à cause du darknet, malgré les nombreux avantages de l'internet, des groupes terroristes, des groupes extrémistes, des organisations de haine et cybercriminels utilisent le darknet pour mener des activités criminelles, promouvoir leur idéologie ou vendre des services ou des biens tels que la drogue, les données de la carte de crédit d'armes, les documents falsifiés, les services de localisation pour le meurtre, les stupéfiants et la pornographie indécente, etc. [23].Toute personne qui veut accéder à tout contenu de l'obscurité, n'a pas besoin de taper des mots-clés dans un navigateur ordinaire, mais devra y accéder de manière anonyme en utilisant le navigateur TOR, qui cache son identité telle que l'adresse IP ou l'emplacement physique. Pour ces raisons, il est difficile pour les organismes chargés de l'application de la loi ou les spécialistes du digital forensic de déterminer l'origine du trafic, de la localisation ou de la propriété d'un ordinateur ou d'une personne sur le réseau noir.

L'implication du darknet est mise en lumière lorsque le Bureau fédéral d'enquête (FBI) a fermé le site web – Silk Road en octobre 2013, qui était un marché noir en ligne et le premier marché darknet moderne pour la vente de drogues illégales [24,25]. Silk Road n'était accessible qu'à travers le réseau TOR et caché du mainstream web. Il y a eu beaucoup de buzz autour de Bitcoin, réseau TOR et dark web parce que la plupart des sites de réseau noir effectué des transactions par l'intermédiaire de la monnaie numérique anonyme, peer à peer, distribué et Bitcoin qui est basé sur la cryptographie principal. Il est très difficile pour les professionnels de la médecine numérique de suivre de telles transactions parce que les utilisateurs et les services sont

anonymes. Le but et l'objectif de ce document est de discuter des techniques forensiques numériques pour traiter de tels crimes de darknet

2.2. Quel est le Darknet ?

2.2.1 Le Web profond, le Web sombre et le Darknet

Il est difficile de comprendre le Darknet sans le différencier du deep web et du dark web, qui sont souvent confondus. Ces réalités restent essentiellement distinctes même si elles se recouvrent partiellement. Le terme "deep web" fait référence aux pages qui ne sont pas référencées par les moteurs de recherche conventionnels tels que Google, Yahoo!, Bing et autres. Il peut s'agir d'informations sécurisées par des mots de passe (comme les comptes bancaires des clients, les e-mails...), d'informations configurées pour ne pouvoir être consultées que par des personnes spécifiques (comme les paramètres de confidentialité des réseaux sociaux), d'intranets ou encore de bases de données. Les données présentes sur le deep web ne sont pas cachées et ne sont pas simplement indexées par les moteurs de recherche, ce qui rend l'accès facile à toute personne visitant le site de l'INSEE ou sa boîte mail. La plupart des internautes utilisant le Web via le filtre des moteurs de recherche ne peuvent pas accéder au Deep Web, sauf s'ils utilisent d'autres moyens.

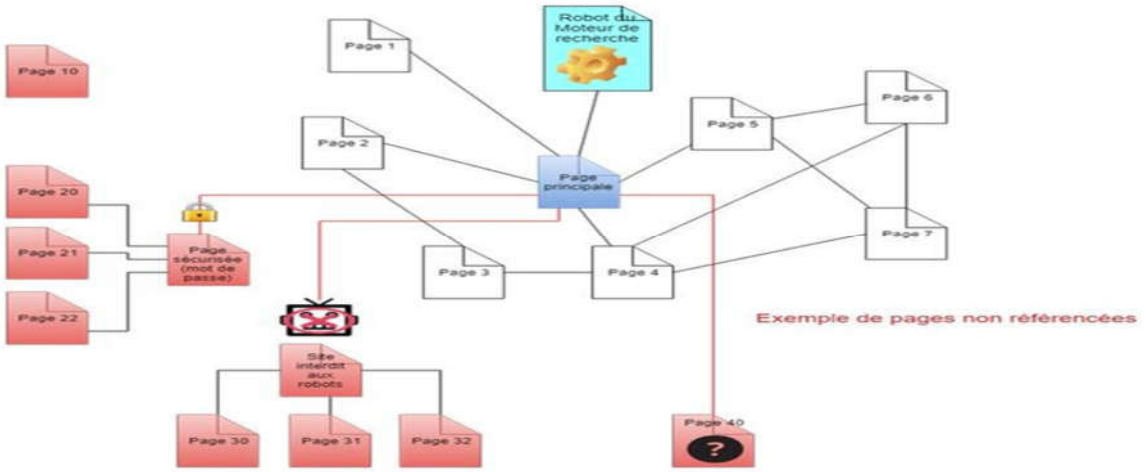


Figure 2.1 : Pages référencées et non référencées

Le deep web se distingue ainsi de surface, qui est la partie du web indexée par les moteurs de recherche classiques. On peut donc raisonnablement douter du sérieux et de l'honnêteté intellectuelle de nombreux articles, parfois publiés sur des plateformes à priori dignes de confiance, décrivant le deep web comme un repaire de criminels en tout genre. Le deep web n'est ni illégal, ni immoral, ni difficile d'accès, ni chiffré ; c'est simplement une couche du web qui ne satisfait pas aux critères de référencement. Il existe de nombreuses raisons pour cela, notamment l'absence d'hyperliens renvoyant vers la page, l'utilisation d'un mot de passe pour accéder à la page, le contenu dynamique, les pages qui sont difficiles à comprendre pour les robots ou encore la demande explicite des administrateurs du site de ne pas le référencer (ce qui est comparable à la décision de mettre son numéro de téléphone sur liste rouge sur Internet).Bien que le Web

indexable soit considérable, il ne représente qu'une petite partie du Web, car il comprend également tous les sites et pages non indexés qui composent le Deep Web. Michael K. Bergman a publié une étude de référence comparant l'envergure du surface web et du deep web, et bien qu'elle soit un peu ancienne, elle est toujours considérée comme faisant autorité par les spécialistes du domaine. Selon cet article, le réseau profond serait 500 fois plus grand que le réseau indexable. Par conséquent, l'image de l'iceberg est souvent utilisée pour représenter la relation entre le deep web et la surface web.

Figure 2.2 : L'image montre la relation entre surface web et deep web

Il est incorrect de dire qu'elle est évocatrice car la partie immergée devrait être 500 fois plus importante que la partie émergée. Le Darknet fait partie du deep web car il n'est pas référencé par les moteurs de recherche classiques. Cependant, il s'en distingue par les protocoles qu'il intègre à la fois nativement et à l'étranger. Il est important de noter que le Darknet ne constitue pas une infrastructure distincte d'Internet, car il utilise les protocoles TCP/IP, qui sont essentiels à son fonctionnement. L'utilisation récurrente de l'article défini suggère que le Darknet n'est pas un seul réseau. Il s'agit plutôt d'un ensemble de réseaux qui peuvent tous être appelés darknets en raison de certaines caractéristiques communes et malgré les différences qu'ils présentent. Quelles sont ces caractéristiques exactement ? Un darknet exige :

- La présence de l'infrastructure Internet (TCP/IP) qui l'accompagne ;
- un protocole spécifique pour créer un réseau superposé ;
- une architecture de pair-à-pair décentralisée ;
- Incorporer des procédures d'anonymisation.

Bien que tout darknet utilise l'infrastructure Internet, il se distingue par la création d'un réseau superposé grâce à un protocole unique : le réseau superposé s'appuie sur l'infrastructure sous-jacente tout en intégrant des fonctions qui lui sont propres. Un darknet est un réseau qui utilise un langage particulier et crée une réalité hermétique, ce qui signifie qu'aucune communication ne peut avoir lieu entre lui et les autres darknets. Contrairement à cela, que les personnes connectées à ce réseau superposé peuvent échanger des informations, sinon on ne parle pas de réseau. Les darknets utilisent une architecture pair-à-pair décentralisée. La plupart des échanges d'informations en ligne se font via une architecture client-serveur. Un logiciel client, installé sur un ordinateur client (le plus souvent un ordinateur personnel), fera une requête auprès d'un logiciel serveur (installé sur un ordinateur beaucoup plus puissant qu'un ordinateur personnel

et dont l'ensemble de la mémoire de calcul est dédié à cette activité). Le client peut ensuite accéder aux données recherchées en répondant au serveur qui stocke les informations nécessaires. Le langage utilisé par un client et un serveur pour échanger des informations sur Internet est le protocole d'échange d'hypertexte HTTP. Un serveur peut répondre à plusieurs clients en même temps grâce à sa capacité de calcul. En raison du fait que l'information est stockée à un endroit précis (sur le serveur), les clients souhaitant accéder à cette information devront tous interroger le même serveur, l'architecture client-serveur est appelée centralisée. En revanche, l'architecture pair-à-pair permet à tout client de devenir serveur, c'est-à-dire d'héberger des données ou de renvoyer les clients vers l'hébergeur. Cependant, elle peut être centralisée car pour accéder aux données sur un ordinateur, le client qui émet la requête devra interroger un serveur qui le redirigera ensuite vers l'ordinateur en question. Dans le cas du pair-à-pair décentralisé sur lequel reposent les darknets, les clients (qui peuvent toujours devenir des serveurs) ne communiquent qu'avec des clients (des pairs) devenant des serveurs pour renvoyer l'information ou transmettre la requête à l'ordinateur censé héberger l'information. L'information n'est plus stockée sur un seul serveur, mais est dispersée sur plusieurs ordinateurs, voire dupliquée, ce qui renforce le système dans son ensemble.

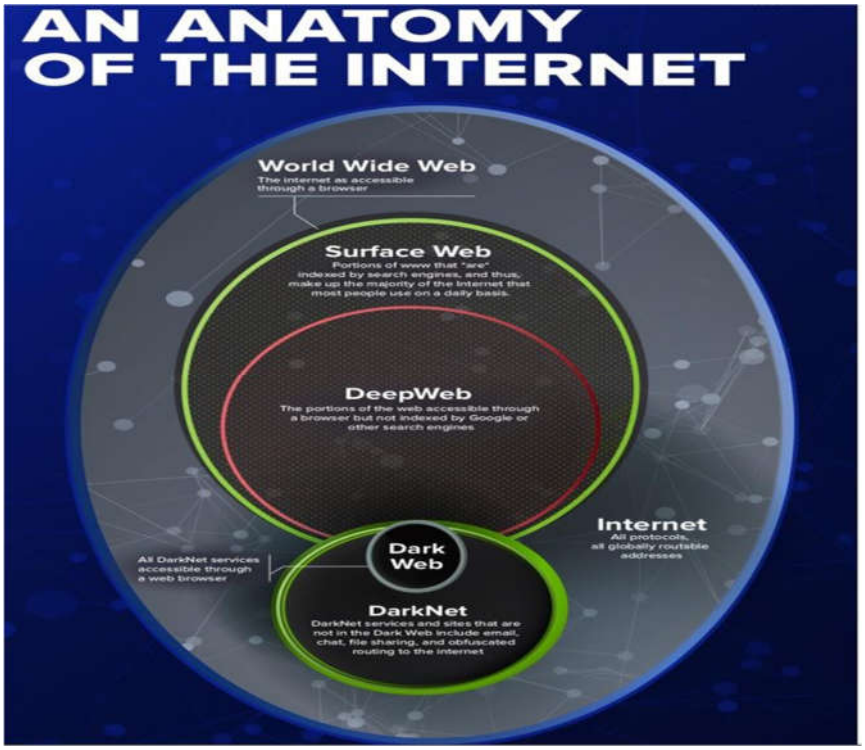


Figure 2.3 : Deep web, dark web et Darknet, des réalités distinctes.

Enfin et surtout, les darknets incluent une fonction d'anonymisation de l'utilisateur. Leur dénomination découle de cette propriété qui leur est consubstantielle. La confidentialité des échanges et la protection de l'anonymat sont leurs principales responsabilités. Afin de comprendre ce que signifient les darknets, il est essentiel de les différencier car ces termes ne

sont pas interchangeables. L'anonymat signifie dissimuler son identité. Nous avons observé que l'identification de l'internaute pouvait être effectuée à l'aide de l'adresse IP : les darknets rendent cette identification sinon impossible, du moins extrêmement difficile. La confidentialité consiste à empêcher que des tiers accèdent aux informations échangées. Si je suis certain que le contenu d'un e-mail ne pourra être lu que par mon correspondant, je pourrai prétendre à la confidentialité dans un échange d'e-mails. Cependant, l'échange ne sera anonyme que si l'origine des e-mails est cachée. Par conséquent, il est possible d'être anonyme sans échanger de données sensibles ou de chiffrer le contenu de ses communications sans dissimuler son identité. Les darknets sont des outils pour protéger la confidentialité et l'anonymat en ligne. Il nous reste un concept à éclaircir avant d'en expliquer le fonctionnement : celui du dark web. Le Darknet ne peut pas se résumer à Internet (il n'en est qu'une couche applicative), tout comme le World Wide Web ne peut pas être identifié à Internet. De nombreuses applications peuvent être considérées comme des darknets sans être liées au web, comme le logiciel RetroShare ou le service d'e-mail Mailpile. Ces applications sont considérées comme des darknets car elles incluent des fonctionnalités de chiffrement des transmissions et d'anonymisation de l'utilisateur, mais elles ne font pas partie du dark web, qui est l'ensemble des sites d'un darknet donné auxquels un navigateur peut accéder.

2.2.2 Darknet et cryptographie : les mixnets

Comme nous l'avons souligné : la caractéristique essentielle d'un darknet est l'anonymisation et la confidentialité des échanges. Pour parvenir à cette fin, les darknets utilisent le chiffrement des données, rendant celles-ci inutilisables par une tierce personne. Comment fonctionne ce chiffrement ? Pour le saisir, il nous faut aborder quelques éléments de cryptographie. La cryptographie est une technique consistant à rendre un message inintelligible afin d'en protéger le contenu. Les méthodes cryptographiques sont aussi nombreuses qu'anciennes et nous ne nous intéresserons ici qu'à deux d'entre elles, largement usitées en informatique et en particulier par les darknets : la cryptographie symétrique et la cryptographie asymétrique. La cryptographie symétrique se développe dans les années 1970 afin d'assurer aux banques et aux entreprises des moyens de communication sécurisés. Elle a depuis été perfectionnée, et sa version la plus aboutie, AES (Advanced Encryption Standard), est encore utilisée aujourd'hui. La cryptographie symétrique consiste à chiffrer un message via une clé, cette même clé permettant également le déchiffrement du message. La clé doit donc être connue non seulement par l'expéditeur, mais également par le destinataire. Cet échange de clé rend l'opération risquée, car si une tierce personne parvenait à intercepter cette clé lors de l'échange, le chiffrement ne serait plus d'aucune utilité. En outre, il est nécessaire de créer une clé suffisamment complexe pour résister à la puissance de calcul des ordinateurs actuels.

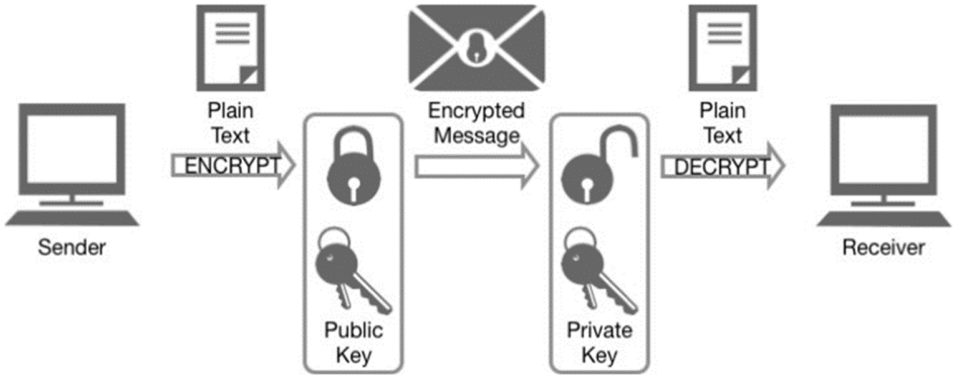


Figure 2.4 : Cryptographie asymétrique : la clé publique sert à chiffrer le message, la cléprivée permet de le déchiffrer.

Cependant, l'avantage de la légèreté de cette technique cryptographique est la principale raison pour laquelle elle est encore largement utilisée sur Internet aujourd'hui.

La cryptographie asymétrique a été développée en réponse aux faiblesses de l'algorithme DES dans sa version 56 bits. En 1976, Winfield Diffie et Martin Hellman suggèrent de diviser la clé en deux parties : l'une serait publique et l'autre serait privée. Il est acceptable que Bob et Alice souhaitent interagir. Alice divise une clé en deux parties. Elle diffuse la clé publique afin que Bob (mais pas seulement) puisse l'utiliser et garder la clé privée dans un endroit sécurisé. Bob utilisera la clé publique d'Alice pour chiffrer le message qu'il veut lui envoyer, et Alice utilisera sa clé privée pour déchiffrer le message de Bob. En plus de permettre des échanges confidentiels sans avoir à échanger une clé secrète, cette méthode de chiffrement présente un avantage supplémentaire : si la clé privée peut déchiffrer un message chiffré avec la clé publique, l'inverse est également vrai. En ajoutant une couche supplémentaire de cryptage, cette propriété permet au destinataire d'authentifier l'émetteur. Si Bob a sa propre clé, il peut chiffrer son message avec sa propre clé privée et la clé publique d'Alice. À la réception, Alice devra d'abord utiliser sa clé privée (déchiffrer la première couche), puis utiliser la clé publique de Bob pour s'assurer que Bob est bien à l'origine du message. Les réseaux privés basés sur la cryptographie asymétrique utilisent des algorithmes pour générer automatiquement des clés de session aléatoires, ce qui garantit la confidentialité de la communication. Cependant, la confidentialité ne signifie pas anonymat, et l'objectif des darknets est de combiner ces deux aspects de la protection de la vie privée.

Pour atteindre cet objectif, les darknets incluent une caractéristique supplémentaire appelée relais, qui est rendue possible par une structure similaire à un réseau mixte. En 1981, David Chaum publie un article fondamental dans lequel il expose un système qui garantit la confidentialité et l'anonymat. Un cryptage à clé publique garantit la confidentialité des échanges. Il est possible d'anonymiser les communications en introduisant un réseau mixte (ou mixnet). Un réseau mixte fonctionne de manière assez simple : les informations ne sont pas transmises directement entre deux ordinateurs (client-serveur), mais plutôt via des relais, des ordinateurs intermédiaires. Chaque relais ne sait que l'identité de son prédécesseur et de son successeur afin

de garantir l'anonymat. Par conséquent, si je passe par trois relais A, B et C, seul le relais A aura accès à mon adresse IP, tandis que le relais B n'aura accès qu'aux adresses A et C, et le relais C ne connaîtra que les adresses de B et du destinataire final. L'information est protégée par plusieurs couches de chiffrement pour garantir la transmission et la confidentialité de la communication. Le destinataire final, à qui nous voulions dissimuler notre identité en premier lieu, ne connaîtra que l'identité du relais C. Le nombre de couches est déterminé par le nombre de relais : afin de préserver la confidentialité de l'échange, les relais ne doivent pas être en mesure de connaître le contenu du message, qui est réservé au destinataire. Admitons que Bob communique avec Alice via un réseau mixte composé de trois relais. Ainsi, l'ordinateur de Bob utilise la clé publique d'Alice pour créer une première couche de chiffrement. Le relais A utilisé sa clé privée pour déchiffrer la première couche du message après l'avoir reçu, puis le relais B utilise sa clé privée pour déchiffrer la deuxième couche. Une fois que le relais A a reçu le message, il utilise sa clé privée pour déchiffrer la première couche. Lorsque le dernier relais, C, transmet le message à Alice, il n'y a plus qu'une couche de chiffrement qui peut être décryptée par Alice avec sa clé privée.

2.3. Les outils utilisés dans le Darknet

Comme nous l'avons souligné, le darknet n'existe pas plus que l'homme et n'est qu'une appellation commode pour regrouper les différents darknets qui existent déjà. Des outils (logiciels) permettent l'accès à un darknet. Nous ne pouvons pas fournir une liste exhaustive, donc nous restons limités aux sites traditionnels tels que Tor et Freenet.

2.3.1. The Onion Router (Tor) :

Le darknet le plus connu, Tor (acronyme de The Onion Router), est celui qui a donné naissance à une réputation sulfureuse dans ce domaine numérique. Le principe de Tor, également connu sous le nom de routage en oignon, a été créé en 1996 par des chercheurs de la Naval Research Laboratory (NRL) dans le but de chiffrer et d'anonymiser les communications militaires. Cependant, sa première version n'a été développée qu'en 2002, six ans plus tard. En 2004, la NRL prend la décision de cesser de financer Tor et publie son code source. L'Electronic Frontier Foundation (EFF) récupère le code et le transforme en The Tor Project.

Le principe de mixnet inventé par David Chaum sert de base au routage en oignon. Tor dissimule le trafic via un ensemble de relais (ou nœuds) au lieu d'établir une connexion directe entre le client et le service recherché. L'ensemble des relais de Tor est renseigné dans un annuaire librement consultable, permettant au fournisseur d'accès à Internet de les détecter facilement. Il est possible de déterminer qu'un internaute utilise Tor, mais il est impossible de savoir ce qu'il fait sur ce réseau. L'établissement d'un circuit Tor se fait en plusieurs étapes : le client choisit un relais d'entrée dans le réseau via l'annuaire des relais, qui sera ensuite élargi au deuxième relais et au troisième relais. En conséquence, un circuit Tor comprend un relais d'entrée, un relais intermédiaire et un relais de sortie. Enfin, le relais de sortie est connecté au serveur de destination. L'ordinateur client détermine ce circuit de manière aléatoire et le change environ toutes les dix minutes.

En tant que réseau mixte, Tor envoie des paquets avec plusieurs couches de chiffrement.

Le client chiffre les données après avoir identifié un circuit en utilisant la clé publique des différents relais. Le routage en oignon est le processus par lequel le paquet est progressivement déchiffré, couche après couche, au cours de son passage par les différents relais. L'anonymat de l'utilisateur est garanti dans chaque relais, ne connaissant que son prédécesseur et son successeur immédiat Seule l'adresse du relais de sortie est connue par le serveur de destination.

Le navigateur Tor vous permet d'accéder à n'importe quel site Web, comme faire des achats sur Amazon, consulter ses e-mails ou regarder des vidéos sur YouTube. Les informations qui voyagent du dernier relais au serveur de destination ne sont pas chiffrées dans cette situation. En effet, lorsque l'on se connecte à des sites du Clearnet, on sort de l'environnement Tor, ce qui ne protège plus les paquets d'informations. Il est impossible pour Tor d'imposer un chiffrement aux serveurs qui ne font pas partie de son réseau. Par conséquent, se connecter à un service via un protocole HTTP simple et fournir son mot de passe revient à s'exposer à une récupération potentielle de données par une personne mal intentionnée, que l'on utilise Tor, Firefox ou Internet Explorer.

En utilisant Tor, il est possible de proposer des services tels que des sites web en cachant l'adresse IP du serveur, comme les services oignon. L'adresse IP de ces sites est composée de 16 lettres et chiffres suivies du nom de domaine. Onion. L'anonymisation du serveur est similaire à celle du client en ce sens que le serveur sélectionne aléatoirement divers chemins d'accès constitués de relais. Parce que le client et le serveur ne sont visibles que par leur troisième relais, leur anonymat est garanti. Le chiffrement est effectif de bout en bout lorsqu'un client et un serveur hébergé sur le réseau Tor communiquent via un protocole de chiffrement asymétrique.

2.3.2 Freenet

Freenet a été créé en 1999 par Ian Clarke, un étudiant en informatique à l'Université d'Édimbourg, et est le darknet le plus ancien encore utilisé aujourd'hui. Ian Clarke visait à atteindre trois objectifs avec Freenet : garantir l'anonymat de ceux qui produisent et consultent l'information, donner à ceux qui stockent l'information la possibilité de nier en avoir connaissance et résister aux tentatives de tiers souhaitant limiter voire supprimer l'accès à l'information. Freenet est un réseau pair-à-pair décentralisé qui utilise tous les nœuds pour stocker et diffuser des données. Même si Freenet était initialement conçu pour le partage de fichiers, il est maintenant doté d'autres fonctionnalités, telles que Freemail, qui permet d'avoir une boîte mail anonyme. Contrairement à Tor, qui gère ses propres relais, Freenet est constitué par les nœuds qui s'y connectent.

Lorsqu'un ordinateur est connecté à Freenet, il reçoit désormais une partie de sa bande passante et de son espace de stockage sur son disque dur. Devenu un nœud, l'ordinateur servira à stocker et à diffuser certains fichiers, contribuant ainsi à l'architecture du réseau. Une table de routage est conservée par chaque nœud du réseau, qui contient une liste des clés et des identifiants des nœuds qui les stockent. Le disque dur contient un stock de fichiers et de clés pour déchiffrer les données. Chaque fichier possède une clé qui permet de l'identifier. Lorsqu'un client souhaite récupérer du contenu, l'ordinateur recherche la clé correspondante et interroge la table de routage pour déterminer si le fichier est stocké sur son propre nœud. Si ce n'est pas le cas, il

transfère la requête au nœud le plus proche qui contient la clé. La requête se poursuit de la sorte jusqu'à ce que le nœud contenant le fichier soit trouvé ou jusqu'à ce que le nombre maximum de sauts soit atteint (un message d'échec est alors renvoyé au premier nœud). Lorsque le fichier est trouvé, il est envoyé au premier nœud en utilisant le chemin inverse de la requête. Le fichier est mis en cache par chaque nœud participant au chemin de retour, c'est-à-dire copié sur son espace dédié. Ce processus permet non seulement de soutenir l'architecture décentralisée de Freenet, car tout fichier chargé est stocké sur plusieurs nœuds, mais il garantit également un déni crédible des utilisateurs. En effet, les fichiers stockés dans un nœud ne sont pas seulement cryptés. Ce qui empêche l'hébergeur de savoir quels fichiers sont stockés sur sa machine, mais chaque nœud qui a participé à l'acheminement d'un fichier peut se présenter comme l'hébergeur, semant le doute quant à l'emplacement original du document. La structure de Freenet rend impossible toute accusation potentielle de détention de fichiers illégaux. Les fichiers les moins utilisés sont supprimés en premier lorsque la limite de stockage de Freenet est atteinte, ce qui permet d'optimiser le réseau selon la demande.

De la même manière que sur Tor, l'anonymat est garanti : chaque nœud ne connaît que son prédécesseur et son successeur. Rien ne permet de distinguer le nœud qui héberge le fichier du nœud qui initie la requête. En transmettant des informations, un nœud ne peut jamais déterminer s'il communique avec un nœud de transfert ou final. Il est important de noter que Freenet peut être configuré pour garantir une protection quasi totale de la vie privée. Le mode réseau invisible permet à l'utilisateur d'utiliser le réseau en sélectionnant uniquement les nœuds auxquels il a confiance. Néanmoins, ce mode demeure peu pratique pour les novices car il est nécessaire de renseigner soi-même les adresses des nœuds de confiance. De plus, ce mode entraîne inévitablement une augmentation de la durée de chargement des pages, qui est déjà cruciale sur Freenet.

2.3.3. Usages du Darknet

Les médias, friands de sensationnalisme, présentent souvent le Darknet sous un jour défavorable : entre pédopornographie, trafic de stupéfiants et services de tueurs à gages, il ne serait qu'un repaire de malfrats. Qu'en est-il réellement ?

Il est bien évidemment difficile de recenser avec précision les usages du Darknet et de les quantifier puisque les différents darknets cherchent par définition à dissimuler leurs utilisateurs. Cependant, il est possible d'analyser le trafic d'un réseau donné sans pour autant compromettre l'anonymat des internautes, car l'analyse de bande passante ne permet pas à elle seule une identification.

Tor étant de loin le premier darknet en termes d'utilisation et de contenu, c'est sur ce réseau que nous avons choisi d'appuyer notre analyse. Afin de conduire celle-ci avec probité, il convient de distinguer les usages du dark web et les usages du Darknet.

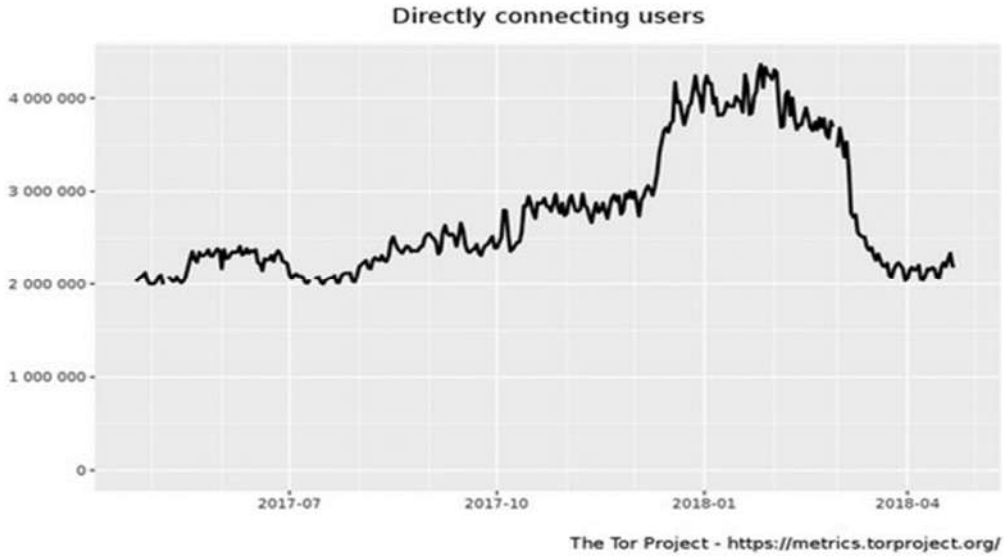


Figure 2.5 : Nombre d'utilisateurs quotidiens de Tor entre avril 2017 et avril 2018

2.4. Fonctionnement de darknet :

Le trafic destiné aux adresses IP du darknet ne devrait pas exister dans un environnement idéal. Cela pourrait être du trafic de scan envoyé par des virus/vers informatiques ou du trafic de rétrodiffusion (backscatter) généré par des attaques DDoS ou DRDoS, Botnet C&C. Les adresses IP qui ne sont pas utilisées peuvent également recevoir du trafic légitime en raison d'un bogue logiciel ou d'une mauvaise configuration.

2.4.1. Activités de scannage :

En effet, un darknet peut être utilisé efficacement pour localiser diverses activités de scanning ou d'analyse du réseau sur Internet [26]. La façon dont le système Darknet capture l'activité de scan est illustrée à la figure 2.7. Il est possible que la machine qui effectue le scan ait été infectée par un ver qui essaie de se propager ou qu'elle soit impliquée dans une analyse automatisée sur Internet. Certains de ces paquets parviennent au télescope.

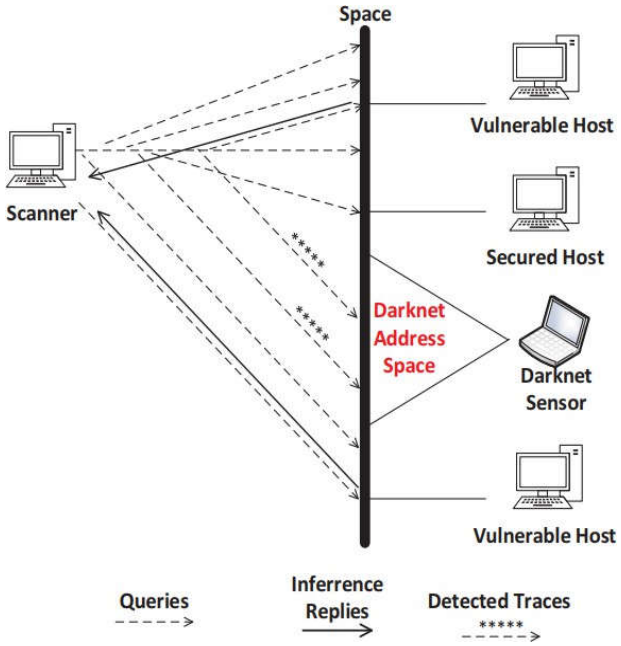


Figure 2.6 : Activités de SCANNAGE

2.4.2 Les attaques DDoS :

Une attaque de déni de service vise à rendre un réseau ou un serveur entier indisponible afin qu'il ne soit plus accessible aux utilisateurs légitimes. L'attaquant crée du trafic de rétrodiffusion lorsque l'attaquant remplace son adresse IP par une adresse IP aléatoire avant d'envoyer les paquets à la victime. Par conséquent, les paquets de réponse sont envoyés à la source usurpée plutôt que à l'attaquant. Il est possible que cette adresse aléatoire soit du bloc d'adresses IP du darknet. Le télescope capte le trafic de rétrodiffusion, comme illustré à la Figure 2.8.

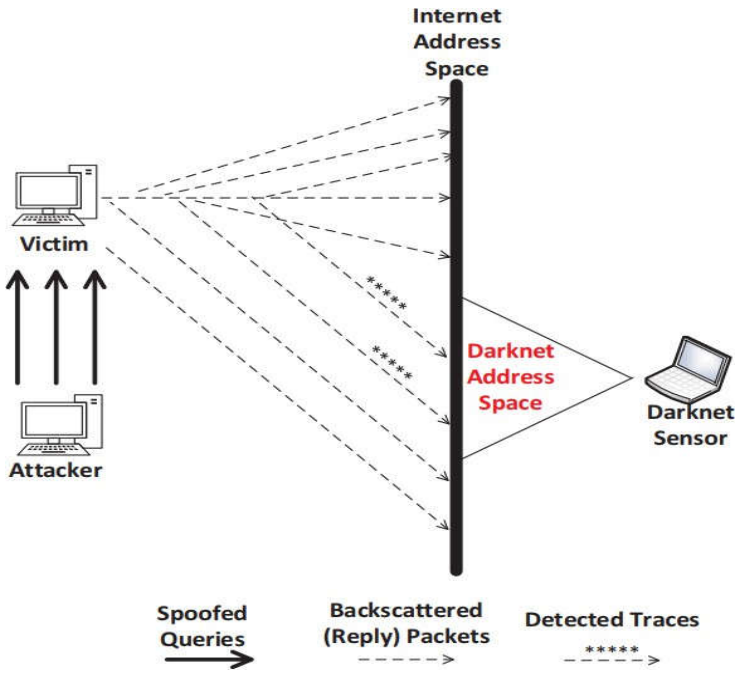


Figure 2.7 : Activités DDoS

2.4.3. Les attaques DRDoS :

Un type unique d'attaque DDoS est l'attaque DRDoS. Pour relayer le trafic d'attaque vers la victime, l'attaquant utilise des tiers, tels que des serveurs Web, des résolvants DNS récursifs ouverts et des serveurs UDP accessibles au public [27]. Les adversaires envoient des requêtes à ces serveurs publics et usurpent l'adresse IP d'une victime. Les serveurs remplissent la victime de réponses valides et épuisent sa bande passante sans qu'elle le sache. Un darknet est utilisé pour déduire les attaques DRDoS [28]. Ce scénario est illustré à la figure 2.9. En règle générale, l'attaquant envoie des requêtes fausses sur Internet dans le but d'atteindre autant d'amplificateurs ouverts que possible pour obtenir un facteur d'amplification élevé. Cela se produit lorsque les attaquants ne connaissent pas les adresses IP des amplificateurs ouverts à

L'avance. Certaines de ces demandes arriveront inévitablement darknet.

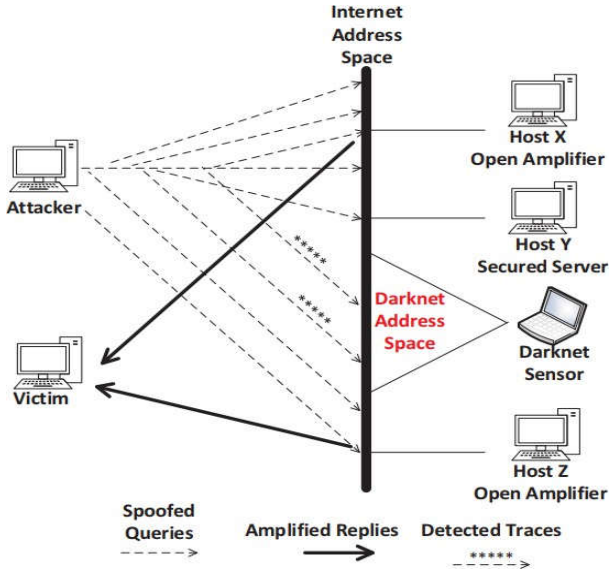


Figure 2.8 : Activités DRDoS

2.5. Données darknet

Les données du darknet peuvent généralement être regroupées dans l'une des trois principales catégories suivantes [29] :

• **Le trafic rétrodiffusé :** Ce trafic est causé par l'utilisation de l'espace adresse surveillée pour créer des identités fausses. La plupart du temps, cela se manifeste sous la forme d'analyses de scans de decoy, d'attaques par déni de service (DoS) ou de résultats d'hôtes mal configurés. Ce trafic est principalement constitué de classes spécifiques de paquets ICMP et TCP contenant des indicateurs RST (réinitialisation), SYN (synchronisation) et ACK (accusé de réception).

• **MISCONFIGURATION :** Ce trafic peut être considéré comme partiellement rétrodiffusé ou potentiellement agressif, et le plus souvent, il provient d'hôtes en ligne mal configurés.

• **AGRESSIF/HOSTILE :** La plupart du trafic observé par le télescope du réseau darknet peut être considéré comme agressif ou potentiellement hostile. Cela comprend les activités DRDoS, les cas évidents d'analyse réseau qui se manifestent à la fois via l'analyse ICMP et TCP, et les paquets avec des charges utiles clairement hostiles.

(Ceux-ci ne sont visibles que dans les exploits basés sur UDP car ils sont sans connexion). Le reste représente le trafic qui peut être classé en différents agents de numérisation automatisés, comme Internet Worms et les logiciels malveillants similaires.

Le tableau 2.1 donne un aperçu de la distribution des protocoles pour comprendre la nature des données darknet. Il a été démontré que les paquets TCP constituent la majorité du trafic darknet. La domination de TCP peut être expliquée par plusieurs faits. Tout d'abord, TCP offre une variété de techniques d'analyse, telles que SYN, Fragmentation et SYN-ACK. Ensuite, générer un scan TCP est généralement plus facile que de générer un scan UDP. Enfin et surtout, les cyberattaques bien connues ciblent les services TCP. Le tableau 2.2 liste également les principaux protocoles d'application utilisés sur darknet.

Les tableaux 2.1 et 2.2 sont le résultat de l'analyse de l'ensemble de données darknet pures capturé au cours d'une période de cinq ans à partir d'un bloc d'adresses [30].

Tableau 2.1. Distribution de protocoles

Count	TCP	UDP	ICMP
Packet	*76.6%*	19,9%	2,8%
Bytes	55,82%	40,82%	2,66%

Tableau 2.2. Les principaux protocoles d'application trouvés

Port	Service
445	*microsoft-ds*
139	*NetBIOS*
4662	*eDonkey*
80	*http*
135	*Endpoint Mapper*

2.6. Déploiement de Darknet :

Le déploiement d'un système de surveillance darknet nécessite une compréhension de la topologie du réseau local. Dans la mesure où un moniteur darknet observe le trafic vers les adresses non utilisées, le routeur en amont doit transférer les paquets non-distribuables au serveur darknet. Cette section présente les techniques de déploiement du darknet et les recherches les plus importantes qui s'y rapportent.

2.6.1 Configuration :

Trois méthodes courantes pour transférer des paquets vers un système de surveillance darknet existent [31] :

• **Envoyer chaque adresse inutilisée au routeur des réponses ARP :** Une méthode simple et utile lorsque vous n'avez pas accès au routeur, mais il faut renvoyer des réponses ARP à intervalles réguliers car un compteur de cache ARP supprime les entrées ARP qui n'ont pas été utilisées pendant une période spécifique, qui varie en fonction des périphériques et des systèmes d'exploitation.

• **Utiliser le routage statique pour un bloc d'adresses :** cette technique est simple mais nécessite que le bloc d'adresses darknet soit mis de côté pour la surveillance. Configurez le routeur pour acheminer statiquement un bloc d'adresses vers le système de surveillance darknet.

• **Le routage statique permet de transférer tous les paquets non utilisés d'un réseau d'organisation au système de surveillance darknet :** Cette méthode est plus adaptable et consiste à rediriger tous les paquets destinés à des emplacements où aucune adresse plus précise n'est configurée (et donc abandonnée). Il est possible de créer une route statique vers le darknet. Seuls les paquets vers des blocs d'adresses non utilisés suivront la route spécifiée pour, et les paquets vers des adresses valides atteindront des préfixes plus précis/16.

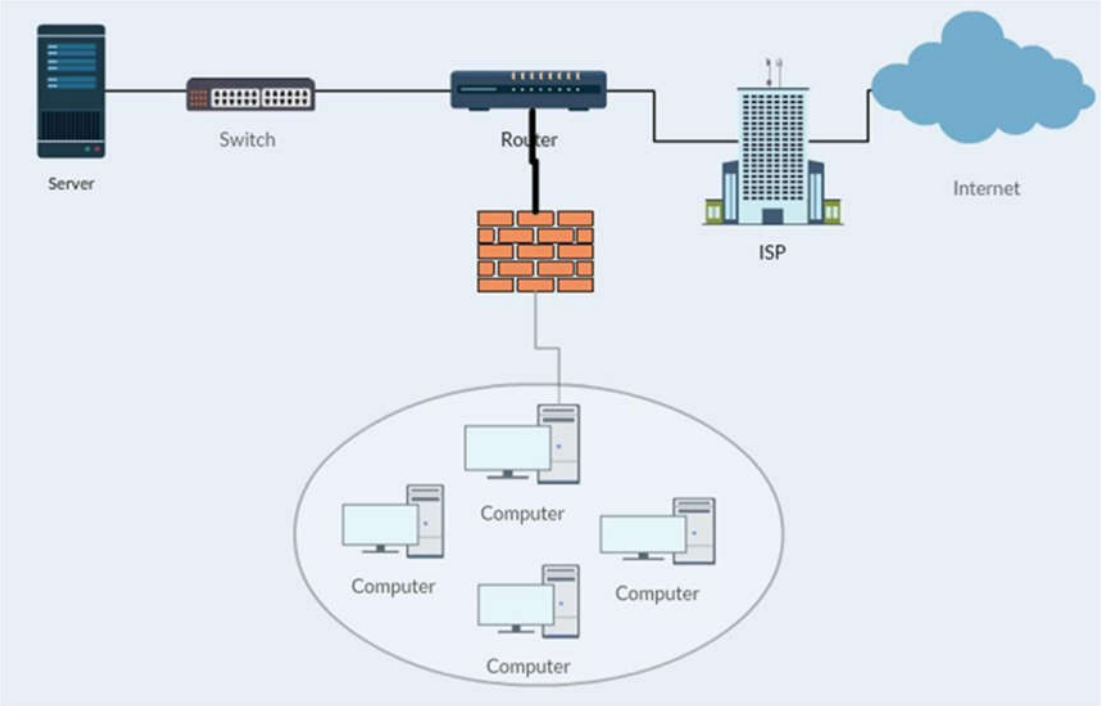


Figure 2.9 : Déploiement d'un serveur Darknet

- Le sous-réseau entre le routeur et le système de surveillance doit utiliser des adresses privées.
- D'autres techniques doivent être utilisées comme le pare-feu, vlan, le routage interne selon l'architecture de réseau, pour garantir que le serveur de surveillance ne répond à aucun paquet.

2.6.2. Variations du Darknet :

Haros et al. [32] utilisent le concept de darknet pour montrer comment un petit nombre d'adresses IP darknet mélangées avec des adresses IP actives peut augmenter l'efficacité de la détection des scans de réseau. Les variantes de darknet sont le déploiement de mécanismes de systèmes de surveillance basés sur des pièges utilisant des techniques semblables à ceux d'un darknet, de l'espace d'adresses IP gris et des moniteurs de darknet. Yu Jin et associés [33] utilisent l'idée de Gray Space IP pour la surveillance passive. Ils identifient les adresses IP Gray Space à l'aide d'un algorithme heuristique. Un système proposé par Polakis et al. [34] permet l'allocation dynamique d'adresses IP non utilisées dans un sous-réseau en utilisant un capteur de surveillance pour les réseaux qui utilisent DHCP.

2.6.3. La visibilité de darknet :

Il est rare que les activités malveillantes observées par deux réseaux darknet de taille identique soient identiques [32]. Les deux principales variables qui influencent ces différences sont l'emplacement d'un darknet et la façon dont un darknet réagit aux paquets entrants. La façon dont les darknets répondent aux paquets entrants détermine grandement la visibilité qu'ils offrent. La meilleure option est de ne pas répondre du tout. Un darknet configuré de manière passive enregistre simplement tous les paquets qu'il observe et ne prend aucune autre mesure. Cependant, car toutes les transactions TCP valides nécessitent une négociation à trois voies qui doit être

effectuée avant l'échange de données au niveau de l'application, le darknet passif n'observe pas les données au niveau de l'application provenant d'hôtes qui tentent de se connecter via TCP. La réponse à un paquet TCP SYN avec des paquets TCP SYN-ACK est une méthode de réponse active simple mais efficace [36].

2.7. Les projets Darknet :

Au niveau mondial, il existe plusieurs projets de tailles et de caractéristiques différentes, certains plus complexes que d'autres, mais qui poursuivent un concept commun : la surveillance de l'activité du trafic réseau à l'aide de Darknet. Voici une vue d'ensemble de certains de ces projets.

2.7.1 Projets à grande échelle dans la darknet :

Certains d'entre eux sont devenus des références importantes pour consulter l'activité du trafic réseau et les tendances sur Internet et pour extraire divers types d'informations sur les cyber menaces en raison de leur taille et de leur capacité.

2.7.1.1 Le télescope réseau :

Les chercheurs du centre d'analyse appliquée des données Internet (CAIDA) ont proposé un système appelé télescopes de réseau UCSD [37, 38] qui inclut un Darknet qui pourrait ressembler à un réseau /8, soit environ 16millions d'adresses IP (l'une des plus grandes infrastructures mondiales de mesure et d'analyse du trafic Internet).

Les attaques par déni de service, la propagation des vers et la détection générale du trafic malveillant créé par les agents automatisés sont tous détectés par ce télescope.

2.7.1.2 Système d'analyse active du niveau de menace (ATLAS) :

Sous la direction d'Arbor Networks [39], ce système de surveillance de réseau analyse collectivement les données traversant un réseau Darknet disparate afin de visualiser les activités malveillantes sur Internet et fournir des informations relatives aux activités malveillantes aux clients d'Arbor services. Les exploits, le phishing, les logiciels malveillants et les botnets font partie des types d'activités malveillantes surveillées par ce système.

2.7.1.3 Le projet Darknet TEAM CYMRU

L'équipe Cymru se concentre sur la recherche sur la sécurité Internet. L'un de ses projets est "The Darknet Project" [40], qui, comme ses pairs, est capable d'identifier les activités malveillantes sur Internet et de générer des statistiques de trafic pour savoir ce qui se passe sur le réseau. Il utilise des outils comme l'analyse des flux et l'analyse du trafic réseau. L'infrastructure de ce projet est constituée de 8 darknets situés dans différentes régions avec un total de 626 944 adresses IP.

2.7.2 Les projets à petite échelle :

Antonatos et al. (une partie du projet NoAH, HoneyHome) [41] propose une plate-forme

de surveillance des adresses IP et des ports inutilisés pour l'extraction d'événements de sécurité à grande échelle. Ce système peu coûteux utilise des capteurs installés sur des utilisateurs réguliers pour surveiller ces adresses IP et ports qui ne sont pas utilisés. De plus, Daedalus [42], qui est basé sur le projet NICTER, vise à détecter les cyberattaques en temps réel. Il y a également d'autres projets qui utilisent la surveillance passive, comme le système d'acquisition de données par numérisation Internet (ISDAS), qui est supervisé par le centre de coordination CERT du Japon [43].

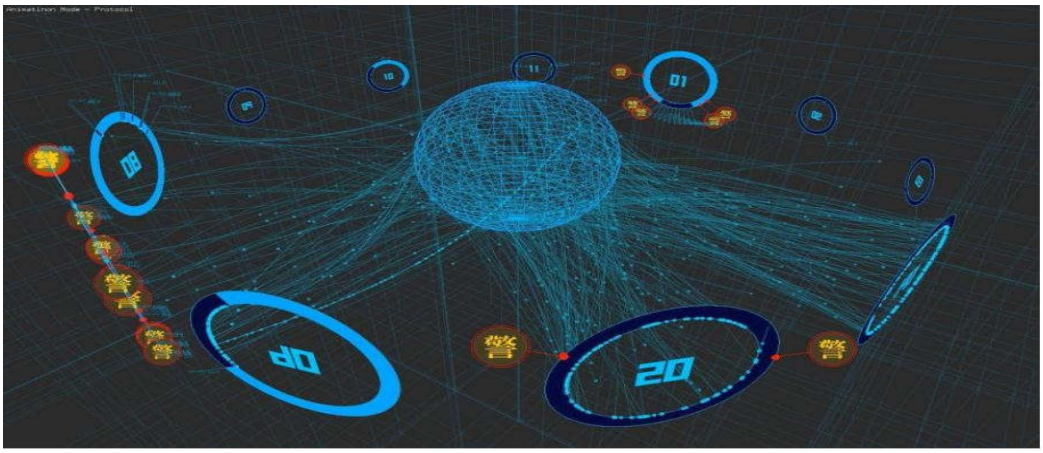


Figure 2.10 : Aperçu DAEDALUS-VIZ

2.7.3 Projets africains :

Comme nous le savons, le projet darknet unique en Afrique a été lancé en 2011 à Rhodes University Network Telescope [44].

2.8. Visualisation Darknet :

Il existe de nombreuses recherches sur l'utilisation du trafic darknet pour détecter les activités malveillantes en utilisant des techniques et des outils de visualisation. Le et al. [45] suggèrent une nouvelle méthode pour déduire le trafic réseau malveillant qui utilise des concepts de la théorie des graphes tels que la distribution des degrés, les mesures de degré maximum et la distance. Les auteurs utilisent la technique des graphiques de dispersion du trafic (TDG) pour modéliser le trafic réseau. Par ailleurs, Joslyn et al. [46] proposent une nouvelle méthode pour faciliter la visualisation et la facilitation des données à grande échelle. L'approche basée sur les graphes utilise les bases de données de routage réseau. Krasser et al. dans un autre travail de visualisation [47] afin de réduire le taux de faux positifs et de faux négatifs, créent un système de visualisation du trafic réseau capable d'analyser des données en temps réel en utilisant des techniques efficaces de visualisation des informations. Fukuda et al [48] proposent une méthode utilisant le traitement d'image pour identifier les activités de balayage dans le trafic darknet. Cette méthode utilise une image en deux dimensions qui représente un trafic indésirable. Un système qui permet le contrôle de réseau collaboratif utilise une technologie de moteur de jeu en 3D, comme le décrit Harrop et Armitage [49, 50]. En traduisant les événements du réseau en activités visuelles, l'approche suggérée utilise des techniques d'interaction simples.

2.9. Conclusion :

Dans ce chapitre nous avons vu de près le déploiement des systèmes Darknet et leur fonctionnement, ainsi que les recherches les plus importantes concernant l'analyse et la visualisation des données Darknet et leurs exploitations pour la cyber intelligence. Nous avons présenté également les projets les plus importants dans ce domaine. Compte tenu de l'absence de projets similaires en Algérie et de l'existence d'un seul projet en Afrique du Sud, l'espace d'adressage Internet en Afrique n'est pas surveillé et peut être considéré comme le plus susceptible d'être utilisé dans des activités malveillantes. Par conséquent, il est nécessaire d'aller vers des projets pour surveiller cet espace d'adressage.

CHAPITRE 03

Résultats du Modèle DNN pour la Détection des Attaques Darknet

3.1. INTRODUCTION

Ce chapitre présente la réalisation d'un modèle de réseaux de neurones profonds (DNN) pour détecter les attaques sur le darknet à l'aide de diverses méthodes et approches d'apprentissage profond. De plus, nous avons évalué les performances de ces modèles en termes de précision, de rappel, de F1-score et d'exactitude. Ce travail vise à améliorer la robustesse de l'infrastructure en utilisant des techniques d'apprentissage profond pour protéger le réseau contre les activités malveillantes. Pour atteindre cet objectif, nous avons utilisé l'ensemble de données CIC-Darknet2020, un ensemble de données complet et largement utilisé dans l'industrie de la sécurité des réseaux.

3.2. Environnement de Travail

3.2.1. Plateforme Utilisée - Kaggle

La plateforme Kaggle, spécialement conçue pour les experts en données et les passionnés d'apprentissage automatique, est un espace communautaire en ligne qui favorise la collaboration avec d'autres utilisateurs. De plus, elle permet de trouver et de publier des ensembles de données. Nous avons choisi Kaggle comme plateforme pour nous aider dans cette recherche en raison de ses nombreux avantages. Elle fournit un accès simple à des ressources de calcul haute performance telles que les CPU, GPU et TPU pour des tâches intensives de traitement. De plus, Kaggle facilite l'importation directe de données en ligne sans avoir à les télécharger sur votre ordinateur personnel, ce qui peut être très utile pour travailler avec de grands ensembles de données. Enfin, la plateforme propose des outils de visualisation et d'analyse des données ainsi que des capacités de collaboration et de partage de projets

3.2.2. Langage Utilisé

3.2.2.1. Python

Le langage Python est un langage de programmation open source, multiplateforme et orienté objet. Grâce à ses bibliothèques spécialisées, Python est utilisé dans de nombreuses situations comme le développement logiciel, l'analyse de données, ou la gestion d'infrastructures [51].

3.2.3 Les bibliothèques

Le tableau 3.1 liste les bibliothèques python les plus utilisées.

Tableau 3.1 : Liste des bibliothèques python.

Bibliothèque	Description
Pandas [52]	Pandas est une bibliothèque Python avec des outils pour la manipulation et l'analyse de données. Il est fréquemment utilisé pour le nettoyage, la transformation, l'exploration et la visualisation des données.
TensorFlow [53]	TensorFlow est une bibliothèque open source qui se concentre sur la formation en réseaux neuronaux profonds et offre des options pour déployer localement ou dans le cloud des modèles d'apprentissage profond. En raison de ses caractéristiques avancées et de son soutien pour la construction et la formation de modèles d'apprentissage profond en utilisant les dernières technologies, c'est un outil indispensable pour le scientifique de données moderne.
Keras [54]	Keras est une bibliothèque open-source d'apprentissage automatique écrite en Python qui permet de créer, entraîner et déployer facilement des réseaux de neurones artificiels.
Numpy [55]	NumPy est une bibliothèque Python qui fournit un support pour les grandes matrices et ensembles multidimensionnels. Il est couramment utilisé dans le calcul scientifique et l'analyse numérique.
Scikit-Learn [56]	Scikit-learn est une bibliothèque Python qui fournit des algorithmes d'apprentissage automatique pour la classification, la régression, le regroupement et la réduction de la dimensionnalité. Il est couramment utilisé dans l'analyse des données et la modélisation prédictive.

3.3 Dataset-CIC-Darknet2020

La création de ce dataset a été réalisée à travers la sélection des ensembles de données ISCXVPN2016[57] et ISCXTor2017[58], qui capturent le trafic régulier, Non-Tor, Non-VPN, VPN et Tor pour sept catégories d'applications : Audio-Streaming, Chat, File-Transfer, Vidéo-Streaming, Email, et VoIP et P2P. En combinant ces ensembles, ils ont créé un nouvel ensemble de données, nommé « Darknet dataset ». Celui-ci comporte deux couches : la première représente le trafic régulier (bénin) et la deuxième, le trafic anonymisé (darknet) lié aux services cachés. Le dataset contient 158 659 enregistrements, dont 134 348 bénins et 24 311 darknet, avec le streaming audio ayant le plus grand nombre d'échantillons.

3.3.1 Description du dataset CIC-Darknet2020

Ce dataset contient dix fichiers CSV représentant dix jours du flux de réseau capturé, avec plus de 15,8 millions d'échantillons. En outre, plus de 83 caractéristiques ont été extraites (voir le table 3.2).

Tableau 3.2 : Caractéristiques du dataset CIC-Darknet2020.

Nom de dataset	CIC-darknet2020
Type de dataset	Multi-classe
Nombre d'observation	158616
Nombre de features	83

3.3.3.1 Types des colonnes dans le dataset CIC-Darknet2020

Le tableau 3.3 liste les différents « Features » ainsi que leurs types appartenant au dataset CIC-Darknet2020.

Tableau 3.3 : Types des colonnes dans le dataset CIC-Darknet2020.

Features	Type	Features	Type
Src Port	int64	Packet Length Min	int64
Dst Port	int64	Packet Length Max	int64
Protocol	int64	Packet Length Max	float64
Flow Duration	int64	Packet Length Mean	float64
Total Fwd Packet	int64	Packet Length Std	float64
Total Bwd Packets	int64	Packet Length Variance	float64
Total Length of Bwd Packet	int64	FIN Flag Count	int64
Fwd Packet Length Min	int64	SYN Flag Count	int64
Fwd Packet Length Mean	float64	RST Flag Count	int64
Fwd Packet Length Std	float64	PSH Flag Count	int64
Bwd Packet Length Max	int64	ACK Flag Count	int64
Bwd Packet Length Min	int64	URG Flag Count	int64
Bwd Packet Length Mean	float64	CWE Flag Count	int64

Bwd Packet Length Std	float64	Down/Up Ra	int64
Flow Bytes/s	float64	Average Packet Size	float64
Flow Packets/s	float64	Fwd Segment Size Avg	float64
Flow IAT Mean	float64	Bwd Segment Size Avg	float64
Flow IAT Std	float64	Fwd Bytes/Bulk Avg	int64
Flow IAT Max	int64	Fwd Packet/Bulk Avg	int64
Flow IAT Min	int64	Fwd Bulk Rate Avg	int64
Fwd IAT Total	int64	Bwd Bytes/Bulk Avg	int64
Fwd IAT Mean	float64	Bwd Packet/Bulk Avg	int64
Fwd IAT Std	float64	Bwd Bulk Rate Avg	int64
Fwd IAT Max	int64	Subflow Fwd Packets	int64
Fwd IAT Min	int64	Subflow Fwd Bytes	int64
Bwd IAT Total	int64	Subflow Bwd Packets	int64
Bwd IAT Mean	float64	Subflow Bwd Bytes	int64
Bwd IAT Std	float64	FWD Init Win Bytes	int64
Bwd IAT Max	int64	Bwd Init Win Bytes	int64
Bwd IAT Min	int64	Fwd Act Data Pkts	int64
Fwd PSH Flags	int64	Fwd Seg Size Min	int64
Bwd PSH Flags	int64	Active Meam	int64
Fwd URG Flags	int64	Active Std	int64
Bwd URG Flags	int64	Active Max	int64
Fwd Header Length	int64	Active Min	int64
Bwd Header Length	int64	Label	Object
Fwd Packets/s	float64	Label.1	Object

Le dataset CIC-Darknet2020 est une collection de trafic réseau utilisée pour l'analyse des activités sur le darknet. Chaque colonne de ce dataset représente une caractéristique spécifique du trafic réseau.

3.4 Implémentation

La figure 3.1 représente une classification schématique des attaques de Darknet utilisant un modèle de Deep Learning.

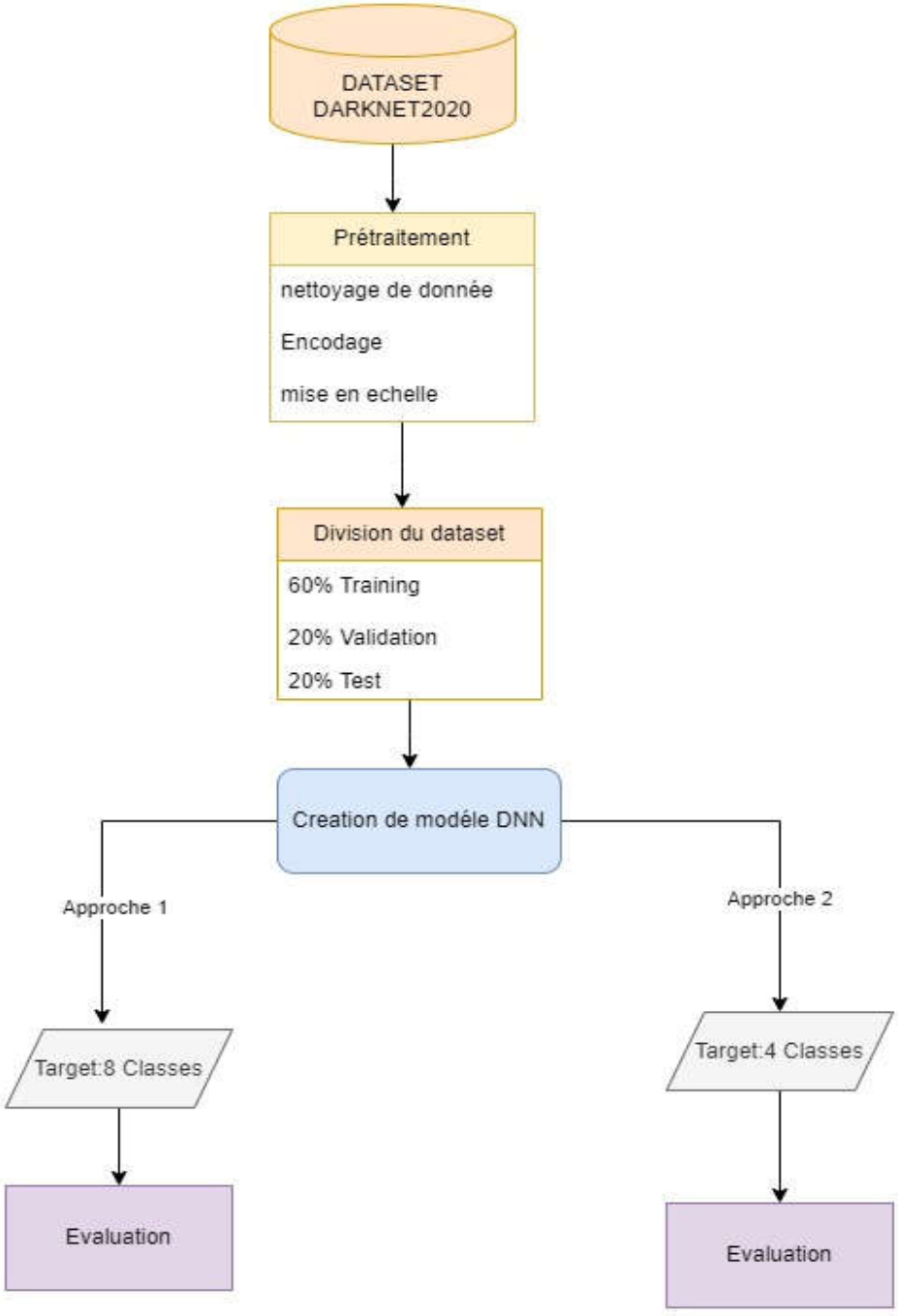


Figure 3.1 : Diagramme de détection des attaques de darknet.

3.5 Prétraitement de dataset

Les algorithmes de machine learning et de deep learning nécessitent des données. Parce que ces données peuvent être incomplètes, dupliquées ou même incorrectes, l'algorithme qui en résultera sera lui-même assez mauvais car il est censé reproduire ce qu'il voit dans les données. Par conséquent, avant d'intégrer ces données dans le modèle d'apprentissage, il est crucial de les prétraiter. Cette étape est appelée prétraitement des données.

3.5.1 Nettoyage des données

Après avoir analysé notre ensemble de données, nous avons éliminé les valeurs manquantes, telles que les valeurs NaN (Not a Number) et INF (Infinity), ainsi que les lignes dupliquées qui pourraient nuire aux performances de notre modèle. Nous avons également identifié et supprimé les caractéristiques non pertinentes pour les attaques (Timestamp, Flow ID, Src IP, idle mean, idle std, idle max, idle min, Dst IP). De plus, nous avons retiré les fonctionnalités contenant des valeurs nulles (Bwd PSH Flags, Bwd URG Flags, Fwd Byts/b Avg, Fwd Pkts/b Avg, FWD Blk Rate Avg, Bwd Bytes/b Avg, Bwd Pkts/b Avg, Bwd Rate Blk Avg) car leur inclusion pourrait introduire des erreurs et diminuer la précision de notre modèle.

3.5.2 Encodage des données

Le dataset peut inclure des variables catégorielles qui doivent être converties en variables numériques pour être utilisables par les algorithmes d'apprentissage automatique. Les techniques de conversion incluent l'encodage par étiquette, l'encodage label ou l'encodage label.1.

3.5.3 Standardisation

La standardisation est une étape clé en prétraitement des données dans l'apprentissage automatique. Elle consiste à redimensionner les données pour que les caractéristiques comportent des valeurs centrées autour de 0 et une variance de 1. Cela est particulièrement utile lorsque les données d'entrée varient sur des échelles très différentes, ce qui pourrait nuire aux performances des algorithmes d'apprentissage automatique, en particulier ceux exploités sur des distances comme les k-plus proches voisins (KNN) ou les régressions linéaires.

3.6 Division du dataset

Le jeu de données a été divisé en trois parties. L'ensemble de test représente 20 % des données, tandis que l'ensemble d'entraînement représente 60 % et l'ensemble de validation représente 20% (voir la figure 3.2).

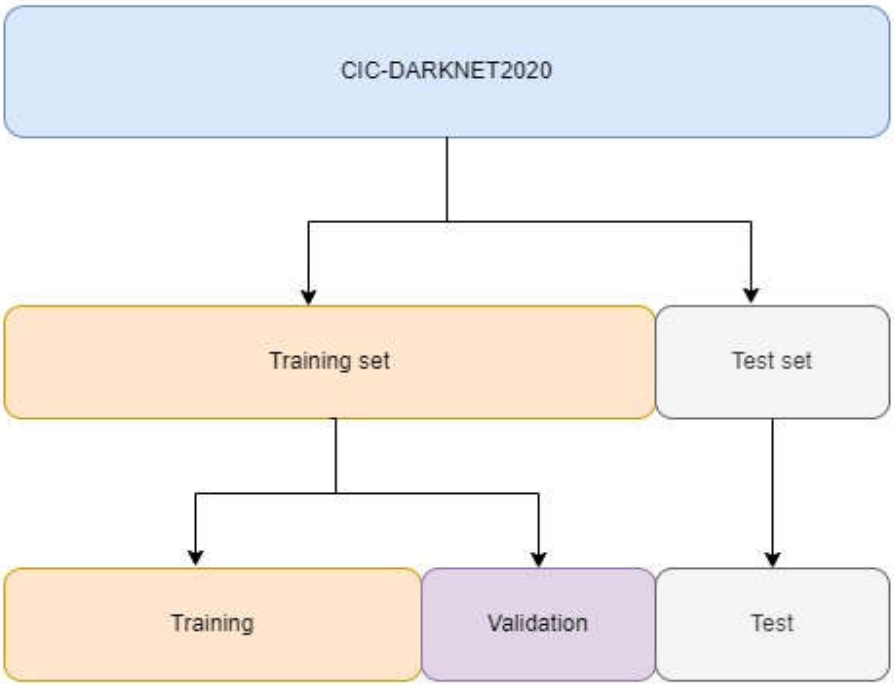


Figure 3.2 : Division du dataset.

3.7 Les hyper-paramètres

Nous avons dû sélectionner les hyper-paramètres appropriés pour entraîner notre modèle afin qu'il fonctionne le mieux sur notre tâche de détection du trafic du darknet. Nous avons commencé par choisir l'algorithme d'apprentissage profond approprié pour notre tâche de classification multi-classe. Pour optimiser les performances de notre modéle, nous avons modifié plusieurs hyper-paramètre. Pour commencer, nous avons choisi le nombre de couches et de neurones à utiliser dans notre réseau de neurones. Nous avons essayé plusieurs configurations et avons finalement choisi une architecture avec sept couches cachées (unités=1024, 256, 128, 64, 32 et 16) avec une fonction d'activation "relu" et une couche de sortie (unités=4 dans la classification du trafic et unités=8 dans la classification des applications) avec une fonction d'activation "sofmax". Pour optimiser notre modèle, nous avons également ajusté le taux d'apprentissage et le nombre d'itérations (époch = 400) (voir la figure 3.3).

Model: "sequential_1"

Layer (type)	Output Shape	Param #
dense_8 (Dense)	(None, 1024)	77,824
dense_9 (Dense)	(None, 512)	524,800
dense_10 (Dense)	(None, 256)	131,328
dense_11 (Dense)	(None, 128)	32,896
dense_12 (Dense)	(None, 64)	8,256
dense_13 (Dense)	(None, 32)	2,080
dense_14 (Dense)	(None, 16)	528
dense_15 (Dense)	(None, 4)	68

Tableau 3.4: Organisation des couches de notre modèle.

Nous avons effectué plusieurs essais pour déterminer les hyper-paramètres les plus appropriés pour notre Modèle. Nous avons utilisé la technique de régularisation L2 qui limite le sur-apprentissage en punissant les coefficients trop élevés. Nous avons fixé le paramètre de régularisation L2 à 0.00001 après plusieurs tests, ce qui a permis d'obtenir des résultats très satisfaisants. Nous avons découvert que ce paramètre de régularisation nous permettait d'obtenir une meilleure généralisation de notre modèle, c'est-à-dire que notre modèle était capable de mieux généraliser ses prédictions sur des données qu'il n'avait pas vues auparavant. Nous avons pu obtenir des résultats satisfaisants tout en réduisant le risque de sur-apprentissage et en améliorant la capacité de généralisation de notre modèle grâce à ce paramètre de régularisation.

3.8 Classification ‘multi-classe’

Nous avons utilisé deux techniques pour l’encodage de la colonne "Label" qui est notre target.

3.8.1 Label Encodage

La méthode de Label Encoding est une technique utilisée pour convertir des labels catégoriels en valeurs numériques, généralement sous la forme d'entiers. Cela est souvent nécessaire car les algorithmes de machine learning travaillent mieux avec des données numériques. Label Encoding attribue un entier unique à chaque catégorie distincte dans les labels.

- **Trafic du Label Encoding**

Un nombre entier. Cette méthode vise à faciliter l'entraînement du modèle.

Dans ce cas, nous utiliserons cette technique pour la variable cible du dataset appelée "Label" qui comprend quatre classes distinctes : Non-Tor, NonVPN et VPN, Tor.

Avec la méthode de "Label Encodage", les catégories Non-Tor, NonVPN, VPN et Tor seront remplacées respectivement par les entiers 0, 1, 2 et 3 (voir la figure 3. 4).

Tableau 3.5: Label Encoding

Label	ID
Non-Tor	0
NonVPN	1
VPN	2
Tor	3

- **Application du label.1 encoding:**

Un nombre entier. Cette méthode vise à faciliter l'entraînement du modèle.

Dans ce cas, nous utiliserons cette technique pour la variable cible du dataset appelée "Label.1" qui comprend huit classe distinctes : P2P, Browsing, Audio-Streaming, Chat, File-Transfer, Vidéo-Streaming, Email, et VOIP.

Avec la méthode de "Label Encodage", les catégories P2P, Browsing, Audio-Streaming, Chat, File-Transfer, Vidéo-Streaming, Email, VOIP seront remplacées respectivement par les entiers 0, 1, 2, 3, 4, 5, 6, 7 (voir le tableau 3.5).

Tableau 3.6: Label.1 Encoding

Label .1	ID
P2P	0
Browsing	1
Audio-Streaming	2
Chat	3
File-Transfer	4
Vidéo-Streaming	5
Email.	6
VOIP	7

Le label encoding est une méthode efficace pour convertir des données catégorielles en valeurs numériques. Dans notre projet de classification du trafic, nous avons utilisé cette méthode pour les cibles "Label" et "Label.1", ce qui a permis de préparer les données pour l'entraînement de notre modèle de machine learning. Cette transformation a facilité l'intégration des données catégorielles dans notre pipeline de machine learning et a contribué à améliorer la performance globale de notre modèle.

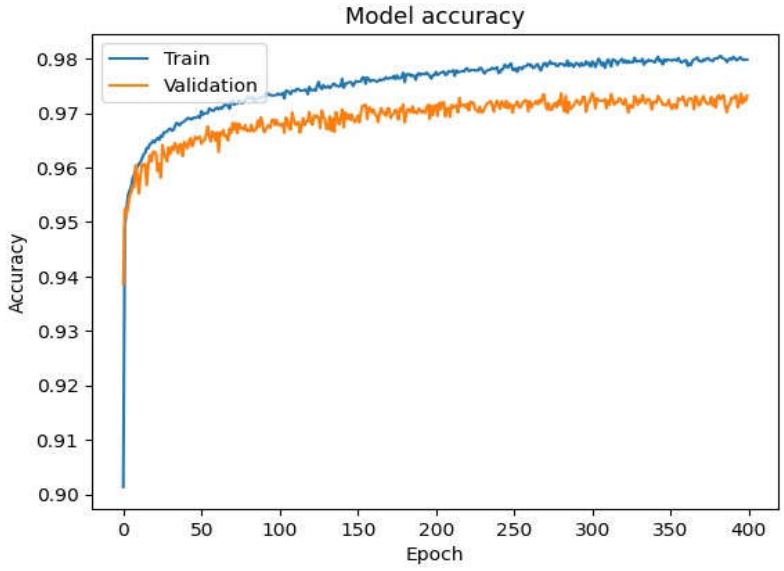
3.9 Résultats et Discussion

Nous avons implémenté un modèle d'apprentissage en profondeur DNN avec deux différentes approches. Cette section présente les résultats des tests réalisés pour la configuration des modèles. Plusieurs tests ont été effectués pour obtenir les bons hyper-paramètres pour chaque modèle. Ces paramètres comprennent, par exemple, le nombre de couches, le nombre de nœuds dans chaque couche, le nombre d'itérations (epochs), type d'optimiseur et la fonction d'activation pour les couches cachées. Nous avons ensuite entraîné chaque modèle sur l'ensemble d'entraînement tout en le validant sur l'ensemble de validation. Nous avons répété ce processus avec diverses valeurs de paramètres jusqu'à ce que nous ayons trouvé les meilleurs résultats. Nous avons enfin confirmé les performances de notre modèle en le testant sur l'ensemble de test.

Le modèle qui présente la plus grande "accuracy" et le plus faible taux d'erreur (loss) est considéré comme le meilleur modèle.

3.9.1 Classification du trafic

La figure 3.5 montre le "Loss" et l'"Accuracy" du modèle en fonction des itérations au cours du processus d'apprentissage.



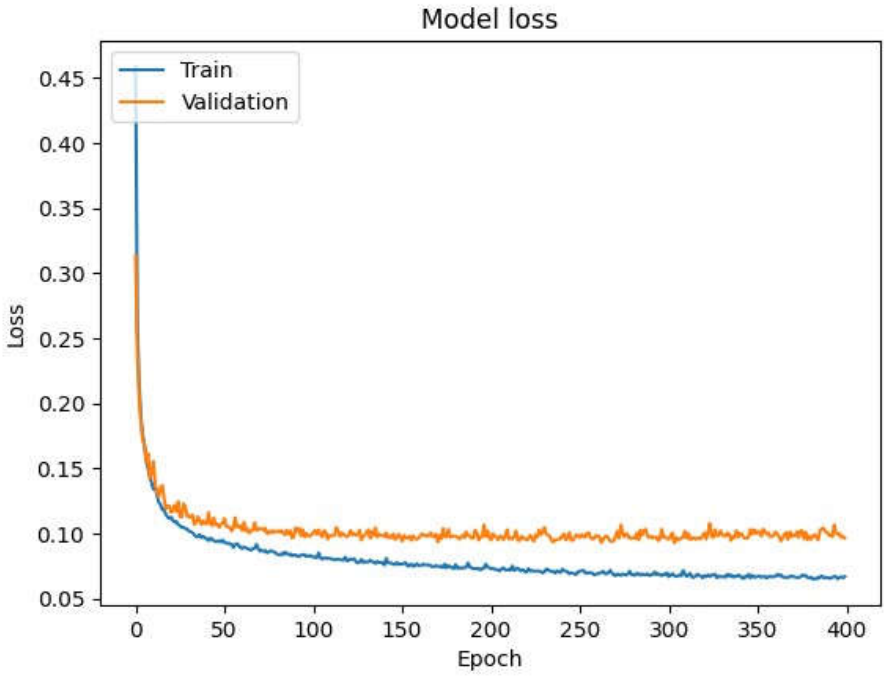


Figure 3.3: Précision et perte (4 classes).

Cette approche a produit des résultats très satisfaisants avec une "Accuracy" de 97.37% pour les données d'entrainement et de 97.32% pour les données de validation.

Les résultats de cette approche montrent également une fiable valeur de "Loss" pour les données de test et de validation, avec une perte de 0.75% pour les données d'entrainement et de 0.90% pour les données de validation.

Le tableau 3.4 affiche les métriques "Accuracy", "Precision", "Recall" et "F1-score" pour les 4 classes Non-Tor, NonVPN, VPN et Tor dans le cas d'une classification du trafic. La matrice de confusion correspondante est donnée par la figure 3.6 pour chaque classe (Non-Tor, NonVPN, VPN, Tor) dans le cas de classification du trafic.

Tableau 3.7: Rapport de classification du trafic

Label	Accuracy	Précision	F1-score	Recall
Non-Tor	99,84%	100%	100%	100%
NonVPN	91,15%	90%	92%	93%
VPN	92,47%	95%	90%	85%
Tor	84,08%	93%	92%	90%

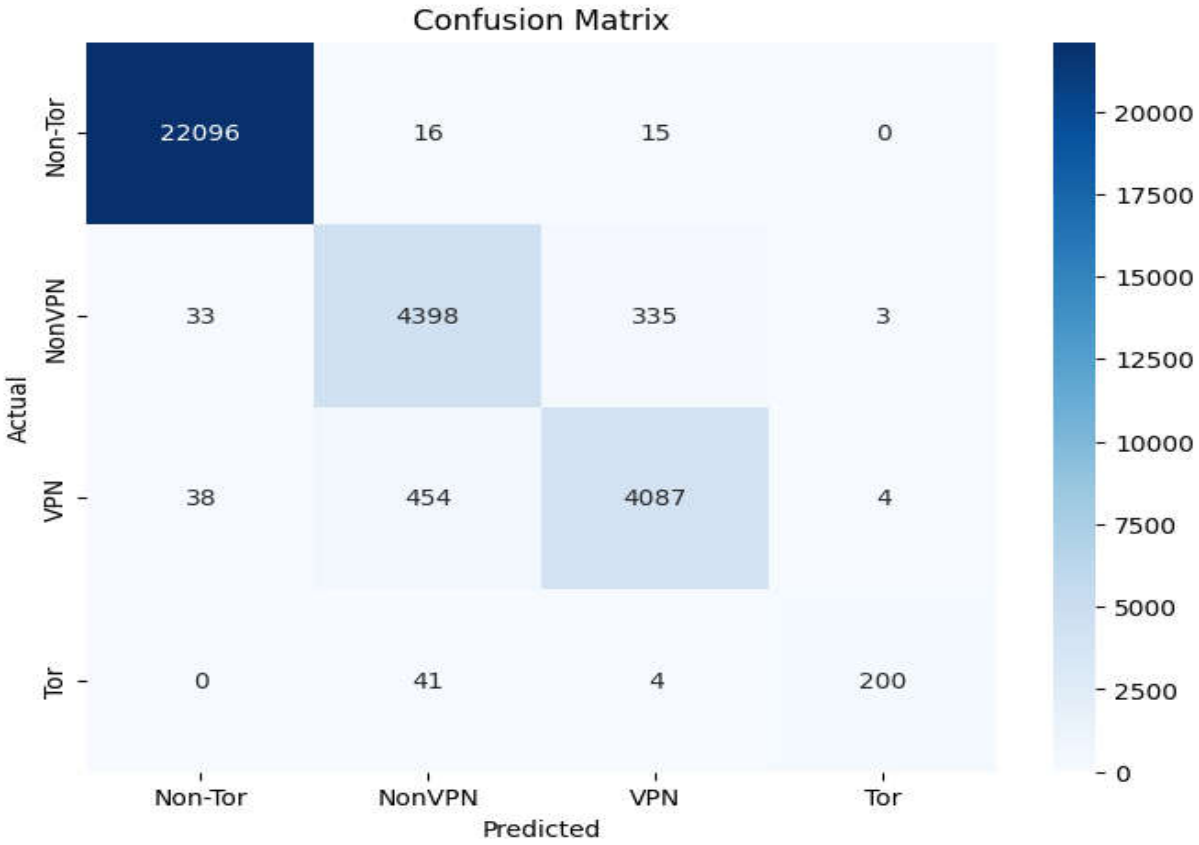


Figure 3.4: Matrice de confusion pour la classification du trafic.

3.9.2 Classification des applications

La figure 3.7 montre le "Loss" et l'"Accuracy" du modèle en fonction des itérations au cours du processus d'apprentissage.

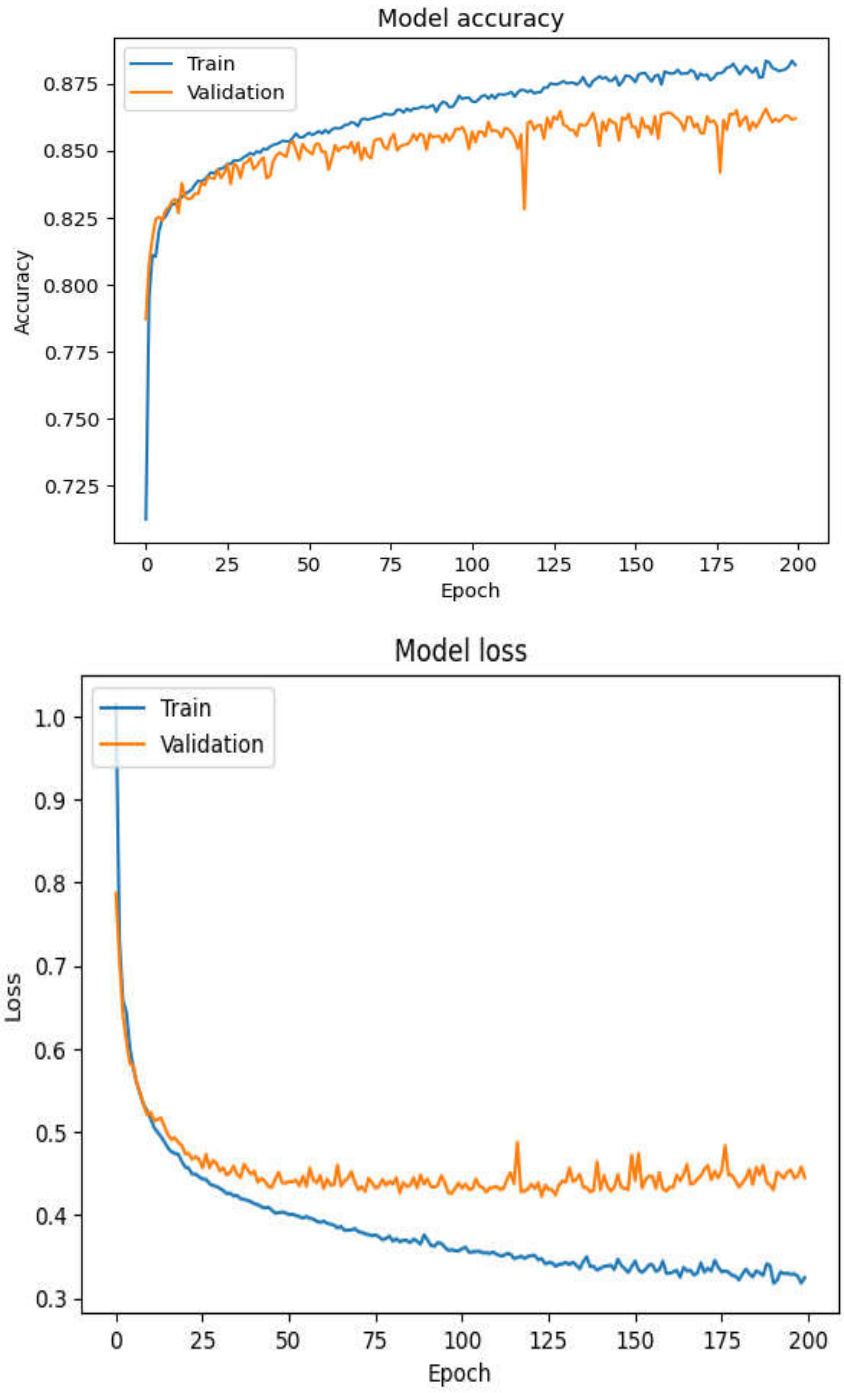


Figure 3.5: Précision et perte (8 Classe).

Cette approche a également produit des résultats très satisfaisants avec une précision de 87.5% pour les données d'entrainement et de 85% pour les données de validation.

Cette méthode a également produit des résultats satisfaisants pour la perte, avec une perte de 0.35% pour les données d'entrainement et de 0.48% pour les données de validation peuvent

être considérées comme bonnes car elles sont assez faibles, ce qui indique que le modèle a bien appris à classer les données.

Le tableau 3.5 affiche les métriques "Accuracy", "Precision", "Recall" et "F1-score" pour les 8 classe P2P, Browsing, Audio-Streaming, Chat, File-Transfer, Vidéo-Streaming, Email, VOIP dans le cas d'une classification multi-classe. La matrice de confusion correspondante est donnée par la figure 3.8 pour chaque classe (P2P, Browsing, Audio-Streaming, Chat, File-Transfer, Vidéo-Streaming, Email, VOIP) dans le cas de classification des applications.

Tableau 3.8: Rapport de classification des applications.

Label	Accuracy	Précision	F1-score	Recall
P2P	**98,97%**	77%	94%	96%
Browsing	**96,47%**	83%	85%	88%
Audio-streaming	**88,02%**	99%	99%	99%
Chat	**88,38%**	61%	72%	88%
File-Transfer	**76,03%**	71%	59%	51%
Vidéo-Streaming	**50,64%**	92%	94%	96%
Email	**33,94%**	77%	77%	76%
VOIP	**24,58%**	92%	94%	96%

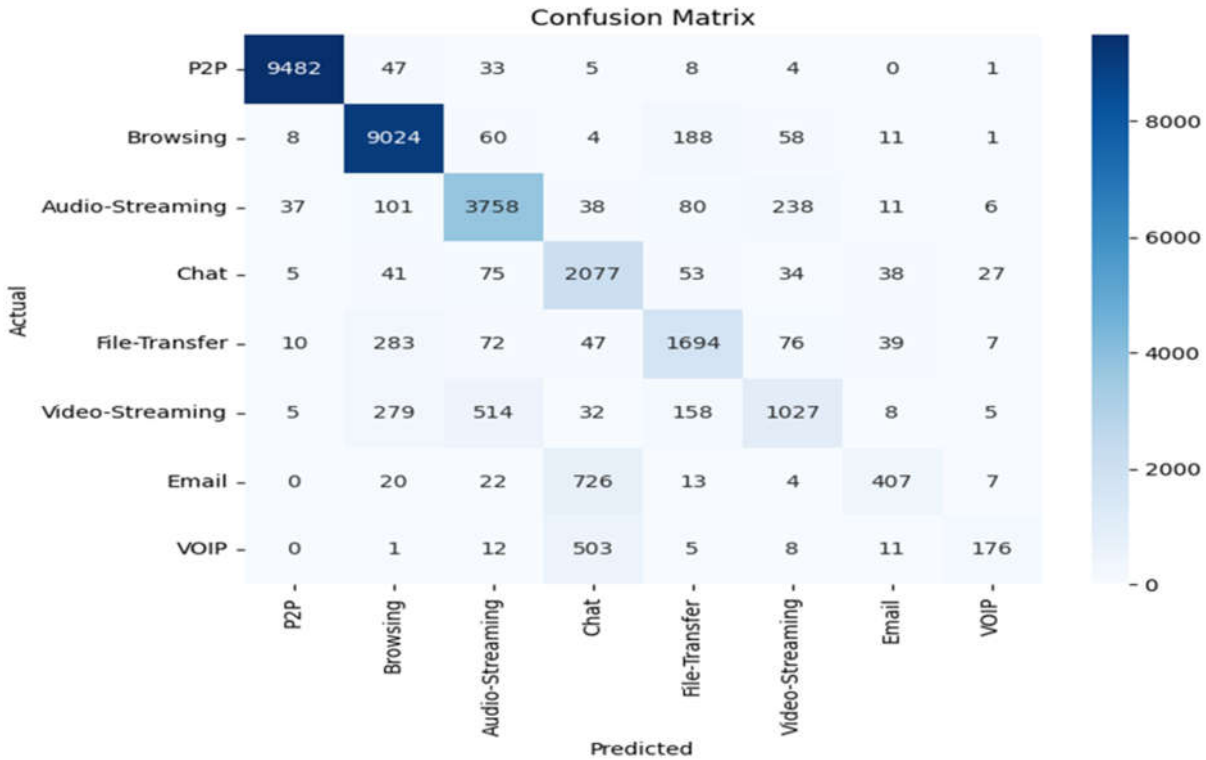


Figure 3.6: Matrice de confusion pour la classification du l'application

3.9.3 Comparaison des algorithmes

Tableau3.9: comparaison des performances des algorithmes (classification du trafic).

Métriques / Algorithme	Accuracy	Précision	F1-score	Recall
Knn	93,79%	52,75%	52,75%	52,75%
GBC	96,75%	52,75%	52,75%	52,75%
XGBoost	98,50%	52,27%	52,27%	52,27%
DNN(6 couches)	97,35%	97,32%	97,32%	97,32%

Tableau 3.10 : Comparaison des performances des algorithmes (classification des application).

Métriques / Algorithme	Accuracy	précision	F1-score	Recall
KNN	88,05%	86 ,98%	86,98%	86,98%
XGBoost	92,03%	21,84%	21,84%	21,84%
GBC	89,58%	21,84%	21,84%	21,84%
DNN	87,14%	87,14%	87,14%	87,14%

Pour la classification du trafic, l'algorithme DNN (6 couches) semble être le meilleur choix, offrant un excellent équilibre entre accuracy, précision, F1-score et rappel.

Pour la classification des applications, l'algorithme KNN et DNN sont tous deux performants, mais KNN a une accuracy légèrement meilleure. Cependant, DNN offre une performance plus équilibrée sur toutes les métriques.

Ces résultats montrent l'importance de choisir un algorithme adapté à la tâche spécifique et de considérer plusieurs métriques de performance pour une évaluation complète.

3.9.4 Analyse des résultats

Les résultats de la classification du trafic et la classification des applications ont produit des résultats très similaires et très performants en termes d'exactitude et de perte, tandis que l'encodage par étiquette a montré des performances légèrement inférieures en termes d'exactitude et de perte.

L'approche classification des applications a produit des résultats impressionnants avec une accuracy parfaite et une très faible valeur de loss. Cependant, il est important de noter que ces résultats peuvent varier en fonction du dataset et des objectifs de l'analyse, et qu'il est donc important de choisir la méthode la plus adaptée.

On peut également remarquer que toutes les valeurs de perte sont très faibles. Techniques, ce qui suggère que les modèles ont réussi à réduire l'erreur de prédiction.

L'objectif de notre étude était de déterminer comment l'équilibrage des données affecte les résultats d'apprentissage et quel offre les meilleurs résultats pour une tâche de classification

multi-classe.

Nos résultats ont montré que la balance des données améliorait significativement la précision, le F1-score et le Recall pour les algorithmes de DL mais que accuracy c'était bon. Cependant, le DL était plus rapide et plus facile à mettre en oeuvre.

3.10 Conclusion

Enfin, notre troisième chapitre présente les résultats de plusieurs essais sur la détection des attaques darknet à l'aide du deep learning. Nous avons comparé diverses mesures de performance, y compris la précision, la f1-score, accuracy, recall. Des résultats encourageants de notre dernier essai démontrent l'efficacité de notre modèle dans la détection d'attaques darknet. De plus, notre étude, comparée à d'autres résultats existants, a confirmé que notre méthode est compétitive. Ces résultats offrent des perspectives prometteuses pour l'utilisation du deep learning dans le domaine de la détection des attaques darknet.

Conclusion générale

L'identification du trafic du darknet est un enjeu majeur en cybersécurité, permettant aux criminels de mener des activités illégales en ligne tout en restant anonymes. Les modèles d'apprentissage automatique supervisé se sont révélés être des outils prometteurs pour relever ce défi, offrant une approche automatisée et efficace pour classifier le trafic réseau et détecter les comportements malveillants. Ce mémoire de master explore l'application de modèles d'apprentissage automatique supervisé pour l'identification du trafic du darknet, tels que les réseaux de neurones artificiels, et leurs performances dans la classification du trafic réseau ont été évaluées Les résultats indiquent que les modèles d'apprentissage automatique supervisé peuvent être extrêmement efficaces pour détecter le trafic du darknet, atteignant des taux de précision élevés même avec des ensembles de données restreints. De plus, ces modèles peuvent s'adapter à l'évolution des menaces, offrant ainsi une solution flexible et évolutive pour la cybersécurité. Cependant, plusieurs défis doivent être relevés pour une adoption plus large de ces modèles dans la lutte contre le trafic du darknet. L'un des principaux obstacles est la nécessité de disposer de données d'entraînement de haute qualité et bien étiquetées. De plus, la complexité de l'interprétation des résultats de ces modèles peut limiter leur acceptation par les experts en cybersécurité. En dépit de ces défis, les modèles d'apprentissage automatique supervisé ont le potentiel de transformer la lutte contre le trafic du darknet. En surmontant les obstacles actuels, ces modèles peuvent devenir des outils essentiels pour protéger les réseaux et les utilisateurs contre les menaces du darknet

Références Bibliographiques

[1] la rédaction de futura. (2018) Qui sont les pionniers de l'intelligence artificielle ? [Online]. Available : https ://www.futurasciences.com/tech/questions-reponses/intelligence-artificielle-sontpionniers-intelligence-artificielle-4907/
[2] A. Cornuéjols, L. Miclet, and V. Barra, Apprentissage artificiel deep learning concepts et algorithmes. Eyrolles, 2018.
[3] A. Cornuéjols, L. Miclet, and Y. Kodratoff, Apprentissage artifi- ciel,concepts et algorithmes. Eyrolles, 2002.
[4] H. Larochelle, "Étude de techniques d'apprentissage non supervisé pour l'amélioration de l'entrainement supervisé de modèles connexionnistes," Ph.D. dissertation, Université de Montréal, 2008.
[5] A. Cornuéjols and L. Miclet, Apprentissage artificiel,concepts et algorithmes. Eyrolles, 2010.
[6] K. Gosalia and R. Lefebvre, Introduction à l'apprentissage automatique. CPA New Brunswick, 2019.
[7] M. Taffar. (2018) Initaition a l'apprentissage automatique. [Online]. Available : https ://www.electronique-mixte.fr/formationpdf/formation-pdf-intelligence-artificielle/cours-53-initiation-alapprentissage-automatique/
[8] S. Kotsiantis. "Supervised machine learning: A review of classification techniques". Informatica Journal, 31 :249–268 (2007).
[9] R. O. Duda et al, "Pattern Classification", chapter : Unsupervised Learning and Clustering. Wiley Inter science (2001).
[10] O. Chapelle, "Semi-supervised Learning". MIT Press (2006) [5] R. Sutton et al, "Reinforcement Learning - An Introduction", MIT Press (2012).
[11] B. Orsier. Etude et application de systémes hybrides neurosymbolique. Thèse de doctorat, Universitié Joseph Fourrier Grenoble, 1995.
[12] H.Mezaache, "Les réseaux de Neurones formels Et Les systèmes Neuro-Flous Pour l'apprentissage par renforcement", Mémoire soumis en vue de l'obtention du Diplôme de Magister, Université El Hadj Lakhdar, Batna, 2008.
[13] P. Campolucci,A circuit Theory Approach to Recurrent Neural Network Architecture and Learning Methods, These Doctorat, Université de Bologne, Février1998.
[14] G.dreyfus, J.-M.martinez, M. Samuelides, M.B.Gordon, S.Thiria, L.Hérault, "réseaux de neurones, méthodologie et applications",2002.
[15] Mourad ABIDI,' Réalisation et implantation d'un réseau de neurones sur une architecture matérielle distribuée à base de réseau sur puce', mémoire de PFE d'ingénieur, Ecole Nationale d'Ingénieurs de Sousse, Juin 2014.
[16] M.R. Alismail, N.Ourchani., "Fusion multimodale des scores pour lareconnaissancedes personnes", Master 2, Université Mohamed Khider Biskra,2011.

[17] https://www.mathworks.com/help/nnet/ug/introduction- to-convolutional-neural networks.html.

[18] [chap2,2018]: « technique neuronales adaptées aux données spatio- temporelles, chapitre II ».

[19] Tseng, Wen-Kung, Lu, and Chung-Sheng, "The system for appraisal of vehicle accident based on radial basis function neural networks," in DOI: 10.1109/ICNC.2011.6022220, 2011, pp. 869–872.

[20] Y. Djeriri. (2017) les réseaux de neurones artificiels. [Online]. Available : https ://www.academia.edu/37046526/Les-RC3A9seaux-de-NeuronesArtificiels

[21] G. Deyfus, "Les reseaux de neurones," 1998. [Online]. Available : https ://www.neurones.espci.fr/Articles-PS/GAMI.pdf

[22] A. Pottiez and M. Duquesnoy. (2018) Intelligence artificielle et apprentissage de réseaux connexionnistes. [Online].Available: http ://math.univlille1.fr/calgaro/TER-2018/wa-files/duquesnoy-pottiez.pdf

[23] Tianjin Fu, Ahmed Abbasi and Hsinchun Chen, "A Focused Crawler for Dark Web Forums", JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, 61(6):1213–1231, 2010

[24] A. Salvail-Bérard. (2012) Réseaux de neurones. [Online].Available: http ://docplayer.fr/84604798-Reseaux-de-neurones-adamsalvail-berard-6-septembre-2012.html

[25] A. Cornuéjols. (2007) Les réseaux de neurones. [Online].Available: https ://www.lri.fr/antoine/Courses/ENSTA/Tr-ensta-nnx9.pdf

[26] Zakir Durumeric, Michael Bailey, and J Alex Halderman. An internetwide view of internet-wide scanning. In USENIX Security Symposium,2014.

[27] Rossow, Christian. (2014). Amplification Hell: Revisiting Network Protocols for DDoS Abuse. 10.14722/ndss.2014.23233.

[28] Xu, Ruomeng & Cheng, Jieren & Wang, Fengkai & Tang, Xiangyan & Xu, Jinying. (2019). A DRDoS Detection and Defense Method Based on Deep Forest in the Big Data Environment. Symmetry. 11. 78. 10.3390/sym11010078.

[29] Irwin, Barry. (2011). A framework for the application of network telescope sensors in a global IP network.

[30] Wustrow, Eric & Karir, Manish & Bailey, Michael & Jahanian, Farnam & Huston, Geoff. (2010). Internet Background Radiation Revisited. Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC. 62-74. 10.1145/1879141.1879149.

[31] Michael Bailey, Evan Cooke, Farnam Jahanian, Andrew Myrick, and Sushant Sinha. Practical darknet measurement. In 40th Annual Conference on Information Sciences and Systems, pages 1496–1501. IEEE, 2006.

[32] Harrop, W & Armitage, G. (2005). Defining and Evaluating Greynets (Sparse Darknets). 344- 350. 10.1109/LCN.2005.46.

[33] Yu Jin, Zhi-Li Zhang, Kuai Xu, Feng Cao, and Sambit Sahu. Identifying and tracking suspicious activities through IP gray space analysis. In Proceedings of the 3rd annual ACM workshop on Mining network data, pages 7–12. ACM, 2007
[34] I. Polakis, G. Contains, S. Ioannidis, and E. P. Markatos. Dynamic Monitoring of Dark IP Address Space (Poster). in International Work shop on Traffic Monitoring and Analysis. Springer, 2011, pp. 193–196.
[35] Michael Bailey, Evan Cooke, Farnam Jahanian, Jose Nazario, and David Watson. The Internet Motion Sensor: A distributed blackhole monitoring system. In Proceedings of Network and Distributed System Security Symposium (NDSS '05), San Diego, CA, February 2005.
[36] CAIDA: The UCSD Network Telescope. http://www.caida.org/projects/ network tele CAIDA. Conficker/Conflicker/Downadup as seen from the UCSD
[37] Network Telescope. http://www.caida.org/research/security/ms08-067/ conficker. xmlscope.
[38] Arbor Networks. ATLAS https://www.netscout.com/product/atlas intelligence-feed-aif
[39] Team Cymru: The Darknet Project https://www.team cymru.com/darknet.html
[40] Spiros Antonatos, Kostas Anagnostakis, and Evangelos Markatos. (2007) Honey@home: a new approach to large-scale threat monitoring. In Proceedings of the ACM workshop on recurring malcode, pages 38– 45.
[41] NoAH project. http://www.fp6-noah.org.
[42] Daisuke Inoue, Mio Suzuki, Masashi Eto, Katsunari Yoshioka, and Koji Nakao. (2009) DAEDALUS: Novel Application of Large-Scale Darknet Monitoring for Practical Protection of Live Networks. In Recent Advances in Intrusion Detection, pages 381–382. Springer.
[43] Japan cert coordination center. http://www.jpcert.or.jp
[44] Barry Vivian William Irwin. (2011) A framework for the application of network telescope sensors in a global IP network. PhD thesis, Rhodes University.
[45] Do Quoc Le, Taeyoel Jeong, H. Eduardo Roman, and James Won-Ki Hong. (2011). Traffic dispersion graph based anomaly detection. In Proceedings of the Second Symposium on Information and Communication Technology, SoICT, pages 36–41, New York, NY, USA.
[46] Cliff Joslyn, Sutanay Choudhury, David Haglin, Bill Howe, Bill Nickless, and Bryan Olsen. (2013) Massive scale cyber traffic analysis: a driver for graph database research. In First International Workshop on Graph Data Management Experiences and Systems, page 3.
[47] Krasser, Sven & Conti, Gregory & Grizzard, Julian & Gribschaw, Jeff & Owen, Henry. (2005). Real-time and forensic network data analyis using animated and coordinated visualization. Proceedings from the 6th Annual IEEE System, Man and Cybernetics Information Assurance Workshop, SMC 2005. 2005. 42 - 49. 10.1109/IAW.2005.1495932.
[48] Fukuda, Kensuke & Fontugne, Romain. (2010). Estimating Speed of Scanning Activities with a Hough Transform. IEEE International Conference on Communications. 1-5. 10.1109/ICC.2010.5502264.
[49] Harrop, Warren & J. Armitage, Grenville. (2006). Real-time collaborative network monitoring and control using 3D game engines for representation and interaction. 31-40. 10.1145/1179576.1179583.
[50] Harrop, Warren & J. Armitage, Grenville. (2006). Modifying first person shooter games to

perform real time network monitoring and control tasks. 10. 10.1145/1230040.1230074.
[51] Python. [En ligne] : https://www.futura-sciences.com/tech/definitions/informatique-python-19349/ [05/5/2024].
[52] “Pandas documentation — pandas 1.5.3 documentation.” (), [En ligne]. Disponible : https://pandas.pydata.org/docs/ (visité le 04/27/2024)
[53] Documentation API | TensorFlow,” TensorFlow. (), [En ligne]. Disponible : https : //www.tensorflow.org/api_docs (visité le 04/27/2024).
[54] W. McKinney, "Data Structures for Statistical Computing in Python," Proceedings of the 9th Python in Science Conference, pp. 51-56, 2010.
[55] NumPy 1.21.6 notes de sortie — NumPy v1.24 manuel.” (), [En ligne]. Disponible : https://numpy.org/doc/stable/release/1.21.6-notes.html (visité le 04/27/2024).
[56] “Scikit-learn : Machine learning in python — documentation de la version 1.2.2 de l’apprentissage en python.” (), [En ligne]. Disponible : https : / / scikit - learn. Org / stable/ (visité le 04/27/2024)
[57] Ali A. Ghorbani. 2016. Characterization of Encrypted and VPN Traffic Using Time-Related Features. In Proceedings of the 2nd International Conference on Information Systems Security and Privacy. 407–414.
[58] Arash Habibi Lashkari, Gerard Draper-Gil, Mamun Seiful Islam, and Ali Ghorbani. 2017. Characterization of Tor Traffic Using Time Based Features. In the proceeding of the 3rd International Conference on Information System Security and Privacy, SCITEPRESS. 253–262.