

الجمهورية الجزائرية الديمقراطية الشعبية
République algérienne démocratique et populaire
وزارة التعليم العالي و البحث العلمي
Ministère de l'Enseignement supérieur et de la Recherche scientifique
جامعة عين تموشنت بلحاج بوشعيب
Université Belhadj Bouchaib Ain Temouchent
Faculté des Sciences et de Technologie
Département D'électronique et des Télécommunications



Projet de Fin d'Études Pour l'obtention du diplôme de Master en
: **Domaine** : SCIENCES ET TECHNOLOGIE
Filière : TELECOMMUNICATIONS
Spécialité : Réseaux et Télécommunications
Thème

Détection d'anomalies dans les signaux radio par
apprentissage non supervisé

Présenté par :

❖ BELKADI Nihad Abir

Devant le jury composé de :

Dr. BOUKHOBZA Abdelkader	MCA UATB.B (Ain Temouchent)	Président
Dr . Yagoub Reda	MCB UATB.B (Ain Temouchent)	Examinateur
Dr.SOUIKI Sihem	MCA UATB.B (Ain Temouchent)	Encadrante

Année universitaire : 2025/2026

Remerciements

Au terme de ce Projet de Fin d'Études, je tiens à exprimer ma profonde gratitude à toutes les personnes qui ont contribué, de près ou de loin, à la réalisation de ce travail.

Je remercie tout particulièrement mon encadrante, Dr SOUIKI.S, pour son accompagnement, sa disponibilité, ses conseils précieux et la qualité de son encadrement tout au long de ce projet. Ses orientations pertinentes, ses remarques constructives et son soutien constant ont été d'une grande aide pour mener à bien ce travail.

J'adresse également mes sincères remerciements aux membres du jury, Dr YAGOUB.R et Dr BOUKHBEZA, pour l'honneur qu'ils me font en acceptant d'évaluer ce travail, ainsi que pour le temps consacré à son examen et les remarques constructives qu'ils apporteront.

Je remercie également l'ensemble des enseignants de ma formation pour la qualité de leurs enseignements et leurs efforts durant tout mon parcours universitaire.

Je remercie profondément mes parents pour leur amour, leurs sacrifices, leur soutien inconditionnel et leurs encouragements constants, qui ont été ma principale source de motivation et de réussite.

Enfin, j'adresse mes remerciements à ma famille, mes amis, et à toutes les personnes qui m'ont soutenue et accompagnée durant ce parcours.

Merci à tous.

Dédicace

Je dédie ce modeste travail à toutes les personnes qui ont contribué, de près ou de loin, à ma réussite au cours de mon parcours universitaire.

À mes très chers parents, **BELKADI Houari** et **AIT ZAUCHE Hafida**, qui m'ont entourée d'amour et de tendresse, et qui n'ont jamais cessé de me soutenir et de se sacrifier pour moi. Je les remercie pour leur patience, leurs encouragements constants et leurs prières sincères, qui ont toujours été ma principale source de force et de détermination. Les mots ne suffisent pas pour exprimer toute ma gratitude et ma reconnaissance envers eux.

À mon cher frère **Youcef Zine Eddine**, mon pilier dans la vie, qui m'a toujours soutenue et encouragée, et qui a été présent dans tous les moments importants de mon parcours.

À mes chères sœurs **Hizia Halima** et **Ismahane Nour El Houda**, qui ont été une source d'amour, de soutien et de motivation tout au long de mes études, et qui m'ont accompagnée avec affection et bienveillance.

Mes chers petits **Sanaa Raghad** et **Anes Zaid**, Je vous souhaite beaucoup de bonheur, de réussite et de beaux moments dans votre vie.

À mes nièces et neveux adorées **Rihem, Nada, Lilia, Karima, Djouri Fares, Mouaad, akram, Ahmed et Ibrahim**, à qui je souhaite un avenir plein de succès et de bonheur.

À mes très chères, que je considère comme mes sœurs **Loukili Ferdaous, Mimoune Rifka, Mimoune Afraa, Belkadi Chahrazed, Belkadi Rania et Ramdane Imane**, avec qui j'ai partagé de précieux moments de ma vie, et qui m'ont apporté un grand soutien moral et affectif.

Enfin, à toutes les personnes qui m'ont soutenue, encouragée et accompagnée tout au long de mon parcours universitaire, je leur adresse mes sincères remerciements et ma profonde gratitude.

Je dédie ce travail avec tout mon amour, ma loyauté et ma reconnaissance à tous ceux qui ont contribué à ma réussite.

Résumé

Ce mémoire porte sur la détection d'anomalies dans les signaux radio à l'aide des techniques de Machine Learning. La notion d'anomalie est définie à partir du rapport signal sur bruit (SNR), indicateur essentiel de la qualité des signaux radio.

L'étude expérimentale utilise la base de données RadioML2016.10a afin de comparer trois méthodes de détection d'anomalies : l'Autoencoder, le One-Class Support Vector Machine (OCSVM) et l'Isolation Forest, en exploitant les composantes In-phase et Quadrature (I/Q) des signaux.

Les résultats montrent que les trois méthodes permettent de détecter efficacement les anomalies. Toutefois, l'Autoencoder a obtenu les meilleures performances grâce à sa capacité à apprendre des représentations profondes et complexes des données. Ces résultats confirment l'intérêt des techniques de Machine Learning pour améliorer la fiabilité et la sécurité des systèmes de communication sans fil.

Mots-clés: Détection D'anomalies, signaux radio, Machine Learning, Auto-encoder, OCSVM, Isolation Forest, RadioML2016.10a, SNR

Abstract

This master's thesis focuses on anomaly detection in radio signals using Machine Learning techniques. The notion of anomaly is defined based on the Signal-to-Noise Ratio (SNR), which is an essential indicator of radio signal quality.

The experimental study uses the RadioML2016.10a dataset to compare three anomaly detection methods: Autoencoder, One-Class Support Vector Machine (OCSVM), and Isolation Forest, by exploiting the In-phase and Quadrature (I/Q) components of radio signals.

The results show that the three methods can effectively detect anomalies. However, the **Autoencoder** achieved the best performance due to its ability to learn deep and complex representations of the data. These findings confirm the potential of Machine Learning techniques to improve the reliability and security of wireless communication systems.

Keywords: Anomaly Detection, radio signals, Machine Learning, Auto-encoder, OCSVM, Isolation Forest, RadioML2016.10a, SNR

الملخص

تتناول هذه المذكرة موضوع كشف الشذوذ في الإشارات الراديوية باستخدام تقنيات التعلم الآلي. وقد تم تعريف مفهوم الشذوذ اعتمادًا على نسبة الإشارة إلى الضوضاء (SNR)، والتي تُعد مؤشرًا أساسيًا لجودة الإشارات الراديوية.

تعتمد الدراسة التجريبية على مجموعة البيانات RadioML2016.10a من أجل مقارنة ثلاث طرق لكشف الشذوذ، وهي: المشفر التلقائي (Autoencoder)، وآلة المتجهات الداعمة أحادية الصنف (OCSVM)، وغابة العزل (Isolation Forest)، بالاعتماد على مركبتي الطور المتوافق (I) والطور المتعامد (Q) للإشارات الراديوية.

أظهرت النتائج أن الطرق الثلاث قادرة على كشف الشذوذ بفعالية. ومع ذلك، حقق المشفر التلقائي (Autoencoder) أفضل أداء بفضل قدرته على تعلم تمثيلات عميقة ومعقدة للبيانات. وتؤكد هذه النتائج أهمية تقنيات التعلم الآلي في تحسين موثوقية وأمن أنظمة الاتصالات اللاسلكية.

الكلمات المفتاحية: كشف الشذوذ، الإشارات الراديوية، التعلم الآلي، المشفر التلقائي، OCSVM، غابة العزل، RadioML2016.10a، نسبة الإشارة إلى الضوضاء (SNR).

Table des matières

Remerciements	II
Dédicace	III
Résumé	IV
Table des matières	VII
Table des figures	XII
Liste des abréviations	XV
Introduction générale.....	2
Chapitre I : Généralités sur la modulation numérique	
Introduction	5
1La modulation.....	5
1.1Modulation analogique	6
1.2La modulation numérique.....	7
2Principe d'une chaîne de transmission numérique	8
3Principe de fonctionnement de la modulation numérique	10
4Les techniques de modulation numérique	11
4.1Phase-shift keying (PSK)	11
4.1.1Constellation.....	12
4.2Frequency-shift keying (FSK)	13
4.3Amplitude-shift keying (ASK) or ON/OFF Keying (OOK)	15
4.4La modulation d'amplitude en quadrature (QAM).....	17
5Performances des modulations numériques	19
5.1Le Bit Error Rate (BER).....	19
5.2SNR (Signal-to-Noise Ratio).....	19
5.3Sensibilité au bruit	20
5.4Efficacité spectrale	22
6Applications de la modulation numérique.....	24
7Conclusion.....	25

Chapitre II : Machine Learning (ML)

Introduction	28
1. L'intelligence artificielle (IA)	28
1.1 Machine learning (ML)	28
1.2 Le Deep Learning (DL).....	29
2. Modèle de Machine Learning	30
3. Pipeline d'un système de Machine Learning	31
3.1 Collecte des données	31
3.2 Nettoyage et prétraitement des données.....	32
3.3 Extraction et sélection des caractéristiques (Feature Engineering).....	32
3.4 Entraînement du modèle.....	32
3.5 Validation et test du modèle.....	32
4. Types d'apprentissage	33
4.1 Apprentissage supervisé (Supervised Learning)	34
4.1.1 La régression:	35
4.1.2 La classification :	35
4.2 Apprentissage non supervisé (Unsupervised Learning).....	36
4.2.1 Clustering (regroupement) :	37
4.2.2 Association (Association Rule Learning) :	38
4.2.3 Détection d'anomalies :	39
4.2.4 Auto-encoders (Auto-encodeurs) :	40
4.2.5 Apprentissage semi-supervisé (Semi-Supervised Learning).....	41
4.3 Apprentissage par renforcement (Reinforcement Learning) :	42
5. Algorithmes de Machine Learning :	44
5.1 Algorithmes de classification :	44
5.1.1 Régression logistique :	45
5.2 Machines à vecteurs de support (SVM) :	45
5.2.1 k plus proches voisins (K-NN).....	45
5.2.2 Les réseaux bayésiens naïfs (Naive Bayes) :	46
5.2.3 Arbres de décision (Decision Tree).....	46
5.2.4 Neural Networks :	47

5.2.5Le Random Forest (RF).....	47
5.3Algorithmes de régression	48
5.3.1Régression linéaire simple et multiple :	49
5.3.2La régression polynomiale :	50
5.3.4La LASSO et la Ridge regression.....	50
5.4Algorithmes de clustering.....	51
5.4.1DBSCAN clustering	51
6.Applications du Machine Learning	52
7.Conclusion.....	53

Chapitre III: étude des méthodes de détection d'anomalie dans les signaux radio

Introduction.....	56
1Définition et types d'anomalies.....	56
1.1Anomalies ponctuelles.....	57
1.2Anomalies contextuelles.....	58
1.3Anomalies collectives.....	58
2La détection d'anomalies dans les signaux radio	59
3Méthodes de détection d'anomalies radio	59
3.1Méthodes de détection traditionnelles	60
3.1.1La détection par énergie	60
3.1.2La détection par filtre adapté	61
3.1.3Détection cyclostationnarité	61
3.2Méthodes de détection par les algorithmes évolutionnaires	61
3.2.1Les algorithmes de colonie de fourmis.....	62
3.2.2Les réseaux immunitaires artificiels	62
3.3Méthode de détection par apprentissage automatique	62
3.3.1K-means.....	63
3.3.2LOF (Local Outlier Factor)	64
3.3.3One-Class SVM.....	66
3.3.4Exemple d'étude 1	66
3.3.5Isolation Forest	67
3.4Apprentissage profond.....	68
3.4.1Autoencodeur	69

Conclusion	70
------------------	----

Chapitre IV : Expérimentaux et discussion des modèles de détection d'anomalie

Introduction	73
1Environnement expérimental	73
2Le RadioML2016.10a	73
3Définition des anomalies	75
4Méthodologie proposée	76
4.1Prétraitement des données	77
4.2Chargement des données	78
4.3Étiquetage des anomalies	78
4.4Mise en forme des données	78
4.5Normalisation	79
4.6Réduction de dimension	79
4.7Modèles de détection d'anomalies	79
4.8Calcul du score d'anomalie	79
4.9Décision de classification	80
4.10Évaluation des performances des modèles de Machine Learning.....	80
4.10.1Matrice de confusion	80
4.10.2La sensibilité (Sensitivity)	81
4.10.3La spécificité (Specificity)	81
4.10.4Le taux de faux positifs (False Positive Rate – FPR).....	81
4.10.5Le taux de faux négatifs (False Negative Rate – FNR).....	82
4.10.6La précision (Precision)	82
4.10.7La valeur prédictive négative (Negative Predictive Value – NPV)	82
4.10.8L'exactitude (Accuracy).....	82
4.10.9Le F1-score.....	82
4.10.10Courbe ROC et AUC.....	82
5Résultats expérimentaux et discussion.....	83
5.1Autoencodeur	83
5.1.1Matrice de confusion	85
5.1.2Courbe ROC	86
5.2Courbe d'apprentissage (Loss).....	87

5.3Distribution des erreurs de reconstruction.....	88
5.4Isolation forset	89
5.4.1Matrice de confusion	91
5.4.2Courbe ROC	92
5.5Distribution des erreurs de reconstruction.....	93
5.6OCSVM(One-Class SVM) :.....	94
5.6.1Matrice de confusion	96
5.6.2Courbe ROC	97
5.6.3Distribution des erreurs de reconstruction.....	98
5.7Comparaison globale des modèles et discussion.....	99
5.7.1Tableau comparatif final des modèles :.....	100
Conclusion.....	100
Conclusion générale.....	103
Travaux futurs	105
Référence.....	107

Table des figures

Chapitre I : Généralités sur la modulation numérique

Figure I.1: Types fondamentaux de modulation	6
Figure I.2: Représentation des signaux modulateurs et des modulations analogiques AM, FM et PM	7
Figure I.3: Principe d'une chaîne de transmission numérique	9
Figure I.4: Représentation temporelle et constellation des modulations BPSK et QPSK	13
Figure I.5: Implémentation de la modulation FSK (Binary Frequency Shift Keying)	15
Figure I.6: Illustration de la modulation d'amplitude (ASK/OOK) appliquée à un signal binaire NRZ	17
Figure I.7: Représentation de la modulation 16-QAM (constellation et signal)	18
Figure I.8: Évolution (BER) en fonction du (SNR) pour les modulations (BPSK, QPSK, 16-QAM et 64-QAM) dans un canal AWGN	20
Figure I.9: Comparaison de la sensibilité au bruit des modulations numériques (BPSK, QPSK et M-QAM) dans les canaux AWGN et Rayleigh en fonction du SNR	22
Figure I.10: Variation de l'efficacité spectrale des modulations numériques	23
Figure I.11: Influence de l'ordre de modulation (M) sur la bande passante pour M-PSK/M-QAM et FSK	24

Chapitre II : Machine Learning (ML)

Figure II.1: Interactions IA ML DL	30
Figure II.2: Entraînement et test d'un modèle de Machine Learning	31
Figure II.3: Pipeline d'un système de Machine Learning	33
Figure II.4: Types d'apprentissage automatique	34
Figure II.5: L'apprentissage supervisé (Supervised Learning)	35
Figure II.6: Classification et régression en apprentissage automatique	36
Figure II.7: Machine Learning-processus d'apprentissage non supervisé	37
Figure II.8: Méthodes d'apprentissage non supervisé en fouille de données	40
Figure II.9: Schéma du processus d'un autoencodeur	41
Figure II.10: L'apprentissage semi-supervisé (Semi-Supervised Learning)	42
Figure II.11: Apprentissage par renforcement (Reinforcement Learning)	43
Figure II.12: Des principaux algorithmes de classification en apprentissage automatique	44
Figure II.13: Les principaux algorithmes de régression de classification en apprentissage automatique	48
Figure II.14: Visualisation des données et des clusters obtenus par l'algorithme DBSCAN	51

Chapitre III : Étude des méthodes de détection d'anomalie dans les signaux radio

Figure III.1: Taxonomie des anomalies	57
Figure III.2: Taxonomie des méthodes de détection d'anomalies radio	60

Chapitre IV : Expérimentaux et discussion des modèles de détection d'anomalie

Figure IV.1: Présentation du dataset RadioML 2016.10a et des signaux IQ pour les différentes modulations	73
Figure IV.2: Schéma général de la méthodologie proposée pour la détection des anomalies dans les signaux radio	75
Figure IV.3: Matrice de confusion utilisée pour l'évaluation des résultats de classification	78
Figure IV.4: Matrice de confusion du modèle Autoencodeur pour la détection d'anomalies dans les signaux radio	82
Figure IV.5: Courbe ROC du modèle Autoencodeur montrant la capacité de discrimination entre signaux normaux et anormaux	83
Figure IV.6: Courbe d'apprentissage du modèle Autoencodeur illustrant la convergence de la fonction de perte MAE	84
Figure IV.7: Distribution des erreurs de reconstruction permettant de distinguer les signaux normaux des anomalies	85
Figure IV.8: Matrice de confusion du modèle Isolation Forest pour la détection d'anomalies des signaux IQ	88
Figure IV.9: Courbe ROC et valeur de l'AUC obtenues pour le modèle d'Isolation Forest	89
Figure IV.10: Distribution des scores d'anomalie de l'Isolation Forest et positionnement du seuil optimal de décision	90
Figure IV.11: Matrice de confusion et performance de classification du modèle OCSVM	92
Figure IV.12: Courbe ROC et performance globale de l'OCSVM pour la détection d'anomalies	93
Figure IV.13: Distribution des scores d'anomalie de l'OCSVM et positionnement du seuil de décision	94

Liste des abréviations

- **RF** : Radio Frequency (Radiofréquence)
- **IQ** : In-phase / Quadrature (Phase et Quadrature)
- **SNR** : Signal-to-Noise Ratio (Rapport Signal sur Bruit)
- **ML** : Machine Learning (Apprentissage automatique)
- **DL** : Deep Learning (Apprentissage profond)
- **AI / IA** : Artificial Intelligence / Intelligence Artificielle
- **AE** : Autoencoder (Autoencodeur)
- **OCSVM** : One-Class Support Vector Machine
- **SVM** : Support Vector Machine
- **IF / IForest** : Isolation Forest
- **LOF** : Local Outlier Factor
- **K-means** : Algorithme de clustering K-moyennes
- **PCA** : Principal Component Analysis (Analyse en Composantes Principales)
- **FFT** : Fast Fourier Transform (Transformée de Fourier Rapide)
- **RMS** : Root Mean Square (Valeur efficace)
- **AWGN** : Additive White Gaussian Noise
- **AM** : Amplitude Modulation
- **FM** : Frequency Modulation
- **BPSK** : Binary Phase Shift Keying
- **QPSK** : Quadrature Phase Shift Keying
- **QAM** : Quadrature Amplitude Modulation
- **RadioML** : Radio Machine Learning Dataset
- **ROC** : Receiver Operating Characteristic
- **AUC** : Area Under Curve
- **TP** : True Positive
- **TN** : True Negative
- **FP** : False Positive
- **FN** : False Negative

- **FPR** : False Positive Rate
- **FNR** : False Negative Rate
- **NPV** : Negative Predictive Value

Introduction générale

Introduction générale

Introduction générale

Les systèmes de communication sans fil occupent aujourd'hui une place essentielle dans les technologies modernes, notamment avec le développement rapide des réseaux mobiles, de l'Internet des objets (IoT) et des applications de transmission de données en temps réel. Cependant, ces systèmes évoluent dans des environnements de plus en plus complexes et dynamiques, où la qualité du signal peut être fortement dégradée par divers facteurs tels que le bruit, les interférences, le fading ou encore les imperfections matérielles.

Dans ce contexte, la détection d'anomalies dans les signaux radiofréquences (RF) devient un enjeu majeur pour garantir la fiabilité et la performance des systèmes de communication. Elle permet d'identifier automatiquement des comportements inhabituels ou dégradés pouvant affecter la transmission de l'information, et contribue ainsi à l'amélioration de la sécurité et de l'efficacité des réseaux sans fil.

Traditionnellement, ces tâches reposaient sur des méthodes statistiques et des techniques classiques de traitement du signal. Toutefois, ces approches montrent leurs limites face à la complexité croissante des environnements radio modernes, caractérisés par des signaux non linéaires, bruités et fortement variables. Pour répondre à ces défis, les approches basées sur l'apprentissage automatique et l'apprentissage profond se sont imposées comme des solutions particulièrement efficaces, grâce à leur capacité à apprendre automatiquement des représentations pertinentes à partir des données.

Dans ce cadre, ce travail propose une étude comparative de plusieurs méthodes de détection d'anomalies appliquées aux signaux radio issus du dataset **RadioML2016.10a**. Trois approches principales sont étudiées : l'Autoencodeur, l'Isolation Forest et le One-Class SVM (OCSVM). Ces modèles permettent de détecter les signaux anormaux définis selon le rapport signal sur bruit (SNR), en distinguant les signaux de bonne qualité des signaux fortement dégradés.

L'objectif principal de ce projet est d'évaluer et de comparer les performances de ces différentes méthodes afin d'identifier l'approche la plus adaptée à la détection d'anomalies dans

Introduction générale

les environnements radio complexes. Cette étude s'appuie sur un pipeline complet incluant le prétraitement des données, l'extraction de caractéristiques, la réduction de dimension et l'évaluation des modèles à l'aide de métriques standards telles que la matrice de confusion, la courbe ROC et le score F1.

Ainsi, ce travail s'inscrit dans une perspective d'amélioration des techniques de surveillance et d'analyse des signaux radio, en exploitant les avancées récentes de l'intelligence artificielle pour renforcer la robustesse et la fiabilité des systèmes de communication sans fil.

Chapitre I : Généralités sur la modulation numérique

Chapitre I : Généralités sur la modulation numérique

Introduction

Ce chapitre a pour objectif de présenter les principes de base de la modulation ainsi que ses différentes formes. Nous aborderons tout d'abord les techniques de modulation analogique, avant de nous intéresser à la modulation numérique, qui constitue aujourd'hui la base des systèmes de communication modernes. Ensuite, les principales techniques de modulation numérique seront détaillées, notamment leurs principes de fonctionnement, leurs avantages et leurs limites. Enfin, nous analyserons les performances de ces techniques à travers des critères essentiels tels que le taux d'erreur binaire, le rapport signal sur bruit et l'efficacité spectrale.

1 La modulation

La modulation est une étape essentielle dans les systèmes de télécommunications. Elle permet de transmettre un signal d'information à travers un canal, que ce soit un support filaire, optique ou sans fil. En pratique, le signal d'information ne peut pas être envoyé directement sur de longues distances, ce qui nécessite l'utilisation d'une onde porteuse. Le principe consiste à modifier certaines caractéristiques de cette porteuse, comme l'amplitude, la fréquence ou la phase, afin de représenter l'information à transmettre. Cette opération permet notamment d'adapter le signal aux conditions du canal et de faciliter sa propagation. On remarque ainsi que la modulation ne sert pas uniquement à transmettre le signal, mais aussi à améliorer la qualité de transmission et à optimiser l'utilisation du spectre [1]. À cet effet, si la variation du paramètre de la porteuse est continue et suit le signal analogique d'entrée, la technique de modulation est appelée modulation analogique. En revanche, si la variation est discrète, elle est alors appelée modulation numérique [1].

Chapitre I : Généralités sur la modulation numérique

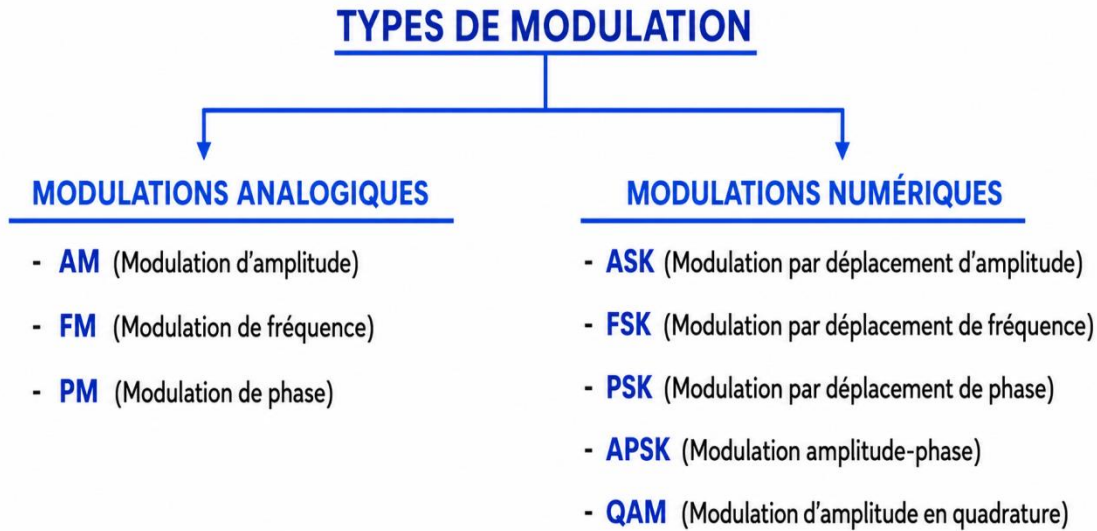


Figure I.1: Types fondamentaux de modulation

1.1 Modulation analogique

Les techniques de modulation analogique comprennent la modulation d'amplitude (AM), la modulation de fréquence (FM) et la modulation de phase (PM). Autrement dit, ces techniques reposent sur la modification d'une onde porteuse continue.

En modulation d'amplitude (AM), l'information est envoyée en changeant l'amplitude de la porteuse. Elle est utilisée en radiodiffusion et peut atteindre de grandes distances grâce à la réflexion ionosphérique, mais elle est sensible aux interférences.

En modulation de fréquence (FM), la fréquence de la porteuse change selon le signal, alors que l'amplitude reste la même. Elle offre une meilleure protection contre le bruit et est utilisée pour la radio FM, la stéréo et certaines transmissions télévisées.

En modulation de phase (PM), la phase de la porteuse est changée selon le signal. Elle est semblable à la FM et est souvent utilisée avec elle dans les systèmes de communication modernes, comme les micro-ondes, le Wi-Fi et le satellite [6].

Chapitre I : Généralités sur la modulation numérique

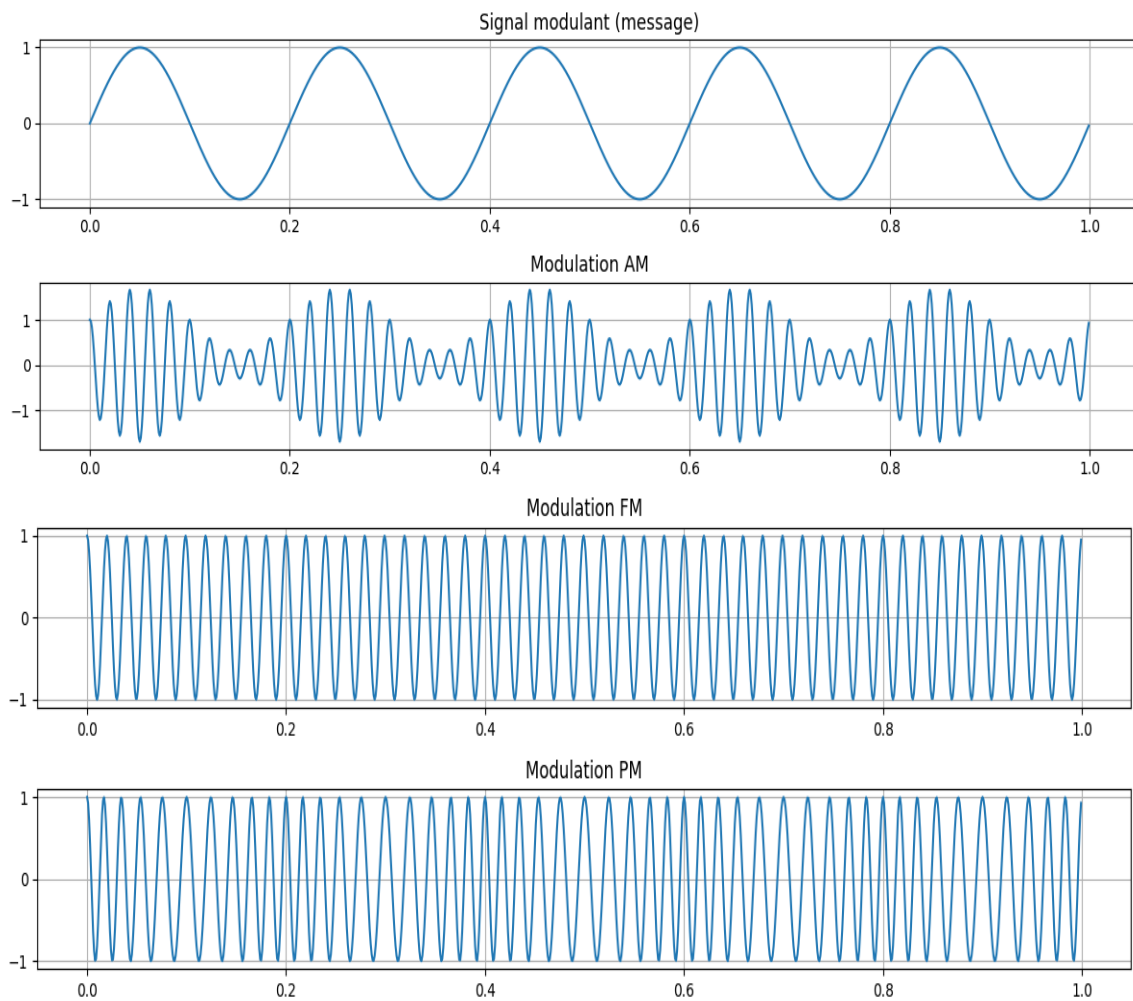


Figure I.2: Représentation des signaux modulateurs et des modulations analogiques AM, FM et PM

1.2 La modulation numérique

La modulation numérique est une technique de communication qui permet de transmettre des données numériques sous forme de signaux discrets, en modifiant les paramètres d'une porteuse sinusoïdale tels que l'amplitude, la fréquence ou la phase. Cette approche assure une transmission efficace sur de longues distances tout en optimisant l'efficacité spectrale. Elle permet également le transport de l'information binaire (0 et 1) à travers un canal de communication. Par rapport aux techniques analogiques, elle offre une meilleure immunité au

Chapitre I : Généralités sur la modulation numérique

bruit et aux interférences, tout en simplifiant certains aspects liés aux équipements de transmission.

Toutefois, elle peut nécessiter une bande passante plus importante selon le type de modulation utilisé. Globalement, cette méthode contribue à une utilisation plus efficace du spectre et joue un rôle essentiel dans les systèmes de communication actuels

L'utilisation des modulations numériques repose principalement sur les principes suivants :

Lorsque la transmission des signaux numériques en bande de base n'est pas possible (par exemple dans l'air ou sur les lignes téléphoniques), il est nécessaire de déplacer le spectre du signal vers les hautes fréquences à l'aide d'une modulation. La transmission numérique consiste à représenter l'information sous forme de bits et à l'envoyer à l'aide d'un signal porteur modulé.

La qualité d'une liaison est mesurée à l'aide du rapport signal/bruit (S/B), et les performances du système sont évaluées par la probabilité d'erreur, qui dépend directement de ce rapport [1]

2 Principe d'une chaîne de transmission numérique

Une chaîne de transmission numérique désigne l'ensemble des étapes et des blocs fonctionnels permettant de transporter une information d'une source vers un récepteur sous forme de signaux numériques. Elle assure la conversion des données, leur adaptation au canal de transmission, puis leur récupération à l'arrivée malgré la présence de bruit, d'interférences et de distorsions. Ce principe repose sur une succession d'opérations bien définies garantissant la fiabilité et la robustesse de la communication.

Chapitre I : Généralités sur la modulation numérique

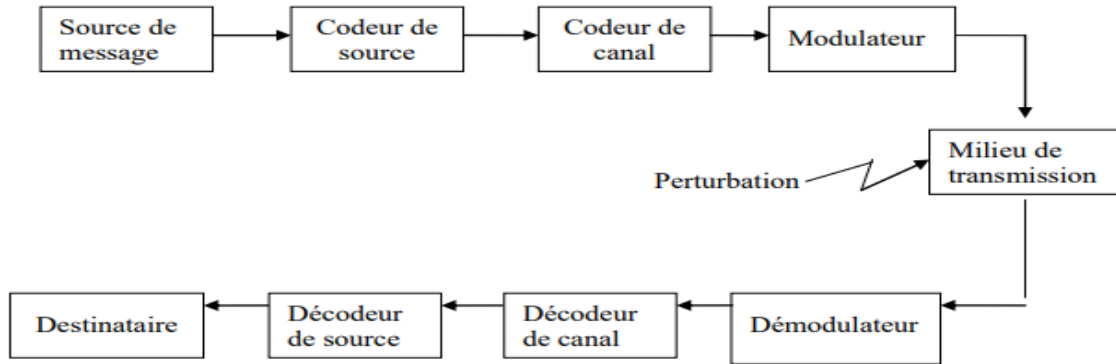


Figure I.3: Principe d'une chaîne de transmission numérique

La figure 3 illustre le schéma de principe d'une chaîne de transmission numérique. Elle permet d'identifier l'origine du message, le milieu de transmission et le destinataire, qui constituent les principaux éléments du système de communication. On distingue également les opérations d'encodage et de décodage source et canal. L'émetteur et le récepteur représentent les fonctions assurées par le système afin de garantir la transmission de l'information.

La modulation numérique présente de nombreux avantages par rapport à la modulation analogique, ce qui explique son utilisation dominante dans les systèmes de communication actuels. Elle permet notamment d'augmenter la capacité de transmission de l'information, d'améliorer la sécurité des données, d'offrir une meilleure qualité de communication et d'assurer une plus grande fiabilité des systèmes.

Parmi ses principaux atouts :

- Une capacité de transmission élevée, permettant l'envoi de grandes quantités de données.
- Une meilleure efficacité spectrale, optimisant l'utilisation de la bande passante disponible.
- Un multiplexage efficace de différents types de signaux, tels que la voix, la vidéo et les données numériques.
- Une forte résistance aux interférences et au bruit, réduisant les erreurs de transmission.
- Une bonne intégrité du signal, limitant les perturbations et les pertes d'information.

Chapitre I : Généralités sur la modulation numérique

- Une mise en œuvre économique et une intégration facile dans les systèmes numériques modernes.
- Un niveau de sécurité élevé grâce aux techniques de cryptage des données [5]

3 Principe de fonctionnement de la modulation numérique

Le fonctionnement de la modulation numérique repose sur un ensemble d'étapes et de principes fondamentaux permettant la transmission efficace de l'information numérique.

Tout d'abord, l'information issue de la source est convertie en une suite de bits (0 et 1) grâce au codage de source. Ensuite, afin de réduire les effets du bruit et des interférences du canal, une redondance contrôlée est ajoutée par le codeur de canal, ce qui améliore la fiabilité de la transmission.

Le signal binaire obtenu est ensuite envoyé vers le modulateur numérique, dont le rôle est d'adapter l'information au canal de transmission. Cette étape consiste à associer les bits à des symboles, puis à modifier les caractéristiques d'une onde porteuse sinusoïdale, telles que l'amplitude, la fréquence ou la phase. On obtient ainsi différents types de modulation comme ASK, FSK ou PSK, ainsi que des techniques combinées comme QAM.

Le signal modulé est ensuite transmis sur le canal de communication, qui peut être filaire (cuivre, fibre optique) ou sans fil. Pendant la transmission, le signal est affecté par le bruit et les interférences, ce qui peut provoquer des erreurs.

À la réception, le démodulateur récupère le signal reçu et estime les symboles transmis. Enfin, le décodeur de canal et le décodeur de source permettent de reconstruire l'information binaire originale.

Ainsi, la modulation numérique vise principalement à assurer une transmission fiable et efficace des données tout en optimisant l'utilisation de la bande passante et en améliorant la résistance au bruit [2]

Chapitre I : Généralités sur la modulation numérique

4 Les techniques de modulation numérique

La modulation numérique constitue un ensemble de techniques essentielles en télécommunications permettant de transmettre l'information binaire à travers une onde porteuse. Selon le paramètre du signal porteur qui est modifié, différentes méthodes de modulation peuvent être utilisées. Les principales techniques de modulation numérique sont présentées ci-dessous.

4.1 Phase-shift keying (PSK)

La modulation par déplacement de phase (PSK) est un type ou technique de modulation numérique qui consiste à transmettre des données en modifiant la phase d'un signal porteur de référence. Elle repose sur l'utilisation d'un ensemble fini de phases, chacune étant associée à une combinaison binaire spécifique. Le principe de la PSK est donc de faire varier la phase de la porteuse en fonction des bits à transmettre, chaque valeur de phase représentant un symbole codé sur un ou plusieurs bits.

Par exemple, en BPSK(Binary Phase-shift keying), deux phases décalées de 180° permettent de représenter un bit par symbole, tandis qu'en QPSK(...), quatre phases différentes sont utilisées afin de coder Deux bits par symbole.

Le signal PSK peut être exprimé de manière générale par :

$$s(t) = E \cos(2\pi f_c t + \varphi_k) \quad (1.1)$$

où :

- E : représente l'amplitude du signal,
- f_c : la fréquence de la porteuse,
- φ_k : la phase associée au symbole transmis.

Les techniques de modulation PSK présentent plusieurs avantages et limites. Parmi les principaux avantages, on note une grande robustesse face au bruit et à la distorsion, notamment pour la modulation BPSK. De plus, certaines variantes comme la QPSK offrent une bonne efficacité spectrale, permettant de transmettre davantage de données dans une même bande

Chapitre I : Généralités sur la modulation numérique

passante. Il est également possible d'augmenter le débit de transmission en utilisant des schémas de modulation d'ordre supérieur.

Cependant, ces techniques présentent aussi certaines limites. La modulation BPSK, bien que robuste, offre un faible débit puisqu'elle ne transmet qu'un bit par symbole, ce qui la rend peu adaptée aux applications à haut débit. Par ailleurs, les schémas de modulation d'ordre supérieur deviennent plus sensibles au bruit et aux interférences.

Enfin, les performances peuvent être limitées lorsque la bande passante disponible est restreinte, en fonction du type de PSK utilisé [2]

4.1.1 Constellation

Une constellation désigne l'ensemble des symboles représentés sous forme de points dans le plan In-phase (I) et Quadrature (Q). Chaque point correspond à un état spécifique du signal modulé et permet de transmettre une séquence de bits. Le choix de la géométrie de la constellation joue un rôle essentiel dans les performances du système en termes de débit, de consommation d'énergie et de robustesse face au bruit [3]

Cette représentation est fondamentale pour l'analyse et la conception des systèmes de modulation numérique, car elle permet d'évaluer la qualité de transmission, notamment en présence de bruit, d'interférences ou de perturbations du canal. Comme le montre la figure4, les constellations BPSK et QPSK permettent de visualiser les différentes phases utilisées pour coder les symboles binaires.

Chapitre I : Généralités sur la modulation numérique

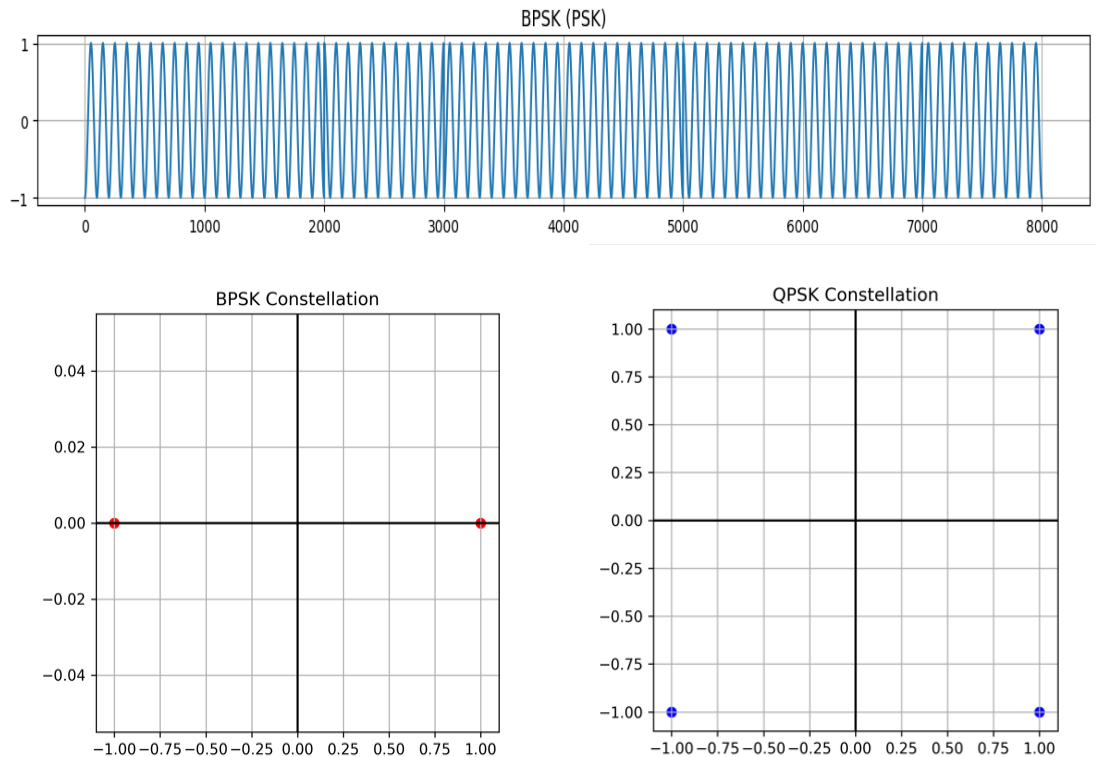


Figure I.4 : Représentation temporelle et constellation des modulations BPSK et QPSK

4.2 Frequency-shift keying (FSK)

La modulation par déplacement de fréquence (FSK) est une **technique de** modulation numérique dans lequel l'information est transmise par des variations discrètes de la fréquence d'une onde porteuse. Le principe de la FSK consiste à représenter les données numériques à l'aide de différentes fréquences. Dans sa forme la plus simple, appelée BFSK (Binary Frequency Shift Keying), deux fréquences distinctes sont utilisées : une fréquence dite *mark* pour représenter le bit 1 et une fréquence dite *space* pour représenter le bit 0. Dans le cas de la MSK (Minimum Shift Keying), l'écart entre les fréquences haute et basse est égal à la moitié du débit binaire, ce qui permet une meilleure efficacité spectrale.

Le signal FSK peut être exprimé de manière générale par :

$$s_0(t) = E \cos(2\pi f_0 t), s_1(t) = E \cos(2\pi f_1 t) \quad (1.2)$$

où :

- E représente l'amplitude du signal,
- f_0 la fréquence associée au bit 0,

Chapitre I : Généralités sur la modulation numérique

- f_1 la fréquence associée au bit 1.

Les techniques de modulation FSK présentent plusieurs avantages ainsi que certaines limites. Parmi leurs principaux atouts, on note la simplicité de la transmission numérique grâce à l'utilisation de fréquences distinctes pour représenter les données. La modulation MSK se distingue également par une meilleure efficacité spectrale, tandis que l'AFSK est particulièrement adaptée aux transmissions via des canaux tels que la radio ou les lignes téléphoniques. Cependant, ces techniques présentent aussi des inconvénients. L'AFSK, par exemple, est moins efficace en termes de puissance et d'utilisation de la bande passante comparée à d'autres techniques de modulation. De plus, elle n'est pas adaptée aux communications à haut débit. Enfin, les performances globales restent limitées par rapport à des méthodes de modulation plus avancées [2]. Comme l'illustre la figure 5, la modulation FSK repose sur la variation de la fréquence de la porteuse afin de représenter les symboles binaires transmis.

.

Chapitre I : Généralités sur la modulation numérique

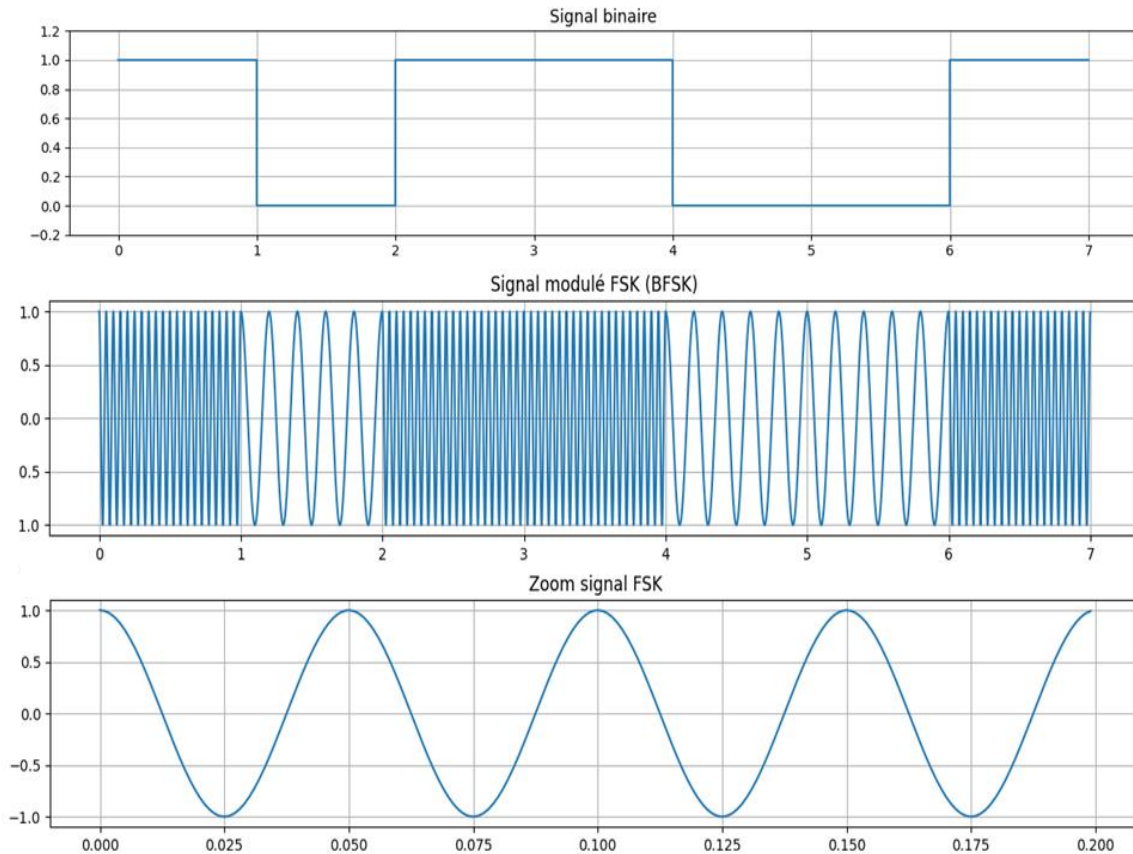


Figure I.5: Implémentation de la modulation FSK (Binary Frequency Shift Keying)

4.3 Amplitude-shift keying (ASK) or ON/OFF Keying (OOK)

La modulation par déplacement d'amplitude (ASK, ou OOK) est une technique de modulation numérique qui consiste à représenter les données binaires par des variations de l'amplitude d'une onde porteuse, tout en maintenant sa fréquence et sa phase constantes. Le principe de l'ASK repose sur le fait que le signal binaire contrôle directement l'amplitude de la porteuse : le bit 1 est généralement représenté par la présence du signal (ou d'une impulsion), tandis que le bit 0 correspond à l'absence de signal. Dans le cas de la BASK, l'amplitude de la porteuse varie en fonction des données transmises. Il s'agit d'une technique souvent cohérente, basée sur la corrélation entre le signal reçu et le signal de référence. L'ASK est notamment utilisée dans certaines applications comme les communications par fibre optique et les liaisons point à point, y compris dans des contextes militaires.

Le signal ASK est donné par l'expression :

Chapitre I : Généralités sur la modulation numérique

$$s(t) = A_K \cos(2\pi f_c t) \quad (1.3)$$

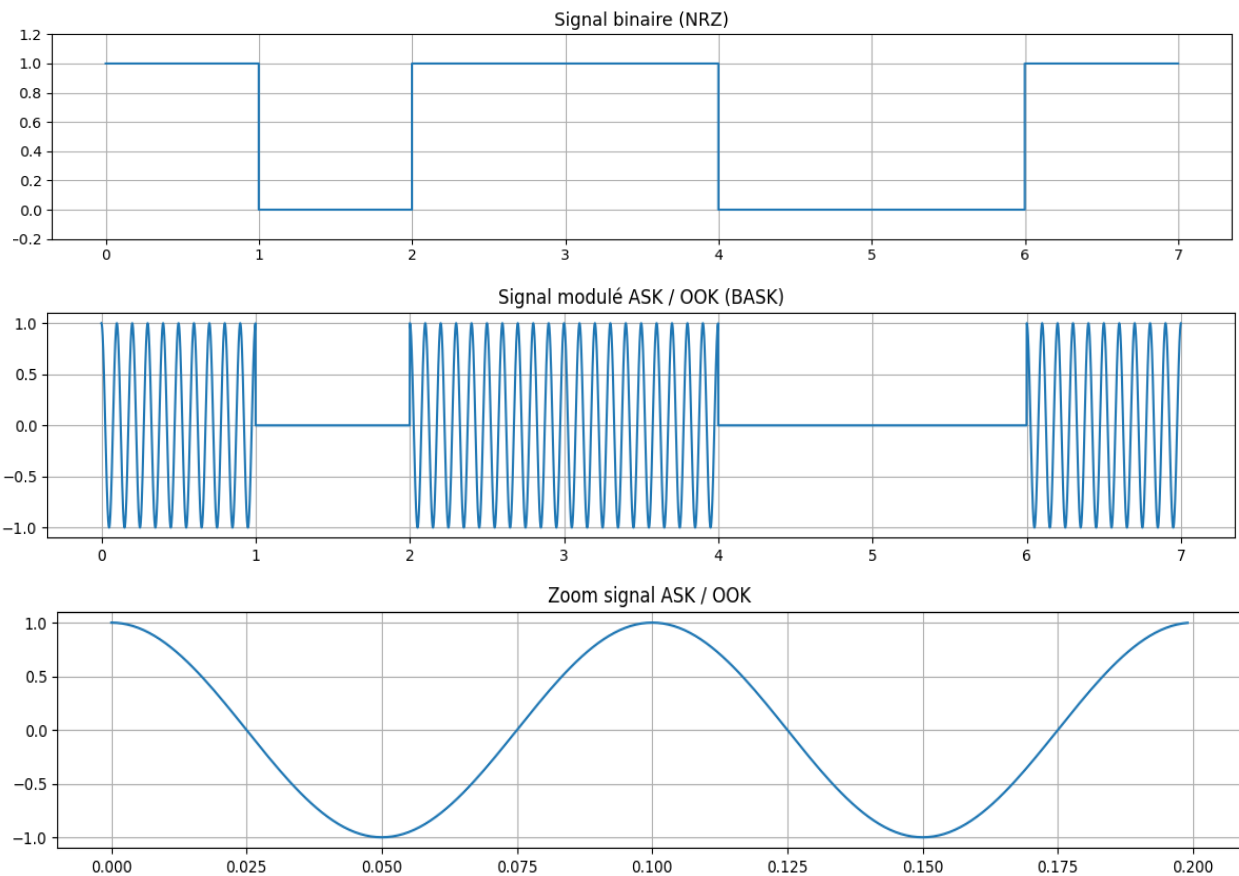
où :

- A_K représente l'amplitude associée au symbole transmis,
- f_c désigne la fréquence de la porteuse.

La modulation ASK présente plusieurs avantages ainsi que certaines limites. Parmi ses principaux atouts, on peut citer sa mise en œuvre simple, ce qui facilite son intégration dans les systèmes de communication. Elle est également utilisée dans certaines applications pratiques, notamment dans les systèmes de transmission par fibre optique. De plus, elle peut être combinée avec la modulation PSK afin de former des techniques plus avancées telles que la QAM et les modulations M-aires ASK, permettant ainsi d'améliorer les performances globales du système.

Cependant, la modulation ASK possède également des inconvénients importants. Elle se caractérise par une très faible efficacité spectrale, ce qui implique une utilisation peu optimale de la bande passante. De plus, elle est très sensible au bruit, ce qui peut affecter la qualité de transmission. Enfin, ses performances restent limitées par rapport à d'autres techniques de modulation plus robustes et plus avancées [2]. Comme le montre la figure 6, la modulation ASK repose sur la variation de l'amplitude de la porteuse afin de représenter les symboles binaires transmis.

Chapitre I : Généralités sur la modulation numérique



4.4 La modulation d'amplitude en quadrature (QAM)

La modulation d'amplitude en quadrature (QAM) est une technique de modulation numérique qui combine la variation de l'amplitude et de la phase d'une onde porteuse afin de permettre la transmission simultanée de deux signaux indépendants sur la même fréquence. Le principe de la QAM repose sur la combinaison de deux composantes orthogonales : une composante en phase (I), basée sur une fonction cosinus, et une composante en quadrature (Q), basée sur une fonction sinus. Ces deux signaux sont modulés en même temps sur la même porteuse, ce qui permet d'augmenter le nombre de symboles transmis par unité de temps. Ainsi, plus l'ordre de modulation M est élevé (par exemple 16-QAM, 64-QAM, etc.), plus le débit binaire est important, mais en contrepartie, la sensibilité du système au bruit et aux interférences augmente également [4]

Le signal QAM peut être exprimé par :

Chapitre I : Généralités sur la modulation numérique

$$s(t) = I(t)\cos(2\pi f_c t) - Q(t)\sin(2\pi f_c t) \quad (1.4)$$

La modulation QAM présente plusieurs avantages ainsi que certaines limites. Elle offre une très haute efficacité spectrale, permettant un débit binaire élevé tout en assurant une utilisation optimale de la bande passante. Grâce à ces performances, elle est largement utilisée dans les systèmes de communication modernes tels que le Wi-Fi, les réseaux câblés et les technologies LTE. De plus, elle permet une bonne adaptation aux conditions du canal, ce qui améliore la flexibilité des systèmes de transmission [4]

Cependant, la modulation QAM présente également des inconvénients. Elle est relativement sensible au bruit, ce qui peut affecter la qualité du signal reçu. De plus, lorsque l'ordre de modulation est élevé, le taux d'erreur binaire (BER) augmente. Elle nécessite également une démodulation plus complexe, impliquant la récupération simultanée de l'amplitude et de la phase. Enfin, elle est moins robuste que les modulations d'ordre faible dans des environnements fortement bruités. Comme le montre la figure 7, la modulation QAM est représentée dans le plan I/Q sous forme de constellation, où chaque point correspond à un symbole défini par une combinaison unique d'amplitude et de phase.

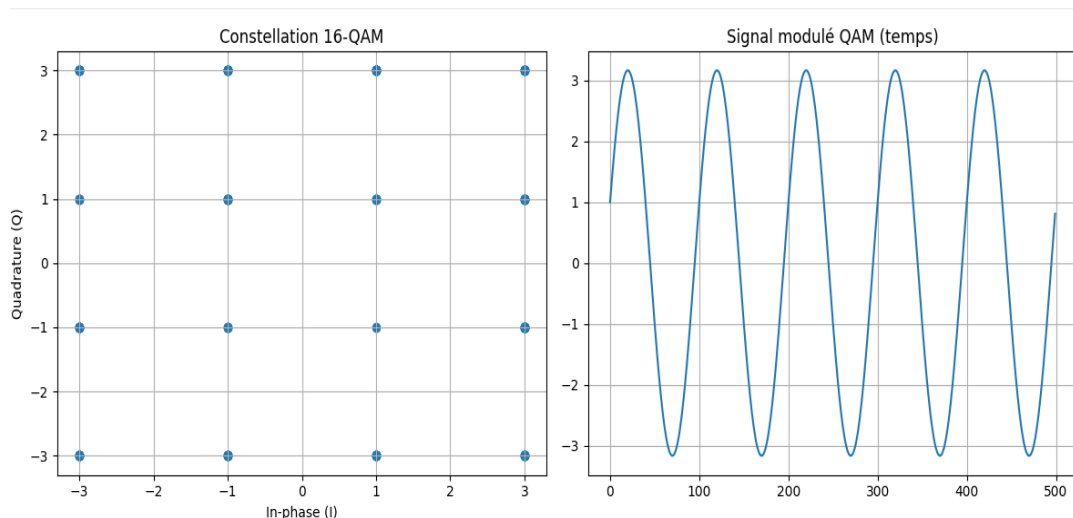


Figure I.7:représentation de la modulation 16-QAM (constellation et signal)

Chapitre I : Généralités sur la modulation numérique

5 Performances des modulations numériques

Les performances des systèmes de modulation numérique sont généralement évaluées à travers plusieurs critères essentiels, notamment le taux d'erreur binaire (BER), le rapport signal sur bruit (SNR), l'efficacité spectrale, la bande passante et l'efficacité énergétique. Ces paramètres permettent d'analyser la fiabilité et la qualité d'un système de communication dans un environnement bruité [5]

5.1 Le Bit Error Rate (BER)

Le Bit Error Rate (BER) est une mesure de performance des systèmes de communications numériques qui quantifie la fiabilité de l'ensemble du système radio de bout en bout, depuis les bits transmis ("bits in") jusqu'aux bits reçus ("bits out"), en incluant les composants électroniques, les antennes et le canal de propagation. Il est défini simplement comme le rapport entre le nombre de bits erronés et le nombre total de bits transmis, selon la formule :

$$BER = \frac{\text{Nombre d'erreurs}}{\text{Nombre total de bits transmis}} \quad (1.5)$$

Le principe du BER repose sur la comparaison entre les bits envoyés et les bits reçus afin de compter les erreurs introduites par le bruit, les imperfections du système et le canal de transmission. Ces erreurs sont influencées par des phénomènes aléatoires tels que le bruit gaussien, le fading de Rayleigh et les erreurs de quantification.

Le BER est utilisé pour évaluer la performance globale des systèmes de communication, comparer différents systèmes, analyser l'impact du bruit et du canal de transmission, ainsi que pour améliorer les performances grâce à des solutions comme l'augmentation du rapport E_b/N_0 ou l'utilisation de techniques de correction d'erreurs afin de réduire le taux d'erreurs [6]

5.2 SNR (Signal-to-Noise Ratio)

Le rapport signal sur bruit (SNR, *Signal-to-Noise Ratio*) est une mesure de la quantité de signal utile divisée par la quantité de bruit reçu. Le SNR est généralement considéré comme

Chapitre I : Généralités sur la modulation numérique

élevé lorsqu'il indique que le signal utile est dominant par rapport au bruit, ce qui est souhaitable, tandis qu'un SNR faible indique une présence importante de bruit. Le SNR est habituellement exprimé sur une échelle logarithmique, ce qui signifie que sa valeur correspond au logarithme du rapport réel entre le signal et le bruit. Il est donné par la relation :

$$\text{SNR} = 10 \log_{10}(S/N) \text{ dB} \quad (1.6)$$

Un bruit élevé (donc un SNR faible) peut entraîner un taux d'erreur binaire (BER) élevé, ce qui provoque des dégradations visibles du signal. Il est important de noter que la probabilité d'erreur (POE) est proportionnelle à E_b/N_0 , où E_b désigne l'énergie par bit transmis et N_0 la densité spectrale de puissance du bruit, et constitue une forme de rapport signal sur bruit [9]. Comme le montre la figure 8

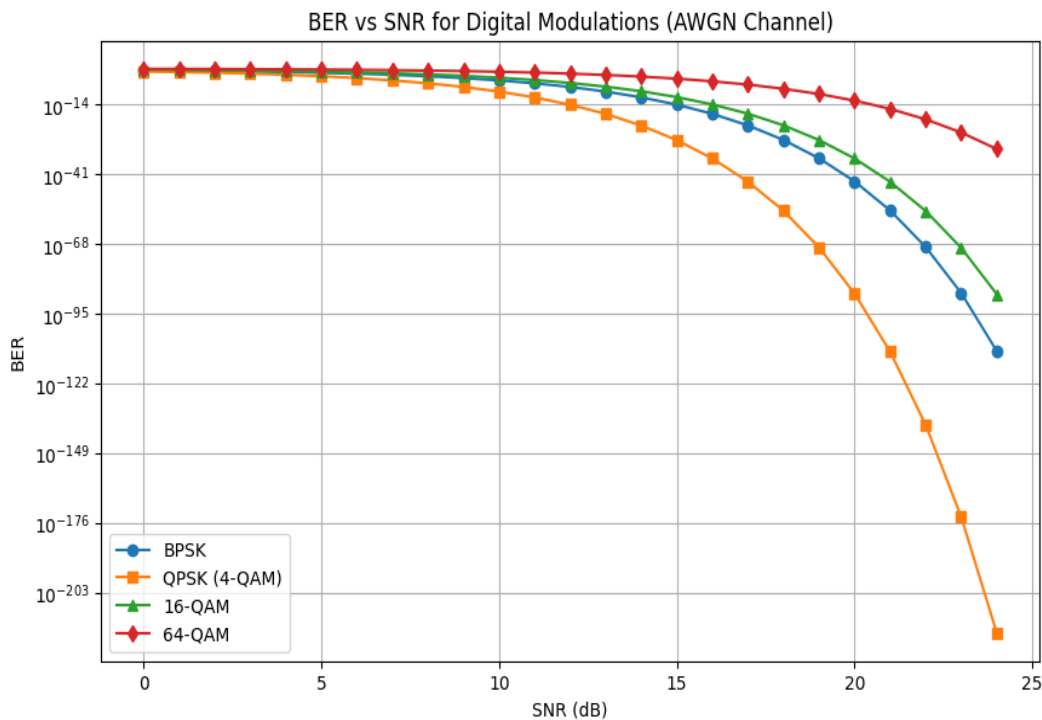


Figure I.8:Évolution (BER) en fonction du (SNR) pour les modulations (BPSK, QPSK, 16-QAM et 64-QAM) dans un canal AWGN.

5.3 Sensibilité au bruit

Les performances des systèmes de modulation numérique sont fortement influencées par la présence de bruit et d'interférences dans le canal de communication. Dans les systèmes de

Chapitre I : Généralités sur la modulation numérique

communication sans fil, le bruit le plus couramment considéré est le bruit additif blanc gaussien (AWGN), qui dégrade le signal transmis en s'ajoutant directement à celui-ci. Ce bruit est aléatoire et suit une distribution gaussienne, ce qui entraîne une distorsion de l'amplitude du signal reçu et augmente la probabilité d'erreur binaire.

La sensibilité au bruit dépend fortement du type de modulation utilisé. Les modulations d'ordre élevé, telles que la QAM-64, offrent un débit de transmission élevé mais sont plus sensibles au bruit et aux interférences, ce qui entraîne une augmentation du taux d'erreur binaire (BER), surtout lorsque le rapport signal sur bruit (SNR) est faible. À l'inverse, les modulations d'ordre inférieur comme BPSK ou QPSK présentent une meilleure robustesse face au bruit, car la distance entre les points de la constellation est plus grande, ce qui réduit la probabilité de décision incorrecte.

Dans le cas des canaux à évanouissement, tels que le Rayleigh et le Rician fading, la dégradation du signal est encore plus importante. Le canal Rayleigh, caractérisé par l'absence de ligne de vue directe, provoque des fluctuations importantes de l'amplitude du signal, ce qui augmente fortement la sensibilité au bruit et dégrade les performances du système. En revanche, le canal Rician, qui inclut une composante en ligne de vue dominante, présente de meilleures performances et une sensibilité au bruit plus faible que le canal Rayleigh [7]

Chapitre I : Généralités sur la modulation numérique

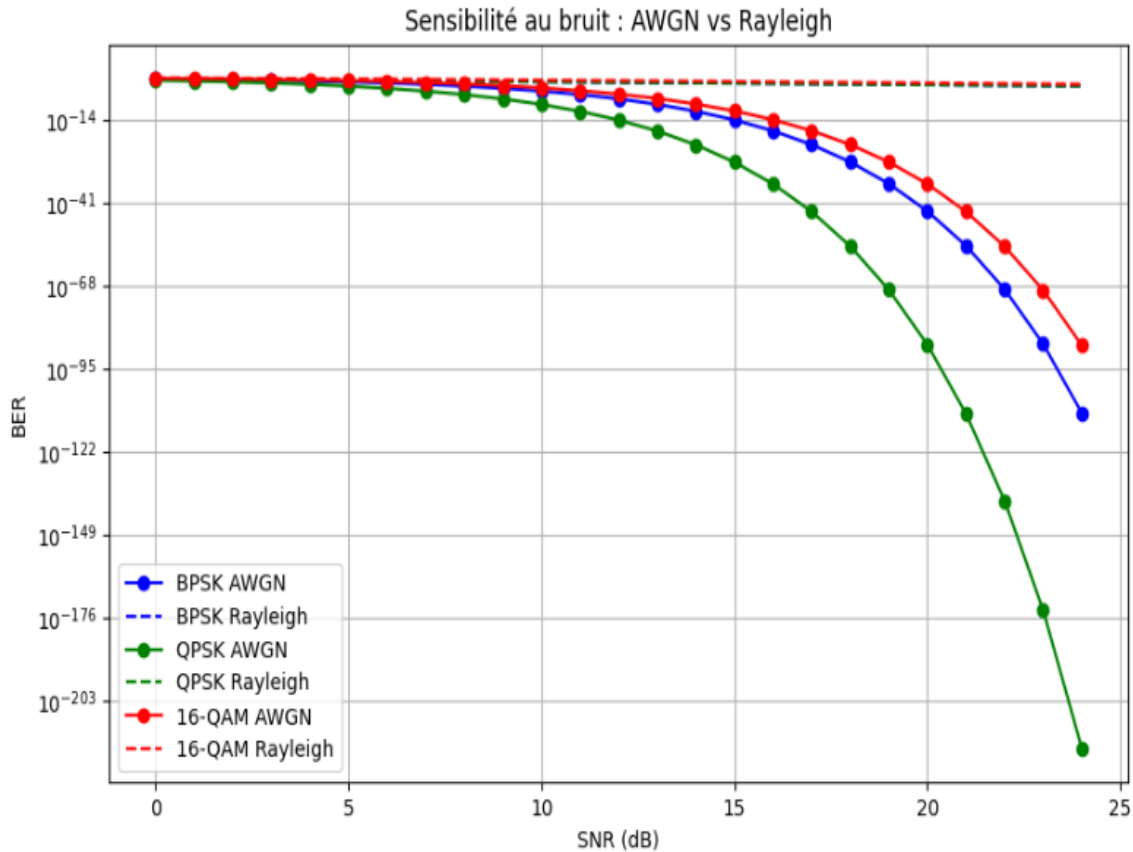


Figure I.9: Comparaison de la sensibilité au bruit des modulations numériques (BPSK, QPSK et M-QAM) dans les canaux AWGN et Rayleigh en fonction du SNR

5.4 Efficacité spectrale

L'efficacité spectrale correspond à la quantité de données transmises par unité de bande passante (bits/s/Hz). Les modulations M-aires permettent d'améliorer cette efficacité en envoyant plusieurs bits par symbole selon la relation $M = 2^n$, où (n) représente le nombre de bits par symbole.

Les modulations comme M-QAM et M-PSK offrent une meilleure efficacité spectrale que les modulations binaires classiques, car elles utilisent respectivement l'amplitude et la phase du signal pour transporter l'information. Ainsi, la M-QAM est très efficace spectralement car elle combine amplitude et phase, tandis que la M-PSK offre également une bonne efficacité, mais avec une énergie plus constante [8]. Comme le montre la figure 11

Chapitre I : Généralités sur la modulation numérique

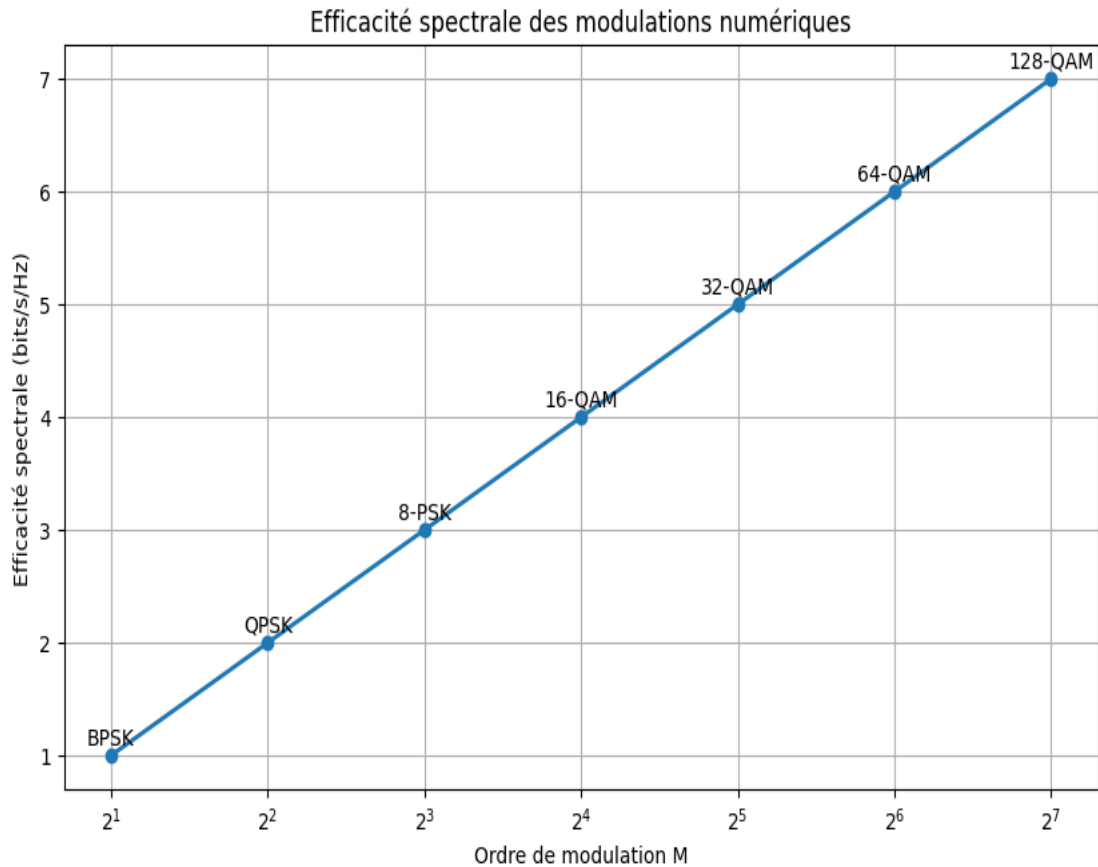


Figure I.10: variation de l'efficacité spectrale des modulations numérique

La bande passante correspond à la ressource fréquentielle nécessaire pour assurer la transmission d'un signal. Les modulations M-aires permettent de réduire la bande passante requise pour un débit donné, en augmentant le nombre de bits transmis par symbole. Dans les systèmes M-PSK et M-QAM, l'augmentation de M permet de transmettre davantage d'informations sans élargir significativement la bande passante, ce qui améliore l'efficacité spectrale. Cependant, cette amélioration s'accompagne d'une augmentation de la complexité du récepteur ainsi que d'une plus grande sensibilité au bruit. En revanche, dans les systèmes M-FSK, l'augmentation de M peut entraîner une extension de la bande passante, ce qui réduit l'efficacité spectrale. Ainsi, il existe un compromis fondamental entre la bande passante, le débit binaire et la robustesse face au bruit [9]. Comme le montre la figure 12

Chapitre I : Généralités sur la modulation numérique

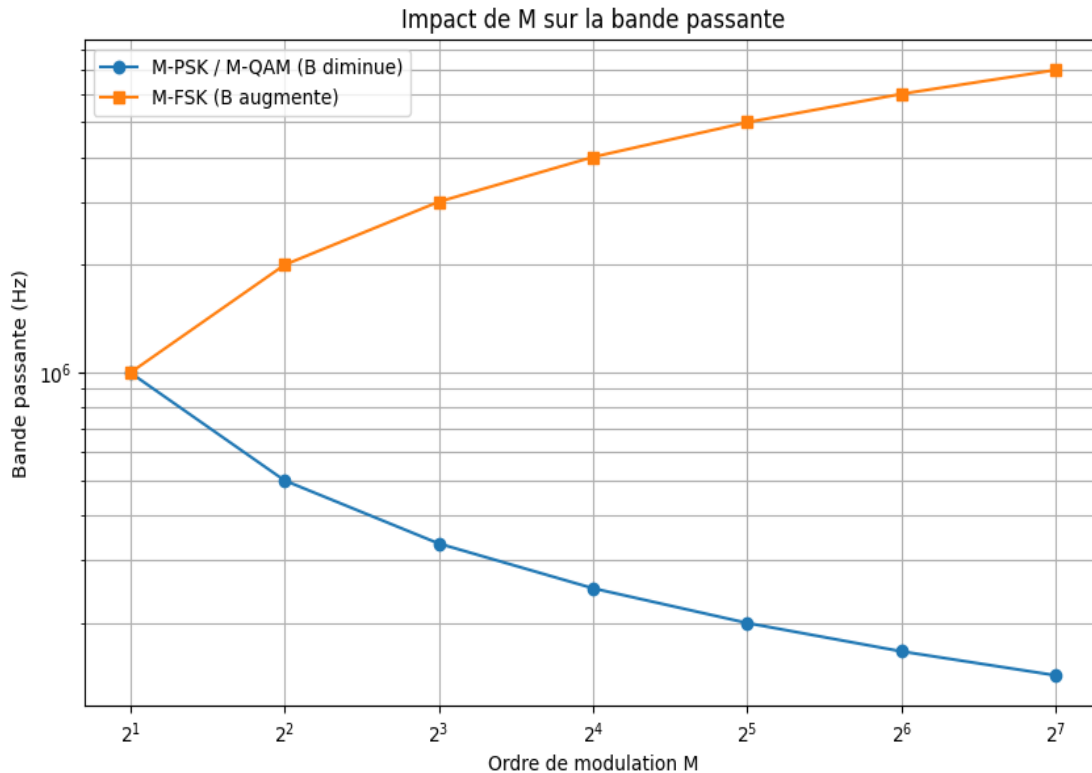


Figure I.11: Influence de l'ordre de modulation (M) sur la bande passante pour M-PSK/M-QAM et FSK

6 Applications de la modulation numérique

La radio intelligente (Cognitive Radio) permet d'analyser l'environnement de communication, de détecter la présence ou l'absence de signal, et d'adapter automatiquement les paramètres de transmission. Par exemple, elle peut changer la modulation en fonction de la qualité du canal afin d'optimiser les performances.

Les systèmes adaptatifs sans fil fonctionnent dans le même principe, en modifiant la modulation en temps réel et en s'adaptant aux conditions du canal. L'objectif principal est de maintenir la qualité et la stabilité de la communication.

Dans les systèmes WiMAX, cette adaptation est également utilisée pour ajuster les caractéristiques du signal et optimiser la transmission. Par ailleurs, dans les applications militaires, ces techniques sont employées pour l'interception de signaux et dans des situations

Chapitre I : Généralités sur la modulation numérique

où aucune information préalable sur le signal reçu n'est disponible, ce qui rend la reconnaissance automatique de la modulation essentielle.

La reconnaissance automatique de modulation est ainsi utilisée pour identifier le type de modulation (QAM, PSK, APSK, etc.) et permettre au récepteur de fonctionner sans information préalable sur le signal transmis [10]

7 Conclusion

Dans ce chapitre, nous avons étudié le rôle fondamental de la modulation dans les systèmes de télécommunications, en mettant en évidence son importance dans la transmission efficace de l'information à travers différents types de canaux. La modulation permet d'adapter le signal d'information aux contraintes du milieu de transmission en exploitant une onde porteuse, dont les paramètres tels que l'amplitude, la fréquence ou la phase sont modifiés pour représenter les données à transmettre.

Nous avons distingué les deux grandes catégories de modulation, à savoir la modulation analogique et la modulation numérique. Alors que les techniques analogiques comme l'AM, la FM et la PM offrent des solutions simples et historiquement importantes, elles présentent certaines limites face aux exigences des systèmes modernes. En revanche, la modulation numérique s'impose aujourd'hui comme une solution incontournable grâce à sa robustesse face au bruit, sa meilleure efficacité spectrale et sa capacité à transporter de grandes quantités de données.

L'analyse des principales techniques de modulation numérique, telles que ASK, FSK, PSK et QAM, a permis de mettre en évidence leurs principes de fonctionnement, ainsi que leurs avantages et leurs limites. Nous avons également examiné les critères de performance essentiels, notamment le taux d'erreur binaire (BER), le rapport signal sur bruit (SNR), l'efficacité spectrale et la bande passante, qui permettent d'évaluer la qualité et la fiabilité des systèmes de communication.

Enfin, ce chapitre a montré que le choix d'une technique de modulation repose sur un compromis entre plusieurs facteurs, tels que le débit, la robustesse au bruit et l'utilisation de la

Chapitre I : Généralités sur la modulation numérique

bande passante. Dans un contexte de croissance continue des besoins en communication, les techniques de modulation numérique jouent un rôle central dans le développement des systèmes modernes, notamment dans les réseaux sans fil, les communications satellitaires et les technologies de nouvelle génération.

Chapitre II : Machine learning ML

Chapitre II : Machine learning ML

Introduction

Ce chapitre est consacré à l'étude des fondements du Machine Learning. Il présente d'abord les notions de base telles que les données, les modèles et le principe d'apprentissage. Ensuite, il décrit les différentes étapes du pipeline d'un système de Machine Learning, depuis la collecte des données jusqu'à l'évaluation des performances. Enfin, il aborde les principaux types d'apprentissage ainsi que les algorithmes les plus utilisés en classification, régression et clustering, tout en mettant en évidence leurs principes de fonctionnement et leurs domaines d'application.

1. L'intelligence artificielle (IA)

L'Intelligence Artificielle (IA) est une technologie qui transforme le secteur des télécommunications en permettant aux opérateurs de concevoir, gérer et fournir des services de manière plus intelligente. Elle repose sur des techniques telles que le machine learning, le traitement du langage naturel et l'analyse prédictive, afin d'automatiser les opérations réseau, améliorer l'expérience client et optimiser la prise de décision.

L'IA est utilisée pour automatiser la détection des pannes, la gestion des ressources, la personnalisation des services et la prévention des fraudes. Elle joue ainsi un rôle essentiel dans l'évolution vers des réseaux plus intelligents, notamment avec l'arrivée de la 5G et de la 6G [11]

1.1 Machine learning (ML)

Le Machine Learning est une branche de l'intelligence artificielle qui permet à un système d'apprendre à partir de l'expérience et des données disponibles.

Son principe repose sur l'utilisation de données brutes (datasets) afin de construire des modèles capables de réaliser différentes tâches comme la classification, la prédiction, l'analyse de données ou la prise de décision, sans être programmés explicitement pour chaque situation.

Chapitre II : Machine learning ML

Ainsi, le Machine Learning a pour objectif de développer des systèmes intelligents capables d'apprendre, de s'adapter et d'améliorer leurs performances au fil du temps grâce aux données et à l'expérience [12]

1.2 Le Deep Learning (DL)

Le Deep Learning (apprentissage profond) est une forme avancée du Machine Learning qui repose sur l'utilisation de réseaux de neurones artificiels à plusieurs couches. Il permet aux systèmes d'apprendre automatiquement à partir des données sans nécessiter une intervention humaine importante pour l'extraction des caractéristiques. Grâce à cette capacité, il est particulièrement efficace pour traiter des données complexes et non structurées, telles que les images, les signaux ou les séries temporelles, et pour modéliser des relations complexes entre les variables [13]

Le Machine Learning (ML) et le Deep Learning (DL) sont deux approches de l'intelligence artificielle utilisées pour analyser les données et faire des prédictions, mais ils se distinguent par leur fonctionnement. Le ML regroupe des algorithmes classiques qui nécessitent généralement une intervention humaine pour sélectionner et extraire les caractéristiques importantes des données. En revanche, le DL, qui est une sous-catégorie du ML, utilise des réseaux de neurones artificiels profonds capables d'apprendre automatiquement ces caractéristiques à partir de données brutes. De plus, le ML est souvent appliqué à des données structurées, tandis que le DL est particulièrement adapté aux données complexes et non structurées comme les images, les signaux ou les séries temporelles, ce qui le rend plus puissant mais aussi plus exigeant en termes de calcul et de donnée [13]. Figure 1 illustre Relations et interactions entre l'Intelligence Artificielle (IA), le Machine Learning (ML) et le Deep Learning (DL).

Chapitre II : Machine learning ML

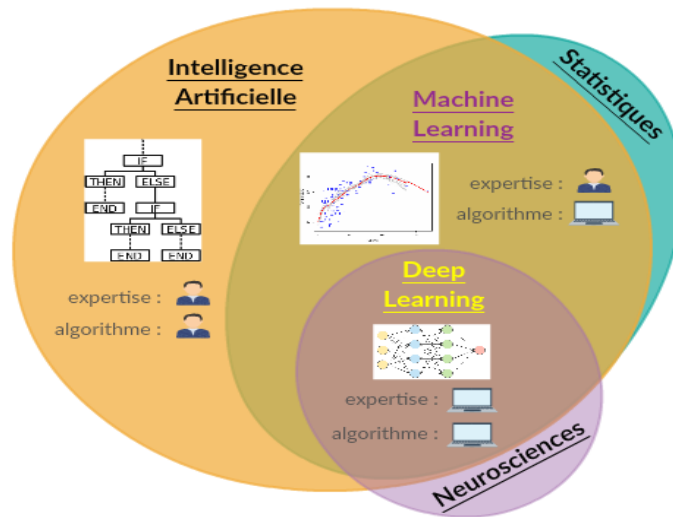


Figure II.1: Interactions IA ML DL

2. Modèle de machine learning

Le machine learning repose sur un processus d'apprentissage automatique permettant à un système informatique d'exploiter des données, appelées données brutes ou *datasets*, afin d'en extraire des connaissances. Ce processus s'organise généralement en deux étapes principales : l'entraînement du modèle et la phase de test, au cours de laquelle le modèle prend des décisions.

Pour réaliser cet apprentissage, deux types de données sont utilisés : les données d'entraînement et les données de test. Les données d'entraînement servent à construire le modèle. Avant cette étape, elles passent par des opérations de prétraitement ainsi que d'extraction de caractéristiques, afin d'améliorer la qualité de l'apprentissage.

Une fois le modèle entraîné, il est appliqué à de nouvelles données, appelées données de test, dont les résultats sont inconnus. Le modèle produit alors des prédictions sur ces données. Par ailleurs, lorsque les résultats réels sont disponibles, il est possible d'évaluer la performance du modèle en mesurant son exactitude comme le montre la figure 2 [12]

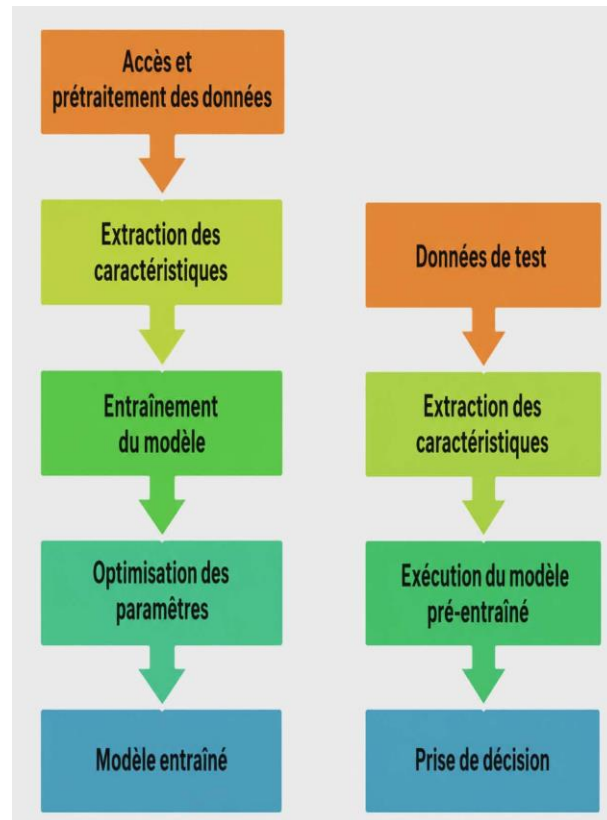


Figure II.2: Entraînement et test d'un modèle de Machine Learning

3. Pipeline d'un système de Machine Learning

Le pipeline d'un système de Machine Learning comprend plusieurs étapes successives allant de la collecte des données jusqu'à l'évaluation du modèle. Chaque étape joue un rôle essentiel dans l'amélioration des performances et de la fiabilité du système. La **Figure 3** illustre les principales phases du pipeline d'un système de Machine Learning.

3.1 Collecte des données

Cette étape consiste à rassembler les données nécessaires à l'apprentissage du modèle. Dans les systèmes ML modernes, les données proviennent de différentes sources et sont utilisées à la fois pour l'entraînement et le test. Les applications de Machine Learning reposent sur de grands volumes de données, ce qui crée des défis importants en termes de stockage et de gestion. Ainsi, la collecte des données est une étape critique du pipeline, car la qualité et la quantité des données influencent directement les performances du modèle.

Chapitre II : Machine learning ML

3.2 Nettoyage et prétraitement des données

Le prétraitement des données comprend plusieurs opérations comme la validation, la transformation et la préparation des données avant leur utilisation par le modèle. Les systèmes ML nécessitent plusieurs étapes de traitement des données telles que le nettoyage, l'analyse et la validation des données. Cette phase permet d'assurer que les données soient cohérentes, exploitables et adaptées au modèle. Elle réduit également les erreurs et améliore la fiabilité des résultats.

3.3 Extraction et sélection des caractéristiques (Feature Engineering)

Cette étape consiste à identifier et construire les variables les plus pertinentes à partir des données brutes. Les pipelines ML incluent des processus comme l'extraction de caractéristiques, qui font partie des étapes essentielles du cycle de vie du Machine Learning. L'objectif est d'améliorer la qualité des données utilisées par le modèle afin d'augmenter sa performance et sa capacité de généralisation. Une bonne sélection de features permet aussi de réduire la complexité computationnelle.

3.4 Entraînement du modèle

L'entraînement est une phase centrale du pipeline ML où le modèle apprend à partir des données préparées. Cette étape nécessite une puissance de calcul importante, souvent fournie par des GPU ou TPU. Les paramètres du modèle (comme les hyperparamètres et les poids des réseaux de neurones) sont ajustés afin de minimiser l'erreur. Les différents modèles profonds tels que ResNet, VGG et MobileNet ont été entraînés avec des architectures et des configurations d'hyperparamètres variées, ce qui met en évidence que cette phase d'apprentissage dépend fortement du choix du modèle et de sa structure.

3.5 Validation et test du modèle

Cette étape consiste à évaluer les performances du modèle à l'aide de données de test et de différentes métriques. L'évaluation des modèles fait partie intégrante du pipeline ML et permet de mesurer leur efficacité. Des métriques comme l'utilisation du GPU, la consommation mémoire et d'autres indicateurs de performance sont utilisées dans l'étude.

Chapitre II : Machine learning ML

L'objectif est de vérifier la capacité du modèle à généraliser sur de nouvelles données et à éviter les problèmes de sur apprentissage. Cette étape permet également de comparer différents modèles et architectures[10].

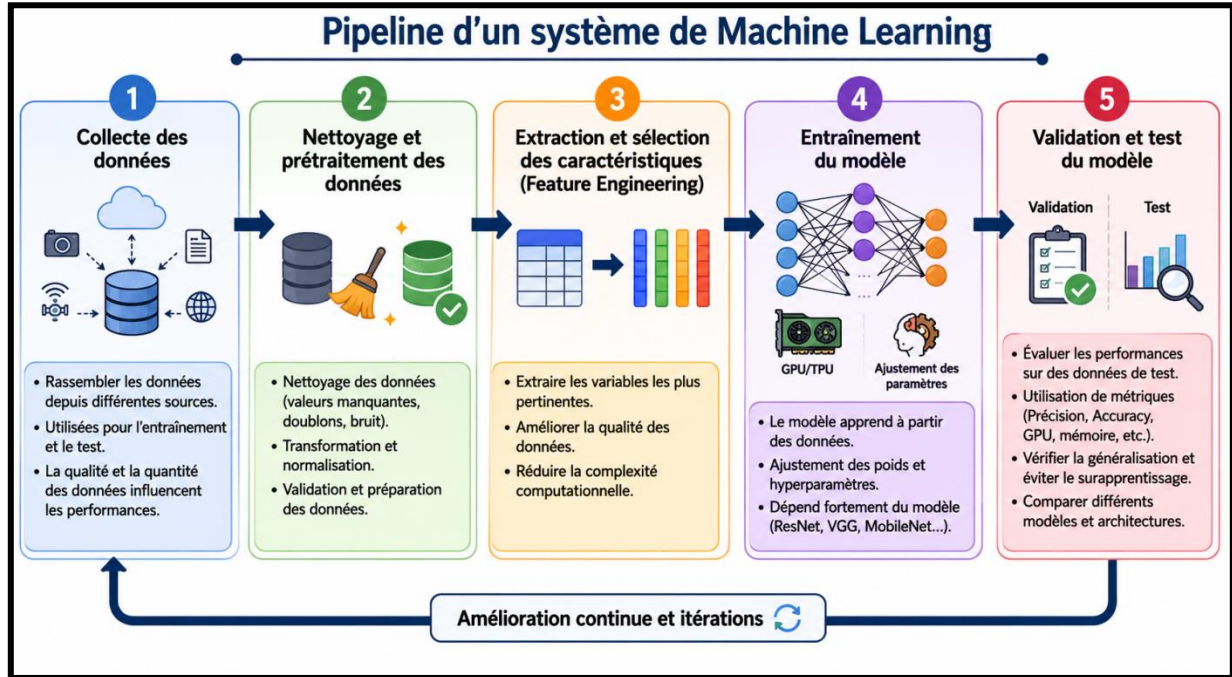


Figure II.3: pipeline d'un système de Machine Learning

4. Types d'apprentissage

Les techniques d'apprentissage automatique se déclinent en plusieurs types, chacun reposant sur une manière spécifique d'exploiter les données afin de permettre au modèle d'apprendre, de prédire ou de prendre des décisions de manière autonome. La Figure 4 suivante illustre les principaux types d'apprentissage.

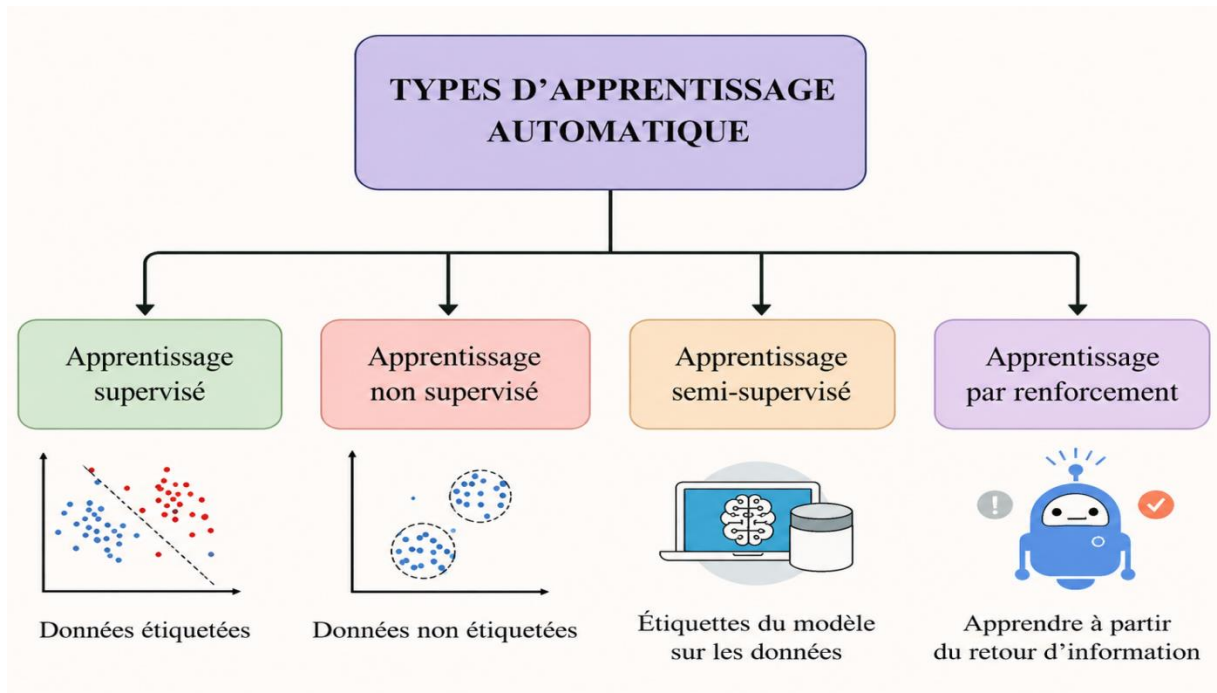


Figure II.4: Types d'apprentissage automatique

4.1 Apprentissage supervisé (Supervised Learning)

L'apprentissage supervisé est une méthode dans laquelle les données sont étiquetées, c'est-à-dire que chaque instance possède déjà une valeur de sortie connue. Dans ce cadre, les données sont généralement divisées en deux ensembles : un ensemble d'entraînement et un ensemble de test. Chaque instance est composée d'attributs (features) et d'une classe (label), et le modèle est entraîné afin d'apprendre la relation entre ces attributs et les classes. L'objectif principal est de construire un modèle capable de prédire des valeurs inconnues à partir de données connues. Cette approche est couramment utilisée pour les problèmes de classification, lorsque les valeurs sont discrètes, et pour les problèmes de régression, lorsque les valeurs sont continues [14] . Ce principe est illustré par la figure 5.

Chapitre II : Machine learning ML

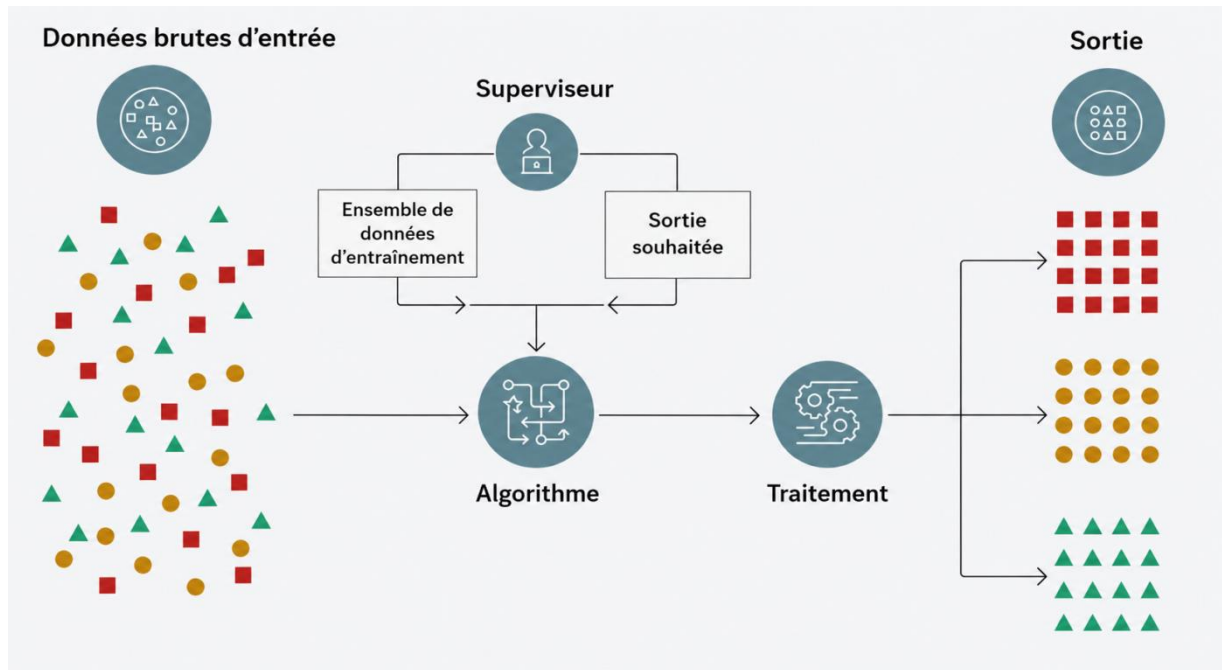


Figure II.5: L'apprentissage supervisé(supervised learning)

4.1.1 La régression:

La régression est une tâche de l'apprentissage supervisé qui consiste à prédire une valeur numérique continue. Par exemple, elle peut être utilisée pour estimer le prix d'une maison, la température ou encore la consommation d'énergie. Dans ce cas, le modèle cherche à établir une relation entre les variables d'entrée et une sortie quantitative. La régression linéaire est l'un des modèles les plus simples utilisés pour cela. Elle repose sur une combinaison linéaire des variables d'entrée afin de minimiser l'erreur entre les valeurs réelles et les valeurs prédites. Ce type de modèle est efficace pour comprendre et prédire des tendances numériques et valeur numérique [15]

4.1.2 La classification :

La classification est utilisée lorsque la sortie appartient à une catégorie précise. Par exemple, elle permet de distinguer un email spam d'un email normal, ou encore de reconnaître une image comme étant un chat ou un chien. Contrairement à la régression, la classification ne prédit pas une valeur numérique mais une classe. Parmi les modèles utilisés pour cette tâche, on trouve la régression logistique et le perceptron. La régression logistique permet d'estimer la probabilité qu'une donnée appartienne à une classe donnée, tandis que le perceptron prend une décision

Chapitre II : Machine learning ML

directe basée sur un seuil. Ces modèles sont largement utilisés dans les problèmes de classification binaire.

En résumé, la régression et la classification sont deux tâches fondamentales de l'apprentissage supervisé. La régression sert à prédire des valeurs continues, tandis que la classification sert à attribuer des catégories. Ces méthodes permettent aux machines d'apprendre à partir de données et de résoudre des problèmes complexes de manière automatique et efficace [15] Comme illustré dans la **Figure 6**, qui présente la différence entre les approches de classification et de régression en apprentissage automatique.

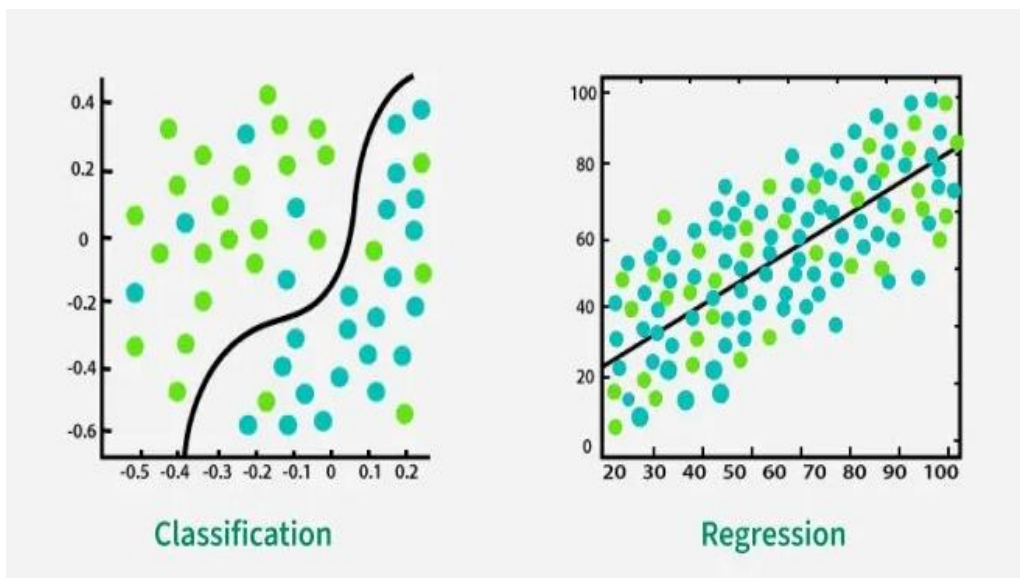


Figure II.6: Classification et régression en apprentissage automatique

4.2 Apprentissage non supervisé (Unsupervised Learning)

L'apprentissage non supervisé est une méthode du machine learning dans laquelle le système est entraîné sur un ensemble de données non étiquetées, c'est-à-dire sans classes ou valeurs de sortie connues. Dans cette approche, seules les données d'entrée sont fournies au modèle, sans supervision humaine directe. L'objectif principal est d'identifier des structures cachées, des relations ou des régularités présentes dans les données. Le système analyse ainsi les similitudes entre les observations afin de les regrouper selon leurs caractéristiques. Ce type d'apprentissage est principalement utilisé pour le clustering, l'analyse d'association, la détection d'anomalies

Chapitre II : Machine learning ML

ainsi que les modèles d'autoencoders, permettant de découvrir des patterns significatifs dans des données non labellisées [16] Ce principe est illustré par la figure 7.

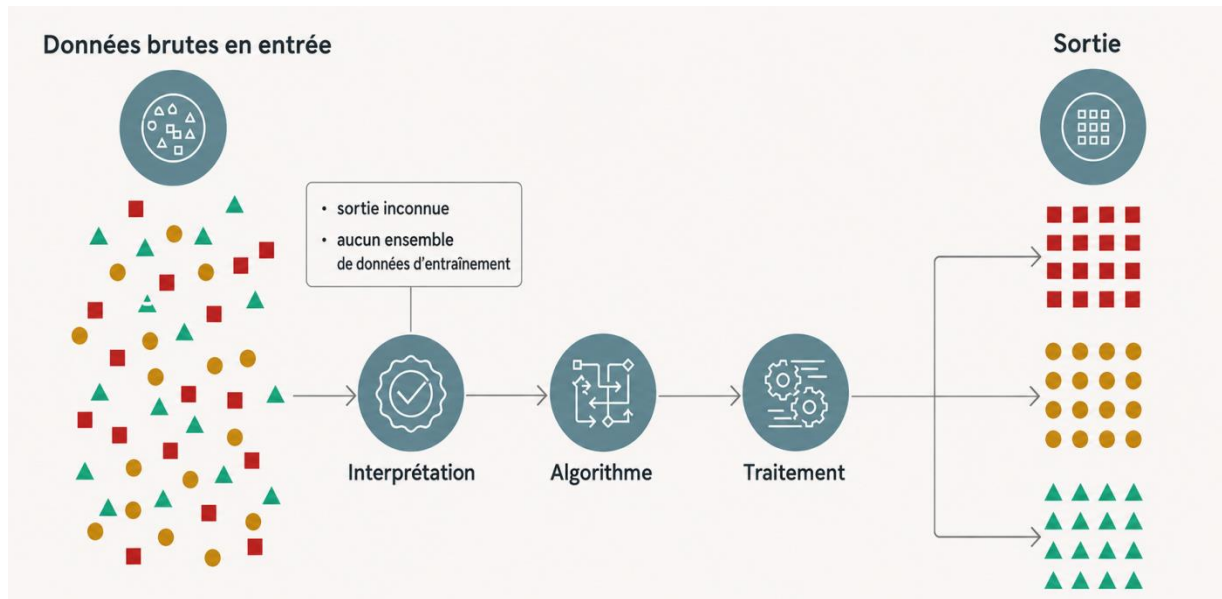


Figure II.7: Machine Learning-processus d'apprentissage non supervise

4.2.1 Clustering (regroupement) :

Le clustering, également appelé analyse de regroupement, correspond au processus de classification des données en différents groupes, appelés *clusters*, en fonction de leur similarité.

Cette approche peut être déclinée en plusieurs formes, notamment le clustering partitionnel, hiérarchique, chevauchant et probabiliste.

Dans le clustering partitionnel, les données sont réparties de manière à ce que chaque élément appartienne à un seul et unique cluster. Ce type de regroupement est aussi désigné par le terme de clustering exclusif. L'algorithme K-means constitue un exemple représentatif de cette approche.

Le clustering hiérarchique repose sur la construction progressive de groupes. Initialement, chaque donnée est considérée comme un cluster individuel. Ensuite, les clusters les plus proches

Chapitre II : Machine learning ML

sont fusionnés de manière itérative, ce qui permet de réduire progressivement le nombre de groupes.

Le clustering chevauchant permet, quant à lui, d'associer une même donnée à plusieurs clusters. Dans ce cas, chaque élément est caractérisé par un degré d'appartenance à chaque groupe, généralement exprimé sous forme de valeur.

Enfin, le clustering probabiliste s'appuie sur des distributions de probabilité pour former les clusters, en associant chaque donnée à une probabilité d'appartenance aux différents groupes.

Selon leur mode de fonctionnement, les méthodes de clustering peuvent être classées en trois grandes catégories : le clustering exclusif, dans lequel chaque élément appartient à un seul cluster ; le clustering hiérarchique, qui organise les données sous forme de structure arborescente ; et le clustering chevauchant, où une donnée peut appartenir simultanément à plusieurs clusters.

4.2.2 Association (Association Rule Learning) :

L'**association**, ou apprentissage par règles d'association, est une méthode d'apprentissage non supervisé qui vise à découvrir des relations entre différentes variables dans de grands ensembles de données.

Cette approche permet d'identifier des liens significatifs entre les éléments, en analysant la manière dont certaines variables apparaissent ensemble dans les données. Elle est particulièrement utile pour comprendre les interactions entre différents attributs, notamment dans des contextes où les données sont nombreuses et variées.

Le principe repose sur l'extraction de règles de type *si/alors*, qui mettent en évidence des associations fréquentes entre les éléments. Par exemple, lorsqu'un produit X est acheté, il est possible d'observer que le produit Y est souvent acheté en même temps.

L'évaluation de ces règles se base principalement sur deux mesures :

Chapitre II : Machine learning ML

- le **support**, qui représente la fréquence d'apparition d'une association dans l'ensemble des données ;
- la **confiance**, qui indique la fiabilité ou la probabilité que la relation soit vérifiée.

Ainsi, l'apprentissage par association permet de détecter des tendances et des relations cachées dans les données, et il est largement utilisé dans des domaines tels que l'analyse des comportements, les systèmes de recommandation et l'exploration de données [16]

4.2.3 Détection d'anomalies :

La détection d'anomalies est une technique qui consiste à identifier des observations inhabituelles ou des valeurs qui s'écartent significativement du comportement normal au sein d'un ensemble de données.

Ces anomalies peuvent représenter des situations rares ou suspectes, telles que des activités réseau anormales, un dysfonctionnement de capteurs, ou encore des données erronées nécessitant un nettoyage avant analyse.

Dans ce contexte, une anomalie correspond à une donnée qui ne suit pas le modèle général ou qui présente un comportement différent de la majorité des observations. Par exemple, un trafic réseau inhabituel peut indiquer une tentative d'intrusion ou le transfert non autorisé d'informations sensibles.

La détection d'anomalies repose sur l'identification ou la prédiction des points de données qui s'écartent du modèle normal. Elle est largement utilisée dans plusieurs domaines, notamment la cybersécurité, la détection de fraude, l'assurance ainsi que la surveillance militaire.

Ainsi, cette approche joue un rôle essentiel dans l'analyse des données, en permettant de repérer des comportements atypiques pouvant avoir un impact important sur les systèmes étudiés [16]

La figure 8 illustre les principales approches de l'apprentissage non supervisé, utilisées pour explorer, structurer et analyser des données non étiquetées à travers trois tâches fondamentales : le clustering, l'apprentissage par règles d'association et la détection d'anomalies.

Chapitre II : Machine learning ML

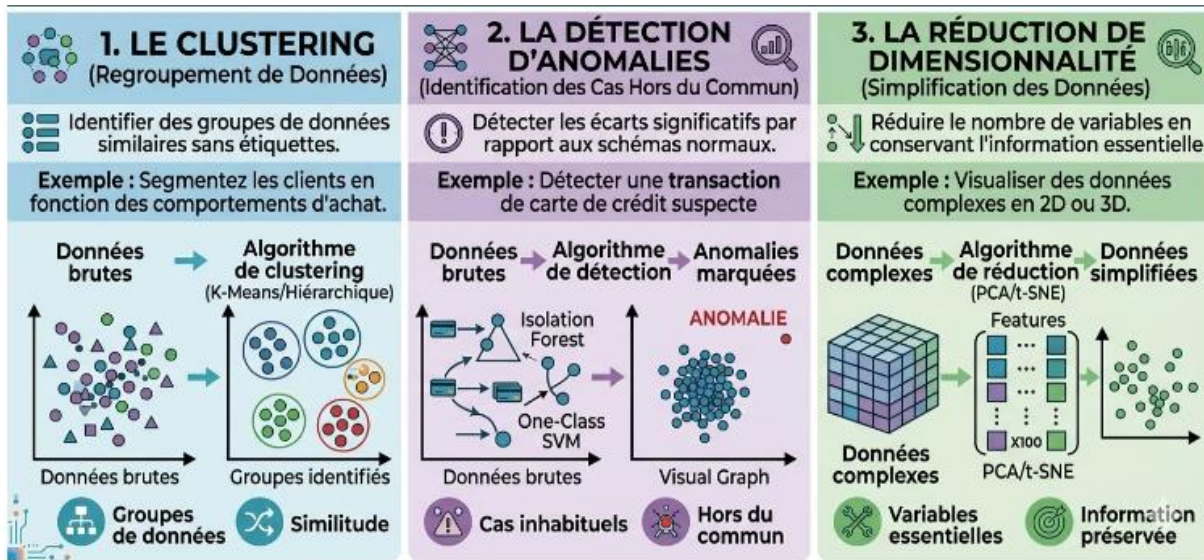


Figure II.8: Méthodes d'apprentissage non supervisé en fouille de données

4.2.4 Auto-encoders (Auto-encodeurs) :

Les auto-encodeurs sont des modèles d'apprentissage non supervisé basés sur les réseaux de neurones, principalement utilisés pour apprendre une représentation compacte et utile des données.

Ils reposent sur une architecture spécifique contenant un goulot d'étranglement (bottleneck), qui force le réseau à réduire la dimension des données d'entrée en une forme plus simple et plus condensée.

Le fonctionnement se déroule en deux étapes : d'abord, les données sont compressées dans une représentation intermédiaire, puis elles sont reconstruites à partir de cette représentation. Cette tâche devient difficile lorsque les variables d'entrée ne sont pas liées entre elles. En revanche, si les données présentent une structure interne ou des relations entre les variables, le modèle peut les exploiter pour mieux apprendre.

Chapitre II : Machine learning ML

Le goulot d'étranglement joue un rôle essentiel, car il empêche le réseau de simplement mémoriser les données. Sans cette contrainte, le modèle pourrait se contenter de reproduire les entrées sans réellement apprendre leur structure.

Ainsi, les autoencodeurs permettent d'extraire des représentations compressées, tout en capturant les informations les plus importantes contenues dans les données. Ce fonctionnement est illustré dans la **Figure 9**, qui présente les différentes étapes du processus d'un autoencodeur, depuis l'encodage des données jusqu'à leur reconstruction [5].

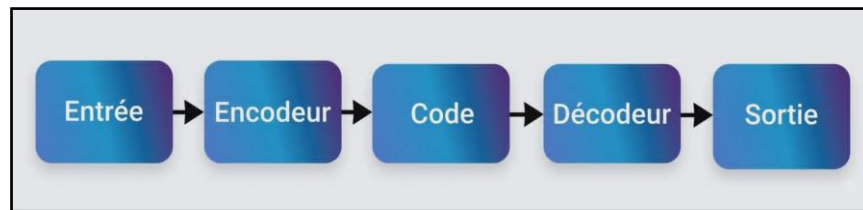


Figure II.9:Schéma du processus d'un autoencoders

4.2.5Apprentissage semi-supervisé (Semi-Supervised Learning)

L'apprentissage semi-supervisé est une méthode de machine learning intermédiaire entre l'apprentissage supervisé et non supervisé. Il repose sur l'utilisation d'un ensemble de données partiellement étiqueté, comprenant à la fois des données annotées et des données non annotées. L'objectif principal de cette approche est de combiner les avantages des deux méthodes afin de réduire le besoin important en données étiquetées, tout en améliorant la qualité des modèles appris.

Dans ce type d'apprentissage, le modèle est principalement entraîné à partir des données étiquetées, tandis que les données non étiquetées sont exploitées pour enrichir et améliorer le processus d'apprentissage. Cela permet de mieux généraliser les résultats et de prédire les étiquettes inconnues avec plus de précision.

Les principales méthodes utilisées en apprentissage semi-supervisé incluent le self-training, le co-training, le multi-view learning, les TSVMs (Transductive Support Vector Machines), ainsi que les approches génératives et discriminatives. Deux formes principales sont souvent distinguées : la classification semi-supervisée, qui permet de classer de grandes quantités de données non étiquetées à partir d'un petit jeu d'exemples annotés, et le clustering semi-

Chapitre II : Machine learning ML

supervisé, qui améliore le regroupement des données en utilisant des contraintes telles que must-link et cannot-link pour guider la formation des clusters [17]. Ce principe est illustré par la figure 10.

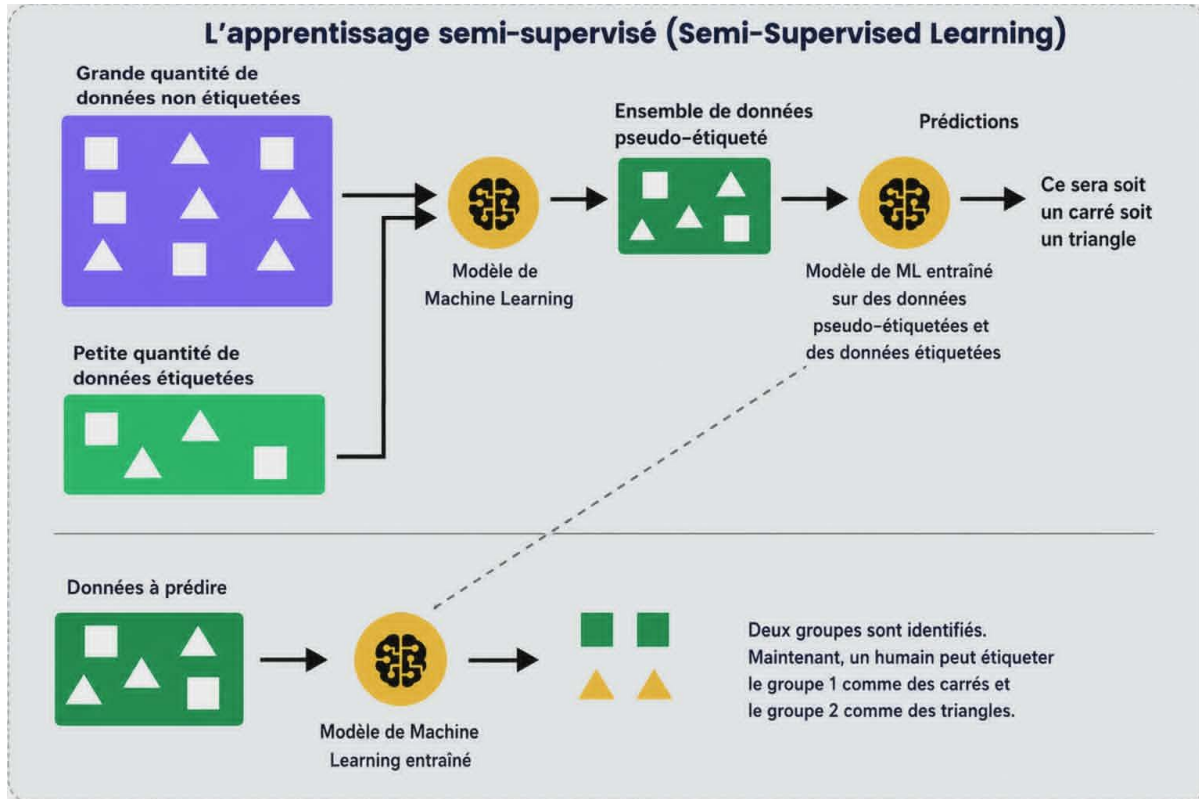


Figure II.10: l'apprentissage semi-supervisé (Semi-Supervised Learning)

4.3 Apprentissage par renforcement (Reinforcement Learning) :

L'apprentissage par renforcement est une technique de machine learning dans laquelle un agent apprend à prendre des décisions en interagissant directement avec un environnement. Contrairement aux autres approches, il ne repose pas sur des données étiquetées, mais sur un système de récompenses qui guide l'apprentissage. L'objectif principal est de maximiser la récompense cumulée sur le long terme en apprenant progressivement à partir des conséquences des actions effectuées.

Ce processus repose sur trois éléments fondamentaux : l'état, qui représente la situation actuelle de l'environnement ; l'action, qui correspond à la décision prise par l'agent ; et la

Chapitre II : Machine learning ML

récompense, qui indique la qualité ou l'efficacité de cette action. L'agent apprend ainsi à associer des situations à des actions optimales afin d'améliorer ses performances.

L'apprentissage est guidé par une politique, qui définit le choix des actions en fonction des états, ainsi que par une fonction de valeur qui permet d'estimer les récompenses futures associées à chaque état. Ces estimations sont mises à jour progressivement à partir de l'expérience et prennent souvent en compte un facteur de réduction des récompenses futures.

Une partie importante de ce type d'apprentissage est l'équilibre entre exploration et exploitation. L'exploration permet de tester de nouvelles actions afin de découvrir de meilleures stratégies, tandis que l'exploitation consiste à utiliser les connaissances déjà acquises pour maximiser les récompenses. Des stratégies comme ϵ -greedy sont souvent utilisées pour gérer cet équilibre.

Enfin, plusieurs méthodes sont utilisées dans l'apprentissage par renforcement, notamment la programmation dynamique, les méthodes de Monte Carlo et les méthodes de différence temporelle (Temporal Difference). Cependant, cette approche présente certaines limites, comme la difficulté de représentation des états, la complexité des calculs et la sensibilité au choix du problème et de son environnement [16]. Ce principe est illustré par la figure 11.

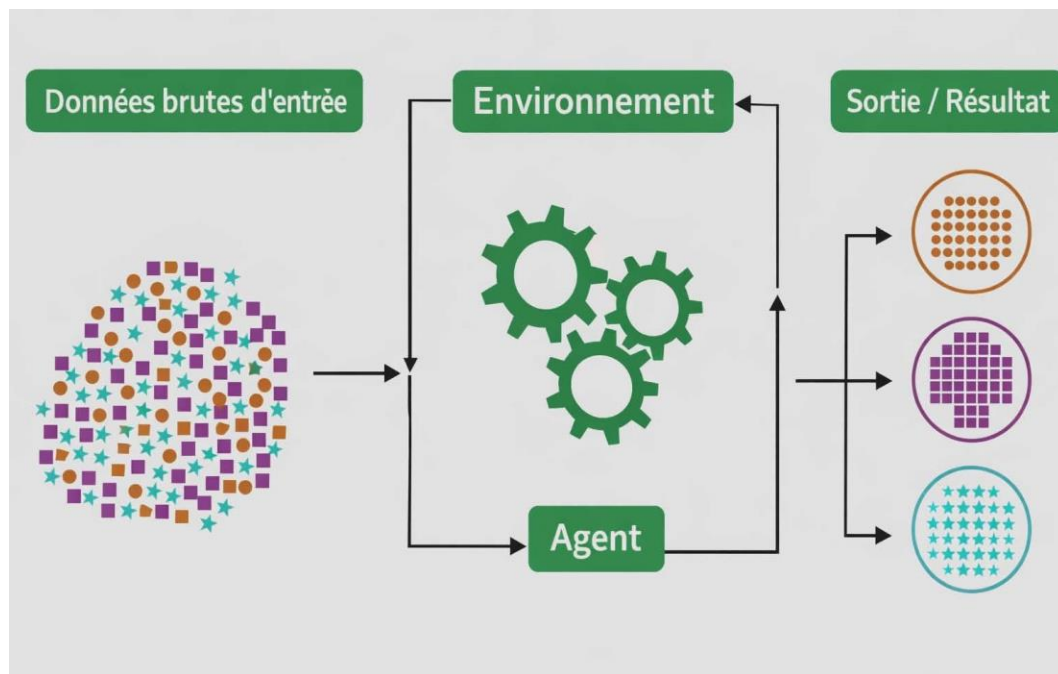


Figure II.11: Apprentissage par renforcement (Reinforcement Learning)

Chapitre II : Machine learning ML

5. Algorithmes de Machine Learning :

5.1 Algorithmes de classification :

Les algorithmes de classification représentent une catégorie importante des méthodes d'apprentissage supervisé en Machine Learning. Ils ont pour objectif d'attribuer une classe ou une étiquette à une donnée à partir de ses caractéristiques. Selon la nature du problème traité, la classification peut être binaire, multi-classes ou multi-label. Plusieurs algorithmes ont été développés dans ce domaine, chacun reposant sur des principes de fonctionnement spécifiques et adaptés à différents types de données et d'applications [18]. La **Figure 12** illustre les principaux algorithmes de classification utilisés en apprentissage automatique, qui seront présentés dans les sections suivantes.

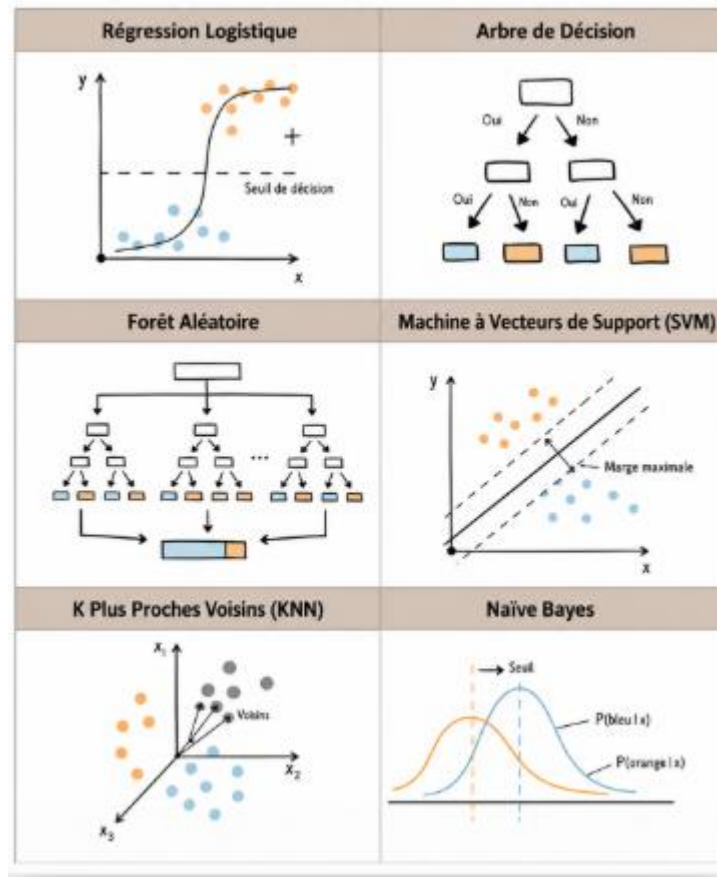


Figure II.12: Des principaux algorithmes de classification en apprentissage automatique

Chapitre II : Machine learning ML

5.1.1 Régression logistique :

La régression logistique est une méthode de classification supervisée qui repose sur un modèle probabiliste permettant d'estimer la probabilité d'appartenance d'un objet à une classe donnée. Elle cherche à déterminer une frontière de décision entre les différentes classes et calcule les probabilités associées à chaque classe en fonction de la distance par rapport à cette frontière. Cette approche est particulièrement adaptée aux problèmes de classification binaire, où le résultat appartient généralement à l'une de deux catégories, comme 0 ou 1 [19]

5.2 Machines à vecteurs de support (SVM) :

Les machines à vecteurs de support (Support Vector Machines, SVM) représentent une technique récente d'apprentissage supervisé en machine learning. Elles sont étroitement liées aux réseaux de neurones perceptron multicouches classiques. Le principe des SVM repose sur la notion de « marge », c'est-à-dire la distance située de part et d'autre d'un hyperplan séparant deux classes de données. L'objectif principal est de maximiser cette marge afin de créer la plus grande distance possible entre l'hyperplan de séparation et les échantillons des deux classes. Cette stratégie permet de réduire une borne supérieure de l'erreur de généralisation, améliorant ainsi la performance du modèle sur de nouvelles données [19]

5.2.1 k plus proches voisins (K-NN)

L'algorithme des k plus proches voisins (k-NN) est particulièrement sensible aux variables non pertinentes en raison de son principe de fonctionnement basé sur le calcul des distances entre les observations. En effet, la présence de caractéristiques inutiles peut fausser ces distances et dégrader significativement la performance du modèle.

Contrairement à certaines méthodes plus complexes, le k-NN ne réalise pas de phase d'apprentissage explicite, mais se base directement sur les données d'entraînement pour effectuer les prédictions. Cela implique que la qualité des résultats dépend fortement de la pertinence des variables utilisées [19]

Chapitre II : Machine learning ML

5.2.2 Les réseaux bayésiens naïfs (Naive Bayes) :

Les réseaux bayésiens naïfs (Naive Bayes) constituent des modèles probabilistes simples appartenant à la famille des réseaux bayésiens. Ils sont représentés sous forme de graphes orientés acycliques comprenant un seul nœud parent, correspondant à la variable non observée (la classe), et plusieurs nœuds enfants représentant les variables observées. Ce modèle repose sur une hypothèse forte d'indépendance conditionnelle entre les attributs, étant donné la classe.

Ainsi, l'approche Naive Bayes s'appuie sur l'estimation des probabilités conditionnelles afin de réaliser la classification. De manière générale, les classificateurs bayésiens présentent des performances parfois inférieures à celles d'algorithmes plus complexes, tels que les réseaux de neurones artificiels. Toutefois, des études comparatives à grande échelle ont montré que le classificateur naïf de Bayes peut, dans certains cas, surpasser des méthodes comme les arbres de décision, l'apprentissage basé sur les instances ou encore l'induction de règles, et ce même en présence de dépendances importantes entre les variables.

Néanmoins, la principale limite de cette méthode réside dans l'hypothèse d'indépendance des attributs. Pour atténuer ce problème, des approches améliorées telles que les estimateurs à dépendance unique moyennée (Averaged One-Dependence Estimators) ont été proposées [14]

5.2.3 Arbres de décision (Decision Tree)

Les arbres de décision (Decision Trees) sont des modèles utilisés pour la classification des instances en les organisant selon les valeurs de leurs caractéristiques. Chaque nœud de l'arbre représente une caractéristique d'une instance à classer, tandis que chaque branche correspond à une valeur possible de cette caractéristique. La classification d'une instance commence à partir du nœud racine et se poursuit en fonction des valeurs de ses attributs jusqu'à atteindre une feuille, qui représente la classe prédite.

L'apprentissage par arbre de décision, largement utilisé en data mining et en apprentissage automatique, repose sur un modèle prédictif qui associe les observations d'un objet à des conclusions concernant sa valeur cible. Ces modèles sont également appelés arbres de classification ou arbres de régression selon le type de problème traité.

Chapitre II : Machine learning ML

En pratique, les classificateurs à base d'arbres de décision utilisent souvent des techniques d'élagage (post-pruning) afin d'améliorer leurs performances et de limiter le surapprentissage. Cette étape consiste à évaluer la performance de l'arbre à l'aide d'un ensemble de validation et à supprimer certains nœuds. Un nœud supprimé est alors remplacé par la classe la plus fréquente parmi les instances d'entraînement qui lui sont associées [14]

5.2.4 Neural Networks :

Les réseaux de neurones (Neural Networks) sont des modèles capables de réaliser plusieurs tâches de régression et/ou de classification, bien que, dans la plupart des cas, un réseau soit conçu pour produire une seule sortie. Dans les problèmes de classification à plusieurs classes, cette sortie peut toutefois être représentée par plusieurs unités de sortie, la phase de post-traitement assurant ensuite la correspondance entre les unités de sortie et les variables finales.

Un réseau de neurones artificiels repose principalement sur trois éléments fondamentaux : les fonctions d'entrée et d'activation des unités, l'architecture du réseau, ainsi que les poids associés aux connexions entre les neurones. Lorsque ces deux premiers éléments sont fixés, le comportement global du réseau dépend essentiellement des valeurs des poids.

Au début de l'apprentissage, ces poids sont initialisés de manière aléatoire. Ensuite, les exemples du jeu de données d'entraînement sont présentés de façon répétée au réseau. Pour chaque instance, les valeurs d'entrée sont propagées dans le réseau afin de produire une sortie, qui est ensuite comparée à la sortie attendue. L'écart entre ces deux résultats permet d'ajuster progressivement les poids dans le sens d'une réduction de l'erreur, afin de rapprocher les sorties du réseau des valeurs désirées. Différents algorithmes peuvent être utilisés pour réaliser cet apprentissage [14]

5.2.5 Le Random Forest (RF)

Le Random Forest est une technique d'apprentissage ensembliste utilisée en machine learning et en data science dans différents domaines. Elle repose sur une méthode de classification basée sur la combinaison de plusieurs arbres de décision afin d'améliorer les performances du modèle global. Cette approche permet d'obtenir une meilleure accuracy et un meilleur contrôle. Ainsi,

Chapitre II : Machine learning ML

un modèle Random Forest composé de plusieurs arbres de décision est généralement plus précis qu'un modèle basé sur un seul arbre de décision.

Pour construire une diversité contrôlée entre les arbres, cette méthode combine le **bootstrap** aggregation (bagging) et la sélection aléatoire des variables. Le Random Forest est également adapté aux problèmes de classification et de régression, et il fonctionne aussi bien avec des données catégorielles que continues [20]

5.3 Algorithmes de régression

Les algorithmes de régression sont des méthodes de Machine Learning utilisées pour prédire des valeurs numériques continues à partir de données d'entrée. Plusieurs techniques de régression ont été développées selon la nature des données et des problèmes traités. La **Figure 13** présente les principaux algorithmes de régression en apprentissage automatique.

Chapitre II : Machine learning ML

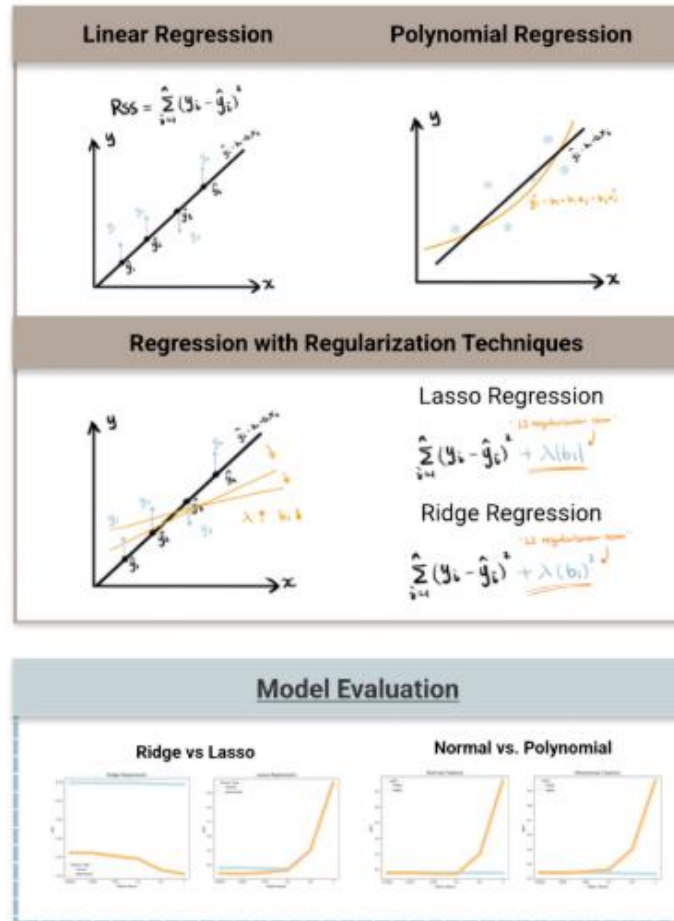


Figure II. 2: Les principaux Algorithmes de régression de classification en apprentissage automatique

5.3.1 Régression linéaire simple et multiple :

La régression linéaire est l'une des techniques les plus utilisées en apprentissage automatique. Elle repose sur l'hypothèse d'une relation linéaire entre la variable dépendante (Y) et une ou plusieurs variables indépendantes (X).

Dans le cas simple, avec une seule variable explicative, le modèle s'écrit : $y = a + b_x + e$ où a représente l'ordonnée à l'origine (intercept), b le coefficient directeur (pente), et e l'erreur du modèle.

La régression linéaire multiple généralise ce modèle à plusieurs variables explicatives :

$$y = a + b_1x_1 + b_2x_2 + b_nx_n \dots \dots + e \quad (2.1)$$

Chapitre II : Machine learning ML

Elle permet donc de prédire la variable cible à partir de plusieurs facteurs simultanément [20]

5.3.2 La régression polynomiale :

est une forme d'analyse de régression dans laquelle la relation entre la variable indépendante x et la variable dépendante y n'est pas linéaire, mais suit un polynôme de degré n en x . L'équation de la régression polynomiale est également dérivée de la régression linéaire (cas de degré 1) et est définie comme suit :

$$y = b_0 + b_1x + b_2x^2 + b_3x^3 + \dots + b_nx^n + e \quad (2.2)$$

Ici, y représente la sortie prédite ou la variable cible, $b_0, b_1, \dots, b_n$ sont les coefficients de régression, x est la variable d'entrée, et e représente l'erreur. En termes simples, lorsque les données ne sont pas distribuées de manière linéaire mais suivent une relation polynomiale de degré n , la régression polynomiale est utilisée pour obtenir la sortie désirée [18]

5.3.3 La LASSO et la Ridge regression

La LASSO et la Ridge regression sont deux techniques de régularisation largement utilisées en apprentissage automatique, surtout lorsqu'il existe un grand nombre de variables explicatives. Elles ont pour objectif principal de réduire le surapprentissage (overfitting) et de limiter la complexité du modèle afin d'améliorer sa capacité de généralisation.

La régression LASSO (Least Absolute Shrinkage and Selection Operator) repose sur la régularisation de type L1. Cette méthode applique une pénalisation basée sur la valeur absolue des coefficients, ce qui a pour effet de réduire certains coefficients à zéro. Ainsi, la LASSO permet non seulement de diminuer l'impact des variables moins importantes, mais aussi de sélectionner automatiquement un sous-ensemble de variables pertinentes pour la prédiction.

De son côté, la régression Ridge utilise la régularisation L2, qui pénalise le carré des coefficients. Contrairement à la LASSO, elle ne met jamais les coefficients à zéro, mais les rend simplement très petits. Cela signifie que la Ridge regression conserve toutes les variables du modèle, tout en réduisant leur influence, ce qui conduit à une solution dense [20]

Chapitre II : Machine learning ML

5.4 Algorithmes de clustering

Les algorithmes de clustering constituent une famille importante de méthodes d'apprentissage non supervisé, permettant de regrouper les données selon leur similarité sans avoir besoin d'étiquettes préalables. Parmi ces méthodes, certaines approches sont basées sur la densité, la distance ou encore la structure des données.

5.4.1 DBSCAN clustering

DBSCAN est un algorithme de clustering basé sur la densité. Il regroupe les points qui sont proches les uns des autres dans des régions de forte densité, séparées par des zones de faible densité. L'algorithme utilise deux paramètres principaux : le rayon de voisinage (Eps) et le nombre minimum de points (MinPts) nécessaires pour former un cluster. Un point est considéré comme appartenant à un cluster s'il possède suffisamment de voisins dans son voisinage défini. Sinon, il est considéré comme un bruit ou un point isolé. DBSCAN est capable de détecter des clusters de formes arbitraires et ne nécessite pas de définir à l'avance le nombre de clusters. Cependant, il peut être moins efficace lorsque la densité des données varie fortement ou lorsqu'il n'y a pas de séparation claire entre les clusters [21]. Ce fonctionnement de DBSCAN est illustré dans la **Figure 15**, qui présente la répartition

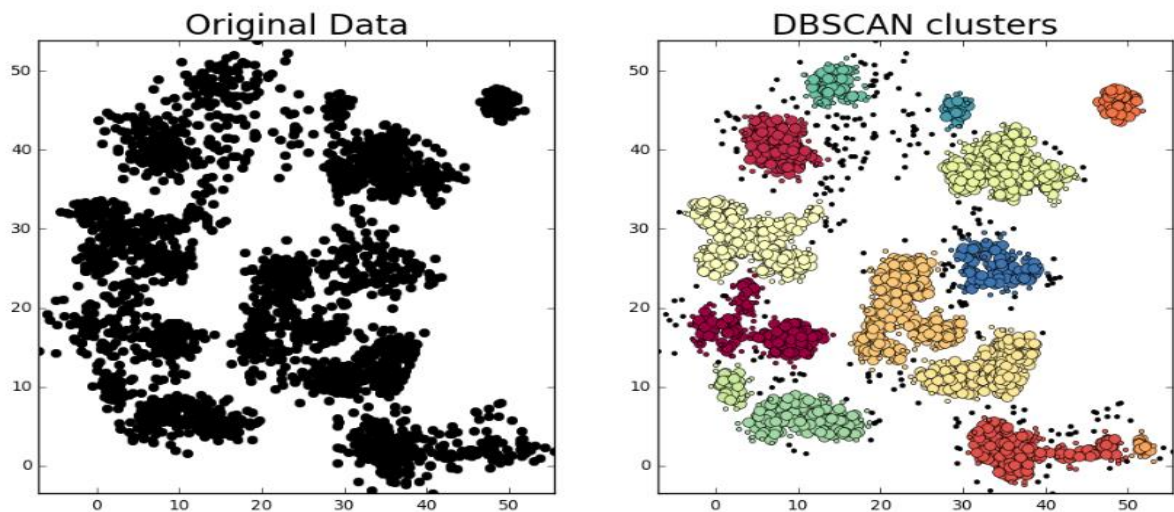


Figure II.14: Visualisation des données et des clusters obtenus par l'algorithme DBSCAN

Chapitre II : Machine learning ML

6.Applications du Machine Learning

À l'ère de la quatrième révolution industrielle (4IR), le *machine learning* s'impose comme une technologie clé dans de nombreux domaines grâce à sa capacité à exploiter les données passées afin de produire des décisions intelligentes. Plusieurs domaines d'application majeurs peuvent être identifiés.

L'analyse prédictive constitue l'un des usages principaux du machine learning. Elle repose sur l'identification de relations entre différentes variables afin d'anticiper des résultats futurs. Cette approche est utilisée, par exemple, pour détecter des fraudes, identifier des criminels ou encore améliorer la gestion des stocks et la logistique dans le commerce électronique. Grâce à des algorithmes tels que les arbres de décision, les machines à vecteurs de support ou les réseaux de neurones, ces systèmes contribuent à améliorer la prise de décision dans divers secteurs comme la finance, la santé ou le marketing.

Dans le domaine de la cybersécurité, le machine learning joue un rôle fondamental en permettant d'analyser de grandes quantités de données afin de détecter des comportements anormaux et des menaces potentielles. Il est ainsi utilisé pour identifier les attaques informatiques, repérer les anomalies ou encore sécuriser les systèmes et les données.

Le développement de l'Internet des objets (IoT) et des villes intelligentes repose également sur le machine learning. Celui-ci permet d'exploiter les données générées par des objets connectés afin d'optimiser les services urbains, tels que la gestion du trafic, de l'énergie ou des infrastructures, et d'améliorer la qualité de vie des citoyens.

Dans le secteur des transports, le machine learning est utilisé pour prédire les conditions de circulation et proposer des solutions adaptées afin de réduire les embouteillages et améliorer l'efficacité des déplacements. Ces systèmes contribuent ainsi à limiter les impacts négatifs liés à la congestion et à la pollution.

En santé, le machine learning permet d'améliorer le diagnostic, le suivi des patients et la prédiction des maladies. Il a notamment été mobilisé dans le contexte de la pandémie de

Chapitre II : Machine learning ML

COVID-19 pour analyser les risques, prévoir la propagation du virus et soutenir les décisions médicales, notamment à travers l'analyse d'images médicales.

Le commerce électronique bénéficie également largement du machine learning, en particulier à travers les systèmes de recommandation. Ceux-ci analysent les comportements des utilisateurs afin de proposer des produits personnalisés, d'optimiser les ventes et d'améliorer l'expérience client.

Le traitement du langage naturel (NLP) constitue un autre domaine important, permettant aux machines de comprendre et d'interpréter le langage humain. Il est utilisé dans des applications telles que les assistants virtuels, les chatbots ou la traduction automatique. L'analyse des sentiments, quant à elle, permet d'identifier les opinions et émotions exprimées dans les textes.

Par ailleurs, le machine learning est largement utilisé dans la reconnaissance d'images, de la parole et des formes, permettant d'identifier des objets, des visages ou des motifs dans des données complexes grâce à différentes techniques d'apprentissage.

Dans le domaine de l'agriculture durable, il contribue à améliorer la productivité et la gestion des ressources en permettant, par exemple, de prédire les rendements, de détecter les maladies des cultures ou d'optimiser l'irrigation.

Enfin, le machine learning est utilisé pour analyser le comportement des utilisateurs et développer des applications mobiles intelligentes. Ces systèmes, dits contextuels, adaptent leurs réponses en fonction des données collectées afin d'offrir des services personnalisés et améliorer l'expérience utilisateur [20]

7.Conclusion

Dans ce chapitre, nous avons présenté les concepts fondamentaux du Machine Learning ainsi que son rôle central dans les systèmes d'intelligence artificielle modernes. Le Machine Learning permet aux systèmes d'apprendre à partir des données et d'améliorer progressivement leurs performances sans intervention explicite, ce qui en fait une approche essentielle dans le traitement de grandes quantités d'informations.

Chapitre II : Machine learning ML

Nous avons également détaillé les principales étapes du pipeline de Machine Learning, depuis la collecte et le prétraitement des données jusqu'à l'entraînement, la validation et l'évaluation des modèles. Ces étapes constituent une chaîne complète indispensable pour garantir la qualité et la fiabilité des systèmes développés.

Par ailleurs, les différents types d'apprentissage ont été étudiés, notamment l'apprentissage supervisé, non supervisé, semi-supervisé et par renforcement. Chacune de ces approches répond à des problématiques spécifiques et permet de traiter différents types de données et d'objectifs.

Enfin, nous avons présenté les principaux algorithmes de Machine Learning en classification, régression et clustering, ainsi que leurs principes de fonctionnement et leurs domaines d'application. Cette étude met en évidence la diversité des méthodes disponibles et leur importance dans la résolution de problèmes complexes dans de nombreux secteurs tels que la santé, la finance, la cybersécurité et les systèmes intelligents.

Chapitre III: étude des méthodes de détection d'anomalie dans les signaux radio

Chapitre III : étude des méthodes de détection d'anomalie dans les signaux radio

Introduction

Ce chapitre présente les fondements théoriques de la détection d'anomalies ainsi que les principales approches utilisées dans le contexte des signaux radio. Il commence par définir les différents types d'anomalies, à savoir les anomalies ponctuelles, contextuelles et collectives, afin de mieux comprendre la nature des comportements anormaux. Ensuite, il introduit le principe de la détection d'anomalies dans les signaux radio et explore les principales catégories de méthodes existantes, allant des techniques traditionnelles aux approches modernes basées sur l'apprentissage automatique et l'apprentissage profond.

Enfin, ce chapitre met en évidence les avantages et les limites de chaque famille de méthodes, ce qui permet de situer les approches récentes dans un contexte plus large et de justifier l'intérêt des techniques intelligentes pour l'analyse des signaux radio complexes.

1 Définition et types d'anomalies

Une anomalie correspond à une observation ou un motif dans un ensemble de données qui s'écarte du comportement général considéré comme normal. Elle représente une information inhabituelle qui ne suit pas les régularités statistiques ou structurelles du reste des données. Dans de nombreux domaines, les anomalies sont particulièrement importantes car elles peuvent indiquer des événements rares mais critiques tels que des pannes, des fraudes, des intrusions ou des dysfonctionnements dans un système.

Dans le domaine de la détection d'anomalies, on distingue généralement trois catégories principales : les anomalies ponctuelles, contextuelles et collectives. Cette classification permet de mieux comprendre la nature des comportements anormaux et d'adapter les méthodes de détection en fonction du type de données et du problème étudié [22]. La **figure1** illustre cette taxonomie des anomalies et résume les trois principales catégories précédemment décrites.

Chapitre III : étude des méthodes de détection d'anomalie dans les signaux radio

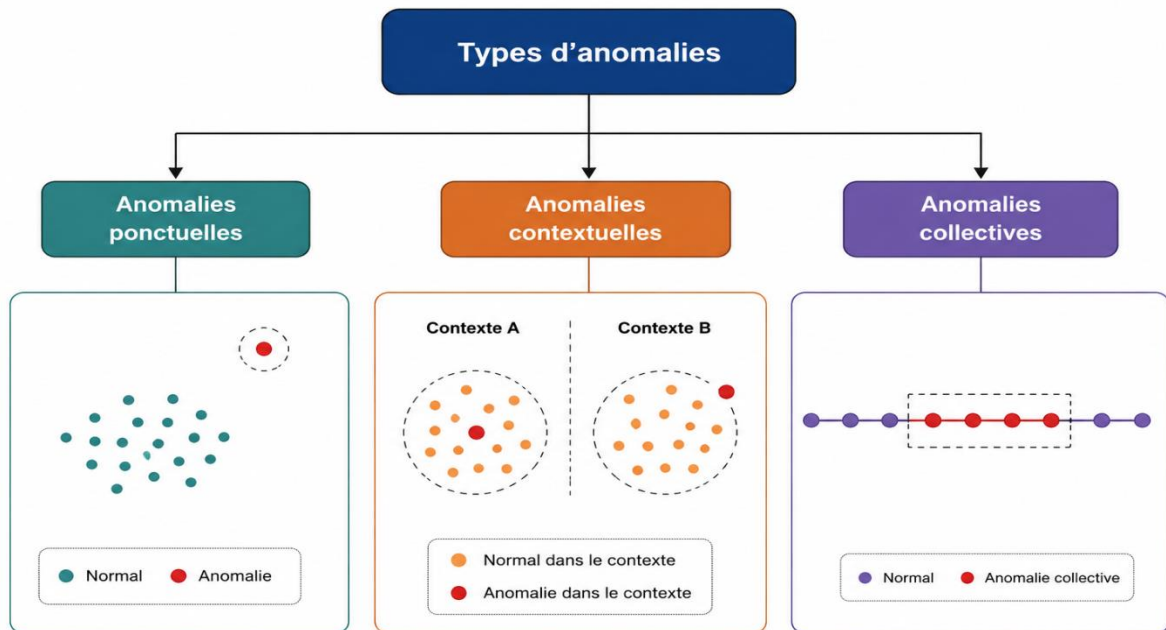


Figure III.1:Taxonomie des anomalies

1.1 Anomalies ponctuelles

Les anomalies ponctuelles correspondent aux cas où une observation individuelle est considérée comme anormale par rapport à l'ensemble des données. Dans ce cas, chaque donnée est analysée indépendamment des autres. Une valeur est jugée anormale si elle est suffisamment éloignée du comportement global du système.

Ce type d'anomalie est le plus simple et le plus étudié. Par exemple, dans des données financières, une transaction d'un montant très élevé par rapport aux habitudes d'un utilisateur peut être considérée comme une anomalie ponctuelle. Dans le cas des signaux numériques, cela peut correspondre à des pics soudains ou des valeurs extrêmes dans un signal I/Q causés par du bruit ou des interférences

Chapitre III : étude des méthodes de détection d'anomalie dans les signaux radio

1.2 Anomalies contextuelles

Les anomalies contextuelles sont des observations qui peuvent être normales dans un certain contexte, mais considérées comme anormales dans un autre. Leur détection repose donc sur la prise en compte du contexte dans lequel les données sont observées.

Chaque donnée est généralement décrite par deux types d'informations : un contexte qui définit la situation d'observation, et des caractéristiques comportementales qui décrivent la valeur elle-même. Une même valeur peut ainsi être normale dans un contexte donné et anormale dans un autre.

Par exemple, une température relativement basse peut être normale en hiver mais anormale en été. Dans les systèmes de communication radio, un signal peut être considéré comme normal dans des conditions de canal propres, mais devenir anormal dans des conditions de forte dégradation ou de faible rapport signal sur bruit. Ce type d'anomalie est particulièrement important dans les environnements dynamiques [23]

1.3 Anomalies collectives

Les anomalies collectives correspondent à des ensembles d'observations qui, prises individuellement, peuvent sembler normales, mais qui deviennent anormales lorsqu'elles sont considérées ensemble. Dans ce cas, l'anomalie est liée à la structure ou à la séquence des données plutôt qu'à une valeur isolée.

Ce type d'anomalie apparaît fréquemment dans les données séquentielles ou temporelles. Par exemple, une suite d'événements dans un système informatique peut être normale prise séparément, mais représenter une activité malveillante lorsqu'elle est observée dans un ordre particulier. De même, dans les signaux biomédicaux ou radio, une anomalie collective peut correspondre à une déformation ou une rupture de comportement sur une fenêtre temporelle complète du signal [23]

Chapitre III : étude des méthodes de détection d'anomalie dans les signaux radio

En Conclusion , La distinction entre anomalies ponctuelles, contextuelles et collectives est essentielle pour comprendre la diversité des comportements anormaux dans les données. Chaque type nécessite des approches de détection adaptées, en particulier dans des domaines complexes comme les communications numériques et les signaux modulés, où les variations peuvent être à la fois individuelles, dépendantes du contexte ou liées à des structures temporelles.

2 La détection d'anomalies dans les signaux radio

La détection d'anomalies dans les signaux radio consiste à identifier des signaux anormaux ou des comportements inhabituels dans l'environnement radiofréquence (RF) afin d'assurer une gestion plus efficace et plus sécurisée des réseaux sans fil. Elle permet de détecter des interférences, des transmissions non conventionnelles ou des activités suspectes pouvant perturber les performances du réseau. Cette détection contribue à une meilleure connaissance de l'environnement radio et aide à optimiser les paramètres du réseau, tels que la sélection des canaux, la puissance de transmission et les technologies d'accès radio utilisées [24]

3 Méthodes de détection d'anomalies radio

Les techniques de détection d'anomalies radio peuvent être globalement classées, selon l'approche utilisée, en trois grandes catégories principales : les méthodes traditionnelles, les méthodes basées sur des algorithmes évolutionnaires, et les méthodes fondées sur l'apprentissage automatique [25] . Une synthèse de ces approches est illustrée dans la Figure 1.

Chapitre III : étude des méthodes de détection d'anomalie dans les signaux radio

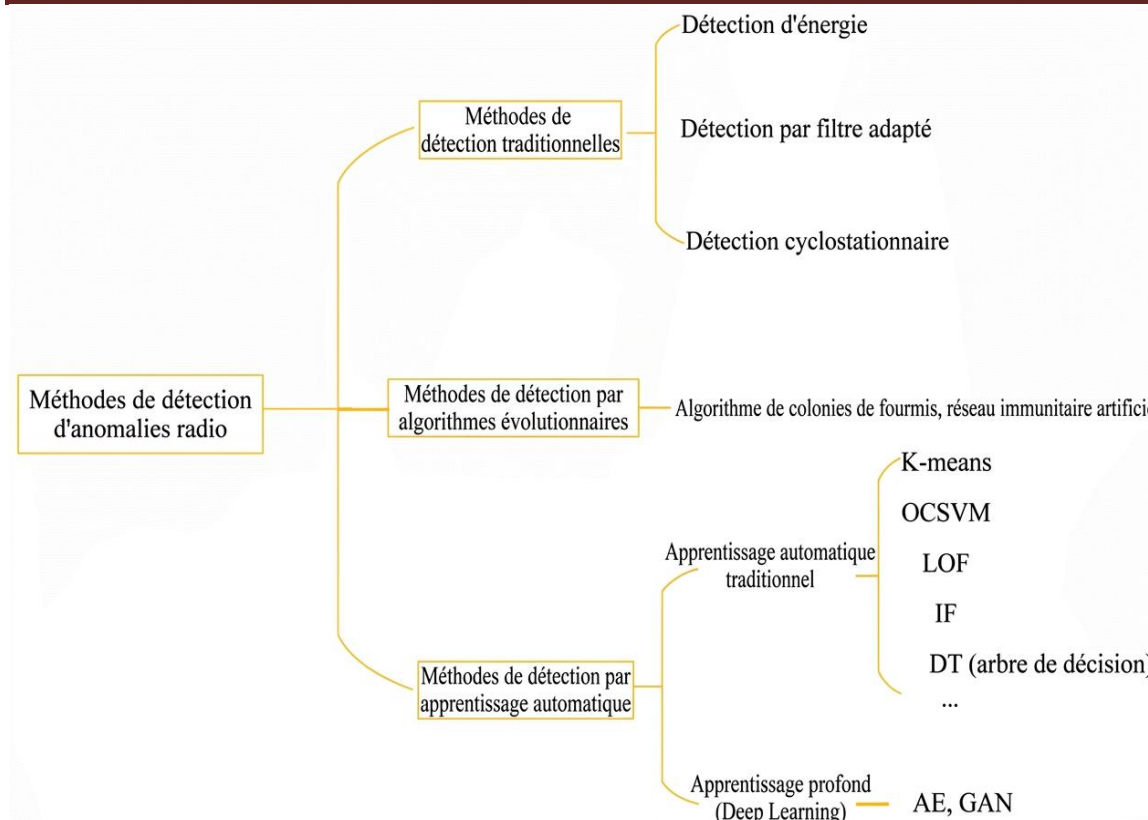


Figure III.2: Taxonomie des méthodes de détection d'anomalies radio

Figure 2 : Classification des principales approches de détection d'anomalies dans les signaux radio, incluant les méthodes traditionnelles (énergie, filtre adapté, cyclostationnarité), les méthodes évolutionnaires (ACO, réseaux immunitaires artificiels) et les approches basées sur l'apprentissage automatique (modèles classiques et apprentissage profond)

3.1 Méthodes de détection traditionnelles

Les méthodes traditionnelles de détection d'anomalies radio regroupent principalement la détection basée sur l'énergie, la détection par filtre adapté (matched filter) et la détection basée sur les caractéristiques de cyclostationnarité.

3.1.1 La détection par énergie

La détection par énergie introduite initialement par Urkowitz, consiste à mesurer l'énergie du signal reçu sur un intervalle de temps donné afin de déterminer la présence ou non d'un signal.

Chapitre III : étude des méthodes de détection d'anomalie dans les signaux radio

Par la suite, des approches dans le domaine fréquentiel ont été proposées afin de détecter non seulement la présence d'interférences, mais aussi d'extraire des caractéristiques telles que la fréquence et l'énergie du signal.

Des améliorations ont été apportées avec des méthodes comme l'algorithme **CME (Consecutive Mean Excision)**, qui calcule de manière itérative une moyenne d'énergie pour établir un seuil dynamique et supprimer les signaux anormaux dépassant ce seuil. Une version améliorée, le **FCME (Forward CME)**, permet de détecter les anomalies même dans des environnements fortement pollués.

3.1.2 La détection par filtre adapté

La **détection par filtre adapté** repose sur la corrélation entre le signal reçu et un signal de référence connu, permettant de détecter la présence du signal utile via des pics en sortie. Cependant, cette méthode nécessite une connaissance préalable complète du signal, ce qui limite son utilisation.

3.1.3 Détection cyclostationnarité

Les méthodes basées sur la **cyclostationnarité** exploitent la périodicité des signaux modulés pour distinguer signal, bruit et interférence en utilisant la fonction de corrélation spectrale. Elles sont robustes en faible SNR mais restent très coûteuses en calcul.

Globalement, ces méthodes traditionnelles sont simples et largement utilisées, mais elles présentent des limites importantes telles que la sensibilité au bruit, la nécessité de réglage de seuil, et une efficacité réduite dans les environnements complexes ou à faible SNR [25]

3.2 Méthodes de détection par les algorithmes évolutionnaires

Les méthodes de détection d'anomalies basées sur les algorithmes évolutionnaires transforment le problème de détection en un **problème d'optimisation**, où l'objectif est de trouver les meilleures solutions (clusters ou chemins) permettant d'identifier les signaux anormaux.

Chapitre III : étude des méthodes de détection d'anomalie dans les signaux radio

Ces méthodes utilisent des mécanismes inspirés de l'évolution naturelle tels que la **sélection**, la **mutation** et le **croisement**, et permettent d'améliorer l'identification des anomalies dans des environnements complexes.

3.2.1 Les algorithmes de colonie de fourmis

Les algorithmes de colonie de fourmis (Ant Colony Optimization – ACO) sont des méthodes d'optimisation inspirées du comportement collectif des fourmis, utilisées pour la détection des signaux anormaux dans les réseaux de communication sans fil. Ils reposent sur l'exploitation de caractéristiques du signal telles que l'intensité, l'état du canal et la fréquence porteuse, afin de guider le processus de recherche. L'efficacité de la détection est améliorée grâce à des mécanismes tels que la coopération multi-colonies, la mise à jour dynamique des phéromones et l'ajustement des poids heuristiques. L'objectif principal de cette approche est de permettre une identification automatique et efficace des signaux anormaux dans des environnements complexes et fortement perturbés.

3.2.2 Les réseaux immunitaires artificiels

Les réseaux immunitaires artificiels (Artificial Immune Systems – AIS) sont des méthodes de détection d'anomalies inspirées du fonctionnement du système immunitaire biologique. Ces approches reposent sur le principe de discrimination entre le “soi” et le “non-soi”, où les signaux normaux représentent le soi et les signaux anormaux correspondent au non-soi. Grâce à ce mécanisme, le système est capable d'identifier et de distinguer les anomalies électromagnétiques par comparaison avec les comportements normaux appris. Ces méthodes sont principalement utilisées pour la surveillance des signaux électromagnétiques et la détection d'anomalies dans des environnements complexes et dynamiques[2].

3.3 Méthode de détection par apprentissage automatique

La détection d'anomalies par apprentissage automatique consiste à utiliser des modèles capables d'apprendre automatiquement les caractéristiques des données afin de distinguer les comportements normaux et d'identifier les déviations comme des anomalies. Le principe repose sur l'apprentissage du comportement normal du système à partir des données disponibles, afin de construire une représentation de référence du trafic ou des observations. Toute donnée qui

Chapitre III : étude des méthodes de détection d'anomalie dans les signaux radio

s'écarter de cette représentation est considérée comme une anomalie. Dans ce cadre, les modèles peuvent être utilisés en apprentissage supervisé, non supervisé ou semi-supervisé selon la disponibilité des données étiquetées. L'objectif principal est de détecter des comportements anormaux dans les données réseau, notamment les intrusions ou attaques inconnues, tout en assurant une surveillance efficace des systèmes et en s'adaptant aux environnements réels où les données d'attaque ne sont pas toujours disponibles [26]

3.3.1 K-means

L'algorithme **K-means** est une méthode d'apprentissage non supervisé utilisée pour regrouper automatiquement un ensemble de données en plusieurs clusters selon leur similarité. Le principe consiste à partitionner les observations en K groupes, de manière à minimiser la distance entre chaque point et le centroïde de son cluster.

Dans le domaine de la détection d'anomalies, K-means est utilisé pour regrouper les comportements similaires et identifier les observations qui s'éloignent fortement des clusters dominants. Ces observations sont considérées comme des anomalies potentielles. Cette méthode est appréciée pour sa simplicité, sa rapidité et sa faible complexité computationnelle, ce qui la rend adaptée aux grands ensembles de données [27]

3.3.1.1 Exemple d'étude 1

Dans ce travail, les auteurs proposent une méthode de détection d'anomalies réseau basée sur l'algorithme K-means. Les données utilisées proviennent des flux réseau (flow records) collectés dans des systèmes de surveillance du trafic. Les auteurs extraient plusieurs caractéristiques statistiques telles que le nombre de paquets, le volume de données échangées et les relations source-destination. Ensuite, K-means est appliqué afin de regrouper automatiquement les comportements similaires du trafic réseau.

Les clusters représentant le comportement normal sont utilisés comme référence, tandis que les observations éloignées des centroïdes sont considérées comme des anomalies. Les résultats montrent que cette approche permet une détection efficace des comportements anormaux tout en conservant une faible complexité de calcul, ce qui facilite son utilisation dans des environnements de surveillance temps réel. Bien que cette étude soit appliquée au trafic réseau,

Chapitre III : étude des méthodes de détection d'anomalie dans les signaux radio

le même principe peut être utilisé dans les signaux radio en exploitant des caractéristiques temporelles et fréquentielles des signaux RF [28]

3.3.1.2 Exemple d'étude 2

Dans ce travail, une méthode de détection d'anomalies basée sur K-means est utilisée dans le cadre de la surveillance des signaux et des données réseau. Les auteurs exploitent des techniques de prétraitement et de clustering afin de séparer automatiquement les comportements normaux des comportements suspects.

L'approche consiste à extraire les caractéristiques principales des données puis à appliquer K-means afin de construire des groupes représentant différents profils de fonctionnement. Les données éloignées des centroïdes des clusters normaux sont identifiées comme anomalies potentielles. L'étude montre que les méthodes de clustering non supervisé comme K-means permettent d'améliorer la détection des comportements anormaux dans des environnements complexes où les données étiquetées sont limitées ou inexistantes. Cette approche peut être adaptée aux signaux radio pour identifier des transmissions inhabituelles, des interférences ou des comportements RF anormaux [29]

3.3.2 LOF (Local Outlier Factor)

Local Outlier Factor est une méthode basée sur l'analyse de la densité locale des données. Elle repose sur la comparaison de la densité d'un point avec celle de ses voisins afin de déterminer son degré d'anormalité. Dans le contexte de la détection d'anomalies radio, les signaux normaux sont généralement regroupés dans des régions de forte densité, tandis que les anomalies apparaissent dans des zones plus dispersées. Ainsi, un point présentant une densité locale significativement plus faible que celle de ses voisins est considéré comme une anomalie. Cette approche est particulièrement efficace pour identifier des anomalies locales dans des environnements non uniformes tels que les signaux radio [30]

1.3.3.1 Exemple d'étude 1

Dans cet article, une approche de détection d'anomalies est proposée en utilisant le Local Outlier Factor (LOF) pour analyser des données issues de systèmes de communication sans fil.

Chapitre III : étude des méthodes de détection d'anomalie dans les signaux radio

Les données exploitées correspondent à des mesures de signaux radio collectés dans des environnements réels caractérisés par du bruit, des interférences et des variations du canal de propagation.

L'approche repose sur le calcul de la densité locale de chaque observation en fonction de ses voisins les plus proches. Chaque échantillon est comparé à son voisinage afin d'évaluer son degré d'isolement dans l'espace des caractéristiques. Les signaux présentant une densité locale nettement inférieure à celle de leurs voisins sont considérés comme des anomalies.

Les résultats expérimentaux montrent que LOF est capable de détecter efficacement les anomalies locales dans des environnements RF complexes où les données ne suivent pas une distribution uniforme. Cette méthode se révèle particulièrement adaptée à l'analyse des signaux radio, notamment pour l'identification des interférences, des perturbations du canal et des comportements de transmission atypiques [31]

1.3.3.2 Exemple d'étude 2

Dans une autre étude récente, le Local Outlier Factor (LOF) est appliqué à la détection d'anomalies dans un système de communication LoRa basé sur des signaux radio réels transformés en images spectrales. L'objectif de cette étude est d'identifier les transmissions falsifiées ou manipulées dans un environnement IoT à grande échelle.

L'approche proposée repose sur la conversion des signaux radio capturés en représentations visuelles (spectrogrammes), permettant ainsi l'extraction de caractéristiques pertinentes pour l'analyse des anomalies. Chaque image représentant un signal est ensuite transformée en vecteur de caractéristiques et analysée par l'algorithme LOF afin d'évaluer sa densité locale par rapport à ses voisins les plus proches.

LOF attribue un score d'anomalie à chaque observation en fonction de son isolement dans l'espace des données. Les signaux présentant une faible densité locale par rapport à leur voisinage sont considérés comme des comportements anormaux, correspondant généralement à des transmissions altérées ou perturbées.

Chapitre III : étude des méthodes de détection d'anomalie dans les signaux radio

Les résultats expérimentaux montrent que LOF atteint une performance élevée avec une précision de 97.78%, surpassant plusieurs méthodes de référence telles que les autoencodeurs, PCA et Isolation Forest. Cette performance confirme l'efficacité de LOF dans des environnements IoT complexes où les données sont bruitées et fortement non linéaires.

Ainsi, cette étude démontre que LOF constitue une solution robuste et efficace pour la détection d'anomalies dans les réseaux LoRa, en particulier lorsqu'il est appliqué à des données issues de signaux radio transformés en images [32]

3.3.3 One-Class SVM

Le One-Class SVM (One-Class Support Vector Machine) est un algorithme d'apprentissage automatique non supervisé utilisé principalement pour la détection d'anomalies et de nouveautés. Son principe consiste à apprendre uniquement les caractéristiques des données normales afin de construire une frontière séparant les observations normales des observations anormales.

Contrairement aux méthodes de classification classiques, le One-Class SVM ne nécessite qu'une seule classe pendant la phase d'apprentissage. L'algorithme cherche à créer un hyperplan ou une frontière englobant la majorité des données normales dans un espace de caractéristiques. Toute nouvelle observation située à l'extérieur de cette frontière est alors considérée comme une anomalie.

Le One-Class SVM utilise généralement des fonctions noyaux, notamment le noyau RBF, afin de traiter des données non linéaires et complexes. Cette méthode est particulièrement adaptée aux applications de détection d'interférences RF, de cybersécurité, de surveillance réseau et d'analyse de signaux radio, où les données anormales sont rares ou difficiles à étiqueter

3.3.4 Exemple d'étude 1

Dans cette étude, les auteurs proposent l'utilisation du One-Class SVM pour la détection d'interférences RF dans des systèmes de communication sans fil. L'approche repose sur

Chapitre III : étude des méthodes de détection d'anomalie dans les signaux radio

l'apprentissage non supervisé à une seule classe afin de distinguer les signaux radio normaux des signaux perturbés par des interférences.

Les données utilisées sont constituées d'images de scalogrammes représentant différents types de signaux RF, notamment les signaux normaux ainsi que plusieurs formes d'interférences radio. Avant l'apprentissage, les données subissent plusieurs étapes de prétraitement comprenant le redimensionnement des images, la transformation en tableaux numériques ainsi que la normalisation des caractéristiques.

L'algorithme One-Class SVM construit une frontière autour des données normales en utilisant une fonction noyau de type RBF. Les nouvelles observations situées à l'intérieur de cette frontière sont considérées comme normales, tandis que celles situées à l'extérieur sont identifiées comme des anomalies ou des interférences RF.

L'étude montre également que le choix des méthodes de normalisation joue un rôle important dans l'amélioration des performances du modèle, notamment dans les environnements RF bruités et complexes. Cette approche est particulièrement adaptée à la détection d'interférences radio dans les systèmes de surveillance spectrale et les réseaux de communication sans fil [33]

3.3.5 Isolation Forest

L'algorithme Isolation Forest (IForest) est une méthode d'apprentissage non supervisé utilisée pour la détection d'anomalies. Contrairement aux méthodes basées sur la distance ou la densité, IForest repose sur un principe d'isolation des observations. L'idée principale est que les anomalies sont plus faciles à isoler que les observations normales, car elles sont rares et différentes du reste des données.

Le fonctionnement de l'algorithme consiste à construire un ensemble d'arbres de décision aléatoires appelés Isolation Trees. Chaque arbre est construit en sélectionnant aléatoirement une caractéristique ainsi qu'une valeur de séparation, afin de partitionner récursivement les données. Les observations qui nécessitent peu de divisions pour être isolées sont considérées comme des anomalies, tandis que celles nécessitant de nombreuses divisions sont considérées comme normales.

Chapitre III : étude des méthodes de détection d'anomalie dans les signaux radio

Dans le domaine de la détection d'anomalies, IForest est largement utilisé pour identifier les observations rares ou inhabituelles dans des ensembles de données volumineux. Il est particulièrement efficace pour les données de grande dimension grâce à sa faible complexité computationnelle et sa capacité à travailler sur des sous-échantillons. Cette méthode est appréciée pour sa rapidité, sa robustesse et sa scalabilité, ce qui la rend adaptée aux applications de grande taille telles que la cybersécurité, la détection de fraude et l'analyse de signaux [34].

3.3.5.1 Exemple d'étude 1

Dans ce travail, les auteurs proposent une approche basée sur l'algorithme Isolation Forest pour la détection d'anomalies dans des ensembles de données volumineux et complexes. L'objectif principal est d'identifier les observations anormales en exploitant le principe d'isolation progressive des points de données.

Les données utilisées dans cette étude proviennent de différents domaines tels que les transactions financières, les systèmes de surveillance réseau et les données industrielles. Les auteurs appliquent IForest en construisant plusieurs Isolation Trees à partir de sous-échantillons aléatoires des données. Chaque observation est ensuite évaluée selon sa longueur de chemin moyenne dans les arbres.

Les résultats montrent que les anomalies présentent des longueurs de chemin plus courtes, ce qui permet leur détection efficace. L'étude conclut que l'Isolation Forest offre de bonnes performances en termes de précision et de rapidité, notamment pour les grands ensembles de données. Cependant, les performances peuvent varier en fonction du choix des hyperparamètres tels que le nombre d'arbres et la taille des sous-échantillons. Bien que cette étude soit appliquée à des données générales, le même principe peut être utilisé pour la détection d'anomalies dans les signaux RF en exploitant les caractéristiques temporelles et fréquentielles des signaux [34].

3.4 Apprentissage profond

Dans le cadre de la détection d'anomalies dans les signaux radio, les approches modernes basées sur l'apprentissage profond non supervisé exploitent principalement des représentations

Chapitre III : étude des méthodes de détection d'anomalie dans les signaux radio

appprises directement à partir des données I/Q ou de leurs transformations spectrales. Selon l'étude de synthèse, plusieurs familles de méthodes sont utilisées pour modéliser le comportement normal des signaux et détecter toute déviation considérée comme anomalie. Les méthodes de type reconstruction, notamment les autoencodeurs (AE), permettent d'apprendre une représentation compacte des signaux dans un espace latent et de reconstruire l'entrée. Une anomalie est alors identifiée lorsque l'erreur de reconstruction est élevée, indiquant un écart par rapport aux distributions apprises. Les variantes comme les Variational Autoencoders (VAE) améliorent cette approche en imposant un espace latent continu et structuré, facilitant la génération de nouveaux échantillons et la modélisation de distributions complexes de signaux radio.

En parallèle, les méthodes génératives telles que les Generative Adversarial Networks (GAN) sont utilisées pour apprendre la distribution sous-jacente des signaux RF et produire des données synthétiques réalistes, utiles notamment pour l'augmentation de données et la robustesse des modèles en environnements contraints. Les approches prédictives, quant à elles, apprennent des représentations discriminantes sans étiquettes en exploitant les relations entre échantillons ou la prédiction de transformations et de comportements futurs des signaux, ce qui permet de capturer des régularités temporelles et spectrales. Enfin, ces méthodes reposent souvent sur des architectures profondes telles que les CNN pour extraire les caractéristiques locales des signaux I/Q, ou des modèles séquentiels comme les LSTM et les Transformers pour capturer les dépendances temporelles. L'ensemble de ces techniques est généralement évalué sur des datasets de référence comme RadioML, dans des conditions variées de SNR et de bruit (AWGN, Rayleigh fading), ce qui permet de valider leur efficacité pour la détection d'anomalies et la reconnaissance de comportements radio anormaux dans des environnements réels ou simulés [35]

3.4.1 Autoencodeur

Autoencoder est un réseau de neurones artificiels conçu pour apprendre une représentation compressée des données. Il est composé de deux parties principales : un encodeur, qui réduit la dimension des données d'entrée, et un décodeur, qui reconstruit les données originales à partir de cette représentation compressée. Dans le cadre de la détection d'anomalies, le modèle est

Chapitre III : étude des méthodes de détection d'anomalie dans les signaux radio

entraîné uniquement sur des données normales afin d'apprendre leur structure. Lorsqu'un signal anormal est introduit, le réseau n'est pas capable de le reconstruire correctement, ce qui entraîne une erreur de reconstruction élevée. Cette erreur est ensuite utilisée comme indicateur d'anomalie dans les signaux radio .

3.4.1.1 Exemple d'étude

L'étude propose une méthode de détection des signaux radio anormaux basée sur un **Adversarial Autoencoder (AAE)** combiné à un réseau CNN. Le modèle apprend de manière non supervisée les caractéristiques des signaux normaux à partir de représentations temps-fréquence. Ensuite, il détecte les anomalies en utilisant l'**erreur de reconstruction** : les signaux normaux sont bien reconstruits, tandis que les signaux anormaux présentent une forte erreur.

Pour améliorer la qualité des représentations latentes, un mécanisme adversarial est ajouté afin de contraindre l'espace latent à suivre une distribution normale. Les expériences réalisées sur des données radio réelles montrent une **précision de détection d'environ 95,5 %**, supérieure aux méthodes classiques comme Isolation Forest ou One-Class SVM[11].

Conclusion

En conclusion, ce chapitre a présenté les bases théoriques de la détection d'anomalies ainsi que les principales approches utilisées dans le contexte des signaux radio. Les différents types d'anomalies, à savoir ponctuelles, contextuelles et collectives, permettent de mieux caractériser la nature des comportements anormaux dans les données.

Chapitre III : étude des méthodes de détection d'anomalie dans les signaux radio

Les méthodes de détection ont ensuite été classées en trois grandes catégories : les approches traditionnelles, les méthodes basées sur des algorithmes évolutionnaires et les techniques d'apprentissage automatique et profond. Il ressort de cette analyse que les méthodes classiques, bien que simples et efficaces dans certains cas, présentent des limites importantes face à la complexité et à la variabilité des environnements radio modernes.

À l'inverse, les approches basées sur l'apprentissage automatique et l'apprentissage profond offrent une meilleure capacité d'adaptation et une performance accrue dans des environnements bruités et non linéaires. Ces méthodes constituent ainsi une solution prometteuse pour la détection d'anomalies dans les systèmes de communication sans fil, notamment lorsque les données étiquetées sont rares ou inexistantes.

Ce chapitre constitue donc une base théorique essentielle pour la suite du travail, qui se concentrera sur la mise en œuvre et l'évaluation des modèles de détection d'anomalies appliqués aux signaux radio.

Chapitre IV : Expérimentaux et discussion des modèles de détection d'anomalie

Chapitre IV : Expérimentaux et discussion des modèles de détection d'anomalie

Introduction

Ce chapitre présente l'ensemble du cadre expérimental ainsi que la méthodologie adoptée pour la détection des anomalies dans les signaux radio issus du dataset RadioML2016.10a. Il décrit d'abord l'environnement de travail utilisé, notamment la plateforme Google Colaboratory, ainsi que les caractéristiques principales du jeu de données exploité, composé de signaux IQ simulés sous différentes modulations et niveaux de SNR.

Ensuite, la notion d'anomalie est définie dans le contexte de cette étude, en se basant sur le rapport signal sur bruit (SNR), permettant de distinguer les signaux normaux des signaux dégradés. Le chapitre détaille également les différentes étapes du pipeline de traitement, allant du prétraitement des données (normalisation, réduction de dimension, extraction de caractéristiques) jusqu'à l'application des modèles de détection d'anomalies.

Enfin, les modèles étudiés, à savoir l'Autoencodeur, l'Isolation Forest et l'OCSVM, sont présentés ainsi que les méthodes d'évaluation des performances utilisées, incluant la matrice de confusion, la courbe ROC et diverses métriques de classification. Le chapitre se termine par une analyse comparative des résultats expérimentaux obtenus et une discussion sur les performances des approches proposées.

1 Environnement expérimental

Les expériences ont été réalisées sur la plateforme Google Colaboratory permettant l'exécution de calculs intensifs dans un environnement cloud.

2 Le RadioML2016.10a

Le dataset **RadioML2016.10a** est une base de données public largement utilisé dans le domaine de la classification automatique de modulation (AMC). Il contient des signaux radio simulés en bande de base complexe, représentés par leurs composantes en phase (I) et en quadrature (Q). Ce dataset comprend 11 types de modulations, incluant des modulations numériques (BPSK,

Chapitre IV : Expérimentaux et discussion des modèles de détection d'anomalie

QPSK, 8PSK, 16QAM, 64QAM, GFSK, CPFSK, PAM4) ainsi que des modulations analogiques (WBFM, AM-DSB, AM-SSB) Comme illustré dans la **Figure 1**. Chaque échantillon de signal est de dimension 2×128 et est généré sous différentes conditions de rapport signal sur bruit (SNR) variant de -20 dB à $+18$ dB avec un pas de 2 dB. Le dataset contient environ 220 000 échantillons au total, avec 1000 signaux par type de modulation et par niveau de SNR. Il est principalement utilisé pour évaluer les performances des modèles de deep learning en classification de modulation et pour tester leur robustesse face aux variations du bruit et des conditions de propagation réalistes des canaux de communication sans fil [36]

Chapitre IV : Expérimentaux et discussion des modèles de détection d'anomalie

RadioML2016.10a

Ce data set contient des signaux radio au format IQ (In-phase et Quadrature), regroupés en 11 modulations : 8 numériques et 3 analogiques.

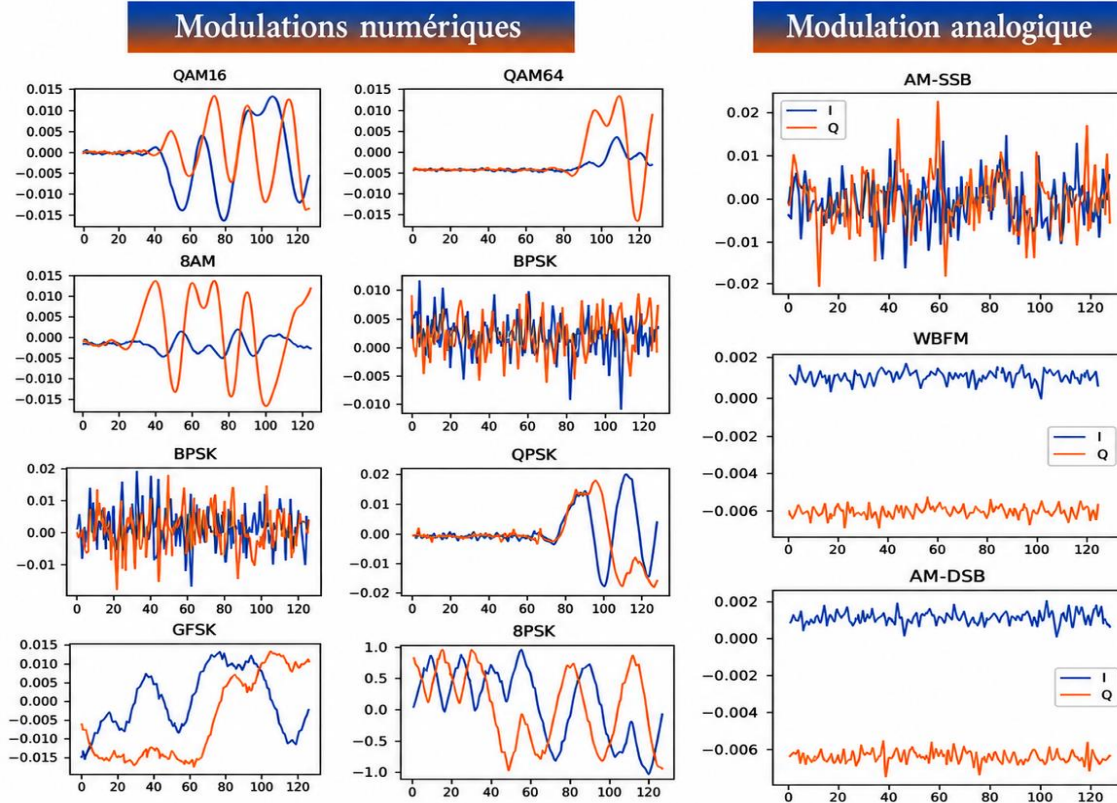


Figure IV.1:Présentation du dataset RadioML 2016.10a et des signaux IQ pour les différentes modulations

3 Définition des anomalies

Dans le cadre de ce travail, la notion d'anomalie est définie à partir du rapport signal sur bruit (SNR), qui constitue un indicateur essentiel de la qualité des signaux radio.

Ainsi, les observations sont classées comme suit :

- **Signaux normaux** : les signaux présentant un SNR supérieur ou égal à 0 dB

Chapitre IV : Expérimentaux et discussion des modèles de détection d'anomalie

-
- **Signaux anormaux** : les signaux présentant un SNR strictement inférieur à 0 dB

Cette définition repose sur le fait que lorsque le SNR devient négatif, le niveau de bruit dépasse celui du signal utile, ce qui entraîne une forte dégradation de l'information transmise. Dans ce cas, le signal devient difficilement exploitable et s'écarte du comportement attendu des signaux de bonne qualité, ce qui justifie son étiquetage en tant qu'anomalie dans le cadre de la détection étudiée

4 Méthodologie proposée

La méthodologie adoptée dans ce travail repose sur une chaîne de traitement structurée permettant la détection des anomalies dans les signaux radio. Le processus global est décrit comme suit :

Chapitre IV : Expérimentaux et discussion des modèles de détection d'anomalie

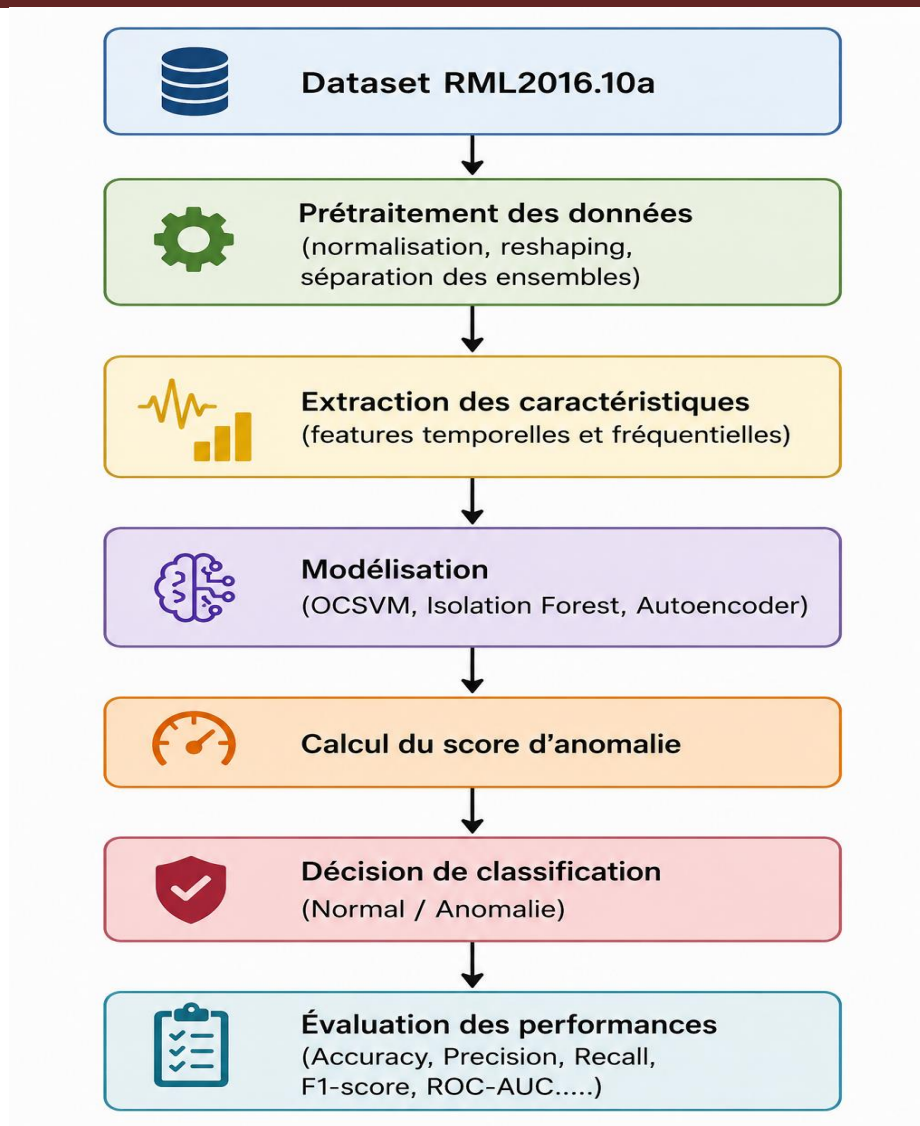


Figure IV.2:Schéma général de la méthodologie proposée pour la détection des anomalies dans les signaux radio

4.1 Prétraitement des données

Le prétraitement des données constitue une étape essentielle avant l'apprentissage des modèles de détection d'anomalies. Il permet de préparer les signaux radio du dataset RadioML2016.10a sous une forme adaptée aux algorithmes de machine learning et de deep learning.

Chapitre IV : Expérimentaux et discussion des modèles de détection d'anomalie

4.2 Chargement des données

Le dataset RadioML2016.10a a été importé depuis Google Drive dans l'environnement Google Colaboratory. Les signaux IQ ainsi que les informations associées aux modulations et aux niveaux de SNR ont été extraits afin de constituer la base de données utilisée dans cette étude.

4.3 Étiquetage des anomalies

Conformément à la définition adoptée dans ce travail, les échantillons présentant un SNR supérieur ou égal à 0 dB ont été considérés comme des signaux normaux, tandis que ceux présentant un SNR inférieur à 0 dB ont été considérés comme des anomalies. Cette étape a permis de construire les classes nécessaires à l'évaluation des modèles.

Afin de garantir une évaluation rigoureuse et une comparaison équitable entre les trois modèles étudiés (Autoencodeur, Isolation Forest et One-Class SVM), un protocole de partitionnement identique a été adopté. Le jeu d'entraînement est constitué exclusivement de 80 % des signaux normaux ($\text{SNR} \geq 0$ dB), soit 88 000 échantillons, permettant aux modèles d'apprendre uniquement le comportement des signaux sains. Le jeu de test, totalement indépendant, comprend les 20 % restants des signaux normaux (22 000 échantillons) ainsi que l'ensemble des 110 000 signaux anormaux ($\text{SNR} < 0$ dB), soit un total de 132 000 échantillons. Cette stratégie garantit l'absence de fuite de données (*data leakage*) entre les phases d'apprentissage et d'évaluation, tout en permettant de mesurer de manière objective la capacité de généralisation de chaque modèle sur des données inédites.

4.4 Mise en forme des données

Les signaux du dataset sont initialement représentés sous la forme de matrices de dimension 2×128 correspondant aux composantes I et Q. Afin de faciliter leur traitement par certains algorithmes, ces matrices ont été transformées en vecteurs unidimensionnels.

Chapitre IV : Expérimentaux et discussion des modèles de détection d'anomalie

4.5 Normalisation

Une normalisation des données a été effectuée à l'aide de la méthode *RobustScaler*. Cette approche de mise à l'échelle permet de limiter l'impact des valeurs aberrantes présentes dans les données tout en préservant les informations essentielles. En réduisant la sensibilité aux outliers, cette technique améliore la stabilité des algorithmes d'apprentissage automatique et favorise une meilleure convergence des modèles lors de la phase d'entraînement [37]

4.6 Réduction de dimension

Afin de réduire la complexité des données et d'améliorer l'efficacité du traitement, une Analyse en Composantes Principales (PCA) a été appliquée comme technique de réduction de dimension. Cette méthode projette les données originales dans un nouvel espace de dimension plus faible, constitué de composantes principales orthogonales représentant la plus grande part de la variance des données. En conservant les caractéristiques les plus informatives et en éliminant les redondances, la PCA permet de diminuer le coût computationnel, d'accélérer le temps d'exécution des algorithmes et d'améliorer les performances de détection des anomalies [38]

4.7 Modèles de détection d'anomalies

- One-Class Support Vector Machine (OCSVM)
- Isolation Forest
- Autoencodeur

4.8 Calcul du score d'anomalie

Chaque modèle produit un score permettant d'évaluer le degré d'anomalie d'une observation. Pour l'OCSVM et l'Isolation Forest, ce score est obtenu à partir de la fonction de décision du modèle. Dans le cas de l'autoencodeur, il correspond à l'erreur de reconstruction calculée entre le signal original et le signal reconstruit.

Chapitre IV : Expérimentaux et discussion des modèles de détection d'anomalie

4.9 Décision de classification

Une étape de seuillage est ensuite appliquée afin de transformer les scores d'anomalie en prédictions binaires. Les observations dont le score dépasse le seuil défini sont classées comme anomalies, tandis que les autres sont considérées comme normales.

4.10 Évaluation des performances des modèles de Machine Learning

L'évaluation des performances des modèles de Machine Learning constitue une étape fondamentale dans le développement et la validation des systèmes d'intelligence artificielle. Elle permet de mesurer la capacité d'un modèle à effectuer des prédictions correctes, à généraliser sur des données non observées et à fournir des résultats fiables dans des conditions réelles. Dans les tâches de classification, cette évaluation repose principalement sur l'analyse des résultats issus de la matrice de confusion et sur le calcul de plusieurs métriques complémentaires permettant une analyse détaillée des performances du modèle.

4.10.1 Matrice de confusion

La matrice de confusion est une représentation synthétique des résultats d'un modèle de classification. Elle compare les prédictions du modèle aux valeurs réelles et se compose de quatre éléments principaux :

		Valeurs prédites	
		Positifs	Négatifs
Valeurs attendues	Positifs	Vrais Positifs (VP)	Faux Négatifs (FN)
	Négatifs	Faux Positifs (FP)	Vrais Négatifs (VN)

Figure IV.3:Matrice de confusion utilisée pour l'évaluation des résultats de classification

Chapitre IV : Expérimentaux et discussion des modèles de détection d'anomalie

Comme l'illustre la Figure 3, les performances d'un modèle de classification peuvent être analysées à partir de quatre catégories de prédictions : les vrais positifs (TP), les vrais négatifs (TN), les faux positifs (FP) et les faux négatifs (FN).

- **TP (True Positives)** : cas positifs correctement prédits
- **TN (True Negatives)** : cas négatifs correctement prédits
- **FP (False Positives)** : cas négatifs classés à tort comme positifs
- **FN (False Negatives)** : cas positifs classés à tort comme négatifs

Ces quatre composantes constituent la base du calcul des différentes métriques d'évaluation.

4.10.2 La sensibilité (Sensitivity)

La sensibilité, également appelée *Recall* ou taux de vrais positifs (*True Positive Rate*), mesure la capacité du modèle à détecter correctement les cas positifs parmi l'ensemble des cas réellement positifs.

$$\text{Sensibilité} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \quad (4.1)$$

4.10.3 La spécificité (Specificity)

La spécificité représente la capacité du modèle à identifier correctement les cas négatifs parmi l'ensemble des cas réellement négatifs.

$$\text{Spécificité} = \frac{\text{TN}}{(\text{TN} + \text{FP})} \quad (4.2)$$

4.10.4 Le taux de faux positifs (False Positive Rate – FPR)

représente la proportion des échantillons négatifs incorrectement classés comme positifs. Cette métrique est également appelée taux de fausses alarmes.

$$\text{FPR} = \frac{\text{FP}}{(\text{TN} + \text{FP})} \quad (4.3)$$

Chapitre IV : Expérimentaux et discussion des modèles de détection d'anomalie

4.10.5 Le taux de faux négatifs (False Negative Rate – FNR)

Le taux de faux négatifs représente la proportion de cas positifs non détectés par le modèle.

$$\text{FNR} = \frac{\text{FN}}{(\text{TP} + \text{FN})} \quad (4.4)$$

4.10.6 La précision (Precision)

La précision mesure la proportion de prédictions positives qui sont réellement correctes.

$$\text{Précision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \quad (4.5)$$

4.10.7 La valeur prédictive négative (Negative Predictive Value – NPV)

La valeur prédictive négative correspond à la proportion de prédictions négatives qui sont réellement correctes.

$$\text{NPV} = \frac{\text{TN}}{(\text{TN} + \text{FN})} \quad (4.6)$$

4.10.8 L'exactitude (Accuracy)

L'exactitude mesure la proportion globale de prédictions correctement classées parmi toutes les observations.

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \quad (4.7)$$

4.10.9 Le F1-score

Le F1-score est une mesure combinant la précision et le rappel sous forme de moyenne harmonique. Il est particulièrement utile dans les problèmes de classification déséquilibrée.

$$\text{F1 - score} = \frac{2\text{TP}}{(2\text{TP} + \text{FP} + \text{FN})} \quad (4.8)$$

4.10.10 Courbe ROC et AUC

La courbe ROC (*Receiver Operating Characteristic*) représente la relation entre le taux de vrais positifs (TPR) et le taux de faux positifs (FPR) pour différents seuils de décision.

Chapitre IV : Expérimentaux et discussion des modèles de détection d'anomalie

L'AUC (*Area Under the Curve*) correspond à l'aire sous la courbe ROC et permet de mesurer la capacité globale du modèle à distinguer entre les classes positives et négatives, indépendamment du seuil. Une valeur proche de 1 indique une excellente performance, tandis qu'une valeur proche de 0.5 correspond à une performance aléatoire [39]

5 Résultats expérimentaux et discussion

5.1 Autoencodeur

Le modèle **Autoencodeur** est une méthode d'apprentissage profond (Deep Learning) utilisée pour la détection d'anomalies à partir des signaux IQ du dataset **RadioML2016.10a**. Son fonctionnement repose sur un apprentissage exclusivement réalisé à partir des signaux normaux ($\text{SNR} \geq 0 \text{ dB}$) afin d'apprendre le comportement normal du système. Les principales étapes du traitement sont les suivantes :

- **Chargement des données** : récupération des **220 000** signaux IQ du dataset RadioML2016.10a.
- **Prétraitement** : réorganisation des signaux sous la forme (128, 2) puis transformation (Flatten) en vecteurs de 256 caractéristiques.
- **Étiquetage des anomalies** : les signaux ayant un $\text{SNR} < 0 \text{ dB}$ sont considérés comme des anomalies, tandis que ceux ayant un $\text{SNR} \geq 0 \text{ dB}$ sont considérés comme des signaux normaux.
- **Séparation des données** : les signaux normaux sont divisés en 80 % pour l'apprentissage et 20 % pour le test, tandis que toutes les anomalies sont utilisées uniquement pour l'évaluation du modèle.
- **Normalisation** : les données sont standardisées à l'aide de StandardScaler afin d'améliorer la stabilité de l'apprentissage et la vitesse de convergence.
- **Ajout de bruit (Denoising)** : un faible bruit gaussien est ajouté aux données d'apprentissage pour améliorer la robustesse et la capacité de généralisation de l'Autoencodeur.
- **Construction de l'Autoencodeur** : le réseau est constitué d'un encodeur qui compresse progressivement les données selon l'architecture $256 \rightarrow 128 \rightarrow 32$, suivi d'un décodeur qui

Chapitre IV : Expérimentaux et discussion des modèles de détection d'anomalie

reconstruit les signaux selon l'architecture $32 \rightarrow 128 \rightarrow 256$. Des couches de Batch Normalization sont intégrées afin de stabiliser l'apprentissage.

- **Apprentissage du modèle** : l'entraînement est réalisé avec l'optimiseur Adam (learning rate $= 10^{-4}$) en minimisant l'erreur absolue moyenne (MAE) entre les signaux d'entrée et leurs reconstructions. Un mécanisme Early Stopping est utilisé afin d'éviter le surapprentissage.
- **Calcul de l'erreur de reconstruction** : après l'apprentissage, chaque signal du jeu de test est reconstruit et son erreur quadratique moyenne (MSE) est calculée.
- **Détermination du seuil optimal** : un seuil de décision est déterminé automatiquement à partir du score F1 maximal obtenu sur la courbe Precision–Recall.
- **Détection des anomalies** : les signaux présentant une erreur de reconstruction supérieure au seuil sont classés comme anomalies, tandis que les autres sont considérés comme normaux.
- **Évaluation des performances** : les résultats sont analysés à l'aide de la matrice de confusion, de la courbe ROC, du score ROC AUC, ainsi que des métriques de précision, rappel, score F1 et exactitude (Accuracy).

Les performances obtenues (**ROC AUC = 0,919** et **Accuracy ≈ 94 %**) montrent que l'Autoencodeur est capable d'apprendre efficacement les caractéristiques des signaux IQ normaux. Grâce à son mécanisme de compression et de reconstruction, il reproduit fidèlement les signaux utilisés lors de l'apprentissage, tandis que les signaux anormaux, caractérisés par un **SNR < 0 dB**, génèrent des erreurs de reconstruction plus élevées. Cette différence permet une séparation efficace entre les données normales et les anomalies.

Chapitre IV : Expérimentaux et discussion des modèles de détection d'anomalie

5.1.1 Matrice de confusion

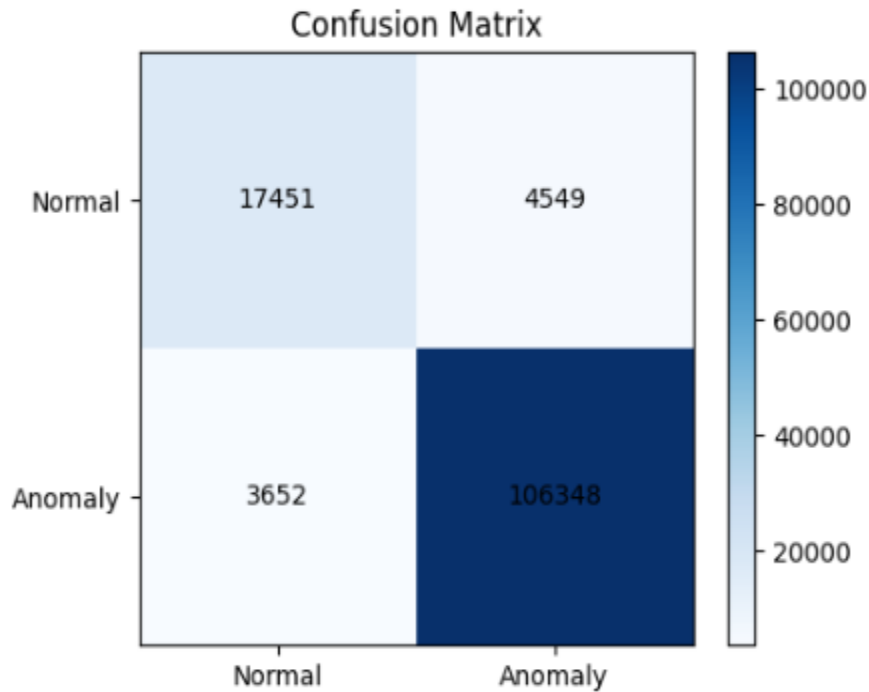


Figure IV.4: Matrice de confusion du modèle Autoencodeur pour la détection d'anomalies dans les signaux radio

À partir de la matrice de confusion fournie, le modèle Autoencodeur présente les résultats suivants : 17 451 vrais négatifs (TN), 4 549 faux positifs (FP), 3 652 faux négatifs (FN) et 106 348 vrais positifs (TP). Ces valeurs indiquent que le modèle parvient à bien distinguer les signaux normaux des anomalies, avec une proportion élevée de bonnes classifications dans les deux classes.

Plus précisément, le nombre important de vrais positifs (106 348) et de vrais négatifs (17 451) montre une bonne capacité globale de détection. Cependant, la présence de 4 549 faux positifs indique que certains signaux normaux sont incorrectement classés comme anomalies, tandis que les 3 652 faux négatifs correspondent à des anomalies qui n'ont pas été détectées par le modèle. Malgré ces erreurs, les performances restent équilibrées, ce qui confirme la robustesse du modèle dans un contexte de détection d'anomalies sur les signaux IQ du dataset.

Chapitre IV : Expérimentaux et discussion des modèles de détection d'anomalie

5.1.2 Courbe ROC

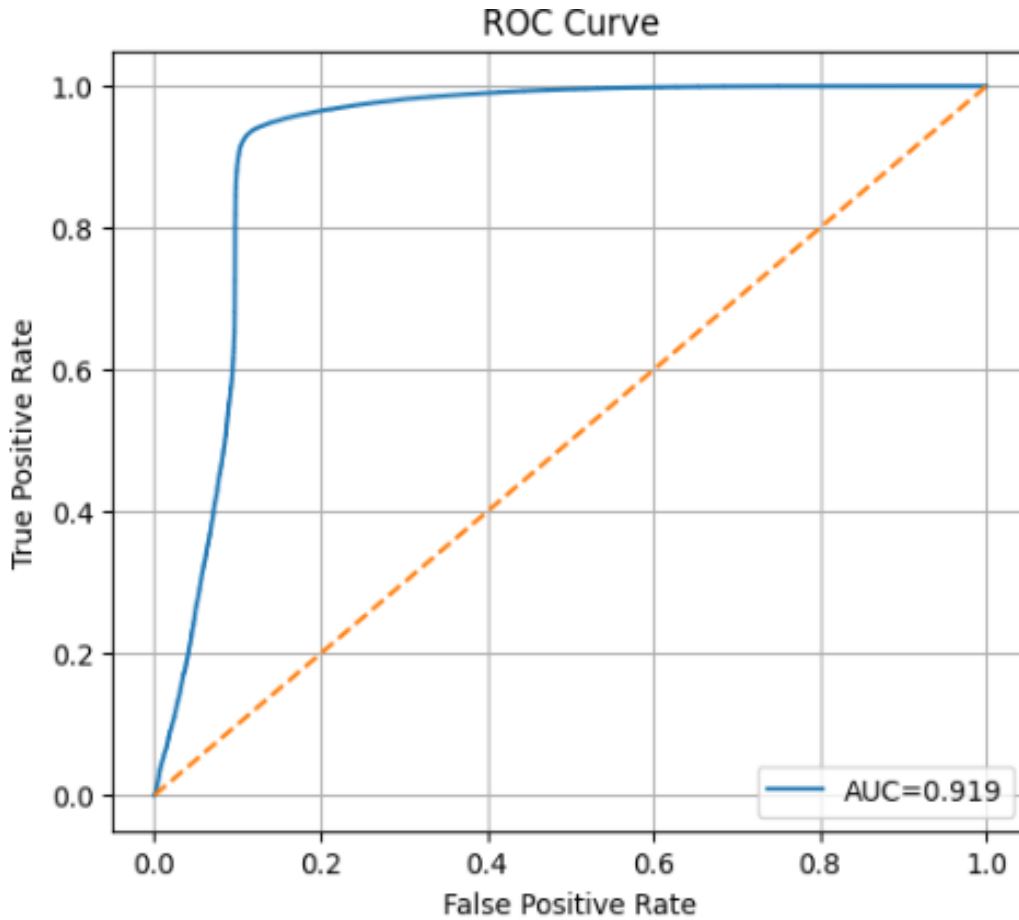


Figure IV.6: Courbe ROC du modèle Autoencodeur montrant la capacité de discrimination entre signaux normaux et anormaux.

La courbe ROC obtenue met en évidence les bonnes performances de l'Autoencodeur dans la détection des anomalies. La courbe est nettement située au-dessus de la diagonale représentant une classification aléatoire, ce qui traduit une forte capacité du modèle à distinguer les signaux normaux des signaux anormaux. De plus, la valeur de l'aire sous la courbe (**AUC = 0,919**) indique un excellent pouvoir de discrimination. Ces résultats montrent que l'Autoencodeur parvient à détecter efficacement les signaux dégradés tout en limitant le nombre de fausses alertes, confirmant ainsi sa pertinence pour la détection d'anomalies dans le dataset.

Chapitre IV : Expérimentaux et discussion des modèles de détection d'anomalie

5.2 Courbe d'apprentissage (Loss)

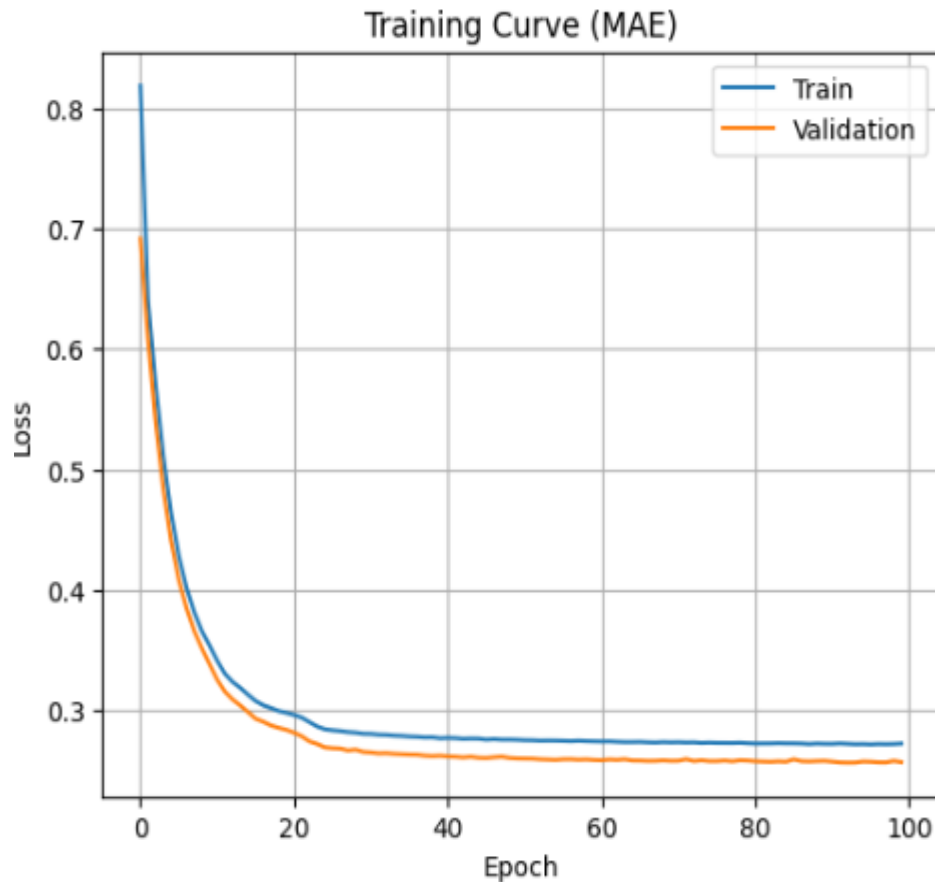


Figure IV.6: Courbe d'apprentissage du modèle Autoencodeur illustrant la convergence de la fonction de perte MAE

La **Figure 6** présente l'évolution de l'erreur moyenne absolue (MAE) au cours de l'apprentissage du modèle **Autoencodeur**, sur les ensembles d'entraînement et de validation. On observe une diminution rapide de la perte durant les premières époques, passant d'environ **0,82 à près de 0,30** pour l'entraînement et de **0,69 à environ 0,29** pour la validation. Cette baisse rapide traduit une phase d'apprentissage efficace, au cours de laquelle le modèle capture rapidement les principales caractéristiques des données. Par la suite, les deux courbes diminuent progressivement jusqu'à se stabiliser autour de **0,27** pour l'entraînement et de **0,26** pour la validation. La faible différence entre les courbes d'entraînement et de validation témoigne d'une bonne capacité de généralisation et ne révèle aucun signe significatif de surapprentissage. En effet, le modèle ne se contente pas de mémoriser les données d'apprentissage, mais apprend des

Chapitre IV : Expérimentaux et discussion des modèles de détection d'anomalie

représentations pertinentes pouvant être appliquées à de nouvelles données. Ainsi, la convergence stable du **MAE** confirme la qualité de l'apprentissage et montre que l'architecture de l'Autoencodeur est bien adaptée à la tâche de détection d'anomalies sur les signaux IQ du dataset.

5.3 Distribution des erreurs de reconstruction

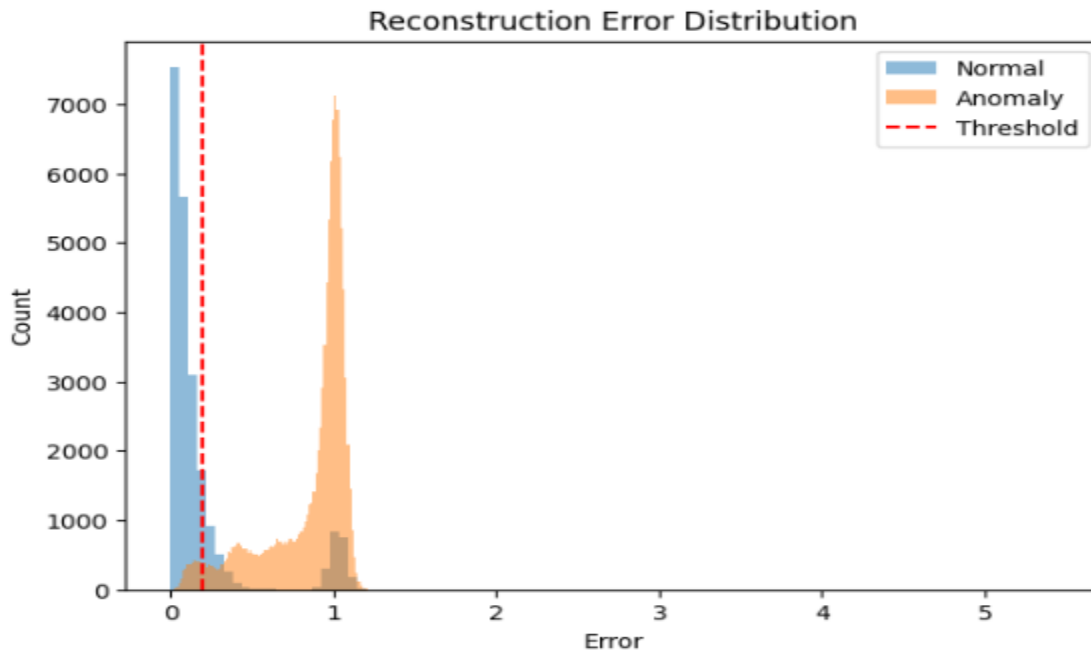


Figure IV.7 : Distribution des erreurs de reconstruction permettant de distinguer les signaux normaux des anomalies

La **Figure 7** présente la distribution de l'erreur de reconstruction calculée par l'Autoencodeur pour les signaux normaux et les signaux anormaux. On observe que les signaux normaux sont principalement concentrés autour de faibles valeurs d'erreur de reconstruction, ce qui montre que le modèle parvient à les reconstruire avec une grande précision. En revanche, les signaux anormaux présentent des erreurs de reconstruction plus élevées, traduisant la difficulté du modèle à reconstruire des données qu'il n'a pas observées lors de l'apprentissage. La ligne rouge pointillée représente le **seuil de décision optimal (0,196)**, déterminé à partir de l'analyse de la courbe ROC afin d'optimiser la séparation entre les deux classes. Les signaux dont l'erreur de reconstruction est supérieure à ce seuil sont considérés comme des anomalies, tandis que ceux dont l'erreur est inférieure sont classés

Chapitre IV : Expérimentaux et discussion des modèles de détection d'anomalie

comme normaux. Bien qu'un léger chevauchement soit observé entre les deux distributions, la séparation reste globalement satisfaisante, ce qui confirme la capacité de l'Autoencodeur à détecter efficacement les anomalies. Cette répartition est cohérente avec les bonnes performances obtenues en termes de précision, de rappel et d'AUC, démontrant la pertinence de l'Autoencodeur pour la détection d'anomalies sur les signaux IQ du dataset.

5.4 Isolation forset

L'algorithme **Isolation Forest** est une méthode non supervisée de détection d'anomalies basée sur le principe d'isolation des observations atypiques. Dans cette étude, il est appliqué aux signaux IQ du dataset **RadioML2016.10a** afin d'identifier les signaux dégradés par le bruit. Les principales étapes de traitement sont les suivantes :

- **Chargement des données** : récupération des 220 000 signaux IQ du dataset RadioML2016.10a.
- **Prétraitement** : réorganisation des signaux sous la forme (128, 2).
- **Étiquetage des anomalies** : les signaux ayant un SNR inférieur à 0 dB sont considérés comme des anomalies, tandis que les autres sont considérés comme normaux.
- **Séparation des données** : les 88 000 signaux normaux sont utilisés pour l'apprentissage, tandis que le jeu de test est constitué de 22 000 signaux normaux et de 110 000 signaux anormaux.
- **Extraction de caractéristiques** : calcul de plusieurs indicateurs statistiques et fréquentiels tels que la moyenne, l'écart-type, la puissance, l'énergie RMS, les caractéristiques spectrales issues de la FFT, le taux de passage par zéro et le rapport crête/RMS.
- **Normalisation robuste** : application du RobustScaler afin de réduire l'influence des valeurs extrêmes.
- **Réduction de dimension par PCA** : compression des caractéristiques extraites en six composantes principales, tout en conservant 99,5 % de la variance des données.
- **Entraînement de l'Isolation Forest** : construction d'une forêt de 600 arbres permettant d'isoler les observations anormales.

Chapitre IV : Expérimentaux et discussion des modèles de détection d'anomalie

-
- **Calcul des scores d'anomalie** : chaque signal reçoit un score reflétant son degré d'anomalie.
 - **Détermination du seuil optimal** : recherche du seuil maximisant le score F1 (0,126).
 - **Classification des signaux** : les signaux dont le score dépasse le seuil optimal sont considérés comme anormaux.
 - **Évaluation des performances** : analyse à l'aide de la matrice de confusion, de la courbe ROC, de la courbe Precision-Recall et des métriques de classification.

Les résultats obtenus montrent que le modèle atteint une **AUC-ROC de 0,840**, une **accuracy de 91 %** et un **score F1 de 94,57 %** pour la classe des anomalies. Le modèle obtient également une **précision de 92 %** et un **rappel de 97 %** pour la détection des anomalies, ce qui indique qu'il identifie correctement la grande majorité des signaux dégradés. En revanche, les performances sont plus faibles pour la classe des signaux normaux, avec une **précision de 79 %**, un **rappel de 60 %** et un **score F1 de 68 %**, ce qui traduit la présence de faux positifs. Malgré cette limitation, l'Isolation Forest présente une bonne capacité de détection grâce à son mécanisme d'isolation des observations atypiques. Toutefois, ses performances restent inférieures à celles de l'Autoencodeur, qui bénéficie de la capacité des réseaux de neurones à apprendre des représentations plus complexes et plus discriminantes des signaux IQ.

Chapitre IV : Expérimentaux et discussion des modèles de détection d'anomalie

5.4.1 Matrice de confusion

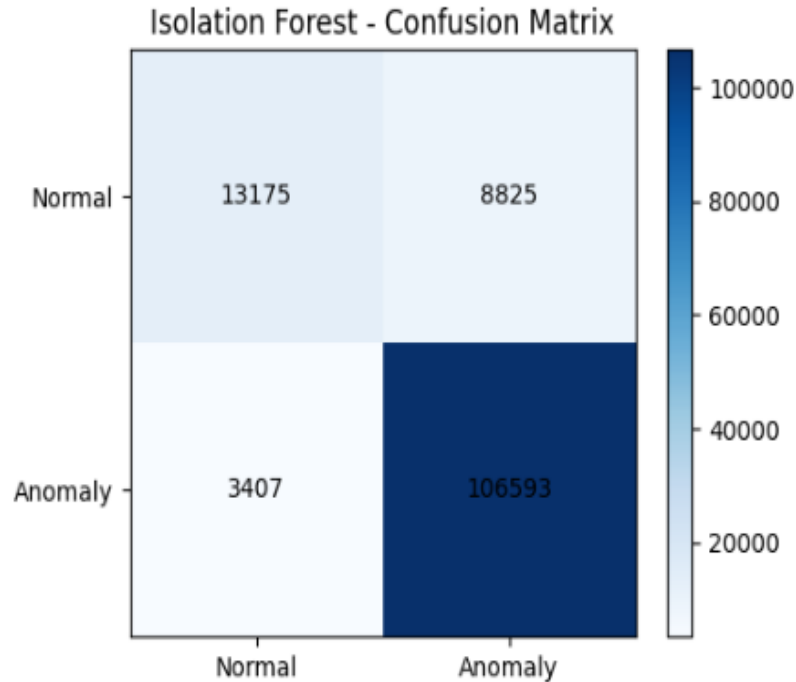


Figure IV.8 : Matrice de confusion du modèle Isolation Forest pour la détection d'anomalies des signaux IQ

À partir de la matrice de confusion obtenue, le modèle Isolation Forest présente les résultats suivants : 13 175 vrais négatifs (TN), 8 825 faux positifs (FP), 3 407 faux négatifs (FN) et 106 593 vrais positifs (TP). Ces résultats montrent que le modèle parvient à distinguer efficacement les signaux normaux des anomalies, avec un nombre élevé de classifications correctes, en particulier pour la classe des anomalies.

Plus précisément, le nombre important de vrais positifs (106 593) et de vrais négatifs (13 175) traduit une bonne capacité globale de détection. Toutefois, la présence de 8 825 faux positifs indique que certains signaux normaux sont incorrectement classés comme des anomalies, tandis que les 3 407 faux négatifs correspondent à des anomalies non détectées par le modèle. Malgré ces erreurs de classification, les performances obtenues restent satisfaisantes, ce qui confirme la robustesse du modèle Isolation Forest pour la détection d'anomalies sur les signaux IQ du dataset.

Chapitre IV : Expérimentaux et discussion des modèles de détection d'anomalie

5.4.2 Courbe ROC

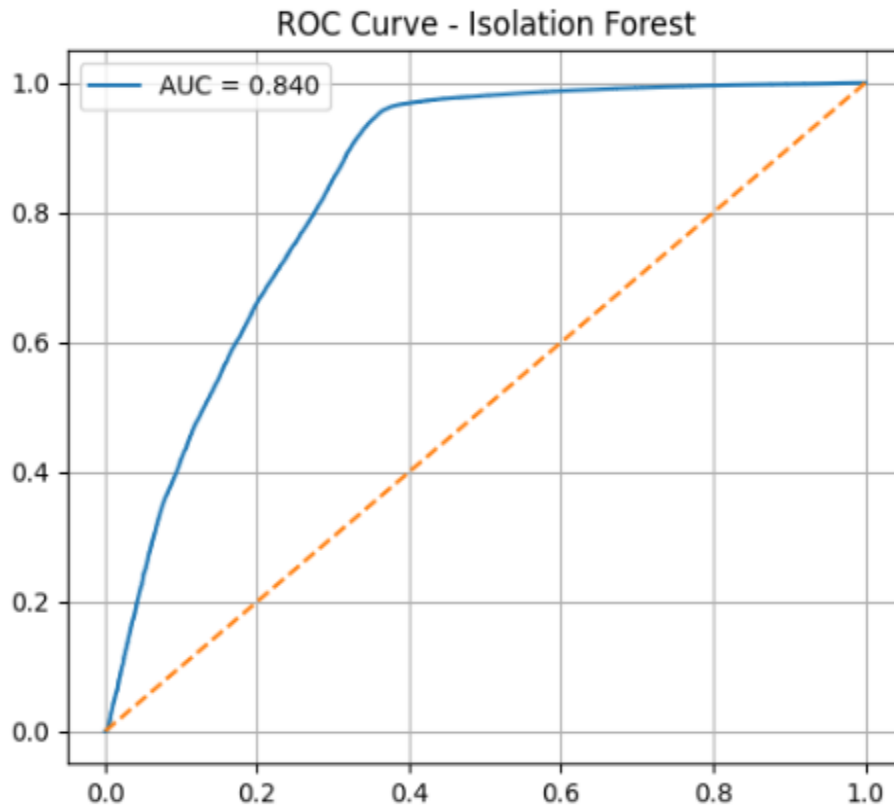


Figure IV.9 : Courbe ROC et valeur de l'AUC obtenues pour le modèle d'Isolation Forest.

La courbe ROC obtenue met en évidence les bonnes performances du modèle **Isolation Forest** dans la détection des anomalies. La courbe est nettement située au-dessus de la diagonale représentant une classification aléatoire, ce qui traduit une bonne capacité du modèle à distinguer les signaux normaux des signaux anormaux. De plus, la valeur de l'aire sous la courbe (**AUC = 0,840**) indique un bon pouvoir de discrimination. Ces résultats montrent que l'Isolation Forest est capable de détecter efficacement les signaux dégradés tout en limitant les erreurs de classification, confirmant ainsi sa pertinence pour la détection d'anomalies sur les signaux IQ du dataset.

Chapitre IV : Expérimentaux et discussion des modèles de détection d'anomalie

5.5 Distribution des erreurs de reconstruction

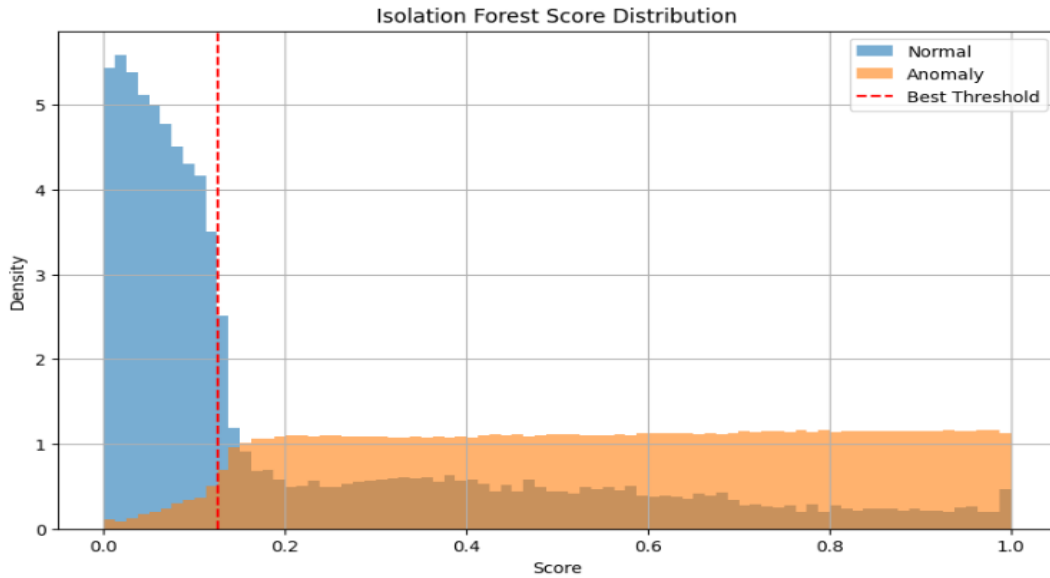


Figure IV.10 : Distribution des scores d'anomalie de l'Isolation Forest et positionnement du seuil optimal de décision.

L'histogramme de distribution des scores obtenu met en évidence les performances du modèle **Isolation Forest** dans la détection des anomalies. Les scores associés aux signaux normaux sont principalement concentrés dans les faibles valeurs, tandis que les scores des anomalies sont majoritairement répartis au-delà du seuil de décision, ce qui traduit une bonne capacité du modèle à distinguer les signaux normaux des signaux anormaux. Le **seuil optimal**, représenté par la ligne pointillée rouge, est fixé à **0,126**. Ce seuil permet de séparer efficacement les deux classes, bien qu'un léger chevauchement subsiste entre leurs distributions, expliquant la présence de quelques erreurs de classification. Dans l'ensemble, ces résultats montrent que l'Isolation Forest détecte efficacement les signaux dégradés tout en maintenant un bon compromis entre le taux de détection des anomalies et le nombre de fausses alertes, confirmant ainsi sa pertinence pour la détection d'anomalies sur les signaux IQ du dataset.

Chapitre IV : Expérimentaux et discussion des modèles de détection d'anomalie

5.6 OCSVM (One-Class SVM) :

Le One-Class Support Vector Machine (OCSVM) est une méthode d'apprentissage non supervisé dédiée à la détection d'anomalies. Son principe consiste à apprendre une frontière englobant uniquement les données normales, de manière à identifier comme anomalies les observations situées en dehors de cette frontière. Dans cette étude, l'OCSVM est appliqué aux signaux IQ du dataset RadioML2016.10a afin de détecter les signaux dégradés par le bruit. Les principales étapes de traitement sont les suivantes :

- **Chargement des données** : récupération des 220 000 signaux IQ du dataset RadioML2016.10a.
- **Prétraitement** : réorganisation des signaux sous la forme (128, 2) et normalisation des amplitudes afin de réduire les variations d'échelle.
- **Étiquetage des anomalies** : les signaux ayant un SNR inférieur à 0 dB sont considérés comme des anomalies, tandis que ceux ayant un SNR supérieur ou égal à 0 dB sont considérés comme normaux.
- **Séparation des données** : les 110 000 signaux normaux sont divisés en 80 % pour l'entraînement (88 000 signaux) et 20 % pour le test (22 000 signaux). L'ensemble de test est complété par les 110 000 signaux anormaux, soit 132 000 échantillons au total.
- **Extraction de caractéristiques** : calcul de 14 caractéristiques statistiques et fréquentielles, notamment la moyenne, l'écart-type, l'asymétrie (*skewness*), la kurtosis, la puissance moyenne, l'écart-type de la puissance, la puissance RMS, l'étendue (*peak-to-peak*), les caractéristiques spectrales issues de la FFT (magnitude maximale, moyenne, écart-type et centroïde spectral), le rapport d'énergie (*energy ratio*) ainsi que le taux de passage par zéro (*zero-crossing rate*).
- **Normalisation des caractéristiques** : application du StandardScaler afin de centrer et réduire les variables.
- **Réduction de dimension par PCA** : compression des caractéristiques en 7 composantes principales, tout en conservant 99,5 % de la variance des données.

Chapitre IV : Expérimentaux et discussion des modèles de détection d'anomalie

- **Entraînement du modèle OCSVM** : apprentissage réalisé uniquement sur les 88 000 signaux normaux à l'aide d'un noyau gaussien (RBF).
- **Recherche des hyperparamètres** : évaluation de plusieurs combinaisons des paramètres gamma et nu afin de sélectionner le modèle offrant la meilleure valeur d'AUC-ROC.
- **Calcul des scores d'anomalie** : chaque signal reçoit un score de décision reflétant son degré d'anomalie.
- **Normalisation robuste des scores** : utilisation de la médiane et de l'intervalle interquartile (IQR) afin de rendre les scores moins sensibles aux valeurs extrêmes.
- **Détermination du seuil optimal** : recherche du seuil maximisant la Balanced Accuracy, conduisant à un seuil optimal de $-0,7115$.
- **Classification des signaux** : les signaux dont le score est supérieur au seuil optimal sont classés comme anomalies.
- **Évaluation des performances** : analyse des résultats à l'aide de la matrice de confusion, de la courbe ROC, de la distribution des scores d'anomalie et des métriques de classification.

Les résultats obtenus montrent que le modèle atteint une AUC-ROC de 0,826, une exactitude (Accuracy) de 86 % et une Balanced Accuracy de 79 %. La matrice de confusion met en évidence 98 621 vrais positifs, 15 032 vrais négatifs, 6 968 faux positifs et 11 379 faux négatifs. Le rappel de 90 % pour la classe anomalie montre que la majorité des signaux dégradés sont correctement détectés. En revanche, la précision de 57 % pour la classe normale indique que certains signaux normaux sont encore confondus avec des anomalies. Malgré cette limitation, l'OCSVM démontre une bonne capacité de discrimination entre les signaux normaux et les signaux dégradés, ce qui en fait une approche pertinente pour la détection d'anomalies dans les signaux IQ du dataset RadioML2016.10a.

Chapitre IV : Expérimentaux et discussion des modèles de détection d'anomalie

5.6.1 Matrice de confusion

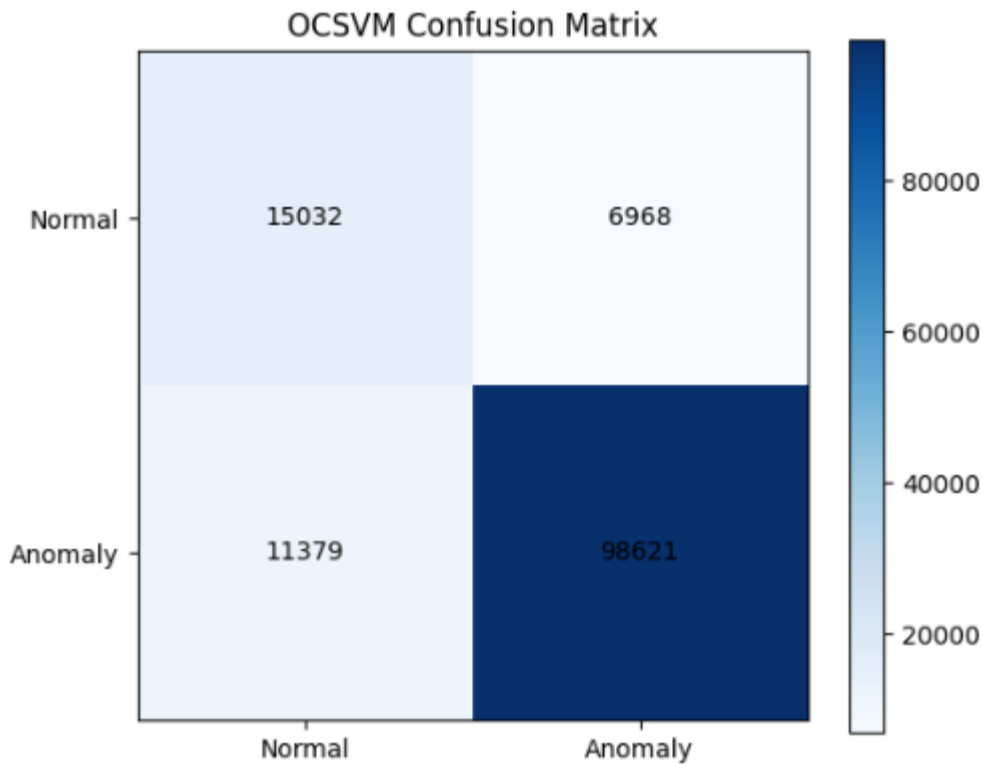


Figure IV.11:Matrice de confusion et performance de classification du modèle OCSVM

À partir de la matrice de confusion fournie, le modèle OCSVM présente les résultats suivants : 15 032 vrais négatifs (TN), 6 968 faux positifs (FP), 11 379 faux négatifs (FN) et 98 621 vrais positifs (TP). Ces valeurs indiquent que le modèle parvient à distinguer efficacement les signaux normaux des anomalies, avec une proportion importante de classifications correctes, en particulier pour la classe des anomalies.

Plus précisément, le nombre élevé de vrais positifs (98 621) montre que le modèle détecte correctement la majorité des anomalies présentes dans les données. En revanche, les 15 032 vrais négatifs indiquent les signaux normaux correctement identifiés. La présence de 6 968 faux positifs signifie que certains signaux normaux sont incorrectement classés comme des anomalies, tandis que les 11 379 faux négatifs correspondent à des anomalies non détectées par le modèle. Malgré ces erreurs de classification, les performances du modèle demeurent

Chapitre IV : Expérimentaux et discussion des modèles de détection d'anomalie

satisfaisantes, ce qui confirme son efficacité pour la détection d'anomalies sur les signaux IQ du jeu de données.

5.6.2 Courbe ROC

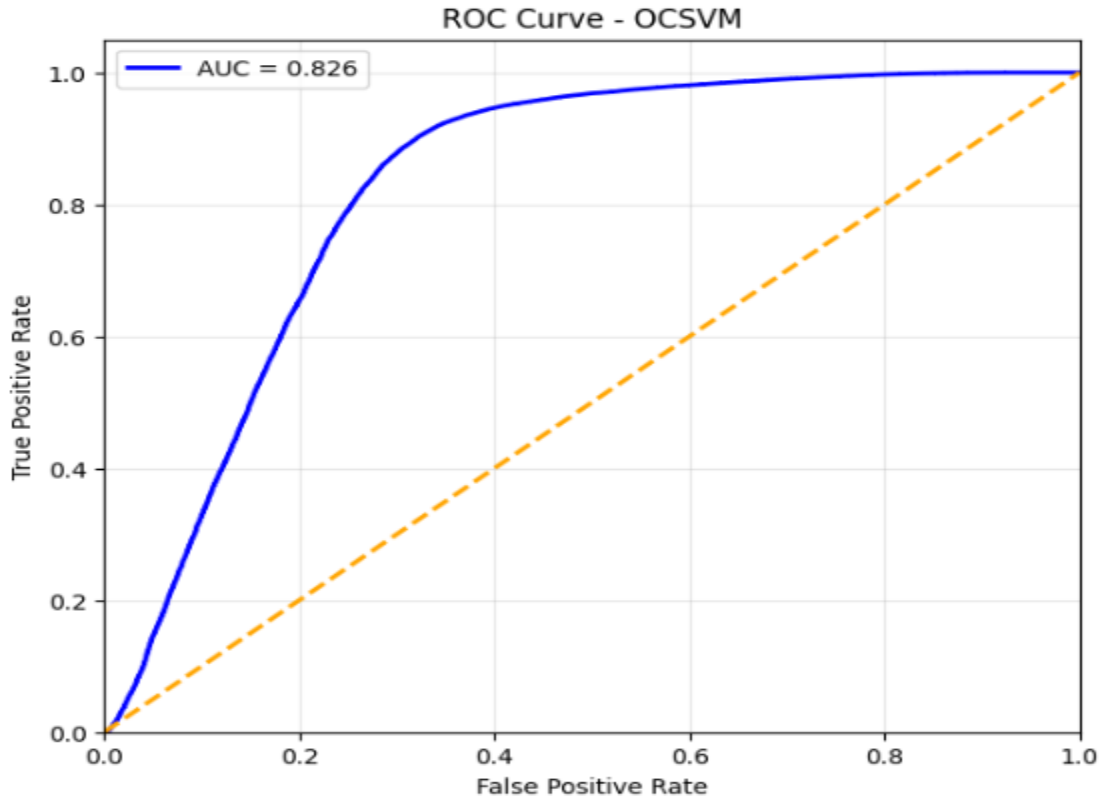


Figure IV.12 : Courbe ROC et performance globale de l'OCSVM pour la détection d'anomalies.

La courbe ROC obtenue met en évidence les bonnes performances de l'**OCSVM** dans la détection des anomalies. La courbe est nettement située au-dessus de la diagonale représentant une classification aléatoire, ce qui traduit une forte capacité du modèle à distinguer les signaux normaux des signaux anormaux. De plus, la valeur de l'aire sous la courbe (**AUC = 0,826**) indique un excellent pouvoir de discrimination. Ces résultats montrent que l'OCSVM parvient à détecter efficacement les signaux dégradés tout en limitant le nombre de fausses alertes, confirmant ainsi sa pertinence pour la détection d'anomalies dans le dataset

Chapitre IV : Expérimentaux et discussion des modèles de détection d'anomalie

5.6.3 Distribution des erreurs de reconstruction

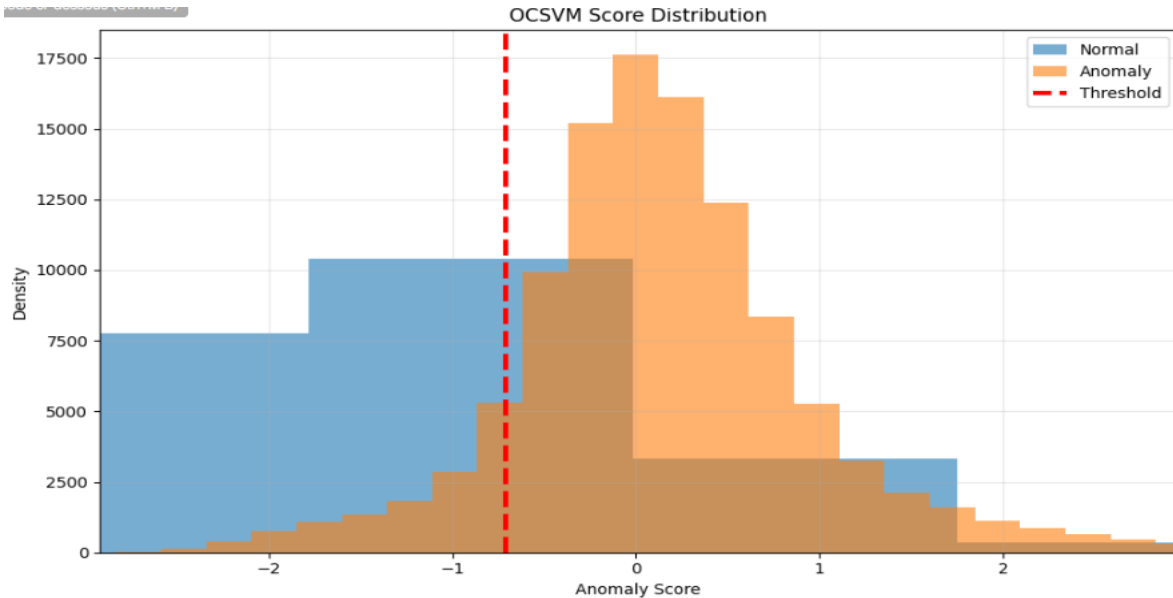


Figure IV.13 : Distribution des scores d'anomalie de l'OCSVM et positionnement du seuil de décision

L'histogramme de distribution des scores de décision de l'OCSVM met en évidence la capacité du modèle à distinguer les données normales des anomalies. Les signaux normaux présentent majoritairement des scores négatifs, tandis que les anomalies sont principalement associées à des scores plus élevés. Toutefois, un chevauchement entre les deux distributions est observé autour de la valeur seuil, ce qui explique la présence de certaines erreurs de classification (faux positifs et faux négatifs).

Le seuil de décision, fixé à **-0,711**, est représenté par la ligne pointillée rouge. Les échantillons dont le score est inférieur à cette valeur sont considérés comme normaux, alors que ceux dont le score est supérieur sont classés comme des anomalies. Malgré le chevauchement des distributions, la majorité des anomalies se situe du côté des scores élevés, tandis qu'une grande partie des données normales reste en dessous du seuil. Ces résultats confirment que l'OCSVM offre une capacité de discrimination satisfaisante pour la détection d'anomalies sur les signaux IQ du dataset.

Chapitre IV : Expérimentaux et discussion des modèles de détection d'anomalie

5.7 Comparaison globale des modèles et discussion

Les résultats expérimentaux obtenus mettent en évidence des différences significatives entre les trois modèles étudiés pour la détection d'anomalies sur les signaux IQ. En comparant leurs performances à l'aide des différentes métriques de classification, il apparaît que l'Autoencodeur constitue l'approche la plus performante parmi les méthodes évaluées.

En effet, l'Autoencodeur atteint une AUC-ROC de 0,919 et une accuracy de 94 %, ce qui témoigne d'une excellente capacité à distinguer les signaux normaux des signaux anormaux. Son apprentissage, basé sur la reconstruction des données normales, lui permet de produire une faible erreur de reconstruction pour les signaux conformes et une erreur plus élevée pour les anomalies. Cette approche se traduit par un faible nombre de faux positifs (4 549) et de faux négatifs (3 652), confirmant ainsi sa robustesse et sa bonne capacité de généralisation.

L'Isolation Forest obtient une AUC-ROC de 0,840 et une accuracy de 91 %. Grâce à son mécanisme d'isolation des observations atypiques, il détecte efficacement les anomalies, comme en témoigne son rappel de 97 % pour la classe des anomalies et son score F1 de 94,57 %. La matrice de confusion montre 13 175 vrais négatifs, 8 825 faux positifs, 3 407 faux négatifs et 106 593 vrais positifs. Bien que le modèle génère davantage de faux positifs que l'Autoencodeur, il conserve une bonne capacité de détection des anomalies, avec des performances globales proches de celles de ce dernier.

Le One-Class Support Vector Machine (OCSVM) atteint une AUC-ROC de 0,826 et une accuracy de 86 %. Bien que son AUC soit légèrement inférieure à celle de l'Isolation Forest, il présente des performances satisfaisantes en matière de détection des anomalies. La matrice de confusion met en évidence 98 621 vrais positifs, 15 032 vrais négatifs, 6 968 faux positifs et 11 379 faux négatifs. Toutefois, le nombre de faux négatifs demeure plus élevé que pour les deux autres approches, ce qui signifie que certaines anomalies ne sont pas détectées. Les performances de l'OCSVM restent également sensibles au choix des hyperparamètres (*gamma* et *nu*) ainsi qu'à la qualité des caractéristiques extraites et de leur représentation après la réduction de dimension par l'analyse en composantes principales (PCA).

Chapitre IV : Expérimentaux et discussion des modèles de détection d'anomalie

Globalement, cette étude montre que l'Autoencodeur est le modèle le plus performant pour la détection d'anomalies. Il présente la meilleure capacité de discrimination ($AUC = 0,919$) tout en conservant une excellente accuracy (94 %) et un faible nombre d'erreurs de classification. L'Isolation Forest constitue une alternative efficace, obtenant une accuracy de 91 % et une très bonne capacité de détection des anomalies, bien que son taux de faux positifs soit supérieur à celui de l'Autoencodeur. Enfin, l'OCSVM offre des performances satisfaisantes, mais reste légèrement en retrait, notamment en raison d'un nombre plus élevé de faux négatifs. Ainsi, parmi les trois méthodes étudiées, l'Autoencodeur apparaît comme la solution la plus adaptée, tandis que l'Isolation Forest représente une alternative intéressante lorsque l'on privilégie un modèle non supervisé simple, rapide et efficace.

5.7.1 Tableau comparatif final des modèles :

Modèle	ROC AUC	Accuracy	Rappel (Anomalies)	Précision	Faux Positifs (FP)	Faux Négatifs (FN)
Autoencodeur	0,919	94 %	97%	96%	4 549	3 652
Isolation Forest	0,840	91 %	97 %	92 %	8 825	3 407
OCSVM	0,826	86 %	90%	93%	6 968	11 379

Le tableau montre que l'Autoencodeur est le modèle le plus performant, avec la meilleure AUC-ROC (0,919) et la meilleure accuracy (94 %). Il présente également le plus faible nombre de faux positifs (4 549) et de faux négatifs (3 652), ce qui traduit une excellente capacité à distinguer les signaux normaux des anomalies.

Chapitre IV : Expérimentaux et discussion des modèles de détection d'anomalie

L'Isolation Forest obtient une AUC-ROC de 0,840 et une accuracy de 91 %. Il détecte efficacement la majorité des anomalies, avec un rappel de 97 % et une précision de 92 %. Toutefois, il génère davantage de faux positifs (8 825) que l'Autoencodeur, ce qui explique des performances légèrement inférieures malgré une bonne capacité de détection.

L'OCSVM atteint une AUC-ROC de 0,826 et une accuracy de 86 %. Bien que son rappel (90 %) soit inférieur à celui de l'Isolation Forest, il présente une précision de 93 % pour la détection des anomalies. En revanche, son nombre de faux négatifs (11 379) est plus élevé, ce qui signifie qu'il laisse passer davantage d'anomalies non détectées.

Conclusion

Ce chapitre a présenté l'ensemble du cadre expérimental et de la méthodologie adoptée pour la détection d'anomalies dans les signaux radio du dataset RadioML2016.10a. Il a permis de décrire en détail l'environnement de travail, les caractéristiques des données utilisées ainsi que la définition des anomalies basée sur le rapport signal sur bruit (SNR).

Le pipeline de traitement a été structuré en plusieurs étapes essentielles, incluant le prétraitement des données, la normalisation, la réduction de dimension et l'extraction de caractéristiques, afin de préparer les signaux pour les modèles d'apprentissage automatique. Trois approches de détection d'anomalies ont été étudiées, à savoir l'Autoencodeur, l'Isolation Forest et l'OCSVM, permettant ainsi une analyse comparative de différentes familles de méthodes.

Les résultats expérimentaux obtenus montrent que les performances varient significativement selon les modèles, avec une supériorité nette de l'Autoencodeur, suivi de l'Isolation Forest, puis de l'OCSVM. Cette comparaison confirme l'importance du choix de l'architecture et de la représentation des données dans la qualité de la détection.

Ainsi, ce chapitre constitue une base essentielle pour la validation des approches proposées et prépare l'interprétation globale des résultats dans la suite du travail.

Conclusion générale

Conclusion générale

Conclusion générale

Ce travail de fin d'études s'inscrit dans le domaine de la détection d'anomalies dans les signaux radio, en utilisant le dataset RadioML2016.10a comme support expérimental. Dans un contexte où les systèmes de communication sans fil deviennent de plus en plus complexes et exposés à des perturbations variées (bruit, interférences et dégradations du canal), la capacité à identifier automatiquement les signaux anormaux constitue un enjeu important pour l'amélioration de la fiabilité et de la qualité des transmissions.

L'objectif de ce travail était d'étudier et de comparer différentes approches de détection d'anomalies basées sur l'apprentissage automatique et l'apprentissage profond, en définissant les anomalies à partir du rapport signal sur bruit (SNR), où les signaux présentant un SNR inférieur à 0 dB sont considérés comme dégradés. Pour atteindre cet objectif, une chaîne de traitement complète a été mise en place, incluant le prétraitement des signaux IQ, la normalisation, l'extraction de caractéristiques pertinentes ainsi que la réduction de dimension.

Trois modèles ont été implémentés et évalués dans ce cadre : l'Autoencodeur, l'Isolation Forest et l'OCSVM. L'évaluation a été réalisée à l'aide de plusieurs métriques standard telles que la matrice de confusion, la courbe ROC, l'AUC ainsi que les indicateurs de précision et de rappel. Les résultats expérimentaux montrent que l'Autoencodeur obtient les meilleures performances globales, avec une AUC de 0,919 et une accuracy d'environ 94 %, ce qui s'explique par sa capacité à apprendre des représentations non linéaires riches et adaptées à la structure complexe des signaux IQ. L'Isolation Forest présente des performances intermédiaires, tandis que l'OCSVM obtient des résultats légèrement inférieurs, notamment en raison de sa sensibilité aux hyperparamètres et de ses limitations face à des distributions fortement non linéaires.

Globalement, cette étude confirme l'intérêt des approches basées sur l'apprentissage profond pour la détection d'anomalies dans les environnements radio complexes et bruités. Elle met également en évidence l'importance du choix de la représentation des données et du modèle utilisé dans la qualité des performances obtenues.

Conclusion générale

Enfin, plusieurs perspectives de recherche peuvent être envisagées. Il serait pertinent d'étendre cette étude à des données issues de systèmes réels de communication, d'explorer des architectures plus avancées telles que les Transformers ou les modèles auto-supervisés, ainsi que d'élargir la notion d'anomalie à d'autres phénomènes radio tels que les interférences multi-sources, le fading sélectif, les décalages fréquentiels ou encore les déséquilibres I/Q.

Travaux futurs

Travaux futurs

Dans le cadre de ce travail portant sur la détection d'anomalies dans les signaux radio à l'aide des techniques de Machine Learning, plusieurs perspectives peuvent être envisagées afin d'améliorer les performances et d'étendre le domaine d'application du système proposé :

- **Étendre l'étude à d'autres jeux de données radio**, tels que RadioML2018.01A ou des données réelles issues d'environnements radio opérationnels, afin d'évaluer la robustesse du modèle dans des conditions plus variées.
- **Explorer des architectures d'apprentissage profond plus avancées**, telles que les réseaux CNN, LSTM, Transformers ou les Auto-encodeurs variationnels (VAE), afin d'améliorer la précision de la détection des anomalies.
- **Étudier la détection d'anomalies dans des environnements radio plus complexes**, caractérisés par des canaux à évanouissement (Rayleigh, Rician) et des niveaux de bruit variables.
- **Développer un système de détection en temps réel**, capable d'analyser continuellement les signaux radio et de détecter instantanément les comportements anormaux.
- **Intégrer des techniques d'apprentissage semi-supervisé ou par transfert d'apprentissage**, afin de réduire la dépendance aux données étiquetées et d'améliorer la généralisation du modèle.
- **Étendre l'approche proposée aux réseaux de radio cognitive et aux systèmes 5G/6G**, où la détection rapide d'anomalies joue un rôle essentiel dans la gestion intelligente du spectre.

Travaux futurs

- **Étudier l'interprétabilité des modèles de Machine Learning**, afin de mieux comprendre les décisions prises par le système et d'accroître la confiance dans les résultats obtenus.

Références

Référence

Référence

- [1] I. Boudjema, «Études des formats de modulations et de démodulations : DPSK et QAM,» Université Aboubakr Belkaïd – Tlemcen, Tlemcen, 2018–2019.
- [2] H. K. Channi, «A Comparative Study of Various Digital Modulation Techniques,» International Journal in IT and Engineering, vol. 4, n° 103, pp. 39-49, 2016.
- [3] P. K. S. P. K. N. B. V. Singya, «A Survey on Higher-Order QAM Constellations: Technical Challenges, Recent Advances, and Future Trends,» IEEE Open Journal of the Communications Society, vol. 2, p. 617–655, 2021.
- [4] C. U. O. O. I. A. E. Ndujiuba, «Comparative Analysis of Digital Modulation Techniques in LTE 4G Systems,» Journal of Wireless Networking and Communications, vol. 5, n° 12, p. 60–66, 2015.
- [5] A. E. N. C. U. O. O. Ibhaze, «Comparative Analysis of Digital Modulation Techniques in LTE 4G Systems,» Journal of Wireless Networking and Communications, vol. 5, n° 12, p. 60–66, 2015.
- [6] G. Breed, «Bit Error Rate: Fundamental Concepts and Measurement Issues,» High Frequency Electronics, vol. 1, n° 12, 2003.
- [7] D. H. H. I. A. A. I. S. Z. S. L. B. M. I. N. A. Al-Nuaimi, «Performance of Feature-Based Techniques for Automatic Digital Modulation Recognition and Classification—A Review,» Electronics, vol. 8, n° 112, 2019.
- [8] D. H. H. I. A. A. I. S. Z. S. L. B. & M. I. N. A. Al-Nuaimi, «Performance of Feature-Based Techniques for Automatic Digital Modulation Recognition and Classification—A Review,» Electronics, vol. 8, n° 112, 2019.

Référence

- [9] V. G. S. S. Goswami, «Performance Analysis of Different M-ary Modulation Techniques over Wireless Fading Channel,» IOSR Journal of Electronics and Communication Engineering (IOSR-JECE), vol. 4, n° %11, p. 32–38, 2012.
- [10] D. M. A. P. J. Le Guennec, «Une nouvelle technique de reconnaissance de modulation numérique utilisant la fiabilité du détecteur de phase,» Colloque GRETSI, p. 513–516, 2011.
- [11] A. R. S. T. D. Garg, «Artificial Intelligence in Telecommunications: Applications, Risks, and Governance in the 5G and Beyond Era,» International Journal of Computer Techniques, vol. 10, n° %11, 2023.
- [12] D. N. K. C. B. Pandey, «Machine Learning Algorithms: A Review,» Machine Learning, vol. 6, n° %12, p. 916–922, 2019.
- [13] L. Prasittisopin, «Machine learning (ML) and deep learning (DL) in sustainable concrete construction: Review, trend and gap analyses,» Journal of Asian Architecture and Building Engineering, 2025.
- [14] I. Y. Z. Muhammad, «Supervised Machine Learning Approaches: A Survey,» ICTACT Journal on Soft Computing, vol. 5, n° %13, p. 946–952, 2015.
- [15] J. Ji, «Regression and Classification in Supervised Learning,» Proceedings of the 2nd International Conference on Computing and Big Data (ICCBD 2019) (dblp.org), p. 99–104, 2019.
- [16] S. A. A. A. S. M. M. Naeem, «An Unsupervised Machine Learning Algorithms: Comprehensive Review,» International Journal of Computing and Digital Systems, vol. 13, p. 911–921, 2023.

Référence

- [17] Y. C. A. P. V. P. R. B. E. Reddy, «Semi-supervised learning: A brief review,» International Journal of Engineering & Technology, vol. 7, n° 11.8, p. 81–85, 2018.
- [18] I. H. Sarker, «Machine Learning: Algorithms, Real-World Applications and Research Directions,» SN Computer Science, 2021.
- [19] I. ., Y. Z. Muhammad, «Supervised Machine Learning Approaches: A Survey,» ICTACT Journal on Soft Computing, p. 946–952, 2015.
- [20] I. ., H. Sarker, «Machine Learning: Algorithms, Real-World Applications and Research Directions,» SN Computer Science, vol. 2, n° 13, 2021.
- [21] K. S. S. H. Kapitanova, «Intelligent Sensor Networks: The Integration of Sensor Networks, Signal Processing and Machine Learning,» CRC Press, 2012, p. 3–29.
- [22] V. B. A. K. V. Chandola, «Anomaly Detection: A Survey,» ACM Computing Surveys, vol. 41, n° 13, 2009.
- [23] V. ., B. A. K. V. Chandola, «Anomaly Detection: A Survey,» ACM Computing Surveys, vol. 41, n° 13, p. 1–58, 2009.
- [24] Q.-V. N. N. T.-T. T. L. B. L. K. H. W.-J. Pham, «Intelligent Radio Signal Processing: A Survey,» IEEE Access, vol. 9, p. 83818–83850, 2021.
- [25] Y. Z. X. C. L. M. Y. Y. M. Zhao, «Research on Abnormal Radio Detection Method Combining Local Outlier Factor and One-Class Support Vector Machine,» Electronics, vol. 14, n° 120, 2025.
- [26] A. Y. M. Çalışkan, «A Machine Learning-Based Approach for Anomaly Detection in IoT Networks,» DergiPark Academic Journal, 2022.

Référence

- [27] G. L. S. C. G. Münz, «Traffic Anomaly Detection Using K-Means Clustering,» GI/ITG Workshop MMBnet, p. 13–14, 2007.
- [28] G. L. S. C. G. Münz, «Traffic Anomaly Detection Using K-Means Clustering,» GI/ITG Workshop MMBnet, p. 13–14, 2007.
- [29] O. Kozlovskiy, «Methods of Anomaly Detection in Information Systems,» Information Technology and Security Journal, 2024.
- [30] ., O. Kozlovskiy, «Methods of Anomaly Detection in Information Systems,» Information Technology and Security Journal, 2024.
- [31] M. M. K. H.-P. N. R. T. S. J. Breunig, «LOF: Identifying Density-Based Local Outliers,» Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, p. 93–104, 2000.
- [32] M. M. K. H.-P. N. R. T. S. J. Breunig, «LOF: Identifying Density-Based Local Outliers,» Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, p. 93–104, 2000.
- [33] L'apprentissage non supervisé pour la détection des signaux d'interférence radio fréquentiels, Montréal, Québec: Université du Québec à Montréal, 2022.
- [34] A. Fadul, «Anomaly Detection based on Isolation Forest and Local Outlier Factor,» ResearchGate, 2023.
- [35] M. M. Breunig, H.-P. Kriegel, R. T. Ng et J. Sander, LOF: Identifying Density-Based Local Outliers, Dallas, Texas: ACM, 2000.

Référence

- [36] N. L. Y. M. L. Y. Y. W. H. Wang, «Multidimensional CNN-LSTM Network for Automatic Modulation Classification,» *Electronics*, vol. 10, n° 114, 2021.
- [37] E. C. G. D. N. L. M. A. A. Ö. D. B. M.-G. Zanj, «Energy Efficiency in Short and Wide-Area IoT Technologies: A Survey,» vol. 9, n° 11, 2021.
- [38] A. George, «Anomaly Detection Based on Machine Learning: Dimensionality Reduction Using PCA and Classification Using SVM,» *International Journal of Computer Applications*, vol. 47, n° 121, p. 5–8, 2012.
- [39] B. J. K. F. Erickson, «Magician's Corner: 9. Performance Metrics for Machine Learning Models,» *Radiology: Artificial Intelligence*, vol. 3, n° 13, 2021.
- [40] D. K. M. S. Sharma, «Analog and Digital Modulation Techniques: An Overview,» *TECHNIA International Journal of Computing Science & Communication Technologies*, vol. 3, n° 11, 2010.
- [41] A. A. Amache Salazar, *L'apprentissage non supervisé pour la détection des signaux d'interférence radio fréquentiels*, Montréal, Québec, Canada: Université du Québec à Montréal, 2022.