

الجمهورية الجزائرية الديمقراطية الشعبية
REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
جامعة سعيدة – د. مولاي الطاهر
UNIVERSITÉ DE SAÏDA – Dr MOULAY TAHAR



Faculté des Sciences et de la Technologie

Département d'Electronique

PROJET DE FIN D'ÉTUDE

Présenté Pour l'Obtention du Diplôme de Master en Electronique

Spécialité : Electronique des Systèmes Embarqués

L'algorithme de classification K-NN : Application dans les systèmes optiques

Présenté par :

ZEGDOU Khaoula

Soutenu le 19/06/2025, devant le jury composé de :

Abdelkader Harzellaoui	MCB	Université de Saïda, Dr Toulay Tahar	Président
Ahmed Garadi	Pr	Université de Saïda, Dr Toulay Tahar	Encadreur
Tourkya Abes	MCB	Université de Saïda, Dr Toulay Tahar	Examineur

Année universitaire : 2024 - 2025

Dédicace

Je rends grâce à Dieu de m'avoir donné le courage et la volonté.

Je dédie ce modeste travail:

À mes très chères :

*À celle qui est toujours à côté de mon cœur, à celle qui m'apprit le vrai
Sens de la vie, à celle qui n'a hésité aucun moment à m'encouragé **Ma**
Chère mère.*

*À celui qui m'a toujours appris comment réfléchir avant d'agir, à celui
Qui m'a soutenu tout au long de ma vie scolaire, à celui qui n'a jamais
Épargné un effort pour mon bien, **mon cher père.***

*A mes sœurs et mon frère : **Maha, Chizou et Nouri.***

*Et bien sûr à toute la famille: " **Zegdou**".*

A tous les enseignants et étudiants de électronique de système embarqué.

Atout mes collègues.

Khaoula

Remerciement

Nous tenons à remercier :

*Dans un premier temps remercier le directeur de mémoire Pr Garadi
Ahmed pour sa disponibilité et surtout ses judicieux conseils.*

*Toutes les personnes qui ont contribué au succès de nos travail et qui
nous avons aidée lors de la rédaction de ce mémoire.*

Toute l'équipe pédagogique de l'université de Dr Molay Tahar.

*Les intervenants professionnels responsables de notre formation,
pour avoir assuré la partie pratique de celle-ci.*

Tous les membres du jury, chacun par son nom.

*A tous ceux et celles qui par leurs conseils et leurs compétences ont
Permis la réalisation de ce mémoire.*

Résumé

Cette recherche vise à améliorer la détection et la classification des pannes dans les réseaux optiques passifs gigabit (GPON) à l'aide de l'apprentissage automatique, en mettant particulièrement l'accent sur l'algorithme des *k plus proches voisins* (K-NN). Le réseau GPON est largement simulé avec le logiciel OptiSystem, où des métriques de performance essentielles telles que la puissance optique (dBm), le taux d'erreur binaire (BER) et le rapport signal sur bruit (SNR) sont analysées dans divers scénarios de panne et d'interférence. Les données recueillies sont ensuite prétraitées et normalisées avant d'être classées à l'aide de l'algorithme K-NN, implémenté dans MATLAB, en utilisant la distance euclidienne comme mesure de similarité.

Les résultats de classification, évalués à l'aide de matrices de confusion, indiquent des taux de précision variant entre 63,16 % et 75,00 % selon les différentes unités de réseau optique (ONU). Les ONU 2 et ONU 8 ont obtenu les meilleures performances avec des précisions respectives de 75,00 % et 72,73 %, tandis que les ONU 1 et ONU 7 ont enregistré les taux les plus faibles, soit 63,64% et 63,16 %.

En parallèle, une analyse détaillée de l'effet de l'atténuation de la fibre sur le BER révèle une dégradation significative du signal à mesure que l'atténuation augmente. Cet effet s'avère particulièrement marqué dans le segment situé entre le répartiteur et les ONUs, comparé au tronçon reliant l'OLT (terminal de ligne optique) au répartiteur.

Ces résultats mettent en évidence l'efficacité des systèmes de diagnostic de pannes basés sur K-NN dans l'automatisation de la détection et l'amélioration de la fiabilité des réseaux GPON, contribuant ainsi à réduire les temps d'arrêt et les coûts d'exploitation. Les travaux futurs pourraient se pencher sur l'utilisation de classificateurs plus avancés, une optimisation du choix des caractéristiques, ainsi que sur des approches de surveillance en temps réel, afin d'améliorer davantage la précision du diagnostic et la résilience du réseau.

Abstract:

This study aims to enhance fault detection and classification in Gigabit Passive Optical Networks (GPON) using machine learning techniques, with a particular focus on the k-nearest neighbors (K-NN) algorithm. The GPON network is extensively simulated using OptiSystem software, where key performance metrics such as optical power (dBm), bit error rate (BER), and signal-to-noise ratio (SNR) are analyzed under various fault and interference scenarios.

The collected data is preprocessed and normalized before being classified using the K-NN algorithm, implemented in MATLAB with Euclidean distance as the similarity metric. The classification results, evaluated through confusion matrices, show accuracy rates ranging from 63.16% to 75.00% depending on the optical network units (ONUs). ONU 2 and ONU 8 achieved the best performance with respective accuracies of 75.00% and 72.73%, while ONU 1 and ONU 7 recorded the lowest at 63.64% and 63.16%.

In parallel, a detailed analysis of the effect of fibre attenuation on BER reveals significant signal degradation as attenuation increases, particularly in the segment between the splitter and the ONUs, compared to the segment between the OLT (Optical Line Terminal) and the splitter.

These findings highlight the effectiveness of K-NN-based fault diagnosis systems in automating fault detection and enhancing GPON reliability, contributing to reduced downtime and operational costs. Future work may explore more advanced classifiers, feature selection optimization, and real-time monitoring approaches to further improve diagnostic accuracy and network resilience

ملخص

تهدف هذه الدراسة إلى تحسين عملية كشف وتصنيف الأعطال في شبكات الألياف البصرية السلبية الجيغابيتية (GPON) باستخدام تقنيات التعلم الآلي، مع التركيز بشكل خاص على خوارزمية أقرب الجيران (K-NN). تم محاكاة شبكة GPON بشكل واسع باستخدام برنامج OptiSystem ، حيث جرى تحليل مؤشرات الأداء الرئيسية مثل القدرة البصرية (dBm) ، ومعدل الخطأ في البتات (BER) ، ونسبة الإشارة إلى الضوضاء (SNR) في سيناريوهات مختلفة للأعطال والتداخل.

تمت معالجة البيانات المجمعة مسبقاً وتطبيقها قبل تصنيفها باستخدام خوارزمية K-NN ، التي تم تنفيذها في بيئة MATLAB، مع اعتماد المسافة الإقليدية كمقياس للتشابه

ظهرت نتائج التصنيف، التي تم تقييمها باستخدام مصفوفات الالتباس، دقة تراوحت بين 63.16% و 75.00% حسب وحدة الشبكة البصرية (ONU). حيث حققت الوحدات 2 و 8 ONU أفضل أداء بدقة بلغت 75.00% و 72.73% على التوالي، في حين سجلت الوحدات 1 و 7 ONU أدنى أداء بدقة 63.64% و 63.16% على التوالي.

بالتوازي، كشفت دراسة تأثير التوهين في الألياف على معدل الخطأ (BER) عن تدهور كبير في الإشارة كلما زاد التوهين، خاصة في الجزء الممتد بين المقسم والوحدات ONU ، مقارنةً بالجزء بين وحدة الإرسال المركزية OLT والمقسم.

في أتمتة كشف الأعطال وتحسين موثوقية K-NN تسلط هذه النتائج الضوء على فعالية أنظمة التشخيص المعتمدة على ، مما يسهم في تقليل فترات التوقف وتكاليف التشغيل. ويمكن أن تركز الأبحاث المستقبلية على استخدام مصنفات GPON شبكات أكثر تطوراً، وتحسين اختيار الخصائص، واعتماد طرق مراقبة فورية لتحسين دقة التشخيص ومرونة الشبكة بشكل أكبر

Sommaire :

Introduction Générale.....	3
I.1 Introduction	3
I.2 Machine Learning:	4
I.2.1 Historique:	4
I.2.2 Définition.....	4
I.2.3 Principe de fonctionnement	5
I.2.4 La différence entre l'IA et le ML.....	5
I.2.5 Domaines d'application du ML	5
I.2.6 Les Avantages et risques de ML	6
I.2.7 Les type de ML	7
I.2.9 L'apprentissage supervisé (Supervised learning):.....	9
I.9.1 Classification	9
I.9.2 Régression	9
I.2.10 L'apprentissage non supervise.....	10
I.2.11 L'apprentissage par renforcement.....	11
I.2.12 Les méthodes de classification non supervisées	12
I.2.13 Les méthodes de classification supervisées	12
I.2.14 Concepts fondamentaux du ML.....	14
I.2.15 Formalisation d'un problème de ML	15
I.2.16 Défis et limites actuels du ML	16
I.2.17 Évaluation des performances d'un modèle de ML	17
I.2.17.1 Pour les problèmes de classification	17
I.2.17.2 Pour les problèmes de régression.....	18
II.2.17.3. Pour les problèmes de clustering (Non supervisé) :.....	18

I.3 Méthode des k plus proches voisins.....	19
I.3.1 Définition d'algorithme des k plus proches voisins (K-NN).....	19
I.3.2 Où utiliser K-NN ?.....	19
I.3.3Le fonctionnement de l'algorithme k-NN.....	19
I.3.3.1 Principe général de fonctionnement.....	19
I.3.4 Avantages et inconvénients de K-NN	20
I.3.5 Choix du paramètre K dans l'algorithme K-NN	21
Étape 3 : Choisir les K plus proches voisins	24
Étape 4 : Prédiction de la classe de Q	24
I.3.6 Comparatif entre l'algorithme K-NN et d'autres algorithmes supervisés.....	24
I.3.9 Limites de K-NN en contexte réel	25
I.4 Conclusion	27
II .1 Introduction :	28
II.2 Définition de Fibre optique	29
II.3 Définition des systèmes optiques :	29
II.4 Définition des réseaux optiques	30
II.5 Réseaux d'accès optiques	30
II.6 Différences entre les réseaux actifs et passifs	31
II. 7 Les technologies d'accès optique	31
II.8 Les différentes technologies PON	32
II.9 GPON.....	34
II.9.1Défintion	34
II.9.2 Pourquoi choisir GPON ?.....	35
II.9.3 Rôle du GPON dans les réseaux d'accès modernes	35
II.9.4Les avantages et les Inconvénients de GPON :	36

II.9.4.1 Les Avantages :	36
II.9.4.2 Les Inconvénients	36
II.9.5 Architecture du réseau optique G-PON	37
II.9.6 Fonctionnement du réseau GPON	38
II.9.7 Sens montant (de l'ONU vers l'OLT)	38
II.9.8 Sens descendant (de l'OLT vers les ONU)	39
II.9.9 Structure de la trame GPON (GTC Frame Format)	40
II.9.10 Gestion de la bande passante dans GPON :	41
II.9.10.1 Principe de la DBA	41
II.9.10.2) Fonctionnement	42
II.9.11 Les caractéristiques principales de GPON	42
II.9.12 Normes de GPON	43
II.10 Applications de la technologie GPON	44
II.11 Sécurité et gestion dans les réseaux GPON	44
II.12 Les pannes dans un réseau GPON	45
II.13 Application du Machine Learning dans PON	46
II.14 Conclusion	48
III.1 Introduction	49
III.2 Présentation du logiciel Optisystem	49
III.3 Configuration étudiée	49
Photodétecteur PIN et Filtre de Bessel :	52
Régénérateur 3R :	52
Analyseur BER (Bit Error Rate):	52
III.4 Résultats de simulations	52
III.5 Conclusion	60

Sommaire

Conclusion générale	61
Référence :	62

Liste des figures:

Figure I.1 : L'IA et ses sous domaines.	3
Figure I.2 :Fonctionnement d'apprentissage machine (ML).	4
Figure I.3: Exemples d'application de Machine Learnig.	6
Figure I.4 : Les 3 types d'apprentissage du machine Learning.	8
Figure I.4 : Apprentissage supervisé.	9
Figure I.5 : Classification vs Régression.....	10
Figure I.6 : Comparaison entre l'apprentissage supervisé et non supervise.	11
Figure I.7 :Schéma général de la formalisation d'un problème d'apprentissage supervise15	15
FigureI.8 : Effet du paramètre K sur la frontière de décision de l'algorithme K-NN.....	22
Figure I.9 : Exemple de Influence du choix de la valeur de K dans l'algorithme K-NN3.....	23
Figure I.10 : l'illustration comparative des frontières de décision pour l'algorithme K-NN et d'autres algorithmes supervisés.....	25
Figure II.1 : Réseau GPON.....	29
Figure II.2 :Fibre optique.	30
Figure II.3 :Architecture d'un réseau d'accès optique.	31
Figure II.4 :Les technologies d'accès optique.....	33
Figure II.5 : Architecture du réseau G-PON.....	38
Figure II.6: Architecture du sens montant.....	40
Figure II.7 : Architecture du sens descendant.	41
Figure II.8: Structure du trame GPON.....	42
Figure III.1 : Liaison GPON uni-directionnelle	52
Figure III.2 : Canal de transmission.....	53
Figure III.3 : Bloc de réception.....	53
Figure III.4 : Le bloc BER.....	55
Figure III.5 : Effet des différentes positions de dommages dans la fibre optique sous des valeurs variables du coefficient d'atténuation	56
Figure III.6 : Effet des dommages à la fibre optique sur les utilisateurs sous différentes valeurs d'atténuation.	58
Figure III.7 : Matrices de confusion du classificateur K-NN pour la prédiction des dommages dans les fibres optiques.....	59

Liste des figures

Figure III.8 : Matrices de confusion du classificateur K-NN pour l'ONU 1	60
Figure III.9 : ROC du classificateur K-NN	61

Liste des tableaux :

Tableau II. 1 :Tableau comparatif détaillé des technologies PON.....	33
Tableau II.1 : la liaison entre l'OLT et le splitter.....	56
Tableau II.2 : la liaison entre splitter et ONU	56

Liste des abréviations

AON	Active Optical Network
APON	ATM Passive Optical Network
ARIA	Adjusted Rand Index
ATM	Asynchronous Transfer Mode
AUC	Area Under the Curve
BER	Bit Error Rate
BPON	Broadband Passive Optical Network
dB	Décibel
EPON	Ethernet Passive Optical Network
FTTH	Fiber To The Home
FTTx	Fiber To The x
IA	Intelligence Artificielle
GPON	Gigabit-capable Passive Optical Network
K-NN	K-Nearest Neighbors
MAE	Mean Absolute Error
ML	Machine Learning
MSE	Mean Squared Error
NG-PON2	Next Generation Passive Optical Network 2
NRZ	Non Return to Zero
OLT	Optical Line Terminal
ONT	Optical Network Terminal
PON	Passive Optical Network
RMSE	Root Mean Squared Error
ROC	Receiver Operating Characteristic
SNR	Signal-to-Noise Ratio
SVM	Support Vector Machine
TDM	Time Division Multiplexing
WDM	Wavelength Division Multiplexing
XG-PON	10-Gigabit-capable Passive Optical Network
XGS-PON	10-Gigabit Symmetrical Passive Optical Network

Introduction générale

Dans un monde où la connectivité numérique constitue un pilier essentiel du développement économique, social et technologique, la demande en services Internet à très haut débit connaît une croissance exponentielle. Cette évolution rapide impose aux infrastructures de télécommunications des exigences accrues en termes de débit, de fiabilité, de faible latence, ainsi que de flexibilité face aux besoins sans cesse changeants des utilisateurs.

Parmi les technologies émergentes capables de relever ces défis, les réseaux optiques passifs à haut débit occupent une place centrale. En particulier, la technologie **GPON** (*Gigabit Passive Optical Network*) s'impose comme une solution incontournable, en offrant une infrastructure à la fois performante et économique pour le transport de données à très haut débit. Grâce à son architecture point-à-multipoint et à l'utilisation d'éléments passifs tels que les diviseurs optiques (*splitters*), le GPON permet de desservir simultanément plusieurs abonnés tout en réduisant les coûts d'installation et de maintenance, tout en maintenant une excellente qualité de service.

Parallèlement, l'essor spectaculaire de l'intelligence artificielle (IA) transforme en profondeur de nombreux secteurs, notamment celui des télécommunications. En particulier, le **Machine Learning** (apprentissage automatique) offre des outils puissants pour analyser de grands volumes de données, détecter des anomalies complexes et prendre des décisions basées sur des modèles statistiques et l'expérience accumulée. Parmi les algorithmes d'apprentissage supervisé, **K-Nearest Neighbors (K-NN)** se distingue par sa simplicité, sa robustesse et son efficacité dans les tâches de classification, ce qui en fait un candidat idéal pour le diagnostic et la gestion intelligente des réseaux.

Le présent travail s'inscrit dans une démarche novatrice visant à intégrer des techniques d'intelligence artificielle à l'analyse et à l'optimisation des performances des réseaux GPON. Plus précisément, il propose d'étudier l'impact des pannes dans un réseau GPON unidirectionnel, modélisé et simulé à l'aide du logiciel **OptiSystem**. Les données issues de ces simulations seront ensuite exploitées à l'aide de l'algorithme K-NN, afin de permettre une détection automatique et une classification précise des défaillances affectant le réseau.

Cette approche multidisciplinaire illustre l'intérêt croissant de l'intelligence artificielle dans le domaine des télécommunications, en vue d'améliorer non seulement la performance et la résilience des réseaux, mais aussi leur capacité à s'auto-diagnostiquer et à anticiper les pannes, ouvrant ainsi la voie à des infrastructures plus intelligentes et autonomes.

Ce mémoire s'articule autour de trois chapitres complémentaires. Le premier chapitre présente les concepts fondamentaux de l'apprentissage automatique, en décrivant les différents types d'apprentissage et en mettant en évidence leurs applications dans divers domaines. Une attention particulière est portée à l'algorithme K-Nearest Neighbors (K-NN), reconnu pour sa simplicité et son efficacité dans les tâches de classification supervisée.

Le deuxième chapitre est consacré à une étude approfondie de la technologie GPON, incluant sa définition, son architecture, son principe de fonctionnement, ses avantages et ses limitations. Une analyse détaillée du budget optique, élément clé pour évaluer la portée et la qualité du signal, y est également développée, accompagnée d'une classification des réseaux GPON en fonction de ce paramètre.

Le troisième chapitre établit enfin un lien clair entre les réseaux optiques et l'intelligence artificielle, en soulignant l'intérêt de l'intégration d'algorithmes tels que le K-NN pour améliorer la surveillance, l'analyse des performances et la détection des anomalies. Le troisième chapitre, quant à lui, analyse l'impact des pannes dans un réseau GPON unidirectionnel en combinant les outils OptiSystem et MATLAB : OptiSystem permet de modéliser le réseau, de simuler divers scénarios de défaillance (ruptures de fibre, pertes de signal, etc.) et d'évaluer la qualité de transmission à l'aide d'indicateurs tels que le facteur Q, le BER ou le SNR, tandis que MATLAB est utilisé pour implémenter et entraîner l'algorithme K-NN sur les données issues des simulations. Ce processus permet de classifier automatiquement les types de défaillances et d'évaluer la performance du modèle (précision, rappel, taux d'erreur), dans le but de concevoir un système de diagnostic prédictif efficace, automatisé et adapté aux spécificités des réseaux GPON.

Chapitre I: Machine learning K-NN

I.1 Introduction :

L'intelligence artificielle (IA) est un domaine de recherche en pleine expansion, visant à développer des systèmes capables d'imiter certaines capacités cognitives humaines, telles que la perception, le raisonnement, l'apprentissage ou la prise de décision. Elle regroupe un ensemble de techniques qui permettent aux machines de résoudre des problèmes complexes, souvent associés à l'intelligence humaine.

Parmi ces techniques, le Machine Learning (ML ou apprentissage automatique) occupe une place centrale. Il s'agit d'un sous-domaine de l'IA qui permet aux machines d'apprendre automatiquement à partir de données, sans être explicitement programmées. Le ML repose sur des algorithmes capables de détecter des régularités dans les données et de faire des prédictions ou des classifications sur de nouvelles données. On distingue principalement trois types d'apprentissage : l'apprentissage supervisé, l'apprentissage non supervisé, et l'apprentissage par renforcement.

Dans le prolongement du ML, le Deep Learning (ou apprentissage profond) est une approche plus récente qui repose sur l'utilisation de réseaux de neurones artificiels profonds. Cette technique permet de traiter de très grandes quantités de données et de résoudre des problèmes particulièrement complexes, notamment dans les domaines de la vision par ordinateur, de la reconnaissance vocale ou du traitement du langage naturel.

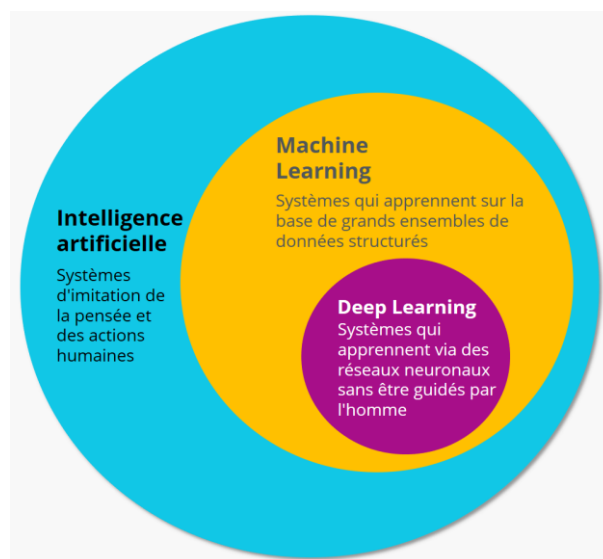


Figure I.1 : L'IA et ses sous domaines.

I.2 Machine Learning:

I.2.1 Historique:

Le ML est apparu en 1959 grâce au mathématicien américain Arthur Samuel, qui a développé un programme capable d'apprendre à jouer aux dames sans intervention humaine. Il définit alors le ML comme « la science qui consiste à amener les ordinateurs à apprendre sans être explicitement programmés » [1]. En 1998, l'Américain Tom Mitchell a proposé une définition plus approfondie de l'apprentissage automatique, en affirmant qu'un système apprend lorsqu'il améliore sa performance dans une tâche donnée au fur et à mesure qu'il acquiert de l'expérience. Ainsi, le ML peut être défini comme la capacité d'un système à déterminer les calculs appropriés pour résoudre un problème en se basant sur les données et les expériences précédentes.

I.2.2 Définition:

ML est un sous-domaine de l'intelligence artificielle (IA). C'est un ensemble de techniques et d'algorithmes permettant à la machine de comprendre ou d'effectuer des tâches jusqu'à maintenant considérées propres à l'humain, comme la reconnaissance d'images, les systèmes de décision, les prédictions ...etc. Les algorithmes d'apprentissage automatique permettent aux ordinateurs de s'entraîner sur des données existantes (données d'apprentissage ou ensemble d'entraînement, ou *training data*), pour produire de nouvelles informations (ou modèles de données). Ces informations détermineront plus tard le comportement du système (décisionnel) vis-à-vis des données saisies pour un problème complexe donné, et permettront ainsi de le résoudre. Le fonctionnement d'un algorithme d'apprentissage est représenté en figure I.2 suivante [2] :

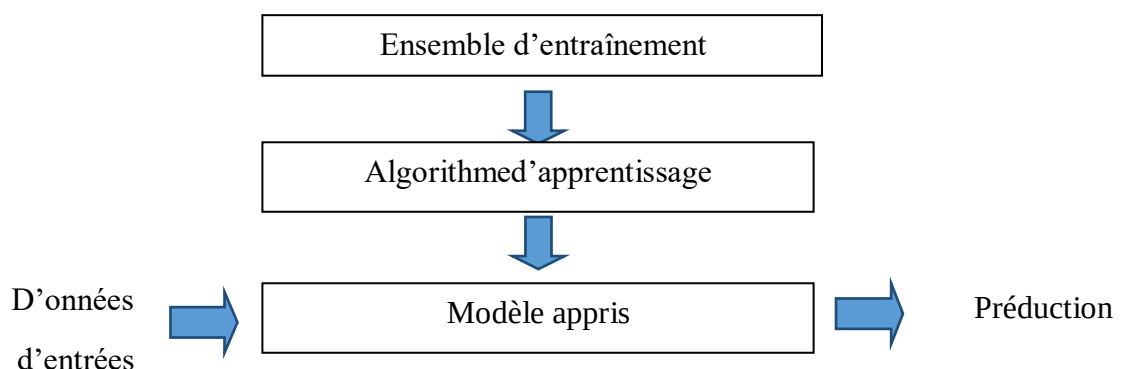


Figure I.2 : Fonctionnement d'apprentissage machine (ML).

I.2.3 Principe de fonctionnement :

Le ML se déroule en deux étapes, la première est la phase d'apprentissage, qui consiste à extraire les informations pertinentes à partir des données étudiées, dans le cadre d'un

processus appelé *entraînement*. Une fois le modèle bien déterminé, vient la phase de déploiement, où de nouvelles données sont introduites pour réaliser la tâche souhaitée.

Les algorithmes d'apprentissage automatique fonctionnent en construisant un modèle à partir d'exemples, afin de faire des prédictions ou de prendre des décisions basées sur des données. L'objectif principal est de minimiser ce qu'on appelle l'erreur, c'est-à-dire l'écart entre les prédictions du modèle et les résultats réels [3].

I.2.4 La différence entre l'IA et le ML:

L'intelligence artificielle (IA) est un domaine de l'informatique qui vise à concevoir des systèmes capables de reproduire certaines fonctions cognitives humaines, telles que le raisonnement, la planification, la compréhension du langage ou la perception visuelle. Elle englobe un large éventail de techniques, allant des approches fondées sur des règles logiques aux méthodes reposant sur l'apprentissage automatique. Le ML, est une sous-discipline de l'IA qui permet aux machines d'apprendre à partir de données, sans être explicitement programmées pour chaque tâche. En analysant de grands ensembles de données, les algorithmes de machine learning détectent des motifs et ajustent leurs modèles pour améliorer leurs performances au fil du temps. Contrairement aux systèmes d'IA traditionnels, ces modèles s'améliorent grâce à l'expérience, c'est-à-dire à l'analyse répétée d'exemples.

I.2.5 Domaines d'application du ML :

Le ML est utilisé dans plusieurs domaines pour extraire de l'information utile à partir de grandes quantités de données et prendre des décisions intelligentes automatiquement. Voici quelques domaines d'application importants :

- **l'e-commerce** (comportement des clients, tendance de vente, ciblage, recommandation de films, voyages, etc),
- **l'énergie** (détection de pannes, d'anomalies, etc),
- **la banque et les assurances** (tarification, mesure du risque crédit, rachat, fraude, cross selling),
- **la santé** (prévention des maladies et amélioration des traitements), le bâtiment (réduction des dépenses inutiles, automatisation du fonctionnement du bâtiment, utilisation de matériaux plus légers),
- **Télécommunication** (filtrage anti-spam, ciblage client, churn, etc), la défense (reconnaissance faciale, biométrie, détection de zones d'agression, etc),

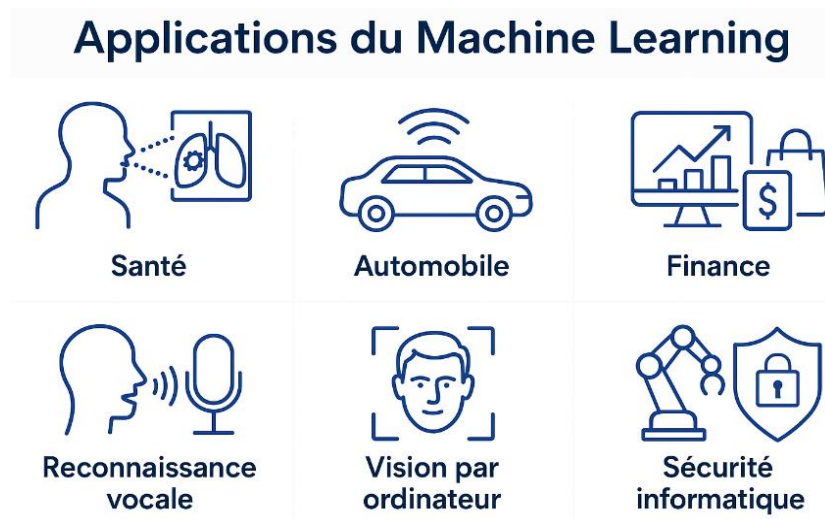


Figure I.3 : Exemples d'application de Machine Learning

I.2.6 Les Avantages et risques de ML:

ML transforme déjà de nombreux aspects de notre quotidien. Il est aujourd'hui utilisé pour développer des modèles capables de détecter des tumeurs cancéreuses dans les images médicales, de repérer des transactions frauduleuses, ou encore d'assister les personnes dans l'apprentissage des langues. Toutefois, comme toute technologie innovante qui bouleverse la société, il présente également certains risques qu'il convient de prendre en compte. Voici un aperçu des principaux avantages et inconvénients potentiels de l'apprentissage automatique :

I.6.1 Avantages :

1. **Diminution des coûts opérationnels** : L'IA et l'apprentissage automatique peuvent aider les entreprises à automatiser certains de leurs emplois, entraînant une diminution des coûts opérationnels globaux.
2. **Amélioration de l'efficacité opérationnelle et de la précision** : Les modèles d'apprentissage automatique sont capables d'effectuer certaines tâches étroites avec une efficacité et une précision extrêmes, garantissant que certaines tâches sont accomplies à un degré élevé en temps opportune.
3. **Amélioration des connaissances** : L'apprentissage automatique a le potentiel d'identifier rapidement des tendances et des modèles dans de grandes quantités de données, ce qui prendrait beaucoup de temps aux humains. Ces informations peuvent apporter aux entreprises, aux chercheurs et à la société dans son ensemble de nouvelles connaissances susceptibles de les aider à atteindre leurs objectifs globaux.

I.6.2 Inconvénients

1. **Besoin de données de qualité** : La ML nécessite des données de qualité pour fonctionner correctement, ce qui peut être difficile à obtenir dans certains cas.

2. **Complexité** : La mise en place de systèmes de ML peut être complexe et coûteuse, nécessitant une expertise technique et un investissement en temps et en ressources.

3. **Interprétabilité** : Les modèles de ML peuvent être difficiles à interpréter, ce qui peut rendre difficile la compréhension des décisions prises par le système.

4. **Sécurité et confidentialité** : Les modèles de ML peuvent être vulnérables à des attaques et peuvent nécessiter des mesures de sécurité supplémentaires pour protéger les données sensibles [7].

ML joue un rôle fondamental dans la transformation numérique des systèmes intelligents. Sa capacité à analyser de vastes ensembles de données et à en extraire des modèles prédictifs permet d'améliorer la performance, l'efficacité et la fiabilité des systèmes modernes. Par exemple, dans les réseaux optiques passifs de type GPON (Gigabit Passive Optical Network), le Machine Learning peut être utilisé pour détecter automatiquement les pannes, prévoir les dégradations de performance ou encore optimiser la gestion des ressources réseau. Grâce à ces techniques, les opérateurs peuvent offrir un service plus stable et proactif [6].

I.2.7 Les type de ML:

Il existe plusieurs méthodes permettant à une machine d'apprendre à partir de données. Chaque approche possède ses propres caractéristiques et s'adapte à des problématiques spécifiques, en fonction des données disponibles et des objectifs visés. Parmi les plus pertinentes, on distingue principalement trois types d'apprentissage : l'apprentissage supervisé, l'apprentissage non supervisé, et l'apprentissage par renforcement (voir Figure 3).

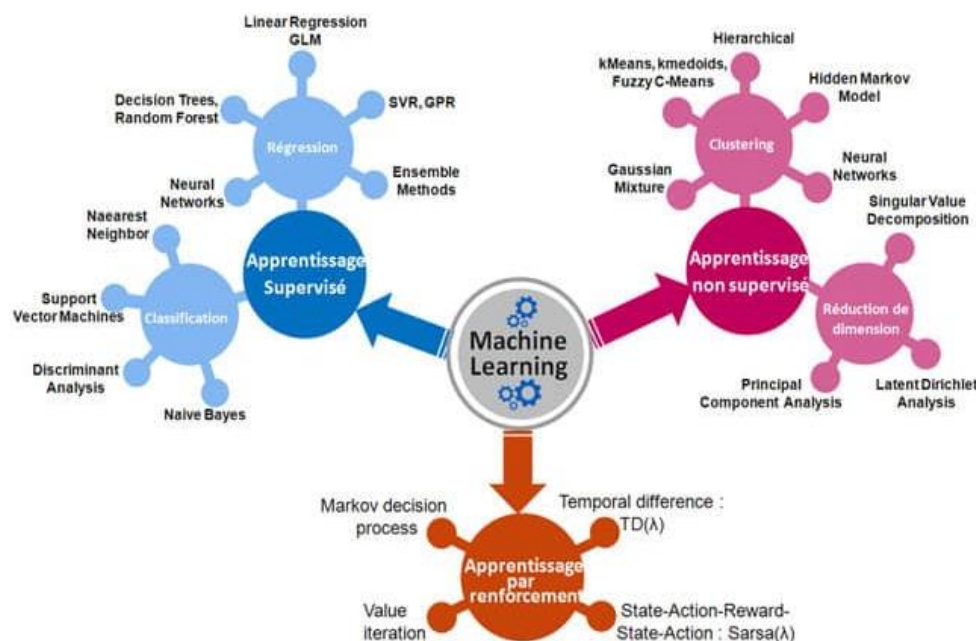


Figure I.4 : Les 3 types d'apprentissage de machine Learning.

La définition du type dépend de la réponse à deux questions clés :

- Est-ce que l'apprentissage nécessite une supervision humaine pour son entraînement et son apprentissage ?

- Est-ce que l'apprentissage utilise une base de données fournie par l'humain ?

A. Si la réponse est oui pour les deux questions, nous avons affaire à un apprentissage supervisé.

B. Si la réponse est non pour la première question, et oui pour la deuxième question, nous parlons d'un apprentissage non supervisé.

C. Si la réponse est non pour les deux questions, il s'agit d'un apprentissage par renforcement.

Dans ce travail, nous nous concentrerons sur l'apprentissage supervisé, qui constitue la méthode choisie pour notre étude dans les réseaux optiques passifs de type GPON (Gigabit Passive Optical Network), nous avons utilisés pour détecter automatiquement les pannes, prévoir les dégradations de performance ou encore optimiser la gestion des ressources réseau. Grâce à ces techniques, les opérateurs peuvent offrir un service plus stable et proactif. Les autres approches seront abordées de manière générale sans entrer dans les détails.

I.2.9 L'apprentissage supervisé :(Supervised learning)

Il est le mode le plus courant en apprentissage automatique. Dans ce mode, on fournit au réseau des données d'entrée ainsi que les sorties correspondantes que l'on souhaite obtenir pour ces entrées. Le réseau doit ensuite se réajuster en recalculant ses poids afin de produire une sortie correspondant le mieux possible à la sortie souhaitée [10].

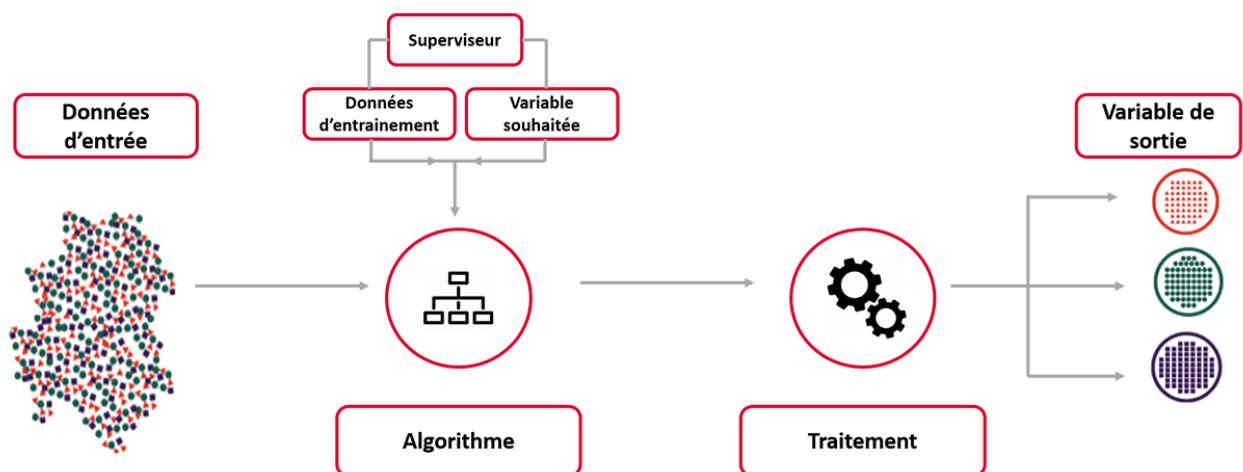


Figure I.4 : Apprentissage supervisé

I.9.1 Classification :

La classification est une tâche d'apprentissage supervisé où chaque donnée d'entrée est associée à une étiquette représentant une classe. L'objectif est d'apprendre un modèle capable d'attribuer à chaque nouvelle donnée la classe à laquelle elle appartient. Lorsque le nombre de classes est limité à deux, on parle alors de classification binaire.

I.9.2 Régression:

Dans le cas de la régression, l'entrée n'est pas liée à une catégorie, mais plutôt à une ou plusieurs valeurs numériques continues. Par exemple, les données d'entrée peuvent représenter les caractéristiques d'une personne, comme son âge, son sexe et son niveau d'études, tandis que la sortie recherchée serait son revenu. L'objectif de la régression est donc d'apprendre une fonction capable de prédire une valeur continue à partir des caractéristiques fournies [6].

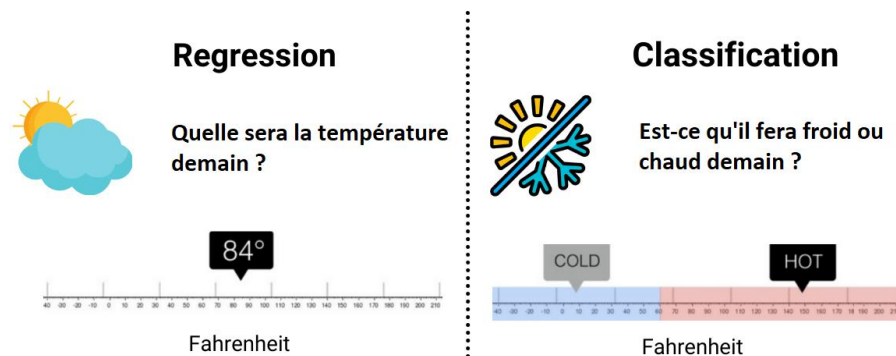


Figure I.5 : Classification vs Régression

Exemple d'apprentissage supervise :

Supposons que l'on souhaite développer un système capable de reconnaître automatiquement les e-mails indésirables (spams). Pour cela, on dispose d'un ensemble de données contenant des e-mails déjà classés manuellement comme « spam » ou « non-spam ». Ces données étiquetées servent à entraîner un algorithme de classification, qui apprendra à reconnaître les caractéristiques communes aux spams (présence de certains mots, liens suspects, etc.). Une fois l'apprentissage terminé, l'algorithme sera capable de prédire si un nouvel e-mail reçu est un spam ou non, en se basant sur les connaissances acquises lors de l'entraînement.

I.2.10 L'apprentissage non supervise:

L'apprentissage non supervisé est un paradigme fondamental de ML dans lequel un algorithme est chargé d'analyser des données d'entrée dépourvues de toute étiquette ou

information sur les résultats attendus. Contrairement à l'apprentissage supervisé, aucun signal externe ne vient guider le processus d'apprentissage. L'algorithme doit alors inférer, de manière autonome, des structures sous-jacentes, des régularités statistiques ou des représentations significatives à partir des seules caractéristiques des données observées.

Ce type d'apprentissage vise principalement à modéliser la distribution des données ou à capturer leur structure latente, souvent à travers des techniques de partitionnement, de regroupement ou de réduction de dimensionnalité. Il repose sur l'hypothèse qu'il existe une organisation implicite au sein des données, que l'on peut mettre en évidence grâce à des mécanismes d'apprentissage automatique sans supervision humaine directe. Cette approche permet notamment de construire des représentations internes compactes, exploitables par d'autres algorithmes ou modèles pour des tâches ultérieures, telles que la classification, la détection ou la génération de données synthétiques.

L'apprentissage non supervisé joue ainsi un rôle central dans la compréhension de grands ensembles de données complexes et dans le développement de systèmes intelligents capables d'extraire et de structurer l'information de manière autonome [7]

Exemple d'apprentissage non supervise :

L'apprentissage non supervisé dans un réseau WDM permet de :

1. Collecter des données de performance des canaux optiques (puissance, bruit, pertes de paquets).
2. Étiqueter les données selon l'état des canaux (fonctionnel, interférence, défaillance).
3. Entraîner un modèle supervisé pour détecter les anomalies dans les canaux.
4. Faire des prédictions en temps réel pour détecter et résoudre rapidement les défaillances.

Cela améliore la gestion et la maintenance des réseaux optiques.

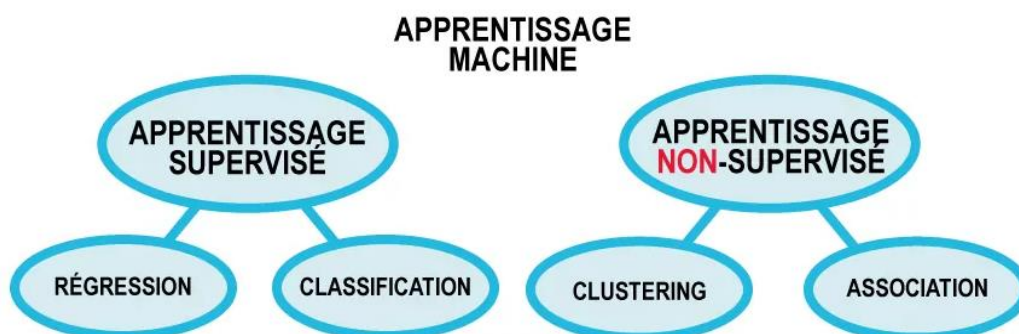


Figure I.6 : Comparaison entre l'apprentissage supervisé et non-supervisé

I.2.11 L'apprentissage par renforcement :

C'est une forme d'apprentissage en interaction dont le but est d'apprendre, à partir des expériences, ce qu'il convient de faire en différentes situations. Pour présenter les problèmes d'apprentissage par renforcement, on considère plongé au sein d'un environnement, un agent autonome qui doit prendre des décisions en fonction de son état courant. En retour, l'environnement procure à l'agent une récompense, qui peut être positive ou négative. L'agent cherche, au travers des expériences itérées, un comportement décisionnel (appelé stratégie ou politique, et qui est une fonction associant à l'état courant l'action à exécuter) optimal. Comme pour l'apprentissage actif, il s'agit de trouver une bonne politique, un compromis entre l'exploitation du modèle courant et l'exploration d'un nouveau meilleur modèle (en réitérant par exemple de nouvelles expériences). C'est le cas par exemple, dans le jeu vidéo FIFA (jeu de football) où, dans une situation donnée, un joueur qui se retrouve avec un ballon possède plusieurs solutions (faire une passe, un tir, ...). Cet apprentissage permet d'atteindre la meilleure solution que va faire le joueur à cette position du terrain pour marquer un but de façon plus sûr [2].

I.2.12 Les méthodes de classification non supervisées :

Les méthodes de classification non supervisées, aussi appelées clustering, visent à regrouper automatiquement des données en groupes (ou clusters) homogènes, sans connaissance préalable des classes. L'objectif est de détecter des structures ou des similarités naturelles au sein des données, en maximisant la cohérence interne des groupes et leur séparation mutuelle[8].

Parmi ces méthodes, on peut citer notamment :

K-means : L'algorithme K-means, proposé par MacQueen en 1967, est une méthode itérative de classification non supervisée. Il consiste à :

1. Choisir un nombre de clusters k à l'avance.
2. Initialiser aléatoirement les centres des clusters.
3. Affecter chaque individu au cluster le plus proche selon une mesure de similarité.
4. Recalculer les centres comme la moyenne des points de chaque cluster.
5. Répéter les étapes jusqu'à stabilisation (plus aucun changement d'affectation).

Limite principale : la qualité des résultats dépend fortement du choix initial des centres et du nombre de clusters[15].

I.2.13 Les méthodes de classification supervisées :

Les méthodes de classification supervisée constituent un pilier fondamental de l'apprentissage automatique. Elles consistent à entraîner un modèle à partir d'un ensemble de données étiquetées, c'est-à-dire des exemples pour lesquels les classes sont connues. L'objectif est de permettre au modèle d'attribuer une classe correcte à de nouvelles données en se basant sur les relations apprises durant l'entraînement [7].

Parmi les méthodes de classification supervisée les plus utilisées, on peut citer :

I.2.13.1 La régression logistique : La régression logistique est un modèle de classification qui prédit la probabilité qu'une instance appartienne à une classe donnée en utilisant la fonction logistique. Bien qu'elle soit linéaire dans ses paramètres, elle est très performante pour des problèmes binaires et facilement généralisable au cas multiclasse [9]

I.2.13.2 Les arbres de décision : Les arbres de décision sont des modèles d'apprentissage supervisé utilisés à la fois pour la classification et la régression. Ils fonctionnent en appliquant une série de tests sur les attributs d'entrée, chaque nœud interne représentant un test, et chaque feuille correspondant à une décision finale. L'apprentissage consiste à identifier les attributs les plus pertinents permettant de diviser les données en groupes homogènes, c'est-à-dire composés majoritairement d'exemples appartenant à la même classe. Pour cela, des mesures de pureté comme l'entropie sont utilisées afin d'évaluer la qualité des divisions. Ces métriques permettent de guider la construction de l'arbre en sélectionnant les attributs les plus discriminants à chaque étape [15].

I.2.13.3 Les machines à vecteurs de support (SVM) : Les SVMs sont des techniques d'apprentissage supervisé utilisées principalement pour la classification. Lorsqu'un jeu de données est linéairement séparable, plusieurs hyperplans peuvent séparer les classes, mais SVM cherche celui qui maximise la marge, c'est-à-dire la distance entre l'hyperplan et les points les plus proches de chaque classe, appelés vecteurs de support. Ce principe permet d'obtenir un modèle ayant une bonne capacité de généralisation. Pour les cas non linéaires, SVM utilise une séparation linéaire devient possible. Les SVM sont largement utilisés dans des domaines tels que la catégorisation de textes, la reconnaissance d'objets et la détection d'actions humaines [14].

I.2.13.4 Le classifieur bayésien naïf (Naïve Bayes) : Les méthodes naïves de Bayes sont des algorithmes d'apprentissage supervisé fondés sur le théorème de Bayes, en supposant une indépendance totale entre les caractéristiques. Le modèle apprend à estimer les probabilités de vraisemblance pour chaque classe à partir des données d'entraînement, puis calcule les probabilités a posteriori pour classer une nouvelle observation. L'exemple est affecté à la

classe ayant la probabilité a posteriori la plus élevée. Cette approche est simple, rapide, et nécessite peu de données, car elle ne demande que l'estimation des moyennes et variances des attributs pour chaque classe [15].

La formule de bayes : avec ;A et B, deux événements :

$$P(A|B) = \frac{P(A|B)P(B)}{P(A)} \quad (I.1)$$

I.13.6 Le classifieur des k plus proches voisins (k-NN) :

Nous accorderons une attention particulière à l'algorithme k-plus proches voisins (k-NN), en examinant en détail son principe de fonctionnement, ses avantages, ses limites, ainsi que ses domaines d'application en classification. Cet algorithme constituera le cœur méthodologique de notre étude, dans laquelle nous l'utiliserons pour analyser les performances du réseau GPON.

I.2.14 Concepts fondamentaux du ML :

I. 2.14.1 "Dataset" :

C'est un ensemble de données qu'on fournit à une machine sous forme d'un couple d'exemples (X,Y) dans l'apprentissage supervisé où X représente les questions et Y les réponses au problème que la machine doit résoudre. Dans l'apprentissage non supervisé, le dataset contient seulement des questions X On ne peut pas démarrer un projet sans avoir de dataset [15]

I. 2.14.2 Modèle et ses paramètres :

Un modèle est une fonction mathématique construite à partir du dataset fourni. Cette fonction peut être linéaire ou non linéaire. Les paramètres du modèle correspondent aux coefficients de cette fonction, ajustés pendant l'entraînement. Ils sont déterminés à partir des prédicteurs X et des annotations Y, que l'on cherche à généraliser à de nouvelles données[13]

I.2.14.3 Les hyperparamètres :

Les hyperparamètres sont des valeurs de réglage fixées avant l'entraînement du modèle. Ils comprennent, entre autres, le nombre d'itérations, la valeur de la graine aléatoire (*seed*), la solution initiale, ainsi que d'autres paramètres spécifiques aux différents modèles testés [15]

I.2.14.4 Fonction Coût :

La phase de validation permet d'évaluer la performance du modèle en termes de taux de faux positifs (fausses alertes) et de faux négatifs (éléments non détectés), que l'on cherche à réduire simultanément. Cette évaluation s'effectue à l'aide d'une fonction coût, qui représente l'ensemble des erreurs que le modèle retourne par rapport au dataset.

La validité du modèle dépend de cette fonction, car c'est en la minimisant que la machine est capable d'identifier les modèles les plus performants [16].

I.2.14.5 La généralisation :

La généralisation est l'étape où le modèle, après avoir été formé sur un ensemble de données d'entraînement, est appliqué à de nouvelles données pour évaluer sa capacité à prédire avec précision en dehors de cet ensemble. Ce processus s'inscrit souvent dans un contexte de traitement de grandes quantités de données, où l'industrialisation des processus devient essentielle, souvent facilitée par la distribution des calculs.

Pour résumer le processus de Machine Learning, on commence par préparer l'échantillon d'entraînement, qui comprend les prédictions et les annotations. Le modèle est ensuite ajusté en fonction des hyper-paramètres définis et entraîné. Une fois l'entraînement effectué, une phase de validation permet de mesurer la performance du modèle en utilisant une fonction coût qui évalue les erreurs. Enfin, pour vérifier sa fiabilité et sa robustesse, le modèle est testé sur des données externes, non utilisées lors de l'entraînement [13].

I.2.15 Formalisation d'un problème de ML:

L'apprentissage automatique peut être formellement modélisé comme le problème d'apprendre une fonction à partir d'exemples. Le but est de trouver une règle de prédiction (ou hypothèse) qui associe correctement une entrée à une sortie [16].

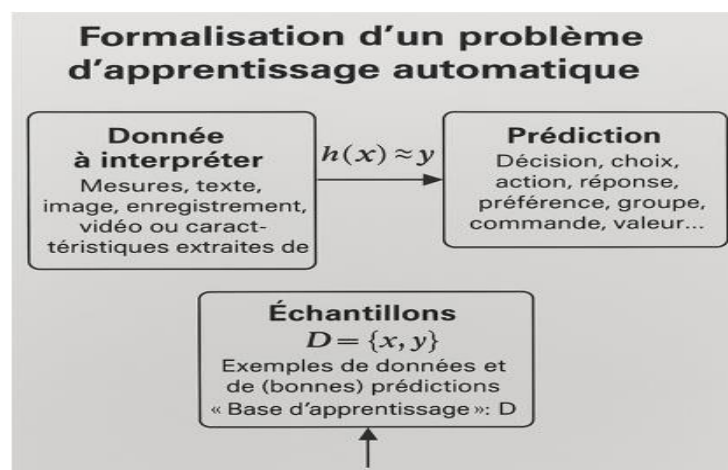


Figure I.7: Schéma général de la formalisation d'un problème d'apprentissage supervisé.

I.2.15.1 Donnée à interpréter ($x \in X$) :

Une entrée x est un vecteur d'attributs représentant un objet ou une situation. Le X est par exemple :

- Une image peut être représentée par un vecteur de pixels

- Un utilisateur par ses préférences numériques
- Un signal réseau par des mesures telles que la puissance, le SNR, etc..

I.2.15.2 Sortie attendue ou prédiction ($y \in Y$) :

La sortie y représente la **réponse correcte** à associer à x . L'ensemble Y dépend du type de tâche :

- En classification : Y est un ensemble fini (ex. : $\{0,1\}$ ou $\{\text{normal, panne}\}$)
- En régression : $Y \in \mathbb{R}$ (valeurs numériques continues)

I.2.15.3 Exemple annoté (x_i, y_i) :

Un exemple est une paire composée d'une entrée et de sa sortie correcte. Ces exemples sont supposés être tirés indépendamment d'une distribution inconnue $\mathcal{D}$ sur $X \times Y$.

I.2.15.4 Ensemble d'apprentissage ($D = \{ (x_i, y_i) \}$) :

C'est un échantillon d'exemples indépendants et identiquement distribués (i.i.d.), utilisé pour entraîner l'algorithme. L'algorithme cherche une fonction $h: X \rightarrow Y$ qui minimise l'erreur de prédiction sur de nouvelles données [16].

I.2.15.5 Objectif : Apprendre une fonction predictive :

L'objectif est de construire une fonction d'hypothèse h appartenant à un ensemble tel que :

$$h(x) \cong y$$

Pour une nouvelle entrée x , inconnue au moment de l'apprentissage [16].

I.2.16 Défis et limites actuels du ML :

Le ML, bien qu'en plein essor et largement utilisé dans de nombreux domaines, fait face à plusieurs défis et limitations qu'il est essentiel de considérer dans toute démarche scientifique ou technique. Ces obstacles peuvent affecter aussi bien la conception des modèles que leur déploiement en conditions réelles [17].

I.2.16.1 Qualité et quantité des données :

Le ML repose sur l'apprentissage à partir de données. Ainsi, la performance du modèle dépend fortement de la qualité, de la quantité et de la représentativité des données disponibles. Des données bruitées, déséquilibrées ou biaisées peuvent fausser les résultats

I.2.16.2 Sur-apprentissage (overfitting) :

Un modèle peut parfois mémoriser les données d'entraînement au lieu de généraliser. Cela entraîne un sur-apprentissage, où le modèle obtient de très bons résultats sur les données vues, mais échoue à prédire correctement sur de nouvelles données.

I.2.17 Évaluation des performances d'un modèle de ML:

L'évaluation des performances d'un modèle de ML est une étape cruciale pour mesurer son efficacité et sa capacité à généraliser sur de nouvelles données. Il existe plusieurs métriques utilisées pour évaluer la performance d'un modèle, et le choix de la métrique dépend du type de problème (classification, régression, etc.) et des objectifs du modèle. Voici un aperçu des différentes approches et des formules courantes [5] :

I.2.17.1 Pour les problèmes de classification

Les modèles de classification attribuent des labels à des données en fonction d'un apprentissage supervisé. Les métriques couramment utilisées pour évaluer ces modèles sont :

a) Exactitude (Accuracy):

L'exactitude mesure la proportion de prédictions correctes parmi toutes les prédictions faites.

$$\text{Exactitude} = \frac{\text{Nombre de prédictions correctes}}{\text{Nombre total de prédictions}} = \frac{TP+TN}{TP+TN+FP+FN} \quad (\text{I.2})$$

- **TP (True Positives)** : Nombre de vrais positifs (prédictions positives correctes).
- **TN (True Negatives)** : Nombre de vrais négatifs (prédictions négatives correctes).
- **FP (False Positives)** : Nombre de faux positifs (prédictions positives incorrectes).
- **FN (False Negatives)** : Nombre de faux négatifs (prédictions négatives incorrectes)

b) Précision (Precision)

La précision mesure la proportion de prédictions positives correctes parmi toutes les prédictions positives faites par le modèle.

$$\text{Précision} = \frac{TP}{TP+FP} \quad (\text{I.3})$$

c) Rappel (Recall):

Le rappel mesure la capacité du modèle à détecter correctement les positifs parmi tous les vrais positifs possibles.

$$\text{Rappel} = \frac{TP}{TP+FN} \quad (\text{I.4})$$

d) Score F1 :

Le score F1 est une moyenne harmonique entre la précision et le rappel, qui prend en compte à la fois les faux positifs et les faux négatifs.

$$F_1 = 2 \times \frac{\text{précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}} \quad (\text{I.5})$$

e) Courbe ROC et AUC :

La courbe ROC (Receiver Operating Characteristic) trace la relation entre le taux de vrais positifs (rappel) et le taux de faux positifs (1 - spécificité). L'AUC (Area Under the

Curve) mesure l'aire sous cette courbe et donne une idée de la capacité du modèle à différencier les classes positives et négatives [5].

I.2.17.2 Pour les problèmes de régression

Les modèles de régression prédisent des valeurs continues. Voici les principales métriques d'évaluation pour ces modèles :

a) Erreur quadratique moyenne (MSE - Mean Squared Error) :

L'erreur quadratique moyenne mesure la moyenne des carrés des erreurs, c'est-à-dire la différence entre les valeurs prédites et les valeurs réelles.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (I.6)$$

- y_i : valeur réelle.
- $\hat{y}_i$: valeur prédite.

b) Racine de l'erreur quadratique moyenne (RMSE - Root Mean Squared Error) :

La racine carrée de la MSE donne une mesure de l'erreur dans les mêmes unités que les données.

$$RMSE = \sqrt{MSE} \quad (I.7)$$

II.2.17.3. Pour les problèmes de clustering (Non supervisé) :

Les modèles de clustering (comme k-means) regroupent des données similaires sans étiquettes. Les métriques d'évaluation pour ces modèles incluent :

a) Indice de Rand ajusté (ARI - Adjusted Rand Index) :

L'ARI mesure la similarité entre deux partitions des données, en ajustant pour les regroupements aléatoires. L'indice varie entre -1 (partition complètement différente) et 1 (partition identique) [5].

b) Indice de Silhouette :

L'indice de silhouette mesure la qualité des clusters en fonction de la proximité des points au sein d'un même cluster et la séparation entre les clusters.

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (I.8)$$

- $a(i)$: distance moyenne entre un point i et tous les autres points dans le même cluster.
- $b(i)$: distance moyenne entre le point i et tous les autres points du cluster le plus proche [18].

I.3 Méthode des k plus proches voisins :

I.3.1 Définition d'algorithme des k plus proches voisins (K-NN) :

L'algorithme **K-Nearest Neighbors** (K-NN), ou **K plus proches voisins**, est l'un des algorithmes d'apprentissage automatique supervisé les plus simples et les plus populaires. Il repose sur le principe selon lequel des points de données proches dans l'une des caractéristiques (features) ont de fortes chances d'appartenir à la même classe. L'hypothèse de base est que des objets similaires ont tendance à se trouver à proximité les uns des autres.

Le K-NN est utilisé aussi bien pour des tâches de classification que de régression, bien qu'il soit plus couramment appliqué à la classification. Il est apprécié pour sa simplicité, sa facilité d'implémentation, et la compréhension intuitive de son fonctionnement. Un exemple d'application de K-NN se retrouve dans les systèmes de recommandation utilisés par les sites de commerce électronique, qui suggèrent des produits similaires en se basant sur les préférences d'utilisateurs aux profils proches [4].

? I.3.2 Où utiliser K-NN

Le K-NN est utilisé à la fois dans les problèmes prédictifs de classification et de régression. Cependant, lorsqu'il est appliqué à des fins industrielles, il est principalement utilisé dans la classification car il est équitable pour tous les paramètres évalués lors de la détermination de l'utilisabilité d'une technique [19]

1. Puissance de prédiction
2. Temps de calcul

I.3.3Le fonctionnement de l'algorithme k-NN:

I.3.3.1 Principe général de fonctionnement

- On définit en entrée un jeu de données étiquetées, une mesure de distance d , et un entier k .
- Pour chaque nouvelle donnée à classer, l'algorithme calcule sa distance avec toutes les données du jeu d'entraînement.
- Il sélectionne les k plus proches voisins (ceux ayant les distances les plus faibles).
- La classe majoritaire parmi ces k voisins est attribuée à la nouvelle donnée.

I.3.3.2Ecriture algorithmique:

On peut schématiser le fonctionnement de K-NN en l'écrivant en pseudo-code suivant :

Début Algorithme

Données en entrée :

- un ensemble de données D .
- une fonction de définition distance d .

- Un nombre entier K

Pour une nouvelle observation X dont on veut prédire sa variable de sortie y Faire :

- 1) Calculer toutes les distances de cette observation X avec les autres observations du jeu de données D
- 2) Retenir les K observations du jeu de données D les proches de X en utilisant la fonction de calcul de distance d.
- 3) Prendre les valeurs de y des K observations retenues :
 - Si on effectue une régression, calculer la moyenne (ou la médiane) de y retenues
 - Si on effectue une classification, calculer le mode de y retenues
- 4) Retourner la valeur calculée dans l'étape 3 comme étant la valeur qui a été prédite par K-NN pour l'observation X [17]

Fin Algorithme

I.3.4 Avantages et inconvénients de K-NN :

I.3.4.1 Les Avantages :

- **Capacité d'adaptation aux nouvelles données :**

L'algorithme K-NN est non paramétrique et paresseux (lazy learning), ce qui signifie qu'il ne nécessite pas d'étape explicite d'apprentissage ou de construction de modèle. Ainsi, il peut s'adapter facilement à l'ajout de nouvelles données d'apprentissage sans avoir à reconstruire un modèle [14].

- **Interprétabilité des résultats :**

La décision prise par l'algorithme est explicable de manière intuitive : la classe attribuée à une nouvelle observation peut être justifiée en montrant les k voisins les plus proches qui ont influencé cette décision. Cela rend l'algorithme particulièrement transparent et interprétable.

I.3.4.2 Les Inconvénients :

- **Temps de classification élevé :**

Comme l'algorithme ne nécessite pas de phase d'apprentissage préalable, tous les calculs sont effectués au moment de la prédiction, ce qui peut entraîner un temps de traitement important, surtout lorsque le jeu de données est volumineux [14].

- **Sensibilité à la dimensionnalité :**

Lorsque le nombre de caractéristiques non pertinentes est élevé par rapport aux caractéristiques pertinentes, les performances de K-NN se dégradent considérablement. C'est ce qu'on appelle la malédiction de la dimension.

- **Dépendance au choix du paramètre k :**

Le bon fonctionnement de l'algorithme dépend fortement de la valeur choisie pour k .

- Un k trop petit peut rendre l'algorithme trop sensible au bruit.
- Un k trop grand peut noyer les cas pertinents dans la majorité.

- **Choix du bon critère de distance :**

L'efficacité de l'algorithme repose sur le choix d'une mesure de distance adaptée (comme la distance euclidienne, etc.). Ce choix dépend du type et de la distribution des données.

I.3.5 Choix du paramètre K dans l'algorithme K-NN:

Le choix de la valeur K à utiliser pour effectuer une prédiction avec K-NN, varie en fonction du jeu de données. En règle générale, moins on utilisera de voisins (un nombre K petit) plus on sera sujette au sous apprentissage (underfitting). Par ailleurs, plus on utilise de voisins (un nombre K grand) plus, sera fiable dans notre prédiction. Toutefois, si on utilise K nombre de voisins avec $K=N$ et N étant le nombre d'observations, on risque d'avoir du overfitting et par conséquent un modèle qui se généralise mal sur des observations qu'il n'a pas encore vu [17].

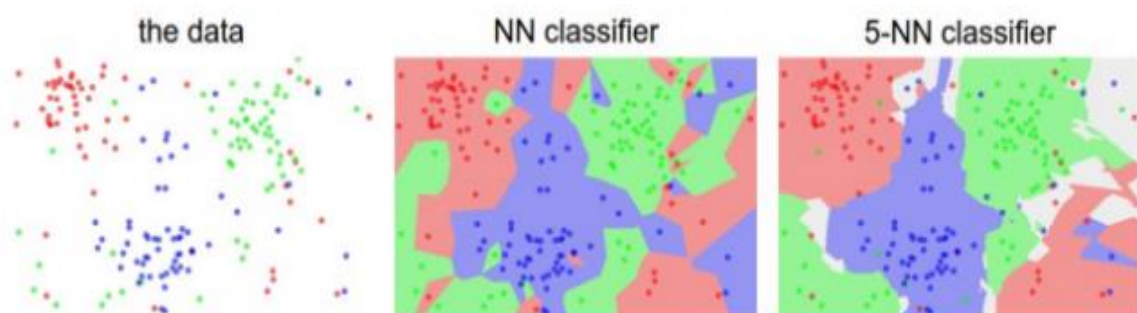


Figure I.8 : Effet du paramètre K sur la frontière de décision de l'algorithme K-NN.

Figure ci-dessus à gauche représente des points dans un plan 2D avec trois types d'étiquetages possibles (rouge, vert, bleu). Pour le 5-NN classifieur, les limites entre chaque région sont assez lisses et régulières. Quant au 1-NN Classifieur, on remarque que les limites sont « chaotiques » et irrégulières. Cette dernière provient du fait que l'algorithme tente de faire rentrer tous les points bleus dans les régions bleues, les rouges avec les rouges etc... c'est un cas d'overfitting [11].

Exemples simple de l'algorithme K-NN:

Exemple 1 : Influencedu choix de la valeur de K dans l'algorithme K-NN

Cette figure illustre l'impact du paramètre K dans la classification d'un point central (représenté par une étoile rouge à l'aide de l'algorithme K des plus proches voisins (K-NN). Les points autour appartiennent à deux classes différentes :

- **Classe A** (en jaune)
- **Classe B** (en violet)

Lorsque **K = 6**, on prend en compte six voisins.

- Cette fois, il y a plus de points de la classe A dans le voisinage.
- Le point central est alors classé comme appartenant à la classe A

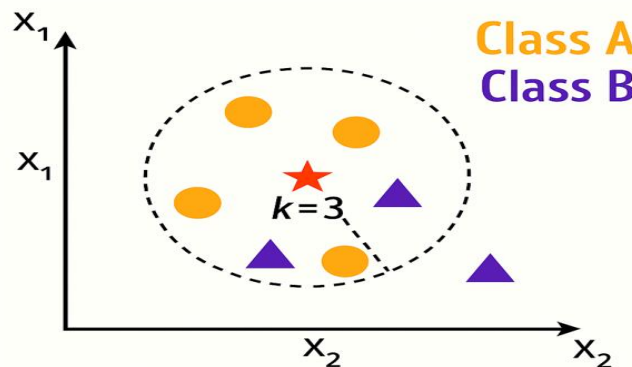


Figure I.9 : Exemple d'Influence du choix de la valeur de K dans l'algorithme K-NN3

Exemple 2 : calcul avec l'algorithme K-NN

Point	X1	X2	classe
P1	1	2	A
P2	2	3	A
P3	3	3	B
P4	6	5	B
P5	7	7	B
P6	8	6	A

Nous voulons prédire la classe du point **Q (5,4)** avec **K=3**

Étape 1 : Calcul des distances euclidiennes

La formule de la distance euclidienne entre deux points P(x1,x2) et Q (x1,y2) est la suivante :

$$d(P, Q) = \sqrt{(x1 - x2)^2 + (y1 - y2)^2}$$

- Distance entre Q(5,4) et P1(1,2) : Étape 2 : Trier les distances

$$d(Q, P1) = \sqrt{(5 - 1)^2 + (4 - 2)^2} = \sqrt{16 + 4} = \sqrt{20} \approx 4.47$$

- Distance entre Q(5,4) et P2(2,3) :

$$d(Q, P2) = \sqrt{(5-2)^2 + (4-3)^2} = \sqrt{9+1} = \sqrt{10} \approx 3.16$$

- Distance entre Q(5,4) et P3(3,3)

$$d(Q, P3) = \sqrt{(5-3)^2 + (4-3)^2} = \sqrt{4+1} = \sqrt{5} \approx 2.24$$

- Distance entre Q(5,4) et P4(6,5) :

$$d(Q, P4) = \sqrt{(5-6)^2 + (4-5)^2} = \sqrt{1+1} = \sqrt{2} \approx 1.4$$

- Distance entre Q(5,4) et P5(7,7) :

$$d(Q, P5) = \sqrt{(5-7)^2 + (4-7)^2} = \sqrt{4+9} = \sqrt{13} \approx 3.61$$

- Distance entre Q(5,4) et P6(8,6) :

$$d(Q, P6) = \sqrt{(5-8)^2 + (4-6)^2} = \sqrt{9+4} = \sqrt{13} \approx 3.61$$

Étape 2 : Trier les distances

Les distances triées par ordre croissant sont :

1. **P4** : 1.41
2. **P3** : 2.24
3. **P2** : 3.16
4. **P5** : 3.61
5. **P6** : 3.61
6. **P1** : 4.47

Étape 3 : Choisir les K plus proches voisins

Pour **K=3**, on prend les 3 points les plus proches de Q :

1. **P4** (distance = 1.41)
2. **P3** (distance = 2.24)
3. **P2** (distance = 3.16)

Les classes associées à ces points sont :

- **P4** : B
- **P3** : B
- **P2** : A

Étape 4 : Prédiction de la classe de Q

Parmi les 3 voisins les plus proches, deux points appartiennent à la classe **B** et un seul à la classe **A**.

L'algorithme K-NN prédit que le point **Q(5,4)** appartient à la classe **B** (majorité des voisins).

I.3.6 Comparatif entre l'algorithme K-NN et d'autres algorithmes supervisés:

L'algorithme K-NN présente une grande simplicité d'implémentation et une intuition claire, ce qui le rend facile à comprendre. Contrairement à des algorithmes comme les SVMs, qui sont particulièrement efficaces pour les données à haute dimension et offrent une bonne généralisation en maximisant la marge entre les classes, K-NN peut devenir lent lorsqu'il faut effectuer des prédictions sur de grands ensembles de données, car il nécessite de calculer la distance entre chaque instance et tous les points d'entraînement. De plus, bien que K-NN soit relativement robuste avec des données non linéaires, il reste sensible au bruit, ce qui peut diminuer ses performances. En revanche, la régression logistique, bien que facile à interpréter et adaptée aux problèmes de classification binaire, ne gère pas aussi bien les relations non linéaires entre les classes. Les arbres de décision, qui sont très interprétables, sont également plus susceptibles au sur-apprentissage, notamment lorsqu'ils sont trop profonds. Cependant, en combinant plusieurs arbres, comme dans les forêts aléatoires, on obtient des modèles robustes et précis, moins sujets au surajustement, mais qui deviennent moins faciles à interpréter. Ainsi, si K-NN offre une solution simple et intuitive, il peut être surpassé en termes de performance par des algorithmes comme les SVM ou les forêts aléatoires sur des ensembles de données complexes et de grande taille [7].

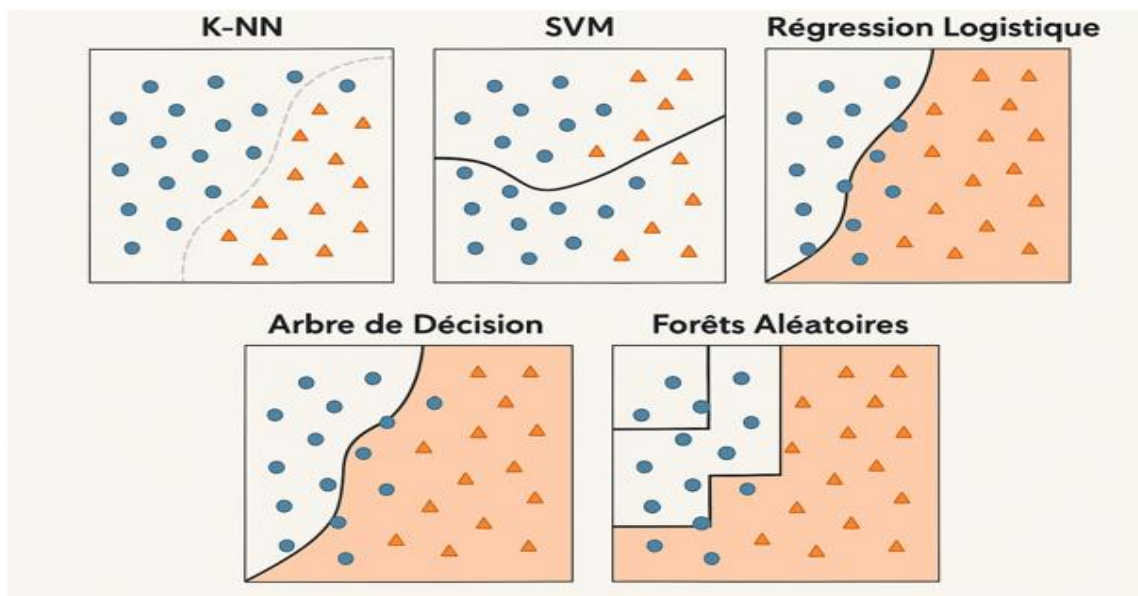


Figure I.10 : L'illustration comparative des frontières de décision pour l'algorithme K-NN et d'autres algorithmes supervisés

I.3.9 Limites de K-NN en contexte réel :

Malgré sa simplicité et son efficacité dans certaines tâches de classification, l'algorithme K-NN présente plusieurs limites lorsqu'il est appliqué à des contextes réels. Premièrement, il est sensible à la dimensionnalité élevée : lorsque le nombre de caractéristiques augmente, les

distances entre les points deviennent moins significatives, ce qui peut dégrader les performances.

Un phénomène connu sous le nom de *malédiction de la dimensionnalité*. Ensuite, K-NN est coûteux en temps de calcul car il nécessite de parcourir l'ensemble du jeu d'entraînement à chaque prédiction, ce qui le rend peu adapté aux grandes bases de données [12].

De plus, l'algorithme ne gère pas bien les données bruitées ou déséquilibrées : des points aberrants ou une forte disparité entre les classes peuvent influencer négativement la décision. Il est aussi important de noter que K-NN ne propose aucun modèle explicite, rendant l'interprétation des résultats plus difficile comparée à d'autres méthodes comme les arbres de décision.

Ces limitations exigent souvent des techniques complémentaires (réduction de dimension, pondération, nettoyage des données) pour rendre K-NN utilisable dans des environnements complexes ou en production [19].

I.4 Conclusion :

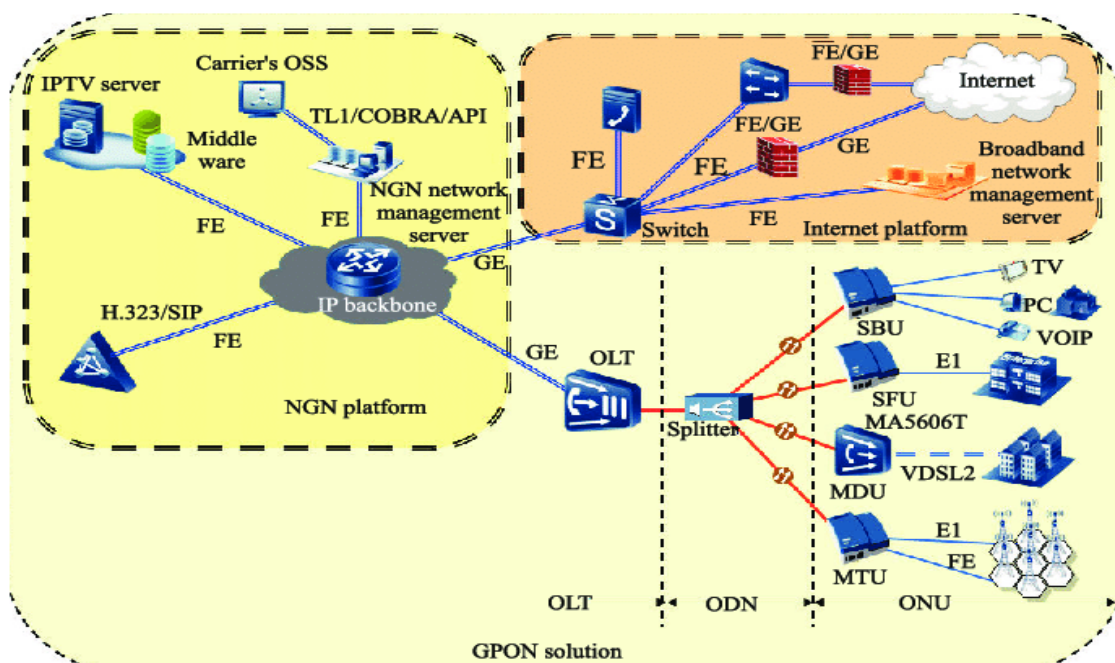
Dans ce chapitre, nous avons présenté les concepts fondamentaux de l'IA, du ML, en mettant en évidence leurs relations hiérarchiques et leurs domaines d'application. Une attention particulière a été portée à l'algorithme K-Nearest Neighbors (K-NN), une méthode d'apprentissage supervisé simple mais puissante, utilisée notamment pour des tâches de classification et de régression. Nous avons détaillé le fonctionnement de l'algorithme K-NN, ses avantages, ses limites, ainsi que l'importance du choix du paramètre K, qui influence directement la précision des résultats obtenus. Ces connaissances constituent une base essentielle pour aborder des applications concrètes de l'apprentissage automatique.

Dans le prochain chapitre, nous allons introduire les principes fondamentaux des réseaux optiques passifs GPON. L'objectif sera d'appliquer les étapes décrites dans ce chapitre pour analyser le comportement du réseau GPON, en utilisant l'algorithme K-NN pour classifier et prédire des phénomènes liés aux performances du réseau, notamment en présence de pannes.

Chapitre II: Le GPON (Gigabit Passive Optical Network)

II .1 Introduction :

Dans un contexte de demande croissante en services Internet à très haut débit, les réseaux en fibre optique se sont imposés comme une solution incontournable pour garantir une qualité de service optimale et une connectivité fiable. Parmi ces technologies, le GPON (*Gigabit Passive Optical Network* – Réseau Optique Passif à haut débit) représente l'une des solutions les plus performantes et rentables dans le domaine des télécommunications modernes. Il permet le transport de données à grande vitesse en s'appuyant sur une infrastructure passive et une architecture de distribution optique partagée.



Error! Bookmark not defined. **Figure II.1** : Réseau GPON.

Le système GPON repose sur une architecture de type "point-à-multipoint", dans laquelle un terminal de ligne optique (OLT – *Optical Line Terminal*) situé au niveau du central est relié à plusieurs équipements chez les abonnés (ONT – *Optical Network Terminal*), via des diviseurs optiques (splitters) passifs. Cette configuration permet de desservir plusieurs utilisateurs tout en réduisant les coûts d'infrastructure, tout en assurant une excellente qualité de transmission. Ce chapitre présente une étude détaillée de la technologie GPON, en commençant par sa définition, son architecture et son principe de fonctionnement. Nous analyserons ensuite ses principaux avantages et limitations. Une attention particulière sera accordée au budget optique, un élément crucial pour évaluer la

portée et la qualité du signal dans le réseau. Nous aborderons également la classification des réseaux GPON selon le budget optique. Enfin, nous établirons un lien avec le thème global de ce mémoire, en mettant en lumière l'intérêt d'appliquer des techniques d'intelligence artificielle, et notamment des algorithmes de classification comme le K-NN, pour analyser les performances des réseaux GPON, détecter les anomalies et anticiper les défaillances. Ce croisement entre réseau optique et apprentissage automatique ouvre la voie vers des infrastructures plus intelligentes, résilientes et autonomes.

II.2 Définition de Fibre optique :

La fibre optique est un fil très fin en verre ou en plastique qui a la propriété de conduire la lumière. Elle est utilisée dans la transmission de données et de lumière. Le signal lumineux permet de transmettre de grandes quantités de données à la vitesse de la lumière sur plusieurs centaines, voire des milliers de kilomètres. La fibre optique offre un débit d'information nettement supérieur à celui des câbles coaxiaux et peut servir de support à un réseau à large bande, permettant ainsi le transit de services tels que la télévision, le téléphone, la visioconférence et les données informatiques[20].

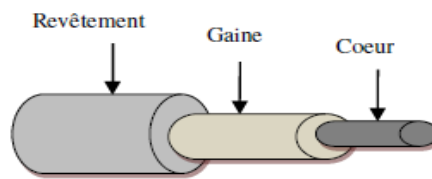


Figure II.2 : Fibre optique

II.3 Définition des systèmes optiques :

Les systèmes optiques sont des dispositifs ou des ensembles de composants conçus pour manipuler et contrôler la lumière afin de réaliser une fonction spécifique. Ces systèmes utilisent des éléments tels que des lentilles, des miroirs, des prismes, des filtres et des fibres optiques pour diriger, focaliser, réfléchir ou modifier les propriétés de la lumière. Les applications des systèmes optiques sont variées et incluent l'imagerie, les télécommunications, la spectroscopie, et l'éclairage.

II.4 Définition des réseaux optiques:

Les **réseaux optiques** sont des systèmes de télécommunication basés sur la transmission de données sous forme de signaux lumineux à travers des fibres optiques. Ces réseaux permettent de transporter de très grandes quantités d'informations sur de longues

distances avec une très faible perte de signal, une bande passante extrêmement élevée, une faible latence, ainsi qu'une forte résistance aux interférences électromagnétiques. Ils sont utilisés dans divers contextes : réseaux longue distance (backbone), réseaux métropolitains (MAN), réseaux de centres de données, et réseaux d'accès comme le FTTH (Fiber To The Home) via des architectures telles que GPON. Les réseaux optiques jouent un rôle central dans le développement des infrastructures numériques modernes, en répondant à la demande croissante de débit imposée par les applications Internet, le cloud computing, la 5G, et l'Internet des objets (IoT) [18].

II.5 Réseaux d'accès optiques :

On définit un réseau d'accès comme l'ensemble des moyens servant à relier les terminaux de télécommunications entre un utilisateur final et un nœud du réseau métropolitain. La distance séparant ces terminaux est souvent de l'ordre de vingt kilomètres (20 km). Le réseau d'accès permet aux utilisateurs de se connecter au cœur du réseau afin de bénéficier des services fournis par ce dernier.

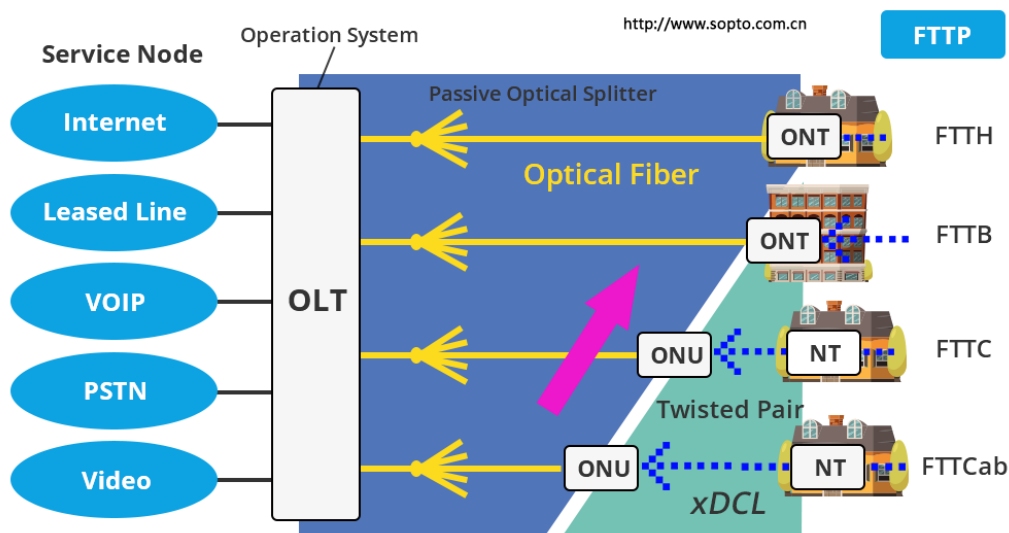


Figure II.3 : Architecture d'un réseau d'accès optique.

Dans ce type de réseau, on distingue deux parties : une partie en fibre optique et une autre en conducteur métallique, qui va jusqu'au terminal de l'abonné. On distingue également les techniques FTTx (Fiber to the x), qui consistent à raccorder la fibre aussi près que possible de l'utilisateur afin d'améliorer la qualité de service et, en particulier, le débit [19].

II.6 Différences entre les réseaux actifs et passifs :

Les réseaux optiques se divisent en deux grandes catégories : les réseaux actifs (AON – Active Optical Network) et les réseaux passifs PON. Les réseaux actifs utilisent des

équipements actifs comme des commutateurs, des routeurs ou des répéteurs pour gérer et diriger le trafic optique entre le fournisseur et les utilisateurs finaux. Cela permet une grande flexibilité dans la gestion du trafic, mais nécessite une alimentation électrique à chaque point intermédiaire, ce qui augmente les coûts d'installation et de maintenance. En revanche, les réseaux passifs, comme GPON, n'utilisent aucun composant actif entre le central et l'abonné. Le signal est simplement divisé à l'aide de coupleurs optiques passifs, ce qui permet de connecter plusieurs utilisateurs à partir d'une seule fibre. Cette architecture réduit considérablement les coûts, tout en permettant une portée pouvant atteindre jusqu'à 20 km. Toutefois, le débit est partagé entre les utilisateurs, contrairement aux réseaux actifs qui offrent un lien dédié à chaque abonné [18].

II. 7 Les technologies d'accès optique:

Les technologies d'accès optique désignent l'ensemble des moyens permettant de connecter les utilisateurs finaux au réseau à l'aide de la fibre optique. Ces technologies visent à répondre à la demande croissante en bande passante due à l'essor d'applications gourmandes en données telles que la vidéo en streaming, le télétravail, la télémédecine et les services en cloud [21].

On distingue principalement plusieurs types d'architectures selon la profondeur du déploiement de la fibre :

- **FTTH (Fiber To The Home)** : la fibre arrive directement au domicile de l'utilisateur final, garantissant les meilleures performances.
- **FTTB (Fiber To The Building)** : la fibre atteint un bâtiment collectif, puis la distribution interne se fait souvent via des câbles en cuivre.
- **FTTC (Fiber To The Curb)** : la fibre arrive jusqu'à un point de proximité, puis la connexion finale se fait par le réseau cuivre existant.

Pour mettre en œuvre efficacement ces réseaux, différentes techniques de transmission optique ont été développées. Parmi elles, les réseaux optiques passifs (PON) jouent un rôle essentiel.

Les PON permettent de partager une seule fibre optique entre plusieurs utilisateurs grâce à des dispositifs passifs (splitters), réduisant ainsi les coûts d'infrastructure. Une des normes les plus répandues dans cette catégorie est GPON (Gigabit Passive Optical Network), qui offre des débits élevés (jusqu'à 2,5 Gbit/s en descendant) tout en assurant une distribution efficace et économique de la bande passante. GPON est ainsi devenu une solution de référence pour le déploiement des réseaux FTTH modernes.

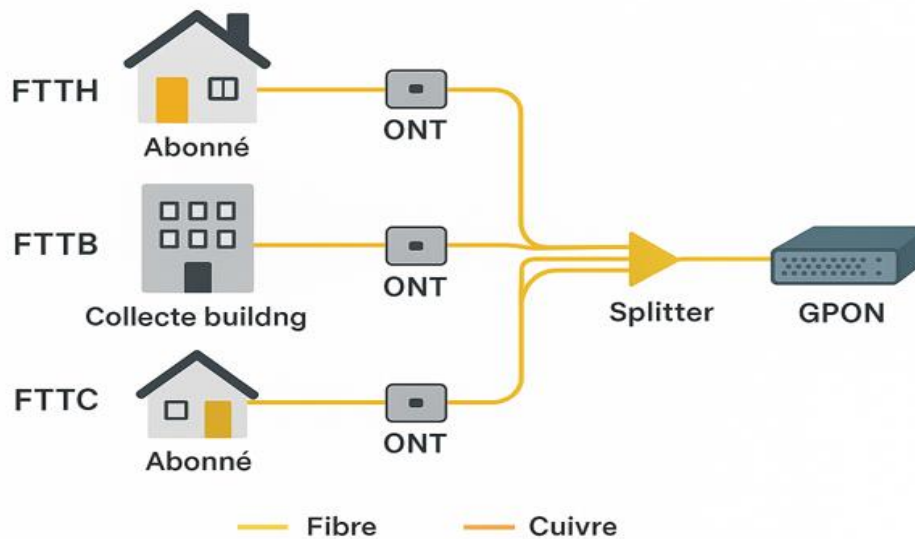


Figure II.4 : Les technologies d'accès optique.

II.8 Les différentes technologies PON:

Les réseaux optiques passifs (PON) sont des architectures essentielles dans le domaine des télécommunications modernes, car ils permettent de fournir des services de haut débit sur de longues distances avec une infrastructure limitée en équipements actifs.

Au fil du temps, plusieurs générations de technologies PON ont été développées pour répondre à l'augmentation constante des besoins en bande passante [22]. Les différentes générations de technologies PON sont citées dans le tableau suivant :

Technologie	Année d'introduction	Débit descendant	Débit montant	Type de transmission	Standard	Particularités
APON	Fin des années 1990	622 Mbit/s	155 Mbit/s	Cellules ATM	ITU-T G.983	Première norme PON, usage limité aux services ATM
BPON	Début des années 2000	622 Mbit/s	155 Mbit/s	ATM + TDM	ITU-T G.983.1	Triple Play, gestion améliorée, sécurité renforcée
GPON	2003	2,5 Gbit/s	1,25 Gbit/s	IP, TDM, ATM	ITU-T G.984	Haute efficacité, compatibilité multi-services
EPON	2004	1 Gbit/s	1 Gbit/s	Ethernet	IEEE 802.3ah	Simplicité IP native, sans ATM
XG-PON	2010	10 Gbit/s	2,5 Gbit/s	IP	ITU-T G.987	Extension de capacité pour usage intensif
XGS-PON	2016	10 Gbit/s	10 Gbit/s	IP	ITU-T G.9807.1	Transmission symétrique, adapté aux entreprises
NG-PON2	2015	≥ 40 Gbit/s	≥ 10 Gbit/s	WDM + TDM	ITU-T G.989	Multiplexage par longueurs d'onde, très haute capacité

Tableau II.1 :Tableau comparatif détaillé des technologies PON.

La première génération, **APON** (ATM-PON), a été conçue pour transporter des cellules ATM (Asynchronous Transfer Mode) à des débits relativement modestes (622 Mbit/s en descendant et 155 Mbit/s en montant). Cette technologie était principalement adaptée aux premiers besoins de transmission de données et aux services d'entreprise [22].

BPON (Broadband PON) est ensuite apparue comme une amélioration de l'APON. Elle a intégré des fonctionnalités de support pour les services Triple Play (voix, vidéo, données) et des mécanismes avancés de gestion et de protection du réseau. Elle utilisait toujours la structure ATM mais visait un public plus large, y compris les utilisateurs résidentiels. Avec l'essor des services numériques et la demande de débits plus élevés, GPON (Gigabit PON) a été introduit. Standardisé sous ITU-T G.984, GPON offre des débits de 2,5 Gbit/s en descendant et 1,25 Gbit/s en montant. Il est capable de transporter simultanément des paquets IP, des cellules ATM et des flux TDM, rendant ainsi la technologie très flexible pour différents types de services.

En parallèle, **EPON** (Ethernet PON), standardisé par l'IEEE (802.3ah), est apparu. Contrairement à GPON, EPON repose directement sur le protocole Ethernet, ce qui simplifie son intégration aux infrastructures de données existantes. EPON propose un débit symétrique de 1 Gbit/s, et offre un modèle de service plus simple, basé uniquement sur IP.

Avec la montée en puissance du cloud, de la vidéo 4K/8K, et des réseaux mobiles 5G, des technologies plus avancées sont devenues nécessaires. **10G-PON** a donc été développé, sous deux formes principales : **XG-PON** (10 Gbit/s descendant / 2,5 Gbit/s montant) et **XGS-PON** (10 Gbit/s symétriques), pour mieux répondre aux besoins de symétrie dans les services d'entreprise.

Enfin, pour les environnements extrêmement exigeants en termes de capacité et de flexibilité, la technologie **NG-PON2** a été introduite. Elle repose sur le multiplexage en longueurs d'onde (WDM) combiné à du multiplexage temporel (TDM), permettant de superposer plusieurs canaux sur une même fibre et d'atteindre des capacités dépassant les 40 Gbit/s, voire davantage avec l'agrégation de plusieurs longueurs d'onde[22]. Dans ce chapitre, l'attention sera portée principalement sur GPON, en raison de sa large adoption mondiale et de son rôle clé dans le déploiement des réseaux FTTH (Fiber To The Home).

II.9 GPON:

II.9.1 Définition :

GPON (Gigabit Passive Optical Network) est une norme de réseau d'accès optique passif à haut débit, définie par l'**ITU-T** dans la série de recommandations **G.984**. Elle utilise

une architecture **point-à-multipoint** pour acheminer des données via des fibres optiques sans composants actifs entre le fournisseur de services et l'utilisateur final.

GPON permet des débits allant jusqu'à **2,5 Gbit/s** en téléchargement (downstream) et jusqu'à **1,25 Gbit/s** en envoi (upstream), tout en assurant une gestion efficace de la bande passante pour supporter simultanément divers types de services : **Internet haut débit, téléphonie (VoIP) et télévision (IPTV)**. Grâce à son efficacité, sa fiabilité et ses coûts de déploiement optimisés, GPON est largement utilisé pour développer les réseaux **FTTH** (Fiber To The Home) [22].

Pourquoi choisir GPON II.9.2?

1. Distance de transmission plus longue :

GPON utilise la fibre optique pour la transmission du signal. La distance de transmission maximale atteint jusqu'à 20 km.

2. Bande passante plus élevée :

Les transmissions en aval du GPON sont synchronisées à 2,5 Gbit/s. Cela répond à la demande toujours croissante de bande passante élevée.

3. Économie de ressources :

Le rapport de division maximum est de 1:128. Après division, une seule fibre optique fournit des services séparés pour plusieurs utilisateurs, économisant ainsi des ressources.

Dans le même temps, GPON prend en charge les services Triple-Play et propose des solutions de service complet pour résoudre le goulot d'étranglement de la bande passante de l'accès par paires torsadées. C'est parfait pour les demandes des utilisateurs en matière de services Internet haut débit, IPTV et VoIP [23].

II.9.3 Rôle du GPON dans les réseaux d'accès modernes :

Le GPON est une technologie d'accès basée sur la fibre optique qui joue un rôle clé dans la transformation numérique des infrastructures télécom. Grâce à ses débits très élevés (jusqu'à 2,5 Gbit/s en descendant et 1,25 Gbit/s en montant), elle répond efficacement aux besoins croissants en bande passante dus à l'explosion des usages numériques : streaming, cloud, télétravail, jeux en ligne, etc. Contrairement aux réseaux traditionnels en cuivre (comme le DSL), GPON repose sur une architecture point-à-multipoint passive, où une seule fibre optique est partagée entre plusieurs utilisateurs via des splitters optiques passifs. Ce modèle permet une réduction considérable des coûts d'installation, de maintenance et de consommation énergétique. GPON prend en charge la transmission simultanée de plusieurs services — Internet, téléphonie (VoIP), télévision (IPTV) — sur un seul réseau unifié, ce qui

facilite la gestion pour les opérateurs tout en améliorant la qualité de service pour les utilisateurs. Avec une portée allant jusqu'à 20 km et une capacité de division allant jusqu'à 128 utilisateurs par fibre, GPON est particulièrement bien adapté à la fois aux zones urbaines à forte densité et aux zones rurales éloignées. C'est pourquoi il est aujourd'hui largement adopté dans les projets FTTH (Fiber To The Home) à l'échelle mondiale.

II.9.4 Inconvénients de GPON Les avantages et les

II.9.4.1 Les Avantages

Le réseau GPON présente plusieurs avantages majeurs qui expliquent son adoption dans les réseaux modernes d'accès optique

- **Architecture passive :**

La structure du GPON repose uniquement sur des dispositifs optiques passifs comme les coupleurs optiques (splitters), ce qui élimine le besoin en alimentation électrique intermédiaire et réduit les coûts de maintenance.

- **Optimisation du génie civil et réduction des coûts :**

- **Partage partiel de l'infrastructure :** Une seule fibre optique est partagée entre plusieurs utilisateurs (jusqu'à 32 ou 64), ce qui diminue le nombre total de fibres nécessaires et permet de réaliser des économies importantes.

- **Architecture favorable à la diffusion :** La topologie point-à-multipoint du GPON est naturellement adaptée pour la diffusion de contenus (comme IPTV ou la télévision numérique) vers plusieurs abonnés de manière efficace.

- **Partage de l'OLT :** Un seul équipement OLT peut desservir plusieurs utilisateurs (par exemple, un duplexeur au central pour 32 clients), ce qui améliore l'efficacité du réseau et réduit les coûts d'équipements [24].

II.9.4.2 Les Inconvénients:

Malgré ses nombreux avantages, le GPON présente également certaines limitations et défis techniques :

- **Débit partagé :** Le débit disponible est commun entre tous les utilisateurs connectés au même splitter optique. Ainsi, la bande passante par utilisateur peut varier en fonction de la charge du réseau.

- **Complexité de la synchronisation en liaison montante :**

Dans le sens montant (de l'utilisateur vers le central), plusieurs ONU doivent transmettre sur la même fibre. Pour éviter les collisions, une synchronisation stricte est nécessaire, ce qui complexifie la gestion du réseau.

- **Sécurité des données :**

Tous les ONT reçoivent l'ensemble du flux émis par l'OLT.

II.9.5 Architecture du réseau optique G-PON :

La technologie du réseau GPON repose principalement sur le principe du multiplexage temporel (TDM – Time Division Multiplexing).

Dans cette architecture, deux longueurs d'onde distinctes sont utilisées :

- une longueur d'onde pour le sens descendant (de l'OLT vers les ONT) ;
- et une autre pour le sens montant (des ONT vers l'OLT).

Le GPON se distingue nettement du BPON (Broadband PON) par sa capacité à transporter de manière native des paquets IP et des trames Ethernet de longueurs variables, sans dépendre d'un encapsulage ATM intermédiaire, ce qui le rend plus flexible et plus adapté aux applications modernes. Sur le plan des performances, GPON offre des débits élevés avec une architecture asymétrique typique :

- Jusqu'à 2,5 Gbit/s en descendant ;
- Et environ 1,2 Gbit/s en montant.

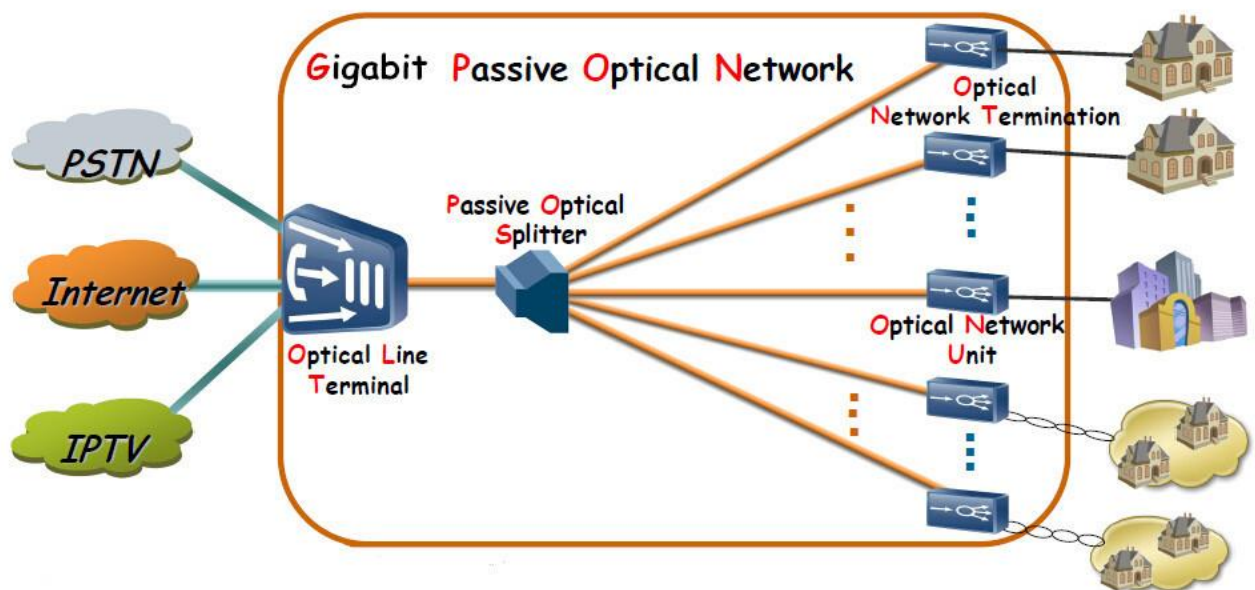


Figure II.5 : Architecture du réseau G-PON.

II.9.6 Fonctionnement du réseau GPON :

1. Architecture point-à-multipoint passive :

Un réseau qui connecte un appareil central à plusieurs utilisateurs via de la fibre optique sans équipements actifs intermédiaires.

2. OLT (Optical Line Terminal) :

L'appareil principal dans le centre du réseau qui envoie et reçoit les données des utilisateurs.

3. Splitter optique passif :

Un dispositif qui divise le signal optique provenant de l'OLT vers plusieurs utilisateurs.

4. ONT/ONU (Optical Network Terminal/Unit) :

Un appareil chez l'utilisateur qui convertit le signal optique en signal électrique pour les équipements.

5. Transmission descendante (broadcast) :

Les données envoyées par l'OLT sont diffusées vers tous les utilisateurs en même temps. La vitesse de transmission des données de l'OLT vers l'ONT peut atteindre 2,5 Gbit/s et La distance maximale peut atteindre 20 km

6. Transmission montante (TDMA) :

Chaque utilisateur envoie ses données à un moment précis pour éviter les interférences. La vitesse de transmission des données de l'ONT vers l'OLT peut atteindre 1,25 Gbit/s.

7. Portée maximale jusqu'à 20 km

La distance maximale entre l'OLT et l'ONT peut atteindre 20 km

8. Partage de la fibre (1:32, 1:64) :

Un seul fibre optique peut être partagée entre 32 ou 64 utilisateurs.

II.9.7 Sens montant (de l'ONU vers l'OLT) :

Dans le sens montant, la transmission des données est réalisée par **multiplexage temporel** au niveau du **coupleur passif**. Toutes les ONU émettent sur **la même longueur d'onde** (1,31 μm) en utilisant des **débits plus faibles** et des **équipements optiques moins coûteux**, tels que des **lasers Fabry-Pérot**. Toutefois, si deux ONU transmettent simultanément, leurs signaux se superposent au niveau du coupleur, générant ainsi un **mélange de données illisible** pour l'OLT.

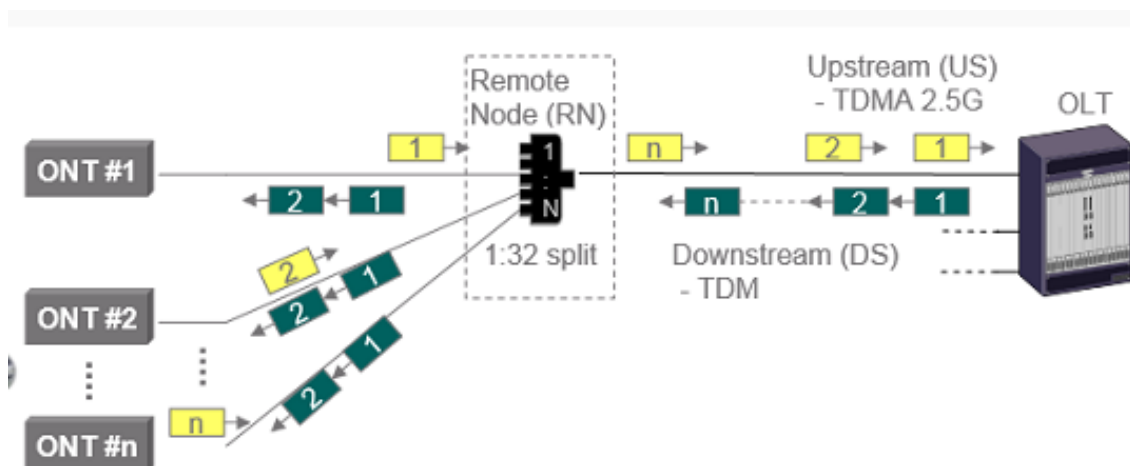


Figure II.6 : Architecture du sens montant.

L'OLT attribue successivement à chaque ONU **un intervalle de temps dédié** (de l'ordre de quelques microsecondes) durant lequel seule cette ONU est autorisée à émettre, afin d'**éviter les collisions** entre les transmissions. Ce mécanisme permet également à l'OLT de **recupérer** et de **synchroniser rapidement l'horloge** avec les données provenant d'utilisateurs situés à **des distances différentes**

II.9.8 Sens descendant (de l'OLT vers les ONU) :

Dans un réseau GPON, le **sens descendant** désigne la transmission des données de l'OLT vers les différentes ONU. Cette transmission utilise une longueur d'onde porteuse de $1,49 \mu\text{m}$ ou $1,55 \mu\text{m}$, choisie pour ses excellentes propriétés optiques, notamment une faible atténuation et une capacité à supporter des débits élevés, tout en assurant un budget de puissance optimisé.

L'OLT diffuse l'ensemble des données vers toutes les ONU en mode broadcast. Chaque paquet de données est marqué par une identification spécifique afin que seule l'ONU destinataire puisse reconnaître et traiter les informations qui lui sont destinées. Bien que toutes les ONU reçoivent le même flux, chacune extrait uniquement ses propres données pour les transmettre au réseau local de l'abonné.

Ce mécanisme assure une transmission efficace, sécurisée et sélective des données. De plus, le débit disponible est partagé dynamiquement entre tous les utilisateurs du réseau selon leurs besoins respectifs.

Dans la partie pratique de ce mémoire, nous allons appliquer notre étude au cas du sens descendant (de l'OLT vers les ONU)

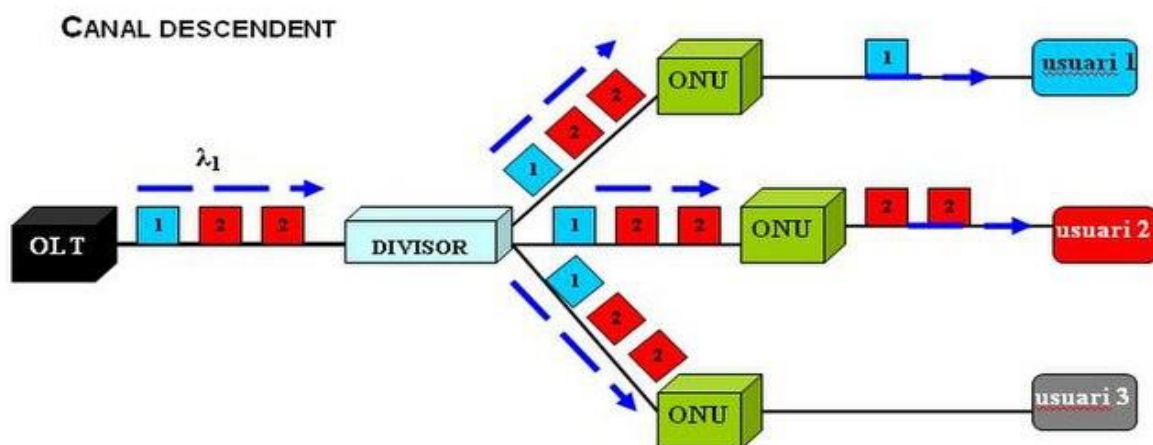


Figure II.7 : Architecture du sens descendant.

II.9.9 Structure de la trame GPON (GTC Frame Format) :

Dans un réseau GPON, la transmission des données repose sur le protocole **GTC (GPON Transmission Convergence)**, qui permet d'unifier différents types de trafic (voix, vidéo, données IP, etc.) dans une structure de trame cohérente.

La trame GTC a une durée fixe de **125microsecondes**, ce qui permet une synchronisation temporelle rigoureuse entre l'OLT et les ONUs. Chaque trame descendante (OLT vers ONUs) se compose principalement de deux parties : un **en-tête GTC (GTC Header)** et une **charge utile GEM (GEM Payload)**.

L'en-tête précise la répartition de la bande passante montante entre les différentes ONUs, ainsi que des champs de synchronisation et de contrôle. La charge utile encapsule les données utilisateur à l'aide de la méthode **GEM (GPON Encapsulation Method)**, permettant le transport efficace de paquets Ethernet, ATM ou autres.

Dans le sens montant (ONU vers OLT), les ONUs envoient leurs trames dans des créneaux temporels attribués dynamiquement par l'OLT selon la table BWmap. Chaque trame montante contient à la fois des données utilisateur (encapsulées dans des trames GEM) et des messages de contrôle.

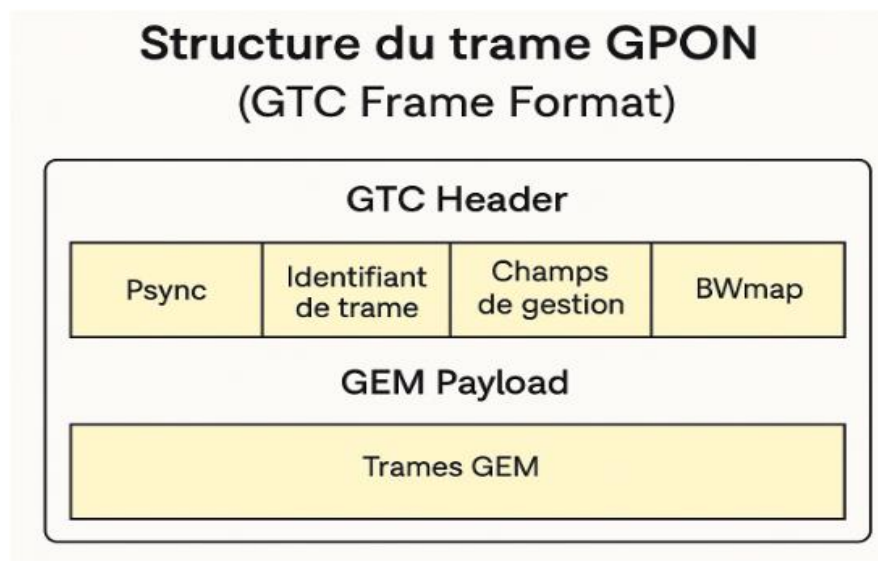


Figure II.8 : Structure de la trame GPON.

Ce mécanisme garantit une communication bidirectionnelle efficace et sans collision, même lorsque plusieurs utilisateurs partagent la même fibre optique.

II.9.10 Gestion de la bande passante dans GPON :

L'un des éléments clés qui confèrent à l'architecture GPON son efficacité est sa capacité à gérer intelligemment la bande passante partagée entre les différents utilisateurs. Cette gestion est assurée par un mécanisme appelé **DBA (Dynamic Bandwidth Allocation)**.

II.9.10.1) Principe de la DBA

Dans un réseau GPON, plusieurs ONT partagent la bande passante descendante (broadcast) et montante (multiplexée en temps). La liaison montante est plus complexe à gérer, car elle nécessite une coordination stricte pour éviter les collisions. C'est ici que le rôle du DBA entre en jeu.

Le **DBA est un algorithme dynamique situé dans l'OLT** qui répartit la bande passante montante selon les besoins de chaque ONT, en tenant compte de divers paramètres tels que :

- Les files d'attente de trafic dans chaque ONT (demande réelle)
- Les types de services (voix, vidéo, données...)
- La priorité du trafic (QoS : Quality of Service)
- Les contrats de service (SLA : Service Level Agreement)

II.9.10.2) Fonctionnement

Le fonctionnement typique du DBA se déroule en plusieurs étapes [25] :

1. **Rapports de demande (BWmap)** : Les ONT envoient périodiquement à l'OLT des rapports (DBRu) indiquant la quantité de données qu'elles ont à transmettre.
2. **Traitement par l'OLT** : L'OLT collecte ces informations et applique un algorithme DBA pour décider de la répartition du temps de transmission (temps d'accès) entre les ONT.
3. **Attribution des créneaux** : L'OLT informe les ONT du temps alloué via des messages de contrôle dans le champ **BWmap** de la trame GPON.
4. **Transmission synchronisée** : Chaque ONT transmet ses données dans le créneau horaire qui lui a été attribué.

II.9.10.3) Avantages de la DBA

- Utilisation efficace de la bande passante montante
- Adaptabilité en temps réel aux besoins des utilisateurs
- Amélioration de la qualité de service (QoS)
- Support de multiples classes de service (T-CONT)

II.9.10.4) Limites :

- Complexité algorithmique côté OLT
- Délai dans l'adaptation dynamique (temps de réponse)
- Performance affectée si les T-CONT ne sont pas bien configurés

II.9.11 Les caractéristiques principales de GPON:

Les caractéristiques principales de GPON sont les suivantes :

1. **Architecture Passive** : GPON utilise des composants passifs (splitters optiques) pour diviser et distribuer le signal optique, ce qui réduit les coûts et la complexité de l'infrastructure.
2. **Débit élevé** : GPON supporte des débits allant jusqu'à 2,488 Gbps en downstream (vers l'utilisateur) et jusqu'à 1,244 Gbps en upstream (depuis l'utilisateur).
3. **Efficacité du partage de bande passante** : La technologie permet de partager la bande passante entre plusieurs utilisateurs via un splitter optique, rendant l'utilisation de la fibre plus rentable.
4. **Distance de transmission** : GPON permet une distance de transmission de jusqu'à 20 km entre l'équipement central (OLT) et les utilisateurs finaux (ONT), avec une possibilité d'extension à 60 km dans des conditions spécifiques.
5. **Support de multiples services** : GPON peut supporter la transmission simultanée de voix, vidéo, données et services Internet haute vitesse.
6. **Scalabilité** : GPON peut être facilement étendu pour répondre à une demande croissante, avec la possibilité d'ajouter de nouveaux utilisateurs sans modification majeure de l'infrastructure existante.
7. **Gestion et surveillance intégrées** : La gestion de réseau est facilitée par des outils de surveillance à distance de l'état des équipements et des performances de la ligne.
8. **Qualité de service (QoS)** : GPON intègre des mécanismes de gestion de la qualité de service pour prioriser les différents types de trafic (ex. : voix sur IP ou vidéo) en fonction des besoins des utilisateurs [22]

II.9.12 Normes de GPON:

Les normes de GPON, définies par la recommandation ITU-T G.984, sont divisées en quatre parties, chacune traitant d'aspects spécifiques du réseau GPON :

- **G.984.1** : Cette partie est dédiée à l'architecture du réseau GPON, aux débits de transmission, à la portée du réseau, au taux de partage du coupleur optique, ainsi qu'aux mécanismes de protection du réseau. Elle définit les spécifications de base

permettant une distribution efficace et fiable des données à travers le réseau passif optique.

- **G.984.2** : Cette partie traite du budget optique et des puissances optiques nécessaires au bon fonctionnement du réseau GPON. Elle décrit les conditions nécessaires pour garantir une transmission de qualité, en tenant compte des pertes dans les fibres, des variations de puissance et de la conception du système optique.
- **G.984.3** : Cette section se concentre sur la transmission entre l'OLT (Optical Line Terminal) et l'ONU (Optical Network Unit), en spécifiant les protocoles et les mécanismes utilisés pour assurer une communication fluide et fiable entre ces deux éléments essentiels du réseau GPON.
- **G.984.4** : Cette partie expose les spécifications concernant les échanges de données entre l'OLT et les différents services, comme la téléphonie, la vidéo, et l'accès à Internet. Elle définit les modalités de gestion de la qualité de service (QoS) et les mécanismes de priorisation du trafic pour assurer un service de haute qualité, même lors de l'usage simultané de multiples applications.

Ces normes sont essentielles pour la conception, l'implémentation et la gestion des réseaux GPON, assurant leur performance, leur fiabilité et leur évolutivité, tout en garantissant la compatibilité entre les différents équipements et services fournis

II.10 Applications de la technologie GPON

- **Accès haut débit résidentiel** : La technologie GPON est couramment utilisée pour fournir un accès Internet à haut débit et d'autres services aux utilisateurs résidentiels. C'est une solution idéale pour les zones où les technologies de haut débit traditionnelles sont indisponibles ou peu fiables.
- **Accès haut débit commercial** : La technologie GPON est également utilisée pour fournir un accès Internet à haut débit et d'autres services aux utilisateurs professionnels, tels que les bureaux, les hôpitaux et les écoles. Elle constitue une solution économique et évolutive pour les entreprises ayant besoin d'un Internet rapide et d'autres services.
- **Services vidéo** : La technologie GPON est aussi utilisée pour fournir des services vidéo, tels que l'IPTV et la vidéo à la demande. Grâce à son débit élevé, elle permet de diffuser des services vidéo de haute qualité aux utilisateurs finaux.

- **Services de téléphonie** :GPON est également utilisée pour fournir des services de téléphonie, comme la VoIP. Elle permet aux utilisateurs de passer des appels vocaux de haute qualité via Internet.

II.11 Sécurité et gestion dans les réseaux GPON :

1. Sécurité :

Les réseaux GPON utilisent une architecture **point-à-multipoint**, où le signal optique émis par l'OLT est distribué à plusieurs abonnés à travers un répartiteur passif (*splitter*). Cette diffusion partagée nécessite la mise en place de mécanismes de sécurité robustes pour garantir la **confidentialité des données**.

Pour cela, GPON utilise le **chiffrement AES (Advanced Encryption Standard) 128 bits** dans le **sens descendant** (de l'OLT vers les ONT). Cela signifie que, même si tous les utilisateurs reçoivent les mêmes données en diffusion, seul l'ONT auquel elles sont destinées peut les déchiffrer. Ainsi, un utilisateur ne peut pas intercepter ou lire les données d'un autre.

2. Gestion :

La gestion des équipements dans un réseau GPON repose sur le protocole **OMCI (ONT Management and Control Interface)**. Ce protocole permet à l'OLT de **contrôler, configurer et surveiller** à distance les équipements ONT chez les abonnés.

Grâce à OMCI, l'opérateur peut :

- Configurer les paramètres des ONT à distance ;
- Mettre à jour les firmwares sans intervention physique ;
- Surveiller la qualité de service (QoS) et les performances du réseau ;
- Diagnostiquer rapidement les pannes et anomalies ;
- Gérer dynamiquement les services (voix, vidéo, données) selon les besoins.

Ce système de gestion centralisée rend GPON **très efficace à grande échelle**, avec un minimum d'intervention humaine.

II.12 Les pannes dans un réseau GPON

Dans le contexte du diagnostic des réseaux, une panne se définit comme la cause racine d'une ou de plusieurs anomalies survenues sur le réseau et observées sous forme d'alarmes ou de paramètres hors du standard de fonctionnement nominal du réseau.

- **Scénarios de défaillance GPON**

Scénario 1 : GPON simple (un seul client est affecté)

Il y a trois défauts potentiels lorsqu'un seul abonné ne peut pas recevoir de service: défaut de la fibre de distribution entre le client et le séparateur le plus proche, défaut de l'équipement ONT ou défaut du câblage du domicile du client.

Scénario 2 : GPON en cascade (tous les clients concernés sont connectés au même répartiteur)

Lorsque tous les clients connectés au même séparateur ne peuvent pas recevoir le service, alors que d'autres connectés à la même terminaison OLT peuvent en être la cause, l'une des deux peut être: une défaillance du dernier séparateur ou une défaillance de la liaison par fibre optique entre les séparateurs en cascade.

Scénario 3 : tous les clients sont affectés (au niveau de la terminaison OLT)

Que le réseau GPON soit en cascade ou non, tous les clients dépendant du même terminal OLT peuvent être affectés. Si tous les clients sont concernés, la cause en est peut-être l'une des trois: défaillance du séparateur le plus proche de la terminaison OLT, défaillance du câble de fibre d'alimentation du réseau ou défaillance de l'équipement OLT.

II.13 Application du Machine Learning dans PON :

- **Détection et localisation des pannes :** Des études ont montré une précision de 97 % pour la détection et une localisation avec erreur moyenne de 0,18 m, Diagnostic de coupures fibre ou ONU défectueuse via l'analyse des traces OTDR (Optical Time Domain Reflectometer) avec des autoencodeurs. Dans les réseaux PON, l'apprentissage automatique permet de détecter et de localiser les anomalies physiques - notamment les ruptures de fibre ou les défaillances d'ONU - à partir de données OTDR souvent bruyantes. Une méthode avancée combine un autoencodeur convolutionnel (ou à base de GRU) pour dénoncer l'anomalie (seuil sur l'erreur de reconstruction), suivi d'un BiLSTM (Bi-GRU) enrichi d'un mécanisme d'attention pour classifier le type de défaut (coupure, connecteur sale...) et en déterminer précisément l'emplacement, avec une précision de l'ordre de 0,2 m. Les résultats issus de bases réelles démontrent une détection avec $F1 \approx 97\%$, un diagnostic à $\approx 98\%$ et une localisation au dixième de mètre. Cette approche permet une automatisation robuste, opérant même avec un faible rapport signal/bruit, et garantit un retour de service rapide sans nécessiter une intervention humaine ni un calibrage manuel intensif.

- **Prédiction des performances du réseau PON:** sur la base de paramètres tels que longueur de fibre, puissance d'émission, nombre de splitters, etc. Des algorithmes comme random forest, regression trees, logistic regression ou model trees permettent de prédire le facteur Q avec 93–95 % de précision. Ou le facteur est une mesure de performance

combinant rapport signal/bruit et dispersion, utilisée pour évaluer la **qualité du signal** transmis.

Des modèles supervisés tels que Random Forest, arbres de régression (model trees) et régression logistique sont entraînés à partir de données issues du plan de contrôle : longueur de fibre, puissance émise par les ONU/OLT, nombre et type de splitters, pertes dues à la dispersion ou aux connecteurs, SNR, OSNR, température, etc.

Le Random Forest Regression excelle dans ce contexte - chaque arbre opère sur un échantillon aléatoire des données (bagging), et la prédiction finale est la moyenne des sorties des arbres, offrant une robustesse élevée aux variables bruitées et un usage direct sur données brutes . Typiquement, ces modèles atteignent un R^2 de 0,93 à 0,95, signifiant une prédiction très fiable du facteur Q, avec une Erreur Quadratique Moyenne (RMSE) basse.

Les model trees enrichissent ces approches en intégrant, dans chaque feuille de l'arbre, une régression linéaire, ce qui améliore la modélisation des relations continues (comme la variation de perte selon la longueur de fibre) tout en conservant une bonne explicabilité.

Enfin, la régression logistique (adaptée pour la classification $Q > \text{seuil}$) permet de filtrer efficacement les scénarios de qualité médiocre : en transformant le facteur Q en une variable binaire (« acceptable / non acceptable »), elle atteint environ 93–95 % de précision, ce qui est utile pour déclencher des alertes automatiques patents.

L'analyse de l'importance des variables dans la forêt aléatoire révèle que la longueur de fibre et le nombre de splitters sont les paramètres les plus influents sur le facteur Q, informant ainsi des leviers d'optimisation opérationnelle.

II.14 Conclusion :

Dans ce chapitre, nous avons étudié en profondeur la technologie GPON, en détaillant son architecture, son mode de fonctionnement, ses avantages ainsi que ses limites. L'accent a également été mis sur l'importance du budget optique, un élément clé dans le dimensionnement et la classification des réseaux GPON, permettant d'adapter l'infrastructure aux contraintes de distance, de performance et de densité d'abonnés.

Grâce à son efficacité, sa capacité à offrir des débits élevés et son architecture passive optimisée, le GPON s'impose aujourd'hui comme une solution privilégiée pour le déploiement des réseaux FTTH à travers le monde.

Dans la continuité de ce travail, et dans le chapitre 3, nous nous pencherons sur l'analyse du réseau GPON dans le but de détecter et classifier les pannes susceptibles d'en perturber le fonctionnement. Cette démarche reposera sur l'utilisation de techniques d'apprentissage automatique (Machine Learning), en particulier l'algorithme K-NN (K-Nearest Neighbors)

Chapitre III : Classification et détection des pannes par l'algorithme K-NN

III.1 Introduction :

Dans ce chapitre, nous analysons l'impact des pannes sur un réseau GPON unidirectionnel en combinant OptiSystem et MATLAB. OptiSystem sert à modéliser le réseau optique, simuler les effets de pannes à différents points (ruptures de fibre, pertes de signal, etc.), évaluer la qualité de transmission (facteur Q, BER, SNR...) et en extraire des jeux de données correspondant à ces cas problématiques. Ensuite, MATLAB est chargé d'implémenter et d'entraîner un algorithme K-Nearest Neighbors (K-NN) sur ces données : il permet de classifier automatiquement les types de défaillances observées. Les résultats (précision, rappel, taux d'erreur) sont alors étudiés pour concevoir un système prédictif de défaillances adapté à notre configuration GPON. L'ensemble fournit ainsi une démarche de diagnostic prédictif, automatisée et validée, pour anticiper et localiser les pannes dans le réseau.

III.2 Présentation du logiciel Optisystem:

Les systèmes de communication optique sont complexes à concevoir et à simuler. Leur conception implique des composants non linéaires ainsi que des sources de bruit non gaussiennes, ce qui complique le travail des ingénieurs. Pour réaliser et optimiser une chaîne de transmission numérique à haut débit sur fibre optique, il est nécessaire d'adopter une méthodologie rigoureuse de caractérisation et de mesure sur la liaison optique.

Dans ce cadre, l'utilisation d'un simulateur, **Optisystem**, de systèmes permet de simplifier cette tâche en offrant un environnement propice à l'analyse et à l'optimisation des performances du réseau optique.

III.3 Configuration étudiée:

Une architecture GPON uni-directionnelle a été simulée à l'aide du logiciel OptiSystem 7, avec une configuration comportant un OLT central et huit ONU répartis sur des branches de longueurs de fibre variables. Le signal est transmis de manière uni-directionnelle depuis l'OLT vers chaque ONU via des coupleurs optiques (Power Splitters). Chaque unité ONU dispose d'une chaîne de réception composée d'un photodétecteur PIN, d'un filtre Bessel et d'un régénérateur 3R pour assurer la qualité du signal. Un analyseur BER permet d'évaluer la performance de chaque liaison.

Les données recueillies sont ensuite exploitées par l'algorithme K-NN pour détecter et classifier les éventuelles pannes dans le réseau. La Figure III.1 illustre la configuration du réseau GPON.

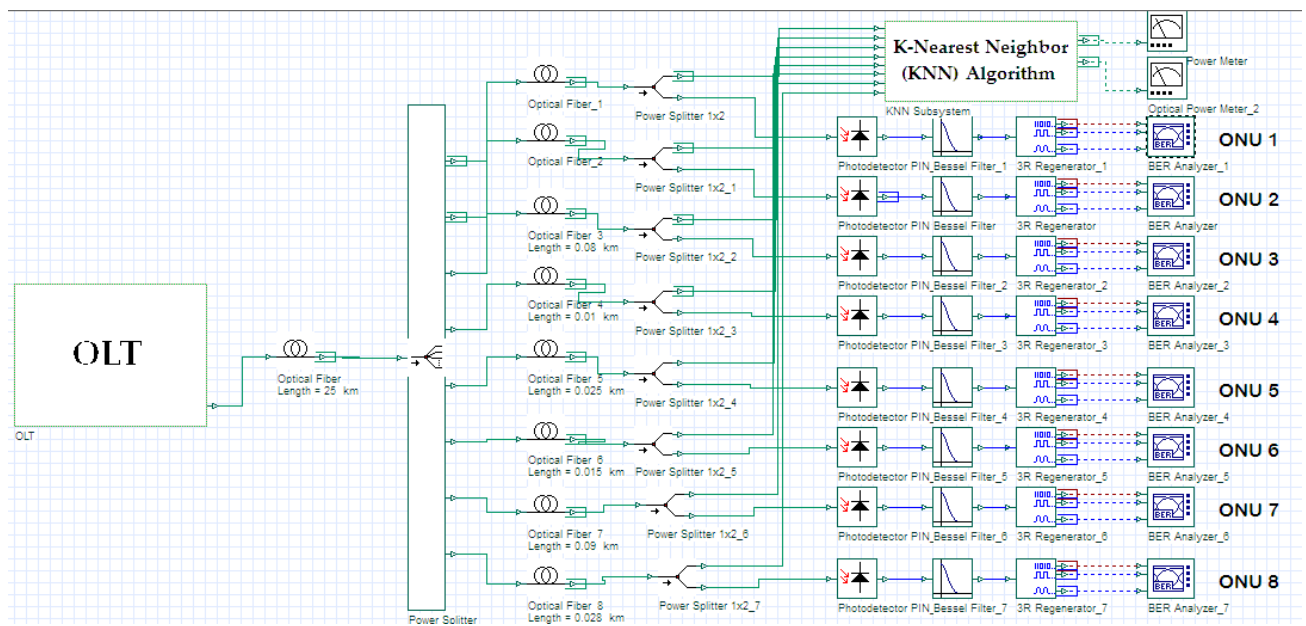


Figure III.1 : Liaison GPON unidirectionnelle.

1. Le bloc d'émission (OLT) : est un circuit intégré. Il fonctionne à une longueur d'onde de **1490 nm** avec un débit binaire de **1.5 Gbits/s**. Le OLT est composé d'un générateur PRBS; ce composant est un générateur de séquences binaires pseudo-aléatoires. Il permet de simuler un flux de données numériques avec un débit binaire de 1.5Gbits/s. Ce flux représente les données à transmettre dans le système GPON. Le deuxième bloc est le générateur NRZ; il s'agit d'un générateur d'impulsions électriques de type NRZ (Non-Return to Zero). Ce bloc code la séquence binaire fournie par le PRBS en une forme adaptée à la modulation, où chaque bit est représenté par une impulsion distincte. Le troisième bloc est la source optique (White Light Source) ; elle génère une lumière continue à une longueur d'onde de 1490 nm, utilisée pour la transmission descendante dans le réseau GPON. Enfin, le dernier bloc est le modulateur Mach-Zehnder ; c'est un modulateur optique externe qui module la lumière émise par la source en fonction du signal NRZ. Le type de modulation utilisé est une modulation d'amplitude avec un taux d'extinction de 30 dB, assurant une bonne qualité de transmission du signal optique.

1. Le bloc de canal : Est composé de deux fibres optiques, chaque ONU est connectée par un segment de fibre optique individuel de longueurs variables (0,028 km à 0,05 km). Ces segments simulent le déploiement réel de la fibre, où des distances différentes peuvent entraîner des niveaux d'atténuation variables. La longueur de la fibre optique entre l'OLT au répartiteur est de 25 km, ce qui correspond à un déploiement GPON réaliste.

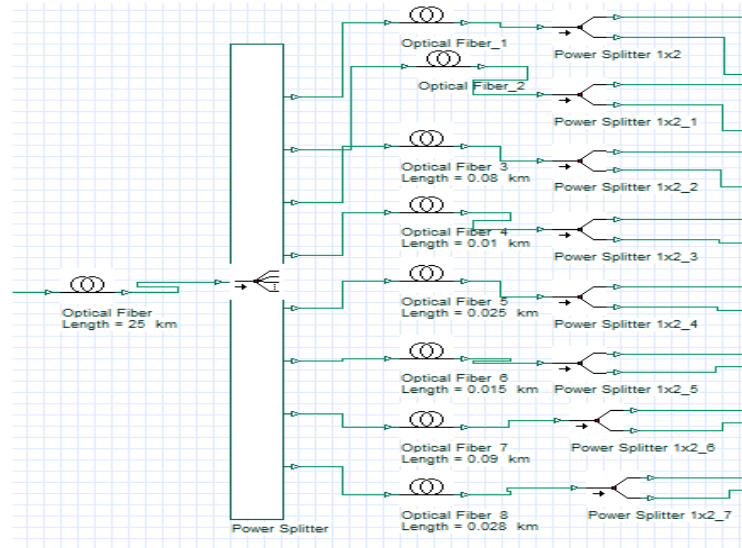


Figure III.2 : Canal de transmission.

Répartiteurs optiques (Power Splitters 1x2) :

Ces dispositifs divisent le signal optique unique en plusieurs branches afin de desservir plusieurs utilisateurs à partir d'une seule source. Ils fonctionnent de manière passive, ce qui signifie qu'ils ne nécessitent aucune alimentation électrique pour opérer ; ils répartissent simplement le signal sans amplification ni traitement.

3. Le bloc de réception (ONU) :

La figure suivante illustre la partie réception au niveau de l'ONU :

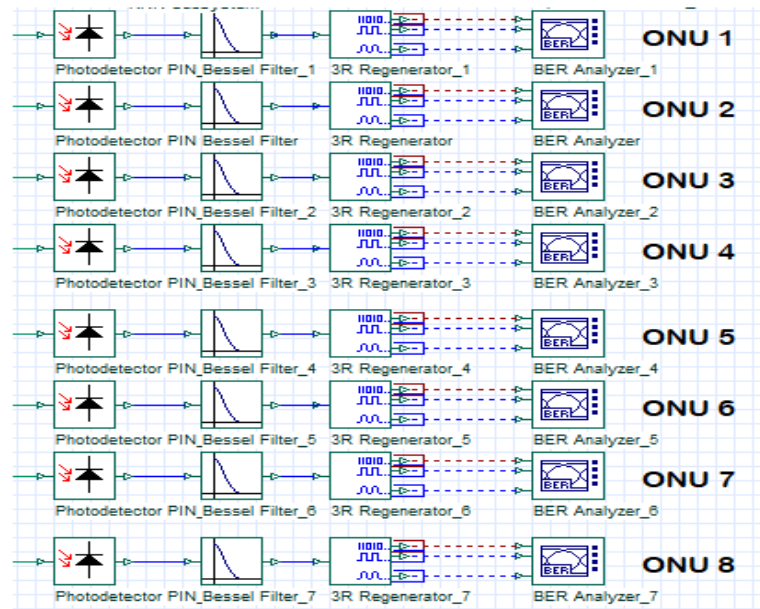


Figure III.3 : Bloc de réception.

Photodétecteur PIN et Filtre de Bessel :

Le **photodétecteur PIN** a pour rôle de **convertir le signal optique reçu en signal électrique** utilisable par les équipements électroniques

Le **filtre de Bessel** est intégré en aval du photodétecteur afin de **nettoyer le signal électrique** en réduisant le bruit hors bande, ce qui améliore la qualité du signal avant les étapes suivantes.

Régénérateur 3R :

Le régénérateur 3R assure trois fonctions clés pour restaurer l'intégrité du signal transmis à travers la fibre optique :

- **Réamplification** : amplification du signal affaibli par la propagation.
- **Reformage** : correction de la forme des impulsions pour réduire les distorsions.
- **Resynchronisation** : réalignement temporel du signal pour compenser les décalages dus à la transmission.

Analyseur BER (Bit Error Rate):

L'analyseur BER est un outil de mesure essentiel dans le système GPON, permettant d'évaluer la **qualité du signal reçu** en calculant le taux d'erreurs sur les bits reçus. Cela sert à vérifier la fiabilité de la transmission et la performance globale du réseau.

Le **BER** est le **rapport entre le nombre de bits erronés reçus** et le **nombre total de bits transmis** sur une certaine période ou sur un certain volume de données.

$$BER = \frac{\text{Nombre de bits erronés}}{\text{Nombre total de bits transmis}} \quad (\text{III.1})$$

4. Sous-système KNN (K-Nearest Neighbor):

Un système d'entraînement pour extraire les datasets utilisés dans Matlab pour l'algorithme KNN, avec 8 entrées issues de chaque ONU, et deux sorties : la première représente la puissance du ONU désiré et la deuxième représente la puissance des ONUs non désirés avec des différentes combinaisons possibles afin d'entraîner l'algorithme KNN avec une large dataset.

III.4 Résultats de simulations:

1. Analyse de la BER en fonction du coefficient d'atténuation :

Dans la première section, nous présentons les résultats de simulation obtenus à l'aide du logiciel OptiSystem, en utilisant la configuration illustrée à la Figure III.1.

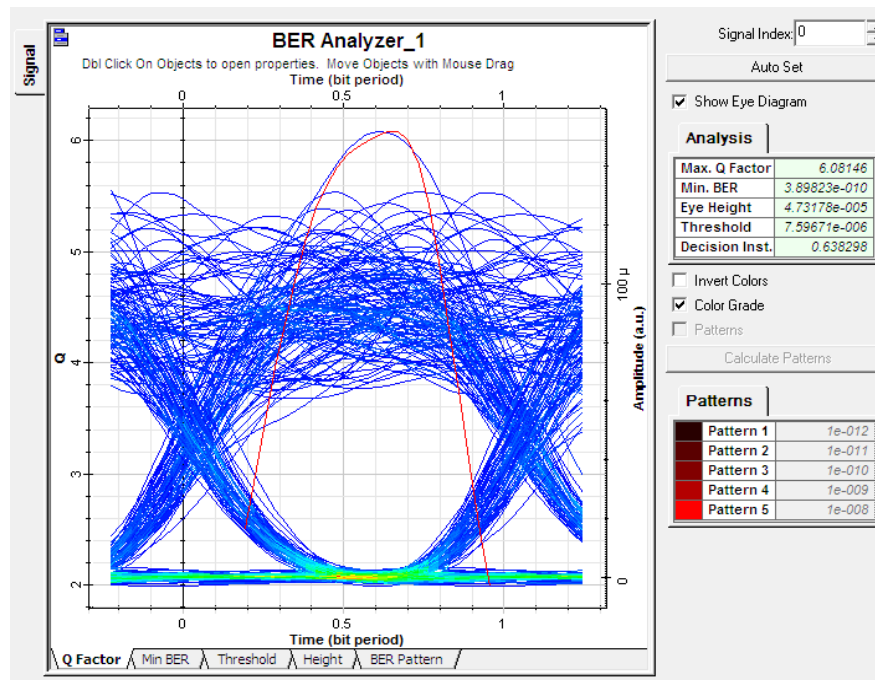


Figure III.4 : Le bloc BER.

La figure ci-dessous indique des performances exceptionnelles du réseau GPON avec un facteur Q élevé de 6,09 et un BER minimal de $3,9 \times 10^{-10}$. Ces résultats confirment que la configuration GPON est acceptable pour des applications critiques comme les liaisons fibre optique, où la qualité du signal est primordiale. Une méthode efficace pour simuler les dommages dans une fibre optique consiste à augmenter le coefficient d'atténuation à certains points ou sur certaines sections de la fibre. Cette approche permet de représenter aussi bien des dommages localisés que distribués. Dans le cas de défauts localisés, une augmentation brusque de l'atténuation est introduite à un point précis de la fibre afin de simuler une fissure, une micro-courbure ou une perte au niveau d'une soudure. À l'inverse, les dommages distribués sont modélisés par une augmentation progressive du coefficient d'atténuation sur une certaine longueur de la fibre, ce qui permet de simuler des effets tels que les courbures, le stress ou la dégradation environnementale affectant une région étendue. Dans le cadre de ce travail, nous nous intéressons particulièrement aux dommages distribués.

2. Évolution du BER en fonction du coefficient d'atténuation pour deux segments du réseau GPON :

α	0.2	0.4	0.6	0.8	1	1.2	1.4	1.6	1.8	2
Log(BER)	-9,46	-9,28	-9,12	-8,98	-8,72	-8,54	-8,40	-8,26	-8,172	-8,15

Tableau II.1 : la liaison entre l'OLT et le splitter.

α	0.2	0.4	0.6	0.8	1	1.2	1.4	1.6	1.8	2
Log(BER)	-9,46	-9,08	-8,82	-8,64	-8,42	-8,24	-8,12	-8,0	-7,92	-7,89

Tableau II.2 : la liaison entre splitter et ONU .

La figure suivante représente l'évolution du taux d'erreur binaire (BER) en fonction du coefficient d'atténuation, pour deux segments distincts du réseau GPON : la liaison entre l'OLT et le splitter, et celle entre le splitter et l'ONU. Elle permet de visualiser l'impact de l'atténuation sur la qualité du signal optique dans chaque partie du réseaux.

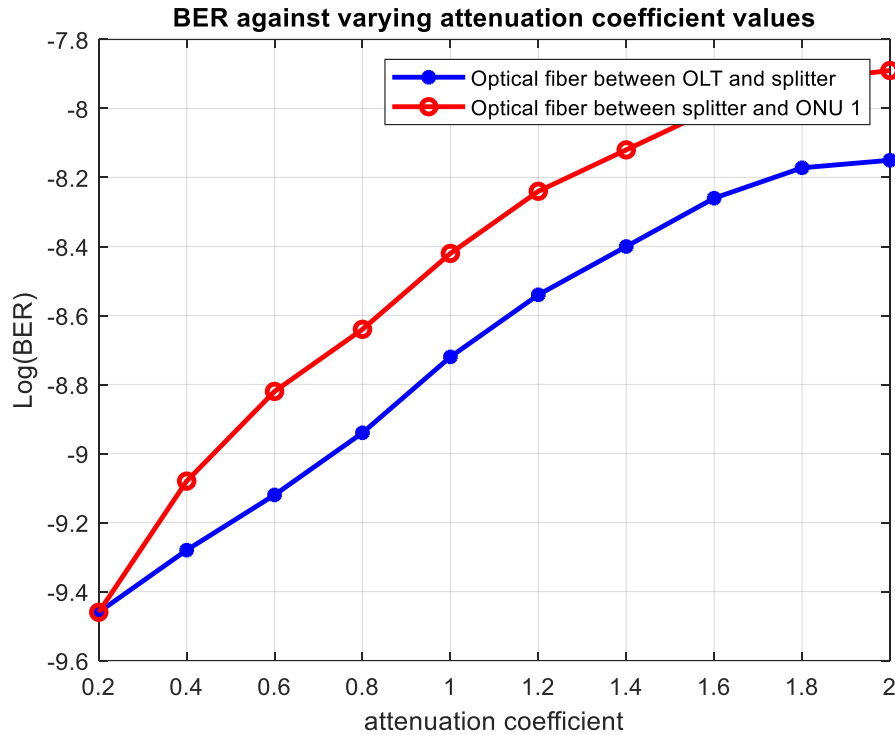


Figure III.5 : Effet des différentes positions de dommages dans la fibre optique sous des valeurs variables du coefficient d'atténuation.

La Figure III.5 illustre l'impact des différentes positions de dommages dans la fibre optique, à savoir entre l'OLT et le répartiteur, ainsi qu'entre le répartiteur et la première ONU, sur les performances du système G-PON proposé, en utilisant l'algorithme K-NN, en termes de taux d'erreur binaire (BER). Il ressort clairement qu'un BER acceptable peut être obtenu lorsque la valeur du coefficient d'atténuation reste inférieure à 0,5 dB/km pour la section de fibre située entre l'OLT et le répartiteur, et inférieure à 0,8 dB/km pour celle située entre le répartiteur et la première ONU. Le graphique met en évidence la relation entre l'augmentation du coefficient d'atténuation et la dégradation des performances du système, mesurée par le BER. On observe une élévation du BER à mesure que l'atténuation augmente, pour les deux segments de fibre optique. Toutefois, la courbe rouge, représentant la fibre entre le répartiteur et l'ONU 1, montre une performance plus dégradée comparée à la courbe bleue, qui correspond à la fibre entre l'OLT et le répartiteur. Cette différence s'explique par les pertes supplémentaires introduites par le processus de répartition ainsi que par la longueur plus importante de la fibre côté ONU. Cela confirme que la section de fibre située entre le répartiteur et l'ONU est plus vulnérable à la dégradation du signal, en raison des

affaiblissements cumulés. Ainsi, cette analyse met en évidence l'importance cruciale de surveiller et d'optimiser les paramètres d'atténuation, notamment dans les sections situées après le répartiteur, pour maintenir la qualité de transmission et garantir la fiabilité du système G-PON.

3. Évaluation des performances de classification des ONUs sous différentes valeurs d'atténuations :

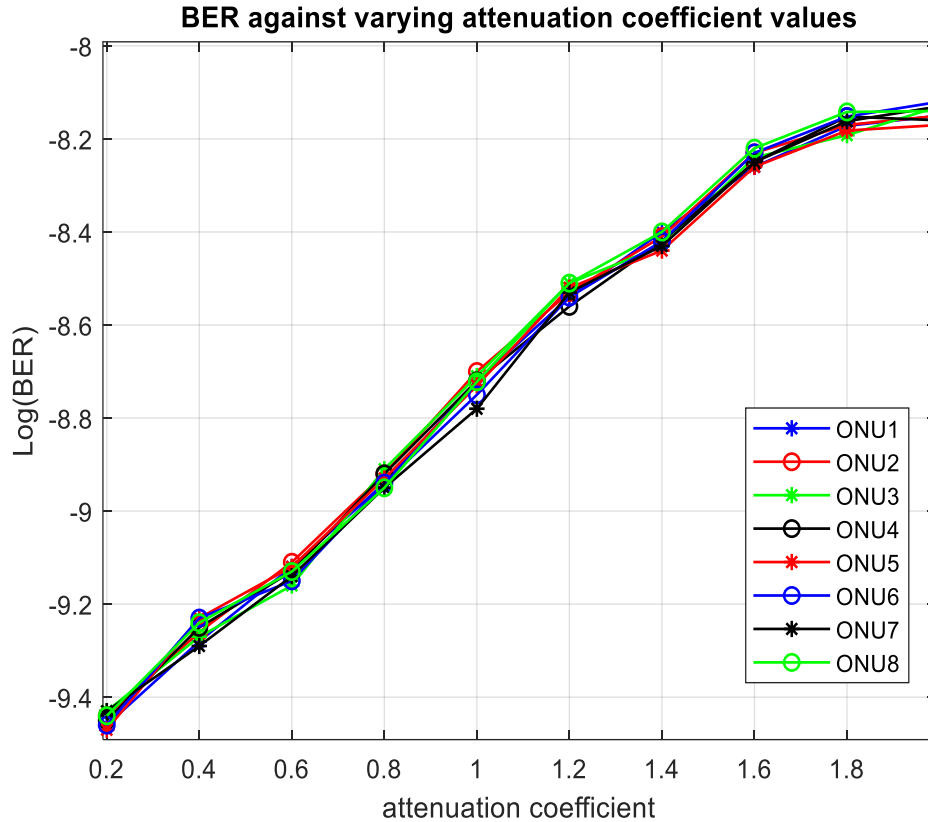


Figure III.6 : Effet des dommages à la fibre optique sur les utilisateurs sous différentes valeurs d'atténuation.

La figure ci-dessus représente une analyse des performances de transmission optique pour les huit unités de réseau optique (ONU1 à ONU8) en fonction du taux d'erreur binaire (BER) et du coefficient d'atténuation. On observe une dégradation progressive du BER avec l'augmentation de l'atténuation, conformément aux attentes théoriques.

Cependant, les différentes ONU présentent des comportements distincts, certaines (comme ONU1 et ONU3) démontrant une meilleure résistance à l'atténuation que d'autres (ONU2 et ONU4). La plupart des ONU maintiennent un BER inférieur au seuil critique de 10^{-9} pour des coefficients d'atténuation inférieurs à ~ 1.0 - 1.2 dB/Km, ce qui suggère une marge opérationnelle acceptable dans ces conditions.

Ces résultats mettent en évidence l'hétérogénéité des performances entre équipements due à la distance de la fibre entre le splitter et les différentes ONU.

2. Application de l'algorithme KNN pour la prédiction des dommages dans les fibres optiques: Avant d'appliquer l'algorithme K-NN, l'extraction et la normalisation des données sont essentielles. La normalisation des données permet de résoudre le problème de la forte variance et de la détection des valeurs aberrantes [3].

Ici, nous extrayons tout d'abord deux paramètres (la puissance reçue par chaque ONU et la puissance totale reçue) à partir de la figure III.1, puis nous les normalisons à l'aide de la méthode du score Z.

La normalisation permet de résoudre ces problèmes en générant de nouvelles valeurs qui préservent la distribution des données tout en maintenant les valeurs dans une échelle commune applicable à l'ensemble des valeurs extraites du modèle. Le score Z est défini comme suit :

$$Z_n = \frac{x_n - \bar{Z}}{\sigma_Z} \quad (\text{III.2})$$

Où Z est l'écart moyen et σ_Z est l'écart type de chaque feature n. Dans cette phase, nous appliquons l'algorithme KNN pour prédire les dommages dans les fibres optiques.

Confusion Matrix for Multi-Class linear K-NN

ONU 1	14	2	1	1		1	1	
ONU 2	2	15	1		1	1		
ONU 3	2	1	14		1			2
ONU 4	1	1	1	14	1		1	1
ONU 5	1	1	1	1	12	1	1	2
ONU 6	1	1		1	1	14	2	
ONU 7	2	1	2	1	1	1	12	
ONU 8			1		1	1	1	16
	ONU 1	ONU 2	ONU 3	ONU 4	ONU 5	ONU 6	ONU 7	ONU 8

Predicted Class

Figure III.7 : Matrices de confusion du classificateur K-NN pour la prédiction des dommages dans les fibres optiques.

L'analyse de la matrice de confusion révèle que le modèle de classification fonctionne correctement, identifiant la majorité des ONUs de manière précise. Les erreurs se produisent principalement entre des ONUs voisines, souvent en raison de caractéristiques de signal

similaires. Par exemple, l'ONU 8 est parfaitement classifiée (16/20), tandis que d'autres subissent de légères erreurs de classification.

Afin d'améliorer la précision globale du système, la stratégie de classification a été spécifiquement ajustée pour distinguer entre l'ONU cible, à savoir ONU #1, et les autres ONUs non souhaitées.

4. Évaluation de la classification binaire ONU #1 / ONU 1 :

La matrice de confusion obtenue met en évidence les performances du modèle dans la différenciation entre ONU #1 et ONU 1, avec une précision globale atteignant 90,63 %. Le modèle a correctement identifié 131 instances correspondant à ONU #1 et 14 instances de l'ONU 1. Toutefois, 9 échantillons appartenant à ONU #1 ont été incorrectement classés comme ONU 1, tandis que 6 échantillons de l'ONU 1 ont été erronément attribués à ONU #1.

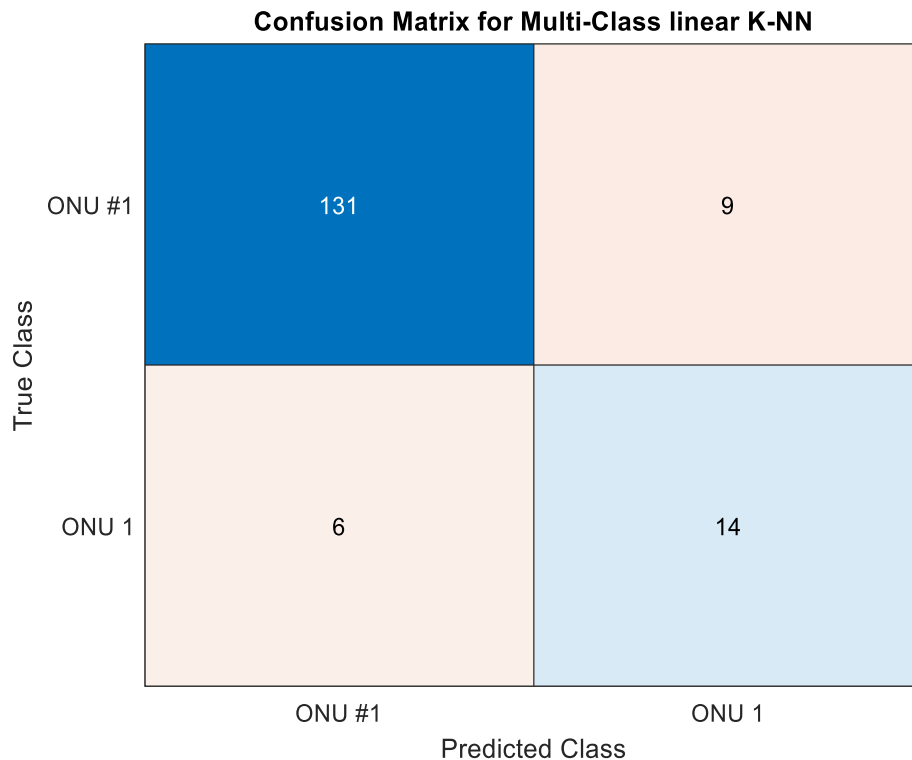


Figure III.8 : Matrices de confusion du classificateur K-NN pour l'ONU 1.

Bien que les résultats obtenus témoignent d'un bon niveau de performance, des erreurs de classification subsistent, en particulier en ce qui concerne ONU 1. Une amélioration des techniques d'extraction de caractéristiques ou une optimisation plus poussée du modèle pourrait permettre d'atteindre une précision encore plus élevée.

La figure III.12 présente une courbe ROC (Receiver Operating Characteristic) pour un classificateur K-NN, utilisée pour évaluer ses performances.

La courbe ROC montre les performances d'un classificateur en traçant le taux de vrais positifs (TPR) en fonction du taux de faux positifs (FPR) pour différents seuils de décision.

Une valeur AUC (Area Under the Curve) de 0,92 indique que le classificateur dispose d'une excellente capacité à discriminer entre les classes positives et négatives (1,0 signifiant une classification parfaite et 0,5 une capacité nulle)

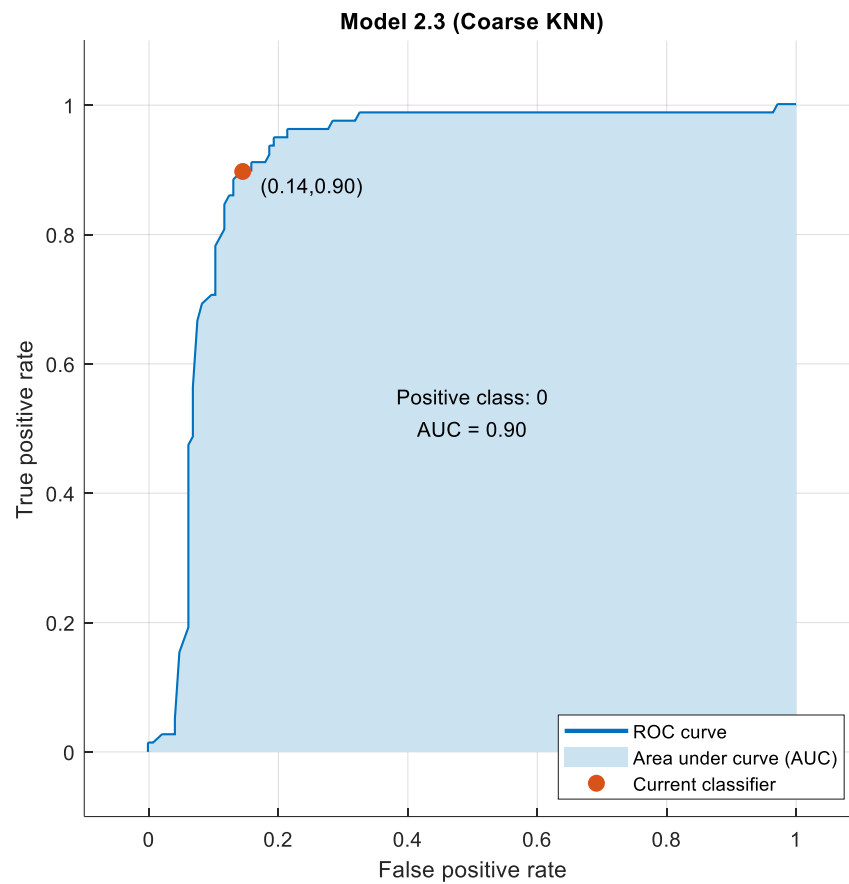


Figure III.9 : ROC du classificateur K-NN

III.5 Conclusion :

Cette étude met en évidence l'efficacité de la détection des pannes dans les réseaux optiques passifs gigabit (GPON) basée sur l'apprentissage automatique, en utilisant l'algorithme des k-plus proches voisins (K-NN). Grâce à des simulations réalisées sous OptiSystem et à une classification des données effectuée dans MATLAB, le système permet de détecter et de classifier avec succès les dégradations du signal affectant plusieurs unités de réseau optique (ONUs).

L'analyse des matrices de confusion révèle une précision de classification variant entre 63,16 % et 75,00 %, avec ONU 2 et ONU 8 affichant les meilleurs taux de précision, tandis que ONU 1 et ONU 7 présentent un taux d'erreurs de classification plus élevé. L'étude souligne également l'impact de l'atténuation sur le taux d'erreur binaire (BER), montrant qu'une augmentation de l'atténuation entraîne une dégradation significative du signal, en particulier dans le segment de fibre situé entre le répartiteur et les ONUs.

Dans l'ensemble, l'intégration de l'algorithme K-NN pour la détection des pannes dans les réseaux GPON améliore la fiabilité du réseau en automatisant le diagnostic des défaillances et en réduisant les temps d'indisponibilité. Toutefois, des améliorations supplémentaires de la performance de classification pourraient être obtenues par une extraction de caractéristiques plus avancée, un réglage optimisé des hyper-paramètres, ou encore l'application de techniques d'apprentissage profond. Les travaux futurs pourraient également explorer l'impact de la surveillance en temps réel et l'utilisation de modèles d'apprentissage adaptatif pour optimiser davantage les performances du réseau dans des conditions variables.

Conclusion générale

À travers ce travail, nous avons exploré l'interconnexion entre les domaines de l'intelligence artificielle, de l'apprentissage automatique (Machine Learning) et des réseaux optiques passifs GPON. Dans un premier temps, nous avons présenté les fondements du Machine Learning, en mettant en avant l'algorithme K-Nearest Neighbors (K-NN), une méthode simple mais puissante pour la classification et la régression. Nous avons souligné l'importance du choix du paramètre K et les implications de ce dernier sur la précision des résultats.

Ensuite, nous avons étudié la technologie GPON en détail, en analysant son architecture, ses principes de fonctionnement, ainsi que ses avantages en termes de débit, de couverture et d'optimisation des ressources. Le rôle critique du budget optique dans le dimensionnement du réseau a été particulièrement mis en évidence.

Enfin, une application concrète de l'algorithme K-NN a été réalisée dans le contexte de la détection et de la classification des pannes dans les réseaux GPON. Les simulations menées sous OptiSystem et MATLAB ont démontré la capacité du modèle à identifier les dégradations du signal avec une précision significative, contribuant ainsi à améliorer la fiabilité et la maintenance des réseaux. Néanmoins, cette approche pourrait être enrichie par des techniques plus avancées telles que l'apprentissage profond ou l'adaptation dynamique en temps réel.

En somme, cette étude met en lumière le potentiel de l'IA dans le domaine des télécommunications, ouvrant la voie à des réseaux plus intelligents, plus performants et plus autonomes.

Conférence Internationale

Zegdou Khaoula, Ahmed Garadi, "Enhancing GPON Reliability Using K-NN Algorithm", 3rd International Conference on Trends in Advanced Research ICTAR 2025 April 04 – 05 in 2025 at Konya/Turkey.

Publication Internationale

Zegdou, K., Garadi, A. (2025), "Intelligent Fault Detection in GPON Networks Using K-NN Algorithm. J Glob Perspect Soc Cult Dev, 1(2), 01-08.

Bibliographie

- [1] BISHOP, C. M. Pattern Recognition and Machine Learning. Springer, 2006, 738 p.
- [2] MUELLER, J. P., and MASSARON, L. Machine Learning for Dummies. WileyPublishing, 2016, 432 p.
- [3] W. A. Imtiaz, H. Y. Ahmed, M. Zeghid, and Y. Sharief, "An optimized architecture to reduce the impact of fibre strands in spectral/spatial optical code division multiple access passive optical networks (OCDMA-PON)," Opt. Fiber Technol., vol. 54, p. 102072, 2020.
- [4] Laouni Bouhezila, Abdelkader Bouarfa, Ahmed Garadi, Abdellah Bensaad, "Cardinality increasing of SAC-OCDMA system using novel decoder design", J. Opt. Commun. January, 2023.
- [5] A. Garadi, A. Djebbari, Taleb-Ahmed Abdelmalik "Exact analysis of signal-to-noise ratio for SAC-OCDMA system with direct detection", Optik, International Journal for Light and Electron Optics, vol. 145, (2017), pp. 89–94.
- [6] Melle MAHDJANE Karima. Détection d'anomalies sur des données biologiques par SVM. Université Mouloud Mammeri de Tizi Ouzou, 14 Octobre 2012.
- [7] Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer. (2006)
- [8] Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. AC Computing Surveys (CSUR), 31(3), 264–323. <https://doi.org/10.1145/331499.331504>
- [9] Alpaydin, E. (2014). Introduction to Machine Learning (3rd ed.). MIT Press.
- [10] Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media
- [11] Aha, D. W., Kibler, D., & Albert, M. K. (1991). *Instance-based learning algorithms*. Machine Learning, 6(1), 37–66. <https://doi.org/10.1007/BF00153759>
- [12] Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.
- [13] Machine Learning et Application en Finance, Université Abderrahmane Mira de B'ejaja Mémoire de fin d'étude.
- [14] Hamouda Djallel, Application des techniques de data mining et de machine learning pour la détection d'intrusions, Université de Jijel, Mémoire de master, 2018.
- [15] Application De L'apprentissage Machine Pour La Reconnaissance Des Activités Humaines, Université Saad Dahlab de Blida 1, mémoire de Master.

Bibliographie

- [16] Shai Shalev-Shwartz & Shai Ben-David, "*Understanding Machine Learning: From Theory to Algorithms*", Cambridge University Press, 2014.
<https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/>
- [17] https://mrmint.fr/introduction-k-nearest-neighbors?utm_source=chatgpt.com
- [18] Stallings, W. (2013). *Data and Computer Communications* (10th ed). Pearson Education.
- [19] Ben abderrahmane, Benamar, "Etude Des Performances Des Réseaux optiques Gpon", universite aboubekr belkaid- Tlemcen, Mémoire de fin d'étude, 2020
- [21] D. Hood, "Gigabit-capable Passive Optical Networks", Wiley, 2012.
- [23] Abdellah Bensaad, Ahmed Garadi, Fatima Zohra Driss-Khodja, "Efficiency enhancement of long-reach passive optical network using the polarization coding", J. Opt. Commun. April, 2023.
- [24] ITU-T G.984.3, "Gigabit-capable Passive Optical Networks (GPON): Transmission convergence layer specification", 2008, <https://www.itu.int/rec/T-REC-G.984.3>
- [25] Zegkou, K. Garadi A (2025). Intelligent Fault Detection in GPON Networks Using K-NN, Algorahon J Glth Perspect Soc Calr Dev 1623, 01-0, doi.org/10.63620/MKJGPSCD.2025.

CERTIFICATE

Of

PARTICIPATION

Zegdou Khaoula

has participated 3rd International Conference on Trends in Advanced Research ICTAR 2025 April 04 – 05 in
2025 at Konya/Turkey.

PAPER TITLE

Enhancing GPON Reliability Using K-NN Algorithm

PRESENTATION TYPE

Oral



All Sciences Academy

Intelligent Fault Detection in GPON Networks Using K-NN Algorithm

Zegdou Khaoula^{1*} & Ahmed Garadi²

¹Department of Electronics, Faculty of Technology, University of Saida Dr Moulay Tahar, Algeria

²Laboratory Technology of Communications, University of Saida Dr Moulay Tahar, Algeria

*Corresponding author: Zegdou Khaoula, Department of Electronics, Faculty of Technology, University of Saida Dr Moulay Tahar, Algeria

Submitted: 31 May 2025 Accepted: 06 June 2025 Published: 13 June 2025

 <https://doi.org/10.63620/MKJGPSCD.2025>.

Citation: Zegdou, K., Garadi, A. (2025). Intelligent Fault Detection in GPON Networks Using K-NN Algorithm. *J Glob Perspect Soc Cult Dev*, 1(2), 01-08.

Abstract

This research aims to improve fault detection and classification in Gigabit Passive Optical Networks (GPON) by utilizing machine learning, focusing on the K-Nearest Neighbors (K-NN) algorithm. The GPON network is extensively simulated using OptiSystem, where essential performance metrics—Optical Power (dBm), Bit Error Rate (BER), and Signal-to-Noise Ratio (SNR)—are analyzed under various fault and interference scenarios. The collected data undergoes preprocessing and normalization before classification with the K-NN algorithm implemented in MATLAB, using the Euclidean distance metric to measure similarity.

Classification results evaluated via confusion matrices show accuracy rates between 63.16% and 75.00% across different Optical Network Units (ONUs). ONU 2 and ONU 8 achieved the highest accuracies of 75.00% and 72.73%, respectively, while ONU 1 and ONU 7 recorded lower accuracies of 63.64% and 63.16%. Additionally, a detailed analysis of fiber attenuation effects on BER reveals significant signal degradation with increased attenuation. This effect is notably more severe in the segment between the splitter and ONUs compared to the path from the Optical Line Terminal (OLT) to the splitter.

These findings highlight the effectiveness of K-NN-based fault diagnosis systems in automating detection and enhancing GPON reliability, thus reducing downtime and operational costs. Future work may explore more advanced machine learning classifiers, improved feature selection, and real-time monitoring techniques to boost detection accuracy and network resilience.

Keywords: G-PONm, K Nearest Neighbor algorithm (K-NN), BER, SNR, ML

Introduction

The ever-increasing demand for high-speed, reliable, and scalable internet access has positioned fiber-optic communication as a fundamental pillar of modern digital infrastructure. Optical systems, leveraging the properties of light for data transmission, have enabled transformative progress in fields such as telecommunications, medical imaging, industrial automation, and scientific instrumentation. Among the leading fiber-based access technologies, Gigabit Passive Optical Networks (GPON) have emerged as a key solution to meet broadband needs for both residential and enterprise environments.

GPON networks utilize a point-to-multipoint topology enabled by passive optical components, allowing a single optical fiber to be shared among multiple users without the need for active elements in the distribution segment. This design reduces operational costs while ensuring high throughput and service quality. Standard GPON implementations operate at downstream rates of 2.5 Gbps and upstream rates of 1.25 Gbps, employing time-sharing protocols such as TDM and TDMA for efficient bandwidth management. However, despite their robustness, GPON systems are not immune to performance degradation resulting from fiber

impairments, component aging, misconfiguration, and environmental factors.

Traditional diagnostic techniques, such as Optical Time-Domain Reflectometry (OTDR), provide vital tools for fault localization. Yet, they often require manual interpretation, lack predictive capability, and may not scale effectively in large, dynamic networks. With the growing complexity of optical infrastructure, there is a critical need for intelligent monitoring and autonomous fault management.

Recent advances in Machine Learning (ML) offer powerful frameworks for enhancing the resilience of optical access networks. By mining data generated from performance monitoring systems, ML algorithms can detect patterns, classify abnormal behaviors, and forecast potential failures. This enables a shift from reactive maintenance toward proactive and even preventive network operation. Techniques such as supervised classification, anomaly detection, clustering, and deep learning have been successfully applied to identify signal impairments, optimize resource allocation, and support self-healing mechanisms in optical networks.

In this context, the K-Nearest Neighbors (K-NN) algorithm is investigated for its effectiveness in GPON fault analysis. K-NN is a non-parametric, instance-based learning method known for its simplicity and versatility. It can be employed to classify various types of optical faults, assess signal degradation, and assist in predictive maintenance workflows by comparing real-time network metrics against historical data patterns. Its ease of implementation and adaptability to diverse data distributions make it an appealing choice for real-world deployment in optical systems.

This work proposes a methodology that integrates K-NN into GPON monitoring systems for fault detection and performance assessment. Section 2 introduces the architecture of GPON, emphasizing upstream and downstream transmission dynamics, along with a mathematical formulation of key metrics such as Signal-to-Noise Ratio (SNR) and Bit Error Rate (BER). Section 3 presents the theoretical basis and application of the K-NN algorithm in the context of network fault classification. Section 4 details the experimental framework, including simulations conducted using MATLAB and OptiSystem platforms, and discusses the performance results. Section 5 concludes with an evaluation of the proposed method's effectiveness and outlines future directions for research in intelligent optical network management.

System Model and Analysis

An optical network typically comprises three core components: the transmitter, the transmission medium, and the receiver. The transmitter section includes a light source—such as a laser diode

or LED—along with a driving circuit that modulates electrical signals into optical ones. The transmission medium mainly consists of optical fiber and may also include additional components like regenerators, splitters, couplers, multiplexers, and connectors, which support signal propagation and enable flexible distribution. At the receiving end, a photodetector—commonly a PIN or avalanche photodiode—converts the incoming optical signal back into an electrical form. This signal is then amplified and processed by a receiver circuit to recover the transmitted data accurately.

Optical Transmission in GPON-Based Access Networks

In fiber-optic access networks—particularly within the "last mile" segment—the infrastructure extends from the service provider's central office to end-user locations such as residences, commercial buildings, and multi-dwelling units. This segment typically adopts a Passive Optical Network (PON) architecture, which utilizes passive optical splitters to distribute signals through a point-to-multipoint tree topology. Such a design enables efficient and cost-effective sharing of optical fiber among multiple subscribers.

Gigabit Passive Optical Network (GPON), standardized under ITU-T G.984, is among the most widely deployed PON technologies. It offers high-capacity broadband access, delivering downstream speeds of up to 2.5 Gbps and upstream speeds of up to 1.25 Gbps. These capabilities support bandwidth-demanding services such as high-definition video streaming, Voice over IP (VoIP), and high-speed internet access.

GPON uses Time Division Multiplexing (TDM) for downstream communication from the Optical Line Terminal (OLT) to Optical Network Units (ONUs), while upstream transmission relies on Time Division Multiple Access (TDMA). This arrangement ensures efficient and fair bandwidth allocation among users. Furthermore, Forward Error Correction (FEC) techniques are incorporated to enhance signal integrity and counteract common transmission impairments such as attenuation and chromatic dispersion.

Thanks to its scalability, reliability, and cost-effectiveness, GPON has become the preferred technology for modern Fiber-to-the-Home (FTTH) and Fiber-to-the-Business (FTTB) deployments. An overview of the general architecture of a GPON system is illustrated in Figure 1.

Signal attenuation in GPON systems can be described using the general optical fiber loss equation:

$$P_{\text{out}} = P_{\text{in}} 10^{-\alpha L/10} \quad (1)$$

Here, P_{out} denotes the output power, P_{in} the input power, α the fiber attenuation coefficient measured in decibels per kilometer (dB/km), and L the fiber length in kilometers (km).

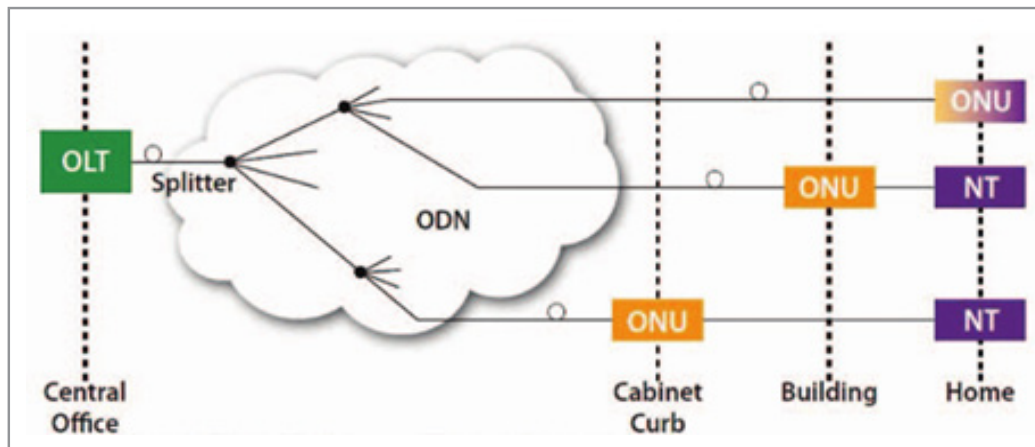


Figure 1: The general structure of a GPON network

Fault detection in Gigabit Passive Optical Networks (GPON)-based Fiber-To-The-Home (FTTH) systems presents a significant challenge due to the absence of physical-layer monitoring at the individual user level. GPON utilizes a point-to-multipoint topology, where numerous Optical Network Units (ONUs) share a single Optical Line Terminal (OLT) through passive splitters. This shared medium limits the operator's ability to isolate faults affecting specific users, unlike active optical networks where each user benefits from a dedicated fiber link and full signal visibility. Conventional diagnostic tools such as Optical Time-Domain Reflectometers (OTDRs) are commonly employed to locate fiber faults; however, their effectiveness diminishes in the shared distribution segment beyond the splitter. Additionally, faults caused by fiber degradation, contaminated connectors, or rogue ONUs inducing signal interference often require labor-intensive field inspections, which escalate operational costs and prolong service outages.

To mitigate these limitations, machine learning offers a promising approach for intelligent, automated fault detection in GPON-based FTTH environments. By monitoring and analyzing key network performance indicators—such as received optical power, bit error rate, latency, and signal-to-noise ratio—data-driven models can detect and classify anomalies with high precision. In this study, we adopt the K-Nearest Neighbors (K-NN) algorithm, leveraging supervised learning techniques trained on historical fault datasets (e.g., fiber cuts, signal degradation, ONU misbehavior). This model enables real-time pattern recognition by comparing live measurements against known fault signatures, thereby facilitating proactive fault localization. The proposed method significantly improves diagnostic accuracy, reduces the reliance on manual intervention, supports timely alerts, and contributes to lower operational expenditures and enhanced network reliability [10]

B.K-Nearest Neighbors (K-NN) Algorithm

The K-Nearest Neighbors (K-NN) algorithm is a widely used supervised machine learning technique applicable to both classification and regression tasks. As a lazy learning method, K-NN

does not involve a separate training phase; instead, it retains the entire training dataset. When a new input instance is introduced, the algorithm identifies the K most similar samples—referred to as the nearest neighbors—based on a predefined distance metric, commonly the Euclidean distance.

For classification problems, the algorithm assigns the input to the most frequent class label among its K nearest neighbors. In regression contexts, it returns the average of the output values corresponding to those neighbors. K-NN's simplicity, interpretability, and effectiveness in handling non-linear decision boundaries make it a popular choice in various practical applications [4-5]

Formally, given a training dataset $\{(x_n, y_n)\}_{n=1}^N$, the algorithm computes the distance between the test sample and each training point, selects the K nearest samples, and makes a prediction accordingly. Since K-NN is non-parametric and instance-based, its performance depends heavily on the choice of K and the structure of the feature space. The Euclidean distance, often used to measure similarity, is calculated as shown in Equation (2) :

$$\text{distance}(x_0, x_n) = \sqrt{(x_0 - x_n)^T (x_0 - x_n)} \quad (2)$$

The query point x is classified based on the majority vote among its k nearest neighbors, as shown in Equation (3) :

$$y_0 = \arg_y \max \sum_{(x_n, y_n)} -\delta(y = y_n) \quad (3)$$

Where: y_0 represents the class label, and y_i corresponds to the class label of the i -th nearest neighbour. The indicator function $\delta(y=y_n)$, returns a value of one if the class y_i of the neighbour x_i matches the class y_0 , and zero otherwise.

G-PON systems are highly recommended because of their straightforward design and strong performance in managing nonlinear data distributions

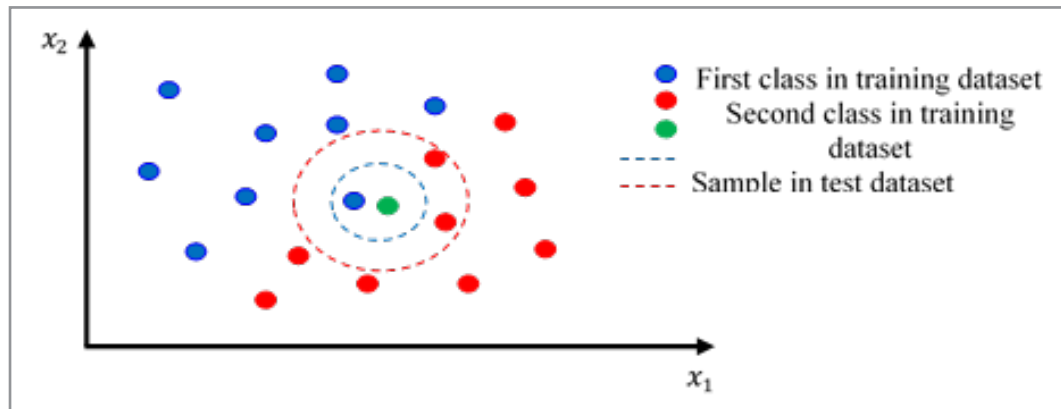


Figure 2:The K-NN algorithm when K=1 and K=3

G-PON System Model Enhanced with K-NN for Fault Detection

Figure 3 illustrates a practical model of a Gigabit Passive Optical Network (G-PON) architecture enhanced with machine learning for proactive fault detection. By integrating the K-Nearest Neighbors (K-NN) classification algorithm, the system effectively identifies signal degradation and predicts potential optical fiber failures. This integration significantly enhances network reliability and optimizes maintenance operations.

The model consists of the following key components

1. **Optical Line Terminal (OLT):** Located at the leftmost part of the architecture, the OLT serves as the central transmis-

sion unit, distributing optical signals to multiple Optical Network Units (ONUs).

2. **Power Splitter:** A cascaded 1×2 splitter divides the optical signal into multiple branches, each directed to a separate ONU. While essential for signal distribution, the splitter inherently introduces attenuation, closely simulating real-world operational conditions.
3. **Optical Fiber Links:** Each ONU connects through a dedicated fiber segment ranging from 0.028 km to 0.05 km. These variations reflect realistic deployment scenarios where distance-dependent losses impact signal integrity. Additionally, the total length between the OLT and the power splitter is fixed at 25 km, aligning with practical G-PON implementations.

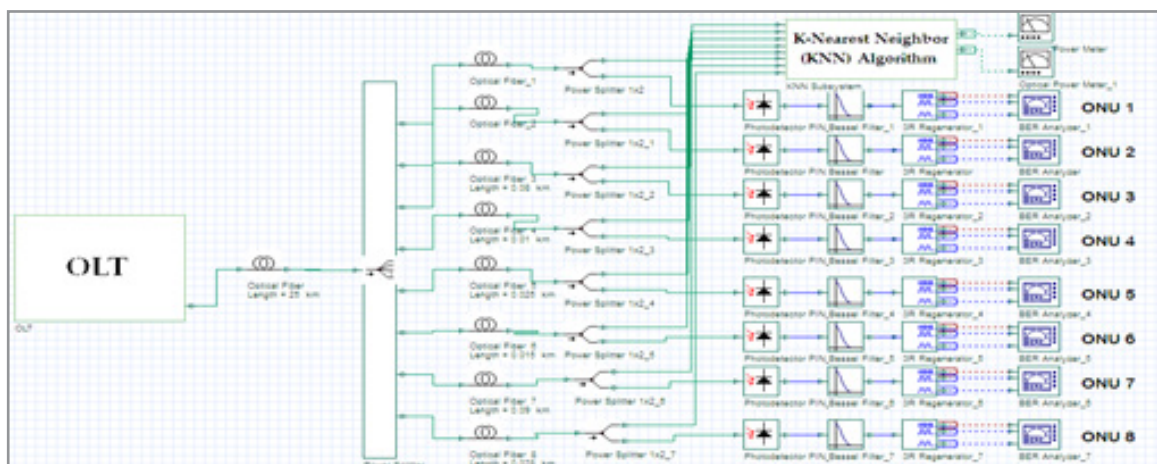


Figure 3: Illustration of the G-PON with K-NN model

The G-PON system integrated with the K-NN algorithm is examined through its classification scheme and performance outcomes. Initially, the input parameters for K-NN, which are the received power values, undergo data preprocessing before applying the machine learning algorithm. The labeled output data (targets) consist of eight classes: ONU1, ONU2, ..., ONU8, each representing different attenuation conditions. Afterward, the K-NN algorithm is applied and evaluated using confusion matrices to determine the accuracy of optical fiber distortion prediction for each user condition.

Prior to applying the K-NN algorithm, data extraction and normalization are essential steps. Data normalization addresses issues related to high variance and outlier detection [3]. In this context, two parameters are extracted from Figure 3: the received power for each individual ONU and the aggregate received power across all ONUs. These parameters are normalized using the Z-score method.

Normalization resolves these issues by producing transformed values that maintain the original data distribution while scaling

all values consistently across the dataset. The Z-score is defined as follows:

$$Z_n = \frac{x_n - \bar{Z}}{\sigma_Z} \quad (4)$$

Where $\bar{Z}$ is the mean deviation and σ_Z is the standard deviation of each feature n .

In this study, since the dataset includes predefined classes corresponding to eight known users, supervised learning techniques are applied. Among various machine learning classifiers, both the K-Nearest Neighbors (K-NN) algorithm and Support Vector Machine (SVM) are selected due to their proven effectiveness in optical communication classification tasks [5]. These classifiers were chosen for their simplicity, computational efficiency, and robustness. Furthermore, their implementation is straightforward and accessible, particularly using software platforms such as MATLAB

Results and Discussion

An effective method for simulating fiber damage involves increasing the attenuation coefficient at specific locations or segments of the optical fiber. This technique allows the modeling of both localized and distributed types of degradation. Localized defects are simulated by applying a sharp attenuation spike at a defined point along the fiber, emulating faults such as micro-bends, cracks, or splice losses. In contrast, distributed damage is modeled by gradually increasing the attenuation coefficient over a certain length of the fiber, reflecting broader physical stress, bending, or environmental aging, which collectively impact a larger region of the fiber. These impairments lead to signal degradation, reduced power levels, and a decline in the signal-to-noise ratio (SNR). In this study, we concentrate on the analysis of distributed damage, aiming to assess its impact on system performance. The simulation results, obtained through Opti-System—a comprehensive optical system design software—are presented in the first section, based on the configuration shown in Figure 3.

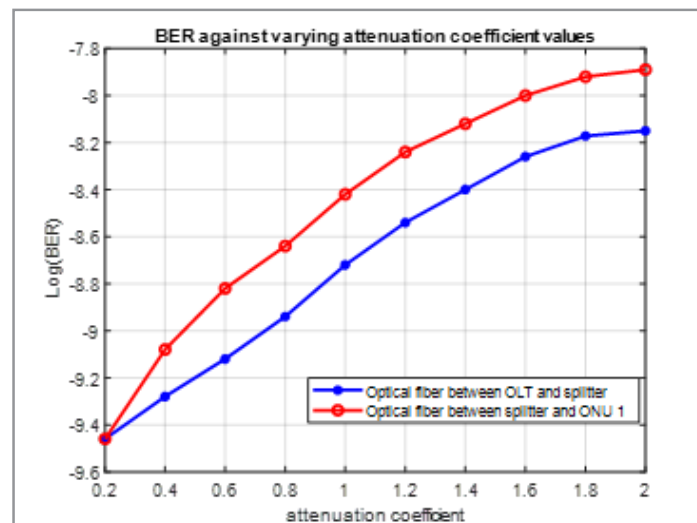


Figure 4. Effect of different optical fiber damage positions under varying attenuation coefficient values

Figure 4 illustrates the effect of fiber damage at two critical segments of the G-PON infrastructure—between the Optical Line Terminal (OLT) and the splitter, and between the splitter and the first Optical Network Unit (ONU)—on system performance when employing the K-Nearest Neighbors (K-NN) classification algorithm. The system's Bit Error Rate (BER) is used as the primary performance metric.

The results demonstrate that acceptable BER levels can be maintained when the attenuation coefficient remains below 0.5 dB/km for the fiber segment between the OLT and splitter, and be-

low 0.8 dB/km for the segment between the splitter and the first ONU.

Moreover, the graph reveals a clear trend: BER deteriorates with increasing attenuation in both segments. Notably, the fiber segment connecting the splitter to ONU 1 (represented by the red curve) exhibits a higher BER compared to the segment between the OLT and splitter (blue curve). This discrepancy is attributed to additional power losses introduced by the optical splitter as well as longer fiber lengths, which collectively contribute to greater signal degradation on the ONU side

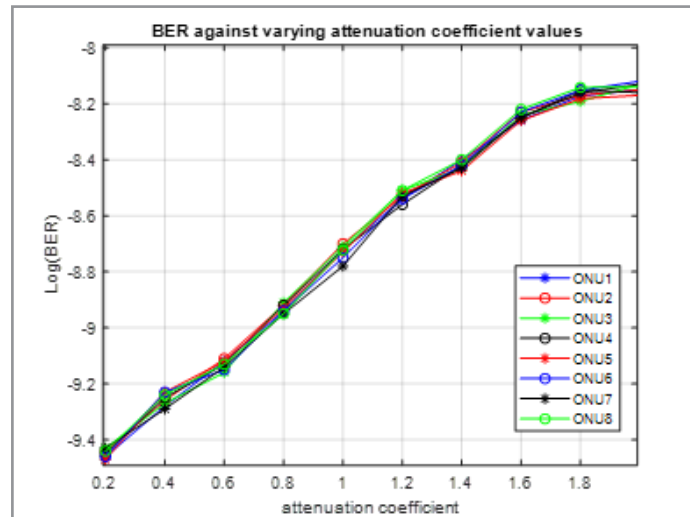


Figure 5: Effect of different optical fiber damage users under varying attenuation values

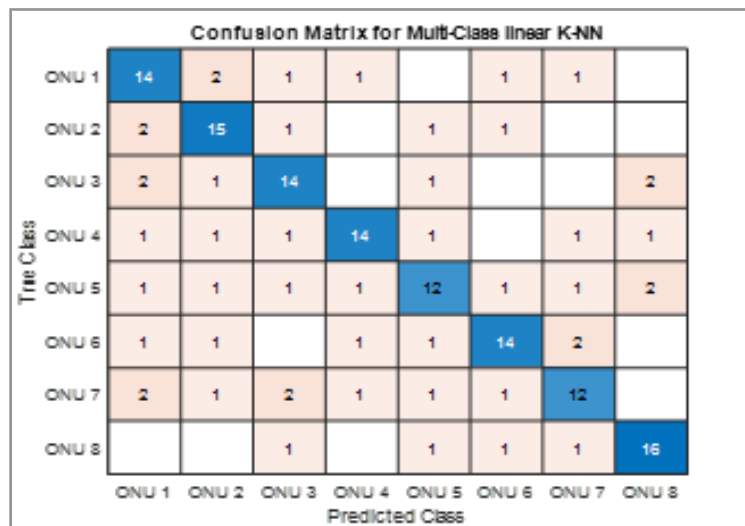


Figure 6: Confusion matrices of K-NN classifier for optical fiber damage prediction

Figure 5 presents comparable BER values for different ONUs under varying attenuation coefficients, with slight variations resulting from the differing distances between the splitter and each ONU. This observation highlights the significant effect of attenuation on the Bit Error Rate (BER) within a multi-ONU optical communication system. As the attenuation increases, the BER correspondingly degrades across all ONUs, following a nearly identical pattern. This consistency indicates that the primary cause of signal degradation is attenuation rather than individual differences among the ONUs.

The confusion matrix further confirms the effectiveness of the classification model, showing accurate identification for the majority of ONUs. Misclassification errors mainly occur between neighboring ONUs, which can be attributed to the similarity in their signal features. Notably, ONU 8 achieved perfect classification with 16 correct identifications out of 20 samples, while other ONUs exhibited minor classification inaccuracies. Overall, the model demonstrates robust performance, although there

is potential for improvement through enhanced feature selection or the adoption of more sophisticated machine learning algorithms to minimize misclassification.

$$\text{Accuracy}_{\text{ONU}_i} = \frac{\text{True Positives (TP)}}{\text{Total Samples for ONU}_i} \quad (5)$$

Where:

- True Positives (TP): Represented by the diagonal elements of the confusion matrix, indicating correct classifications.
- Total Samples for ONU: The sum of all entries in the respective row, representing the total number of classification attempts for that ONU.

The classification performance across the eight ONUs demonstrates generally good accuracy, ranging from 63.16% to 75.00%. ONU 2 achieved the highest accuracy (75.00%), followed closely by ONU 8 (72.73%), suggesting that these ONUs

exhibit more distinct signal features, making them easier for the model to identify. In contrast, ONU 1 (63.64%) and ONU 7 (63.16%) showed the lowest classification accuracy, indicating a higher rate of confusion with other ONUs. The remaining ONUs recorded intermediate accuracies between 66.67% and 70.00%, reflecting moderate classification reliability.

The confusion matrix reveals a noticeable degree of misclassification, likely due to overlapping feature distributions or signal similarities among neighboring ONUs. Despite this, the model demonstrates reasonable overall performance. However, further enhancements—such as more discriminative feature extraction or the implementation of advanced classification algorithms—could improve accuracy, particularly for ONUs with lower classification rates

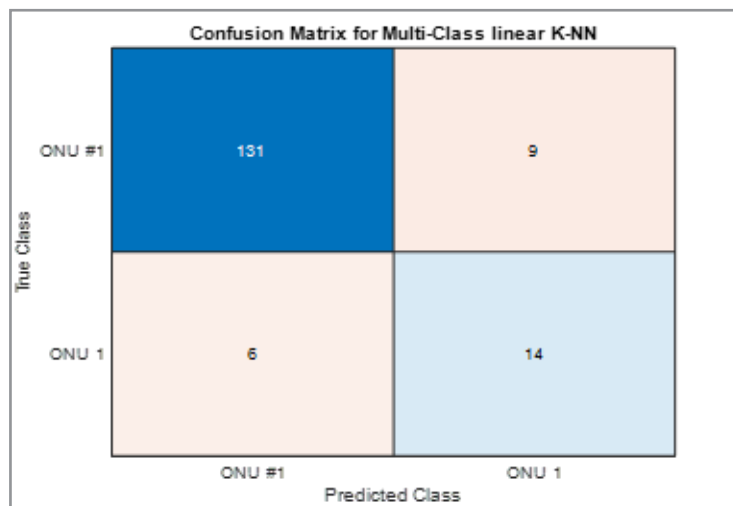


Figure 7: Confusion matrices of K-NN classifier for ONU 1

To enhance the accuracy of the system, the classification was refined to distinguish only between the desired ONU and the other undesired ONUs. The confusion matrix demonstrates the classification performance between ONU #1 and ONU 1, achieving an overall accuracy of 90.63%. The model correctly identified 131 instances of ONU #1 and 14 instances of ONU 1, while 9 ONU #1 samples were misclassified as ONU 1, and 6 ONU 1 samples were misclassified as ONU #1. Although the classifier performs well, some misclassifications remain, particularly affecting ONU 1. Improving feature extraction or optimizing the model could further enhance classification accuracy.

Conclusion

This study highlights the effectiveness of using machine learning-based fault detection in Gigabit Passive Optical Networks (GPON) through the application of the K-Nearest Neighbors (K-NN) algorithm. By conducting simulations in OptiSystem and performing data classification in MATLAB, the system efficiently detects and classifies signal degradations across multiple Optical Network Units (ONUs). Analysis of the confusion matrices indicates classification accuracy ranging from 63.16% to 75.00%, with ONU 2 and ONU 8 achieving the highest accuracy levels, while ONU 1 and ONU 7 are more susceptible to misclassification. Furthermore, the study underscores the impact of attenuation on the Bit Error Rate (BER), demonstrating that increased attenuation causes significant signal degradation, especially in the fiber segment between the splitter and the ONUs.

Overall, integrating the K-NN algorithm for fault detection in GPON significantly improves network reliability by automating fault diagnosis and minimizing downtime. Nevertheless, there

remains room for enhancement in classification performance through more sophisticated feature extraction methods, hyperparameter optimization, and the incorporation of advanced deep learning techniques. Additionally, future research could investigate the benefits of real-time network monitoring and adaptive machine learning models to dynamically optimize GPON performance under varying operational conditions. These advancements have the potential to further strengthen fault resilience and service quality in optical access network

References

1. Kalera, R., Kaler, R. S. (2011). Simulation of fiber to the home at 10 Gbit/s 10-GPON architecture. *Optik - International Journal for Light and Electron Optics*, 122(17), 1362-1366. <https://doi.org/10.1016/j.ijleo.2010.06.019>
2. Wang, J., Wang, G., Zhang, L., Li, H. (2017). Ground simulation method for arbitrary distance optical transmission of a free-space laser communication system based on an optical fiber nanoprobe. *IEEE/OSA Journal of Optical Communications and Networking*, 9(12), 1131-1137. <https://doi.org/10.1364/JOCN.9.001131>
3. Shumate, W. P. (2008). Fiber-to-the-home. *Journal of Lightwave Technology*, 26(9), 1093-1103. <https://doi.org/10.1109/JLT.2008.920769>
4. Sun, J. W., Du, W., Shi, N. (2018). A survey of kNN algorithm. In *Information engineering and applied computing* 416-420. Springer. https://doi.org/10.1007/978-3-319-93846-2_57
5. Cunningham, P., Delany, S. J. (2007). k-nearest neighbour classifiers. *ACM Computing Surveys*, 39(1), 3. <https://doi.org/10.1145/1216370.1216375>

6. Suyal, M., Goyal, P. (2022). A review on analysis of K-Nearest Neighbor classification machine learning algorithms based on supervised learning. *International Journal of Engineering Trends and Technology*, 70(7), 43-48. <https://doi.org/10.14445/22315381/IJETT-V70I7P206>
7. Wang, H. (2002). Nearest neighbours without k: A classification formalism based on probability (Technical report). University of Ulster, Faculty of Informatics. https://link.springer.com/chapter/10.1007/3-540-32370-8_12
8. Karthikeya, K., Sudarshan, K. H. (2019). Prediction of agriculture crops using KNN algorithm. *International Journal of Innovative Science and Research Technology*, 5(5), 1422-1424.
9. Krishnamoorthy, N., Umarani, N. (2021). Diabetes prediction in healthcare using KNN algorithm. *International Journal of Multidisciplinary Educational Research*, 10(5), 36-39.
10. Chen, Y., Zhang, Y., Liu, H. (2021). Machine learning-based fault detection in GPON fiber access networks. *IEEE Access*, 9, 118034-118045. <https://doi.org/10.1109/ACCESS.2021.3108530>